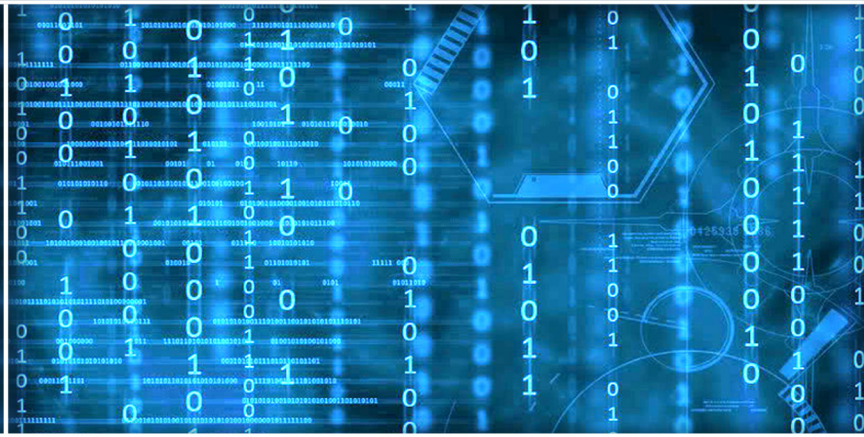


Volume 9 Issue 1

January 2018



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)



[www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)

# Editorial Preface

## *From the Desk of Managing Editor...*

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

**Thank you for Sharing Wisdom!**

**Managing Editor**  
**IJACSA**  
**Volume 9 Issue 1 January 2018**  
**ISSN 2156-5570 (Online)**  
**ISSN 2158-107X (Print)**  
**©2013 The Science and Information (SAI) Organization**

# Editorial Board

## Editor-in-Chief

**Dr. Kohei Arai - Saga University**

*Domains of Research: Technology Trends, Computer Vision, Decision Making, Information Retrieval, Networking, Simulation*

---

## Associate Editors

**Chao-Tung Yang**

**Department of Computer Science, Tunghai University, Taiwan**

*Domain of Research: Software Engineering and Quality, High Performance Computing, Parallel and Distributed Computing, Parallel Computing*

**Elena SCUTELNICU**

**"Dunarea de Jos" University of Galati, Romania**

*Domain of Research: e-Learning, e-Learning Tools, Simulation*

**Krassen Stefanov**

**Professor at Sofia University St. Kliment Ohridski, Bulgaria**

*Domains of Research: e-Learning, Agents and Multi-agent Systems, Artificial Intelligence, Big Data, Cloud Computing, Data Retrieval and Data Mining, Distributed Systems, e-Learning Organisational Issues, e-Learning Tools, Educational Systems Design, Human Computer Interaction, Internet Security, Knowledge Engineering and Mining, Knowledge Representation, Ontology Engineering, Social Computing, Web-based Learning Communities, Wireless/ Mobile Applications*

**Maria-Angeles Grado-Caffaro**

**Scientific Consultant, Italy**

*Domain of Research: Electronics, Sensing and Sensor Networks*

**Mohd Helmy Abd Wahab**

**Universiti Tun Hussein Onn Malaysia**

*Domain of Research: Intelligent Systems, Data Mining, Databases*

**T. V. Prasad**

**Lingaya's University, India**

*Domain of Research: Intelligent Systems, Bioinformatics, Image Processing, Knowledge Representation, Natural Language Processing, Robotics*

## Reviewer Board Members

- **Aamir Shaikh**
- **Abbas Al-Ghaili**  
Mendeley
- **Abbas Karimi**  
Islamic Azad University Arak Branch
- **Abdelghni Lakehal**  
Université Abdelmalek Essaadi Faculté  
Polydisciplinaire de Larache Route de Rabat, Km 2 -  
Larache BP. 745 - Larache 92004. Maroc.
- **Abdul Razak**
- **Abdul Karim ABED**
- **Abdur Rashid Khan**  
Gomal University
- **Abeer Elkorany**  
Faculty of computers and information, Cairo
- **ADEMOLA ADESINA**  
University of the Western Cape
- **Aderemi A. Atayero**  
Covenant University
- **Adi Maaita**  
ISRA UNIVERSITY
- **Adnan Ahmad**
- **Adrian Branga**  
Department of Mathematics and Informatics,  
Lucian Blaga University of Sibiu
- **agana Becejski-Vujaklija**  
University of Belgrade, Faculty of organizational
- **Ahmad Saifan**  
yarmouk university
- **Ahmed Boutejdar**
- **Ahmed AL-Jumaily**  
Ahlia University
- **Ahmed Nabih Zaki Rashed**  
Menoufia University
- **Ajantha Herath**  
Stockton University Galloway
- **Akbar Hossain**
- **Akram Belghith**  
University Of California, San Diego
- **Albert S**  
Kongu Engineering College
- **Alcinia Zita Sampaio**  
Technical University of Lisbon
- **Alexane Bouënard**  
Sensopia
- **ALI ALWAN**  
International Islamic University Malaysia
- **Ali Ismail Awad**  
Luleå University of Technology
- **Alicia Valdez**
- **Amin Shaqrah**  
Taibah University
- **Amirrudin Kamsin**
- **Amitava Biswas**  
Cisco Systems
- **Anand Nayyar**  
KCL Institute of Management and Technology,  
Jalandhar
- **Andi Wahyu Rahardjo Emanuel**  
Maranatha Christian University
- **Anews Samraj**  
Mahendra Engineering College
- **Anirban Sarkar**  
National Institute of Technology, Durgapur
- **Anthony Isizoh**  
Nnamdi Azikiwe University, Awka, Nigeria
- **Antonio Formisano**  
University of Naples Federico II
- **Anuj Gupta**  
IKG Punjab Technical University
- **Anuranjan misra**  
Bhagwant Institute of Technology, Ghaziabad, India
- **Appasami Govindasamy**
- **Arash Habibi Lashkari**  
University Technology Malaysia(UTM)
- **Aree Mohammed**  
Directorate of IT/ University of Sulaimani
- **ARINDAM SARKAR**  
University of Kalyani, DST INSPIRE Fellow
- **Aris Skander**  
Constantine 1 University
- **Ashok Matani**  
Government College of Engg, Amravati
- **Ashraf Owis**  
Cairo University
- **Asoke Nath**

St. Xaviers College(Autonomous), 30 Park Street,  
Kolkata-700 016

- **Athanasios Koutras**
- **Ayad Ismaeel**  
Department of Information Systems Engineering-  
Technical Engineering College-Erbil Polytechnic  
University, Erbil-Kurdistan Region- IRAQ
- **Ayman Shehata**  
Department of Mathematics, Faculty of Science,  
Assiut University, Assiut 71516, Egypt.
- **Ayman EL-SAYED**  
Computer Science and Eng. Dept., Faculty of  
Electronic Engineering, Menofia University
- **Babatunde Opeoluwa Akinkunmi**  
University of Ibadan
- **Bae Bossoufi**  
University of Liege
- **BALAMURUGAN RAJAMANICKAM**  
Anna university
- **Balasubramanie Palanisamy**
- **BASANT VERMA**  
RAJEEV GANDHI MEMORIAL COLLEGE, HYDERABAD
- **Basil Hamed**  
Islamic University of Gaza
- **Basil Hamed**  
Islamic University of Gaza
- **Bhanu Prasad Pinnamaneni**  
Rajalakshmi Engineering College; Matrix Vision  
GmbH
- **Bharti Waman Gawali**  
Department of Computer Science & information T
- **Bilian Song**  
LinkedIn
- **Binod Kumar**  
JSPM's Jayawant Technical Campus, Pune, India
- **Bogdan Belean**
- **Bohumil Brtnik**  
University of Pardubice, Department of Electrical  
Engineering
- **Bouchaib CHERRADI**  
CRMEF
- **Brahim Raouyane**  
FSAC
- **Branko Karan**
- **Bright Keswani**  
Department of Computer Applications, Suresh Gyan  
Vihar University, Jaipur (Rajasthan) INDIA
- **Brij Gupta**

University of New Brunswick

- **C Venkateswarlu Sonagiri**  
JNTU
- **Chanashekhhar Meshram**  
Chhattisgarh Swami Vivekananda Technical  
University
- **Chao Wang**
- **Chao-Tung Yang**  
Department of Computer Science, Tunghai  
University
- **Charlie Obimbo**  
University of Guelph
- **Chee Hon Lew**
- **Chien-Peng Ho**  
Information and Communications Research  
Laboratories, Industrial Technology Research  
Institute of Taiwan
- **Chun-Kit (Ben) Ngan**  
The Pennsylvania State University
- **Ciprian Dobre**  
University Politehnica of Bucharest
- **Constantin POPESCU**  
Department of Mathematics and Computer  
Science, University of Oradea
- **Constantin Filote**  
Stefan cel Mare University of Suceava
- **CORNELIA AURORA Gyorödi**  
University of Oradea
- **Cosmina Ivan**
- **Cristina Turcu**
- **Dana PETCU**  
West University of Timisoara
- **Daniel Albuquerque**
- **Dariusz Jakóbczak**  
Technical University of Koszalin
- **Deepak Garg**  
Thapar University
- **Devena Prasad**
- **DHAYA R**
- **Dheyaa Kadhim**  
University of Baghdad
- **Djilali IDOUGH**  
University A.. Mira of Bejaia
- **Dong-Han Ham**  
Chonnam National University
- **Dr. Arvind Sharma**

- Aryan College of Technology, Rajasthan Technology University, Kota
- **Duck Hee Lee**  
Medical Engineering R&D Center/Asan Institute for Life Sciences/Asan Medical Center
  - **Elena SCUTELNICU**  
"Dunarea de Jos" University of Galati
  - **Elena Camossi**  
Joint Research Centre
  - **Eui Lee**  
Sangmyung University
  - **Evgeny Nikulchev**  
Moscow Technological Institute
  - **Ezekiel OKIKE**  
UNIVERSITY OF BOTSWANA, GABORONE
  - **Fahim Akhter**  
King Saud University
  - **FANGYONG HOU**  
School of IT, Deakin University
  - **Faris Al-Salem**  
GCET
  - **Firkhan Ali Hamid Ali**  
UTHM
  - **Fokrul Alom Mazarbhuiya**  
King Khalid University
  - **Frank Ibikunle**  
Botswana Int'l University of Science & Technology (BIUST), Botswana
  - **Fu-Chien Kao**  
Da-Y eh University
  - **Gamil Abdel Azim**  
Suez Canal University
  - **Ganesh Sahoo**  
RMRIMS
  - **Gaurav Kumar**  
Manav Bharti University, Solan Himachal Pradesh
  - **George Pecherle**  
University of Oradea
  - **George Mastorakis**  
Technological Educational Institute of Crete
  - **Georgios Galatas**  
The University of Texas at Arlington
  - **Gerard Dumancas**  
Oklahoma Baptist University
  - **Ghalem Belalem**  
University of Oran 1, Ahmed Ben Bella
  - **gherabi noreddine**
  - **Giacomo Veneri**  
University of Siena
  - **Giri Babu**  
Indian Space Research Organisation
  - **Govindarajulu Salendra**
  - **Grebenisan Gavril**  
University of Oradea
  - **Gufan Ahmad Ansari**  
Qassim University
  - **Gunaseelan Devaraj**  
Jazan University, Kingdom of Saudi Arabia
  - **GYÖRÖDI ROBERT STEFAN**  
University of Oradea
  - **Hadj Tadjine**  
IAV GmbH
  - **Haewon Byeon**  
Nambu University
  - **Haiguang Chen**  
ShangHai Normal University
  - **Hamid Alinejad-Rokny**  
The University of New South Wales
  - **Hamid AL-Asadi**  
Department of Computer Science, Faculty of Education for Pure Science, Basra University
  - **Hamid Mukhtar**  
National University of Sciences and Technology
  - **Hany Hassan**  
EPF
  - **Harco Leslie Henic SPITS WARNARS**  
Bina Nusantara University
  - **Hariharan Shanmugasundaram**  
Associate Professor, SRM
  - **Harish Garg**  
Thapar University Patiala
  - **Hazem I. El Shekh Ahmed**  
Pure mathematics
  - **Hemalatha SenthilMahesh**
  - **Hesham Ibrahim**  
Faculty of Marine Resources, Al-Mergheb University
  - **Himanshu Aggarwal**  
Department of Computer Engineering
  - **Hongda Mao**  
Hossam Faris
  - **Huda K. AL-Jobori**  
Ahlia University
  - **Imed JABRI**

- **iss EL OUADGHIRI**
- **Iwan Setyawan**  
Satya Wacana Christian University
- **Jacek M. Czerniak**  
Casimir the Great University in Bydgoszcz
- **Jai Singh W**
- **JAMAIAH HAJI YAHAYA**  
NORTHERN UNIVERSITY OF MALAYSIA (UUM)
- **James Coleman**  
Edge Hill University
- **Jatinderkumar Saini**  
Narmada College of Computer Application, Bharuch
- **Javed Sheikh**  
University of Lahore, Pakistan
- **Jayaram A**  
Siddaganga Institute of Technology
- **Ji Zhu**  
University of Illinois at Urbana Champaign
- **Jia Uddin Jia**  
Assistant Professor
- **Jim Wang**  
The State University of New York at Buffalo,  
Buffalo, NY
- **John Sahlin**  
George Washington University
- **JOHN MANOHAR**  
VTU, Belgaum
- **JOSE PASTRANA**  
University of Malaga
- **Jui-Pin Yang**  
Shih Chien University
- **Jyoti Chaudhary**  
high performance computing research lab
- **K V.L.N.Acharyulu**  
Bapatla Engineering college
- **Ka-Chun Wong**
- **Kamatchi R**
- **Kamran Kowsari**  
The George Washington University
- **KANNADHASAN SURIYAN**
- **Kashif Nisar**  
Universiti Utara Malaysia
- **Kato Mivule**
- **Kayhan Zrar Ghafoor**  
University Technology Malaysia
- **Kennedy Okafor**  
Federal University of Technology, Owerri
- **Khalid Mahmood**  
IEEE
- **Khalid Sattar Abdul**  
Assistant Professor
- **Khin Wee Lai**  
Biomedical Engineering Department, University  
Malaya
- **Khurram Khurshid**  
Institute of Space Technology
- **KIRAN SREE POKKULURI**  
Professor, Sri Vishnu Engineering College for  
Women
- **KITIMAPORN CHOOCHOTE**  
Prince of Songkla University, Phuket Campus
- **Krasimir Yordzhev**  
South-West University, Faculty of Mathematics and  
Natural Sciences, Blagoevgrad, Bulgaria
- **Krassen Stefanov**  
Professor at Sofia University St. Kliment Ohridski
- **Labib Gergis**  
Misr Academy for Engineering and Technology
- **LATHA RAJAGOPAL**
- **Lazar Stošić**  
College for professional studies educators  
Aleksinac, Serbia
- **Leanos Maglaras**  
De Montfort University
- **Leon Abdillah**  
Bina Darma University
- **Lijian Sun**  
Chinese Academy of Surveying and
- **Ljubomir Jerinic**  
University of Novi Sad, Faculty of Sciences,  
Department of Mathematics and Computer Science
- **Lokesh Sharma**  
Indian Council of Medical Research
- **Long Chen**  
Qualcomm Incorporated
- **M. Reza Mashinchi**  
Research Fellow
- **M. Tariq Banday**  
University of Kashmir
- **madjid khalilian**
- **majzoob omer**
- **Mallikarjuna Doodipala**  
Department of Engineering Mathematics, GITAM  
University, Hyderabad Campus, Telangana, INDIA

- **Manas deep**  
Masters in Cyber Law & Information Security
- **Manju Kaushik**
- **Manoharan P.S.**  
Associate Professor
- **Manoj Wadhwa**  
Echelon Institute of Technology Faridabad
- **Manpreet Manna**  
Director, All India Council for Technical Education,  
Ministry of HRD, Govt. of India
- **Manuj Darbari**  
BBD University
- **Marcellin Julius Nkenlifack**  
University of Dschang
- **Maria-Angeles Grado-Caffaro**  
Scientific Consultant
- **Marwan Alseid**  
Applied Science Private University
- **Mazin Al-Hakeem**  
LFU (Lebanese French University) - Erbil, IRAQ
- **Md Islam**  
sikkim manipal university
- **Md. Bhuiyan**  
King Faisal University
- **Md. Zia Ur Rahman**  
Narasaraopeta Engg. College, Narasaraopeta
- **Mehdi Bahrami**  
University of California, Merced
- **Messaouda AZZOUZI**  
Ziane Achour University of Djelfa
- **Milena Bogdanovic**  
University of Nis, Teacher Training Faculty in Vranje
- **Miriampally Venkata Raghavendra**  
Adama Science & Technology University, Ethiopia
- **Mirjana Popovic**  
School of Electrical Engineering, Belgrade University
- **Miroslav Baca**  
University of Zagreb, Faculty of organization and  
informatics / Center for biometrics
- **Moeiz Miraoui**  
University of Gafsa
- **Mohamed Eldosoky**
- **Mohamed Ali Mahjoub**  
Preparatory Institute of Engineer of Monastir
- **Mohamed Kaloup**
- **Mohamed El-Sayed**  
Faculty of Science, Fayoum University, Egypt
- **Mohamed Najeh LAKHOUA**  
ESTI, University of Carthage
- **Mohammad Ali Badamchizadeh**  
University of Tabriz
- **Mohammad Jannati**
- **Mohammad Alomari**  
Applied Science University
- **Mohammad Haghighat**  
University of Miami
- **Mohammad Azzeh**  
Applied Science university
- **Mohammed Akour**  
Yarmouk University
- **Mohammed Sadgal**  
Cadi Ayyad University
- **Mohammed Al-shabi**  
Associate Professor
- **Mohammed Hussein**
- **Mohammed Kaiser**  
Institute of Information Technology
- **Mohammed Ali Hussain**  
Sri Sai Madhavi Institute of Science & Technology
- **Mohd Helmy Abd Wahab**  
University Tun Hussein Onn Malaysia
- **Mokhtar Beldjehem**  
University of Ottawa
- **Mona Elshinawy**  
Howard University
- **Mostafa Ezziyyani**  
FSTT
- **Mouhammd sharari alkasassbeh**
- **Mourad Amad**  
Laboratory LAMOS, Bejaia University
- **Mueen Uddin**  
University Malaysia Pahang
- **MUNTASIR AL-ASFOOR**  
University of Al-Qadisiyah
- **Murphy Choy**
- **Murthy Dasika**  
Geethanjali College of Engineering & Technology
- **Mustapha OUJAOURA**  
Faculty of Science and Technology Béni-Mellal
- **MUTHUKUMAR SUBRAMANYAM**  
DGCT, ANNA UNIVERSITY
- **N.Ch. Iyengar**  
VIT University
- **Nagy Darwish**



Department of Computer and Information Sciences,  
Institute of Statistical Studies and Researches, Cairo  
University

- **Najib Kofahi**  
Yarmouk University
- **Nan Wang**  
LinkedIn
- **Natarajan Subramanyam**  
PES Institute of Technology
- **Natheer Gharaibeh**  
College of Computer Science & Engineering at  
Yanbu - Taibah University
- **Nazeeh Ghatasheh**  
The University of Jordan
- **Nazeeruddin Mohammad**  
Prince Mohammad Bin Fahd University
- **NEERAJ SHUKLA**  
ITM UNiversity, Gurgaon, (Haryana) Inida
- **Neeraj Tiwari**
- **Nestor Velasco-Bermeo**  
UPFIM, Mexican Society of Artificial Intelligence
- **Nidhi Arora**  
M.C.A. Institute, Ganpat University
- **Nilanjan Dey**
- **Ning Cai**  
Northwest University for Nationalities
- **Nithyanandam Subramanian**  
Professor & Dean
- **Noura Aknin**  
University Abdelamlek Essaadi
- **Obaida Al-Hazaimeh**  
Al- Balqa' Applied University (BAU)
- **Oliviu Matei**  
Technical University of Cluj-Napoca
- **Om Sangwan**
- **Omaima Al-Allaf**  
Asesstant Professor
- **Osama Omer**  
Aswan University
- **Ouchtati Salim**
- **Ousmane THIARE**  
Associate Professor University Gaston Berger of  
Saint-Louis SENEGAL
- **Paresh V Virparia**  
Sardar Patel University
- **Peng Xia**  
Microsoft

- **Ping Zhang**  
IBM
- **Poonam Garg**  
Institute of Management Technology, Ghaziabad
- **Prabhat K Mahanti**  
UNIVERSITY OF NEW BRUNSWICK
- **PROF DURGA SHARMA ( PHD)**  
AMUIT, MOEFDRE & External Consultant (IT) &  
Technology Tansfer Research under ILO & UNDP,  
Academic Ambassador for Cloud Offering IBM-USA
- **Purwanto Purwanto**  
Faculty of Computer Science, Dian Nuswantoro  
University
- **Qifeng Qiao**  
University of Virginia
- **Rachid Saadane**  
EE departement EHTP
- **Radwan Tahboub**  
Palestine Polytechnic University
- **raed Kanaan**  
Amman Arab University
- **Raghuraj Singh**  
Harcourt Butler Technological Institute
- **Rahul Malik**
- **raja boddu**  
LENORA COLLEGE OF ENGINEERNG
- **Raja Ramachandran**
- **Rajesh Kumar**  
National University of Singapore
- **Rakesh Dr.**  
Madan Mohan Malviya University of Technology
- **Rakesh Balabantaray**  
IIIT Bhubaneswar
- **Ramani Kannan**  
Universiti Teknologi PETRONAS, Bandar Seri  
Iskandar, 31750, Tronoh, Perak, Malaysia
- **Rashad Al-Jawfi**  
Ibb university
- **Rashid Sheikh**  
Shri Aurobindo Institute of Technology, Indore
- **Ravi Prakash**  
University of Mumbai
- **RAVINA CHANGALA**
- **Ravisankar Hari**  
CENTRAL TOBACCO RESEARCH INSTITUE
- **Rawya Rizk**  
Port Said University

- **Reshmy Krishnan**  
Muscat College affiliated to Stirling University.U
- **Ricardo Vardasca**  
Faculty of Engineering of University of Porto
- **Ritaban Dutta**  
ISSL, CSIRO, Tasmania, Australia
- **Rowayda Sadek**
- **Ruchika Malhotra**  
Delhi Technological University
- **Rutvij Jhaveri**  
Gujarat
- **SAADI Slami**  
University of Djelfa
- **Sachin Kumar Agrawal**  
University of Limerick
- **Sagarmay Deb**  
Central Queensland University, Australia
- **Said Ghoniemy**  
Taif University
- **Sandeep Reddivari**  
University of North Florida
- **Sanskriti Patel**  
Charotar University of Science & Technology,  
Changa, Gujarat, India
- **Santosh Kumar**  
Graphic Era University, Dehradun (UK)
- **Sasan Adibi**  
Research In Motion (RIM)
- **Satyena Singh**  
Professor
- **Sebastian Marius Rosu**  
Special Telecommunications Service
- **Seema Shah**  
Vidyalankar Institute of Technology Mumbai
- **Seifedine Kadry**  
American University of the Middle East
- **Selem Charfi**  
HD Technology
- **SENGOTTUVELAN P**  
Anna University, Chennai
- **Senol Piskin**  
Istanbul Technical University, Informatics Institute
- **Sérgio Ferreira**  
School of Education and Psychology, Portuguese  
Catholic University
- **Seyed Hamidreza Mohades Kasaei**  
University of Isfahan
- **Shafiqul Abidin**  
HMR Institute of Technology & Management  
(Affiliated to GGSIP University), Hamidpur, Delhi -  
110036
- **Shahanawaj Ahamad**  
The University of Al-Kharj
- **Shaidah Jusoh**
- **Shaiful Bakri Ismail**
- **Shakir Khan**  
Al-Imam Muhammad Ibn Saud Islamic University
- **Shawki Al-Dubae**  
Assistant Professor
- **Sherif Hussein**  
Mansoura University
- **Shriram Vasudevan**  
Amrita University
- **Siddhartha Jonnalagadda**  
Mayo Clinic
- **Sim-Hui Tee**  
Multimedia University
- **Simon Ewedafe**  
The University of the West Indies
- **Siniša Opic**  
University of Zagreb, Faculty of Teacher Education
- **Sivakumar Poruran**  
SKP ENGINEERING COLLEGE
- **Slim BEN SAOUD**  
National Institute of Applied Sciences and  
Technology
- **Sofien Mhatli**
- **sofyan Hayajneh**
- **Sohail Jabbar**  
Bahria University
- **Sri Devi Ravana**  
University of Malaya
- **Sudarson Jena**  
GITAM University, Hyderabad
- **Suhail Sami Owais Owais**
- **Suhas J Manangi**  
Microsoft
- **SUKUMAR SENTHILKUMAR**  
Universiti Sains Malaysia
- **Süleyman Eken**  
Kocaeli University
- **Sumazly Sulaiman**  
Institute of Space Science (ANGKASA), Universiti  
Kebangsaan Malaysia

- **Sumit Goyal**  
National Dairy Research Institute
- **Supareerk Janjarasjitt**  
Ubon Ratchathani University
- **Suresh Sankaranarayanan**  
Institut Teknologi Brunei
- **Susarla Sastry**  
JNTUK, Kakinada
- **Suseendran G**  
Vels University, Chennai
- **Suxing Liu**  
Arkansas State University
- **Syed Ali**  
SMI University Karachi Pakistan
- **T C.Manjunath**  
HKBK College of Engg
- **T V Narayana rao Rao**  
SNIST
- **T. V. Prasad**  
Lingaya's University
- **Taiwo Ayodele**  
Infonetmedia/University of Portsmouth
- **Talal Bonny**  
Department of Electrical and Computer Engineering, Sharjah University, UAE
- **Tamara Zhukabayeva**
- **Tarek Gharib**  
Ain Shams University
- **thabet slimani**  
College of Computer Science and Information Technology
- **Totok Biyanto**  
Engineering Physics, ITS Surabaya
- **Touati Youcef**  
Computer sce Lab LIASD - University of Paris 8
- **Tran Sang**  
IT Faculty - Vinh University - Vietnam
- **Tsvetanka Georgieva-Trifonova**  
University of Veliko Tarnovo
- **Uchechukwu Awada**  
Dalian University of Technology
- **Udai Pratap Rao**
- **Urmila Shrawankar**  
GHRCE, Nagpur, India
- **Vaka MOHAN**  
TRR COLLEGE OF ENGINEERING
- **VENKATESH JAGANATHAN**
- ANNA UNIVERSITY
- **Vinayak Bairagi**  
AISSMS Institute of Information Technology, Pune
- **Vishnu Mishra**  
SVNIT, Surat
- **Vitus Lam**  
The University of Hong Kong
- **VUDA SREENIVASARAO**  
PROFESSOR AND DEAN, St.Mary's Integrated Campus, Hyderabad
- **Wali Mashwani**  
Kohat University of Science & Technology (KUST)
- **Wei Wei**  
Xi'an Univ. of Tech.
- **Wenbin Chen**  
360Fly
- **Xi Zhang**  
illinois Institute of Technology
- **Xiaojing Xiang**  
AT&T Labs
- **Xiaolong Wang**  
University of Delaware
- **Yanping Huang**
- **Yao-Chin Wang**
- **Yasser Albagory**  
College of Computers and Information Technology, Taif University, Saudi Arabia
- **Yasser Alginahi**
- **Yi Fei Wang**  
The University of British Columbia
- **Yihong Yuan**  
University of California Santa Barbara
- **Yilun Shang**  
Tongji University
- **Yu Qi**  
Mesh Capital LLC
- **Zacchaeus Omogbadegun**  
Covenant University
- **Zairi Rizman**  
Universiti Teknologi MARA
- **Zarul Zaaba**  
Universiti Sains Malaysia
- **Zenzo Ncube**  
North West University
- **Zhao Zhang**  
Deptment of EE, City University of Hong Kong
- **Zhihan Lv**

Chinese Academy of Science

- **Zhixin Chen**  
ILX Lightwave Corporation
- **Ziyue Xu**  
National Institutes of Health, Bethesda, MD

- **Zlatko Stacic**  
University of Zagreb, Faculty of Organization and  
Informatics Varazdin
- **Zuraini Ismail**  
Universiti Teknologi Malaysia

# CONTENTS

Paper 1: Novel Methods for Resolving False Positives during the Detection of Fraudulent Activities on Stock Market Financial Discussion Boards

*Authors: Pei Shyuan Lee, Majdi Owda, Keeley Crockett*

PAGE 1 – 10

Paper 2: Inferring of Cognitive Skill Zones in Concept Space of Knowledge Assessment

*Authors: Rania Aboalela, Javed Khan*

PAGE 11 – 17

Paper 3: Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine

*Authors: Apapan Pumsirirat, Liu Yan*

PAGE 18 – 25

Paper 4: Voice Detection in Traditionnal Tunisian Music using Audio Features and Supervised Learning Algorithms

*Authors: Wissem Ziadi, Hamid Amiri*

PAGE 26 – 31

Paper 5: Predicting Fork Visibility Performance on Programming Language Interoperability in Open Source Projects

*Authors: Bee Bee Chua*

PAGE 32 – 37

Paper 6: Cadastral and Tea Production Management System with Wireless Sensor Network, GIS based System and IoT Technology

*Authors: Kohei Arai*

PAGE 38 – 44

Paper 7: LOD Explorer: Presenting the Web of Data

*Authors: Karwan Jacksi, Subhi R. M. Zeebaree, Nazife Dimililer*

PAGE 45 – 51

Paper 8: Agent-Based System for Efficient kNN Query Processing with Comprehensive Privacy Protection

*Authors: Mohamad Shady Alrahhah, Maher Khemakhem, Kamal Jambi*

PAGE 52 – 66

Paper 9: A Multiclass Deep Convolutional Neural Network Classifier for Detection of Common Rice Plant Anomalies

*Authors: Ronnel R. Atole, Daechul Park*

PAGE 67 – 70

Paper 10: Improve Mobile Agent Performance by using Knowledge-Based Content

*Authors: Tarig Mohamed Ahmed*

PAGE 71 – 78

Paper 11: Kit-Build Concept Map with Confidence Tagging in Practical Uses for Assessing the Understanding of Learners

*Authors: Jaruwat Pailai, Warunya Wunnasri, Yusuke Hayashi, Tsukasa Hirashima*

PAGE 79 – 91

Paper 12: Method for Detection of Foreign Matters Contained in Dried Nori (Seaweed) based on Optical Property

Authors: Kohei Arai

PAGE 92 – 98

Paper 13: On P300 Detection using Scalar Products

Authors: Monica Fira, Liviu Goras, Anca Lazar

PAGE 99 – 104

Paper 14: Implicit and Explicit Knowledge Mining of Crowdsourced Communities: Architectural and Technology Verdicts

Authors: Husnain Mushtaq, Babur Hayat Malik, Syed Azkar Shah, Umair Bin Siddique, Muhammad Shahzad, Imran Siddique

PAGE 105 – 111

Paper 15: Web-Based COOP Training System to Enhance the Quality, Accuracy and Usability Access

Authors: Amr Jadi, Eesa A. Alsolami

PAGE 112 – 118

Paper 16: Standard Intensity Deviation Approach based Clipped Sub Image Histogram Equalization Algorithm for Image Enhancement

Authors: Sandeepa K S , Basavaraj N Jagadale, J S Bhat

PAGE 119 – 124

Paper 17: Quality Ranking Algorithms for Knowledge Objects in Knowledge Management Systems

Authors: Amal Al-Rasheed, Jawad Berri

PAGE 125 – 134

Paper 18: The Effect of Music on Shoppers' Shopping Behaviour in Virtual Reality Retail Stores: Mediation Analysis

Authors: Aasim Munir Dad, Andrew Kear, Asma Abdul Rehman, Barry J. Davies

PAGE 135 – 145

Paper 19: A Web Service Composition Framework based on Functional Weight to Reach Maximum QoS

Authors: M.Y. Mohamed Yacoab, Abdalla AlAmeen, M. Mohemmed Sha

PAGE 146 – 150

Paper 20 Encrypted Fingerprint into VoIP Systems using Cryptographic Key Generated by Minutiae Points

Authors: Mohammad Fawaz Anagreh, Anwer Mustafa Hilal, Tarig Mohamed Ahmed

PAGE 151 – 154

Paper 21: General Characteristics and Common Practices for ICT Projects: Evaluation Perspective

Authors: Abdullah Saad AL-Malaise AL-Ghamdi, Farrukh Saleem

PAGE 155 – 163

Paper 22: Identification of Toddlers' Nutritional Status using Data Mining Approach

Authors: Sri Winiarti, Herman Yuliansyah, Aprial Andi Purnama

PAGE 164 – 169

Paper 23: A Comparative Study on Steganography Digital Images: A Case Study of Scalable Vector Graphics (SVG) and Portable Network Graphics (PNG) Images Formats

Authors: Abdulgader Almutairi

PAGE 170 – 175

**Paper 24: Detection of Violations in Credit Cards of Banks and Financial Institutions based on Artificial Neural Network and Metaheuristic Optimization Algorithm**

*Authors: Zarrin Monirzadeh, Mehdi Habibzadeh, Nima Farajian*

**PAGE 176 – 182**

**Paper 25: Data Exfiltration from Air-Gapped Computers based on ARM CPU**

*Authors: Kenta Yamamoto, Miyuki Hirose, Taiichi Saito*

**PAGE 183 – 190**

**Paper 26: A Seamless Network Database Migration Tool for Insititutions in Zambia**

*Authors: Mutale Kasonde, Simon Tembo*

**PAGE 191 – 199**

**Paper 27: Software Engineering: Challenges and their Solution in Mobile App Development**

*Authors: Naila Kousar, Muhammad Sheraz Arshad Malik, Aramghan Sarwar, Burhan Mohy-ud-din, Ayesha Shahid*

**PAGE 200 – 203**

**Paper 28: Analysis of Valuable Clustering Techniques for Deep Web Access and Navigation**

*Authors: Qurat-ul-ain, Asma Sajid, Uzma Jamil*

**PAGE 204 – 211**

**Paper 29: Pre-Trained Convolutional Neural Network for Classification of Tanning Leather Image**

*Authors: Sri Winiarti, Adhi Prahara, Murinto, Dewi Pramudi Ismi*

**PAGE 212 – 217**

**Paper 30: Iteration Method for Simultaneous Estimation of Vertical Profiles of Air Temperature and Water Vapor with AQUA/AIRS Data**

*Authors: Kohei Arai*

**PAGE 218 – 223**

**Paper 31: A Robust System for Noisy Image Classification Combining Denoising Autoencoder and Convolutional Neural Network**

*Authors: Sudipta Singha Roy, Sk. Imran Hossain, M. A. H. Akhand, Kazuyuki Murase*

**PAGE 224 – 235**

**Paper 32: A New Healthcare Context Information: The Social Context**

*Authors: Isra'a Ahmed Zriqat, Ahmad Mousa Alftamimi*

**PAGE 236 – 239**

**Paper 33: Brainwaves for User Verification using Two Separate Sets of Features based on DCT and Wavelet**

*Authors: Loay E. George, Hend A. Hadi*

**PAGE 240 – 246**

**Paper 34: FARM: Fuzzy Action Rule Mining**

*Authors: Zahra Entekhabi, Pirooz Shamsinejadbabki*

**PAGE 247 – 252**

**Paper 35: A Secured Interoperable Data Exchange Model**

*Authors: A. Bahaa, A. Sayed, L. Elfangary*

**PAGE 253 – 260**

**Paper 36: Iterative Removing Salt and Pepper Noise based on Neighbourhood Information**

*Authors: Liu Chun, Sun Bishen, Liu Shaohui, Tan Kun, Ma Yingrui*

**PAGE 261 – 265**

**Paper 37: Attendance and Information System using RFID and Web-Based Application for Academic Sector**

*Authors: Hasanein D. Rjeib, Nabeel Salih Ali, Ali Al Farawn, Basheer Al-Sadawi, Haider Alsharqi*

**PAGE 266 – 274**

**Paper 38: Social Network Link Prediction using Semantics Deep Learning**

*Authors: Maria Ijaz, Javed Ferzund, Muhammad Asif Suryani, Anam Sardar*

**PAGE 275 – 283**

**Paper 39: Matrix Clustering based Migration of System Application to Microservices Architecture**

*Authors: Shahbaz Ahmed Khan Ghayyur, Abdul Razzaq, Saeed Ullah, Salman Ahmed*

**PAGE 284 – 296**

**Paper 40: DoS/DDoS Detection for E-Healthcare in Internet of Things**

*Authors: Iffikhar ul Sami, Maaz Bin Ahmad, Muhammad Asif, Rafi Ullah*

**PAGE 297 – 300**

**Paper 41: Truncated Patch Antenna on Jute Textile for Wireless Power Transmission at 2.45 GHz**

*Authors: Kais Zeouga, Loffi Osman, Ali Gharsallah, Bhaskar Gupta*

**PAGE 301 – 305**

**Paper 42: A Hybrid Approach for Feature Subset Selection using Ant Colony Optimization and Multi-Classifer Ensemble**

*Authors: Anam Naseer, Waseem Shahzad, Arslan Ellahi*

**PAGE 306 – 313**

**Paper 43: Efficient Smart Emergency Response System for Fire Hazards using IoT**

*Authors: Lakshmana Phaneendra Maguluri, Tumma Srinivasarao, Maganti Syamala, R. Ragupathy, N.J. Nalini*

**PAGE 314 – 320**

**Paper 44: An Information Theoretic Analysis of Random Number Generator based on Cellular Automaton**

*Authors: Amirahmad Nayyeri, Gholamhossein Dastghaibifard*

**PAGE 321 – 329**

**Paper 45: Requirement Elicitation Techniques for Open Source Systems: A Review**

*Authors: Hafiza Maria Kiran, Zulfiqar Ali*

**PAGE 330 – 334**

**Paper 46: A Parallel Community Detection Algorithm for Big Social Networks**

*Authors: Yathrib AlQahtani, Mourad Ykhlef*

**PAGE 335 – 340**

**Paper 47: Comparison between Two Adaptive Controllers Applied to Greenhouse Climate Monitoring**

*Authors: Mohamed Essahafi, Mustapha Ait Lafkih*

**PAGE 341 – 346**



**Paper 48: A Predictive Model for Solar Photovoltaic Power using the Levenberg-Marquardt and Bayesian Regularization Algorithms and Real-Time Weather Data**

*Authors: Mohammad H. Alomari, Ola Younis, Sofyan M. A. Hayajneh*

**PAGE 347 – 353**

**Paper 49: Improving Energy Conservation in Wireless Sensor Network Using Energy Harvesting System**

*Authors: Abdul Rashid, Faheem Khan, Toor Gul, Fakhr-e-Alam, Shujaat Ali, Samiullah Khan, Fahim Khan Khalil*

**PAGE 354 – 361**

**Paper 50: A New Motion Planning Framework based on the Quantized LQR Method for Autonomous Robots**

*Authors: Onur Sencan, Hakan Temeltas*

**PAGE 362 – 374**

**Paper 51: Bearing Fault Classification based on the Adaptive Orthogonal Transform Method**

*Authors: Mohamed Azergui, Abdenbi Abenaou, Hassane Bouzahir*

**PAGE 375 – 380**

**Paper 52: Improving Security of the Telemedicine System for the Rural People of Bangladesh**

*Authors: Toufik Ahmed Emon, Uzzal Kumar Prodhan, Mohammad Zahidur Rahman, Israt Jahan*

**PAGE 381 – 390**

**Paper 53: Nonlinear Model Predictive Control for Ph Neutralization Process**

*Authors: Hajer Degachi, Wassila Chagra, Moufida Ksouri*

**PAGE 391 – 398**

**Paper 54: An Efficient Participant's Selection Algorithm for Crowdsensing**

*Authors: Tariq Ali, Umar Draz, Sana Yasin, Javeria Noureen, Ahmad shaf, Munwar Ali*

**PAGE 399 – 404**

**Paper 55: An Energy-Efficient User-Centric Approach for High-Capacity 5G Heterogeneous Cellular Networks**

*Authors: Abdulziz M. Ghaleb, Ali Mohammed Mansoor, Rodina Ahmad*

**PAGE 405 – 411**

**Paper 56: Lifetime Maximization on Scalable Stable Election Protocol for Large Scale Traffic Engineering**

*Authors: Muhammad Asad, Arsalan Ali Shaikh, Soomro Pir Dino, Muhammad Aslam, Yao Nianmin*

**PAGE 412 – 418**

**Paper 57: Comparative Analysis of Raw Images and Meta Feature based Urdu OCR using CNN and LSTM**

*Authors: Asma Naseer, Kashif Zafar*

**PAGE 419 – 424**

**Paper 58: An Empirical Evaluation of Error Correction Methods and Tools for Next Generation Sequencing Data**

*Authors: Atif Mehmood, Javed Ferzund, Muhammad Usman Ali, Abbas Rehman, Shahzad Ahmed, Imran Ahmad*

**PAGE 425 – 431**

**Paper 59: Combinatorial Double Auction Winner Determination in Cloud Computing using Hybrid Genetic and Simulated Annealing Algorithm**

*Authors: Ali Sadigh Yengi Kand, Ali Asghar Pourhaji Kazem*

**PAGE 432 – 436**

**Paper 60: QoS-based Cloud Manufacturing Service Composition using Ant Colony Optimization Algorithm**

*Authors: Elsoon Neshati, Ali Asghar Pourhaji Kazem*

**PAGE 437 – 440**

**Paper 61: Envisioning Internet of Things using Fog Computing**

*Authors: Urooj Yousuf Khan, Tariq Rahim Soomro*

**PAGE 441 – 448**

**Paper 62: A Group Decision-Making Method for Selecting Cloud Computing Service Model**

*Authors: Ibrahim M. Al-Jabri, Mustafa I. Eid, M. Sadiq Sohail*

**PAGE 449 – 456**

**Paper 63: Prediction of Stroke using Data Mining Classification**

*Authors: Ohoud Almadani, Riyadh Alshammari*

**PAGE 457 – 460**

**Paper 64: Hardware Implementation for the Echo Canceller System based Subband Technique using TMS320C6713 DSP Kit**

*Authors: Mahmud. A. Al Zubaidy, Sura Z. Thanoon*

**PAGE 461 – 467**

**Paper 65: SME Cloud Adoption in Botswana: Its Challenges and Successes**

*Authors: Malebogo Khanda, Srinath Doss*

**PAGE 468 – 478**

**Paper 66: Survey Paper for Software Project Team, Staffing, Scheduling and Budgeting Problem**

*Authors: Rizwan Akram, Salman Ihsan, Shaista Zafar, Babar Hayat*

**PAGE 479 – 484**

**Paper 67: Real-Time Experimentation and Analysis of Wifi Spectrum Utilization in Microwave Oven Noisy Environment**

*Authors: Yakubu S. Baguda*

**PAGE 485 – 491**

**Paper 68: Deep Learning Technology for Predicting Solar Flares from (Geostationary Operational Environmental Satellite) Data**

*Authors: Tarek A M Hamad Nagem, Rami Qahwaji, Stan Ipson, Zhiguang Wang, Alaa S. Al-Waisy*

**PAGE 492 – 498**

**Paper 69: Cyber-Security Incidents: A Review Cases in Cyber-Physical Systems**

*Authors: Mohammed Nasser Al-Mhiqani, Rabiah Ahmad, Warusia Yassin, Aslinda Hassan, Zaheera Zainal Abidin, Nabeel Salih Ali, Karrar Hameed Abdulkareem*

**PAGE 499 – 508**

**Paper 70: Measuring Quality of E-Learning and Desaire2Learn in the College of Science and Humanities at Alghat, Majmaah University**

*Authors: Abdelmoneim Ali Mohamed, Faisal Mohammed Nafie*

**PAGE 509 – 512**

**Paper 71: TSAN: Backbone Network Architecture for Smart Grid of P.R China**

*Authors: Raheel Ahmed Memon, Jianping Li, Anwar Ahmed Memon, Junaid Ahmed, Muhammad Irshad Nazeer, Muhammad Ismail*

**PAGE 513 – 520**

**Paper 72: Data Synchronization Model for Heterogeneous Mobile Databases and Server-side Database**

*Authors: Abdullahi Abubakar Imam, Shuib Basri, Rohiza Ahmad, Abdul Rehman Gilal*

**PAGE 521 – 531**

**Paper 73: Data Mining Techniques to Construct a Model: Cardiac Diseases**

*Authors: Noreen Akhtar, Muhammad Ramzan Talib, Nosheen Kanwal*

**PAGE 532 – 536**

**Paper 74: Fuzzy Logic based Approach for VoIP Quality Maintaining**

*Authors: Mohamed E. A. Ebrahim, Hesham A. Hefny*

**PAGE 537 – 542**

**Paper 75: Conceptual Modeling of a Procurement Process**

*Authors: Sabah Al-Fedaghi, Mona Al-Otaibi*

**PAGE 543 – 553**

**Paper 76: Average Link Stability with Energy-Aware Routing Protocol for MANETs**

*Authors: Sofian Hamad, Salem Belhaj, Muhana M. Muslam*

**PAGE 554 – 562**

**Paper 77: Agent based Architecture for Modeling and Analysis of Self Adaptive Systems using Formal Methods**

*Authors: Natash Ali Mian, Farooq Ahmad*

**PAGE 563 – 567**

**Paper 78: Reverse Engineering State and Strategy Design Patterns using Static Code Analysis**

*Authors: Khaled Abdelsalam Mohamed, Amr Kamel*

**PAGE 568 – 576**

**Paper 79: OpenMP Implementation in the Characterization of an Urban Growth Model Cellular Automaton**

*Authors: Alvaro Peraza Garzón, René Rodríguez Zamora, Wenseslao Plata Rocha*

**PAGE 577 – 582**

# Novel Methods for Resolving False Positives during the Detection of Fraudulent Activities on Stock Market Financial Discussion Boards

Pei Shyuan Lee

School of Computing, Mathematics  
& Digital Technology  
The Manchester Metropolitan  
University, Chester Street,  
Manchester, M1 5GD, UK

Majdi Owda

School of Computing, Mathematics  
& Digital Technology  
The Manchester Metropolitan  
University, Chester Street,  
Manchester, M1 5GD, UK

Keeley Crockett

School of Computing, Mathematics  
& Digital Technology  
The Manchester Metropolitan  
University, Chester Street,  
Manchester, M1 5GD, UK

**Abstract**—Financial discussion boards (FDBs) have been widely used for a variety of financial knowledge exchange activities through the posting of comments. Popular public FDBs are prone to being used as a medium to spread false financial information due to larger audience groups. Although online forums are usually integrated with anti-spam tools, such as Akismet, moderation of posted content heavily relies on manual tasks. Unfortunately, the daily comments volume received on popular FDBs realistically prevents human moderators to watch closely and moderate possibly fraudulent content, not to mention moderators are not usually assigned with such task. Due to the absence of useful tools, it is extremely time consuming and expensive to manually read and determine whether each comment is potentially fraudulent. This paper presents novel forward and backward analysis methodologies implemented in an Information Extraction (IE) prototype system named FDBs Miner (FDBM). The methodologies aim to detect potentially illegal Pump and Dump comments on FDBs with the integration of per-minute share prices in the detection process. This can possibly reduce false positives during the detection as it categorises the potentially illegal comments into different risk levels for investigation purposes. The proposed system extracts company's ticker symbols (i.e. unique symbol that represents and identifies each listed company on stock market), comments and share prices from FDBs based in the UK. The forward analysis methodology flags the potentially Pump and Dump comments using a predefined keywords template and labels the flagged comments with price hike thresholds. Subsequently, the backward analysis methodology employs a moving average technique to determine price abnormalities and backward analyse the flagged comments. The first detection stage in forward analysis found 9.82% of potentially illegal comments. It is unrealistic and unaffordable for human moderators or financial surveillance authorities to read these comments on a daily basis. Hence, by integrating share prices to perform backward analysis can categorise the flagged comments into different risk levels. It helps relevant authorities to prioritise and investigate into the higher risk flagged comments, which could potentially indicate a real Pump and Dump crime happening on FDBs when the system is being used in real time.

**Keywords**—Financial discussion boards; financial crimes; pump and dump; text mining; information extraction

## I. INTRODUCTION

The internet has become the number one source for information. Unsurprisingly, this includes financial advice and investor sentiments. There are many online forums where likeminded people can hold conversations in the form of posted messages. Financial Discussion Boards (FDBs), also known as Financial Message Boards or Financial Forums allows investors to exchange knowledge, information, experience and opinions about the investment opportunities. There are a few popular share price based FDBs based in the UK which specifically allows investors to discuss share prices. These FDBs include the London South East<sup>1</sup>, Interactive Investor (III)<sup>2</sup> and ADVFN<sup>3</sup>.

Normally, forum content is moderated by human moderators when it is discovered or reported for breaching forum rules such as racism, sexism, hatred, foul language, third party advertisements and so on. Although online forums seem to be a useful source of information, not all information shared on the forums is accurate or truthful. Even anti-spam plugins such as Akismet<sup>4</sup> can only prevent spammers from registering or posting generic spam messages. There is little to no measurements taken by forum moderators or financial surveillance authorities to monitor and detect potential crimes on the FDBs, such as comments indicative of Pump and Dump (P&D).

P&D can happen if an organised group of false investors decided to attack shares by buying and selling a specific share in a scheduled time frame and giving the market false statements about the share throughout the process. Textual comments such as “This is the right time let's start pumping this share” can reveal a hidden potential illegal activity of P&D on these FDBs. Novice investors can be easily deceived and make huge losses during the “dump” while the fraudsters take huge profits. Without a tool, manual monitoring and detection of potentially illegal activities on popular and active FDBs can

<sup>1</sup> <http://www.lse.co.uk>

<sup>2</sup> <http://www.iii.co.uk>

<sup>3</sup> <http://uk.advfn.com>

<sup>4</sup> <https://akismet.com>

cost significant time and money, which is impracticable in the long run.

There has been research conducted around the area of share price based FDBs associated with P&D financial crimes [1]-[6]. Research from recent years highlighted that the comments on FDBs were found manipulative and positively related to the market returns, volatility and trading volumes [7]-[11]. However, there has been very little attempt [5], [6] made to build tools for monitoring and detection of potential financial crimes on share price based FDBs. Furthermore, other than the initial work presented in [12], none of the other existing research take share prices into account when designing a financial surveillance tool for detection of potentially illegal FDB comments.

FDBs contain semantically understandable artefacts (i.e. FDBs' artefacts that can be processed by computers) such as stock ticker symbols, date, time, prices, comment author usernames and comments. Information Extraction (IE) is defined as the process of extracting information automatically into a structured data format from an unstructured or semi-structured data source [13]. Therefore, IE techniques are used in this research to extract and analyse these data. IE has been used in other areas such as accounting [14] and search engine [15]. However, other than the initial work described in [6] and [16], there is very little usage of IE techniques in FDBs' financial crimes related research.

Two novel methodologies, i.e. forward analysis and backward analysis, are introduced in this paper are implemented in a prototype system named FDBs Miner (FDBM). The methodologies are used to detect potential P&D crimes on FDBs by flagging potentially illegal comments and reducing false positives (i.e. errors present in evaluation processes or scientific tests that are mistakenly found) during the detection process. FDBM could significantly support financial surveillance authorities to regulate by enabling real-time monitoring and alerting based on fraudulent risk levels.

In the forward analysis methodology, all the potentially illegal comments will first be highlighted and flagged. This is done by analysing the comments against the predefined P&D IE keywords template. Next, the method matches and appends the price figure to the flagged comments which share the same or closest date and time based on same ticker symbol. Subsequently, the forward analyser takes each flagged comment's price as a base price and calculates  $\pm 2$  days' worth of prices to check if there is any price hike 5%, 10% and 15% more than the base price. Finally, it appends the price hike threshold labels to these flagged comments. By doing so, a relevant authority can pick the comments belonging to any threshold depending on the severity for investigations. Although the forward analysis in this research has drastically reduced the number of comments needed to be read by relevant authorities, the amount of categorised flagged comments could still be somewhat large to read on a daily basis. Thus, a backward analysis methodology is designed to overcome this issue.

In the backward analysis methodology, a simple moving average method is used to calculate and highlight the price

hikes. Any price hikes that hit certain price hike thresholds will be matched backwards to the flagged comments found in the forward analysis stage. Such matches are done so that the already flagged comments can be further classified to reduce false positives and allow investigators to quickly examine the higher and highest risked flagged comments before everything else.

Section II describes some examples of FDBs related financial crimes and reviews the background and usage of Information Extraction (IE) and Text Mining. Section III presents the architecture overview of the FDBs Miner (FDBM) prototype system and an overview of the FDBs dataset (FDB-DS). This followed by Section IV and V introducing the two novel methodologies (i.e. forward analysis and backward analysis) respectively and discussing the findings. Lastly, Section VI concludes the research and proposes some future work.

## II. BACKGROUND

This section first provides a few related and significant examples of financial crimes on share price based FDBs, followed by the literature review related to IE and text mining which are the techniques used in this research for locating meaningful information, and collection and formation of datasets respectively. Lastly, Pump and Dump (P&D) and FDBs related literature review will also be presented.

### A. Financial Crimes on Share Price based FDBs

Generally, there are many P&D financial crimes which are actively investigated and dealt with by the Security Exchange Commission (SEC) for many years. However, P&D crimes on FDBs are loosely monitored by FDB moderators and relevant authorities. There were several popular FDB related P&D financial crimes in the early years, which are highlighted as follows:

- 15-year-old Jonathan Lebed was the first minor to involve in a stock market fraud in 2000 [3]. Lebed earned a total revenue of US\$800,000 by pumping the share price through Yahoo! Finance Message Board over half a year and charged by Security Exchange Commission (SEC) [3], [4].
- In 2000, two people were being charged for pumping the price of a share by 10,000% by posting on the Raging Bull message board and then dumped millions of shares which the profit made were at least US\$5 million [3].
- In addition, in 2009, eight participants were charged by the Security Exchange Commission (SEC)<sup>5</sup> for being involved in penny stock (i.e. stock prices that are less than a dollar) manipulation throughout the year of 2006 and 2007. These wrongdoers met each other through a popular penny stock message board.

Based on the above FDBs related P&D financial crimes, there is certainly a need to create methods and tools for detection of potentially illegal FDB comments in real time.

<sup>5</sup> <http://www.sec.gov/litigation/litreleases/2009/lr21053.htm>

This is instead of investigating the crimes after being committed – which is probably too late as the harm has been done.

### B. Information Extraction and Text Mining

This research makes use of Information Extraction (IE) and Text Mining. IE is defined [17] as the process of extracting information automatically into a structured data format from an unstructured or semi-structured data sources. It was suggested [18] that there is a need for systems that extract information automatically from text data. IE is not Information Retrieval (IR) [19]. The difference between IE and IR is that IE extracts information that fits predefined templates or databases and then presents the information to the users, whereas IR finds data and presents the information to the users. IE systems are knowledge-intensive as these systems extract only snippets of information that will fit predefined templates (fixed format) which represent useful and relevant information about the domain then display to the user.

IE is divided into two fundamental classes i.e. the Knowledge Engineering (KE) approach and the automatic training approach. The KE approach is also called as the rule-based approach since it requires rules to be developed by the human expertise. Rule-based approach is usually ignored in the research community, but it is mostly favourable in the commercial market even by the large vendors such as IBM (for text analysis systems) and Microsoft (enterprise search platform) [20]. Rule-based IE systems are easy to maintain and comprehend as well as errors being traced and fixed easily. On the other hand, although the automatic training approach, also known as machine learning approach, requires less manual efforts, the approach requires pre-labelled data and retraining for adaptation [20]. This paper focuses on IE implementation since it is designed to support the financial market surveillance authorities.

Text mining was described [21] as the process to extract useful information from unformatted textual data or natural language text into a form of meaningful knowledge for processing. According to [22], the internet users have been seeking and sharing opinions and information using the Internet more easily than ever and this raises concerns about the credibility of the information sources. This means the likelihood of getting deceptive information is also significant. Similarly, on popular share price based FDBs that receive a significant amount of comments in each day, novice investors who seek investment advice could also be deceived easily. Also, a text mining based study was conducted [23] on a Twitter dataset and its relationship to be able to predict stock prices. In addition, stock price trends were also being successfully forecasted via press releases using text mining techniques [24].

In this paper, text mining is used alongside IE rule-based technique to extract and analyse FDB artefacts such as comments, prices and stock ticker symbols.

### C. Pump and Dump and Share Price Based FDBs

Traditionally, Pump and Dump (P&D) happens mostly through word of mouth. But with the existence of the Internet,

it becomes so common that the fraudsters commit crimes through various channels such as emails, discussion boards and social media.

The use of spam emails is one of the older tactics. Regulators like Securities and Exchange Commission (SEC) has been actively taking actions against P&D spam campaign fraudsters. Email spam filters are also constantly being improved by Internet services such as Google and Symantec. In research conducted in [25], a total of 1,299 suspicious stock recommendation emails was obtained. It involved 221 stocks recommended in 252 advertising campaigns. An event study and a sentiment analysis have been conducted on whether P&D involving the internet is still an issue in today's world. Unsurprisingly, the research empirically proved that the internet still plays a major role in enhancing this type financial crime. Due to the limitations in spam emails, newer tactics such as social media and discussion boards were adopted mainly because these channels allow more freedom of speech. Other researchers [7]-[11] have found the relation between FDB comments and market performance. FDB comments can be manipulative and affect the share prices.

In [5], the authors introduced a novel classification technique for a classifier training in order to automate moderation tasks on online discussion sites (ODSs). A partially labelled corpus is used for the training purpose and then attempt to moderate the inappropriate content on ODSs using the technique. The authors implemented and tested the technique on a corpus of comments posted on a popular Australian FDB named HotCopper<sup>6</sup>. The results indicated that the classification technique is helpful and can be used to decrease the number of comments that need to be moderated by human moderators. However, this system is not yet a fully automated moderation system due to the use of partially labelled corpus. According to the authors, the misclassification errors remain too significant. Besides, the research takes only comments into account and no prices involved during the classification of comments.

A system named Financial Discussions Detection System (FDDS), an initial work to this research, was proposed by the authors in [6] to flag potentially illegal comments made on FDBs. The system allows users to create and modify predefined templates (i.e. lists of potentially illegal keywords that commenters may or frequently use on FDBs), download comments from FDBs and matches the downloaded comments against the potentially illegal keywords created in earlier steps. By looking only at the comments during the detection processes appear to be insufficient in terms of accuracy. Thus, this paper introduces the novel methodologies in attempt to reduce false positives by integrating share prices in the detection process.

The authors in [11] examined whether the messages posted on the largest stock message board in Australia, HotCopper, has an impact on the Australian Stock Exchange (ASX) market. Results show that the FDB messages have impacts on the small capitalisation stocks but not affecting the large stocks.

---

<sup>6</sup> <https://hotcopper.com.au>

In [26], the authors introduced a software prototype (FMS-DSS) to support decision making in financial market surveillance. FMS-DSS consists of three components i.e. data, models and user interface. The system collects both unstructured and structured data of the selected listed companies. The models take into account of attributes such as market segment, market capitalisation, trading volume, age of company and so on. Subsequently, attribute scales ranging from very low to very high were defined by the regulatory authority members. The scales were then used for aggregation to determine whether there is suspicious activity happening.

In the research presented in this paper there is an attempt to resolve what was missing in existing research. Share prices are taken into account when flagging potentially illegal comments, accompanied by two key novel built-in methodologies (namely, the forward analysis and the backward analysis) for resolving false positives during the comments flagging process.

### III. ARCHITECTURE OVERVIEW

This section presents the FDBM architecture which consists of several key components. These key components are the data crawler, data transformer, FDB dataset (FDB-DS), IE keyword template, forward analyser and the backward analyser (Fig. 1). Fundamentally, FDBM collects data, transform unstructured data into structured data format and analyse the data using both forward and backward analysers. The forward analyser and backward analyser components are used within the novel methodologies introduced in this paper attempt to resolve false positives during the process of detection of potentially illegal comments.

#### A. Overview

Fig. 1 provides an overview of the FDBM architecture of the prototype system.

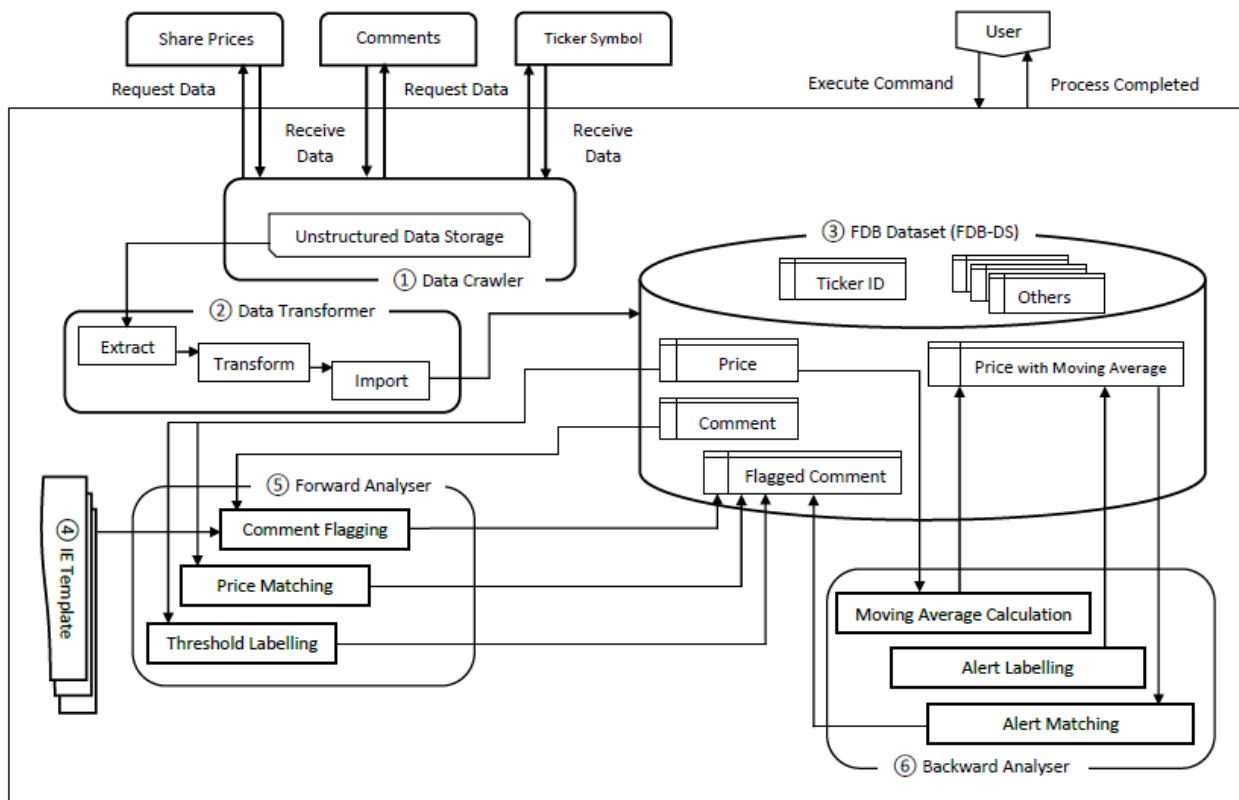


Fig. 1. Architecture overview diagram.

Each component in the architecture diagram is described as follows:

1) **Data Crawler:** The data crawler is responsible for automatically collecting unstructured data from the three FDBs (i.e. LSE, III and ADVFN) at different time intervals for a period of 12 weeks (from 23<sup>rd</sup> September 2014 to 22<sup>nd</sup> December 2014). These unstructured and semi-structure data consist of 941 ticker symbols that were listed on London Stock Exchange (LSE), FTSE100 and FTSE AIM All-Share, 1-minute bar price figures for all the 941 companies and all the available FDB comments belong to the 941 companies.

FTSE100 index consists of the first hundred companies with the highest market capitalisation listed on LSE, whereas FTSE AIM All-Share consists of all the UK and non-UK companies listed on the Alternative Investment Market (AIM). As an effort for potential future work, director deals data and broker ratings data were also collected. Table I in Section B summarises the total sum of collected data.

2) **Data Transformer:** Once the data collection is done by the data crawler, the data transformer extracts and converts the collected unstructured data in various formats such as HTML, CSV and XML into structured data.

3) **FDB Dataset (FDB-DS):** After the collected data is being processed by the data transformer, the structured data such as price figures, comments, comment author usernames, date and time of comments and prices are stored in the FDB-DS accordingly. For example, the ticker symbols are parsed into `ticker` table, price data are parsed into `price` table and comment data are parsed into `comment` table. The FDB-DS is also responsible to store additional data produced from research analysis.

4) **IE Templates:** The Pump and Dump IE keyword template has been created and saved locally in the prototype system in a text (TXT) file format. It can be easily modified whenever needed. The IE keyword template consists of a series of keywords and phrases that were thoroughly researched [2], [27]-[29] and has been validated by experts in the relevant field. The IE keyword template will be used by the forward and backward analysers for the comments

flagging process. Section C shows a sample list of the keywords and phrases.

5) **Forward Analyser:** The forward analyser matches the Pump and Dump IE keyword template against the comments in order to flag potentially illegal FDB comments, followed by matching the prices to the flagged comments, calculating and labelling price thresholds. The novel methodology used in this component is further discussed in Section IV.

6) **Backward Analyser:** Backward analyser performs the calculation and labelling of price hikes using a price moving average technique i.e. simple moving average (SMA). SMA is calculated by adding the prices for a specific time period and divide by the number of the time period. This calculation is applied against a total of 29 million price figures which belong to 941 companies. Subsequently, price hike SMA alerts will be matched back towards the initially flagged comments in forward analysis process. This methodology is further elaborated in Section V.

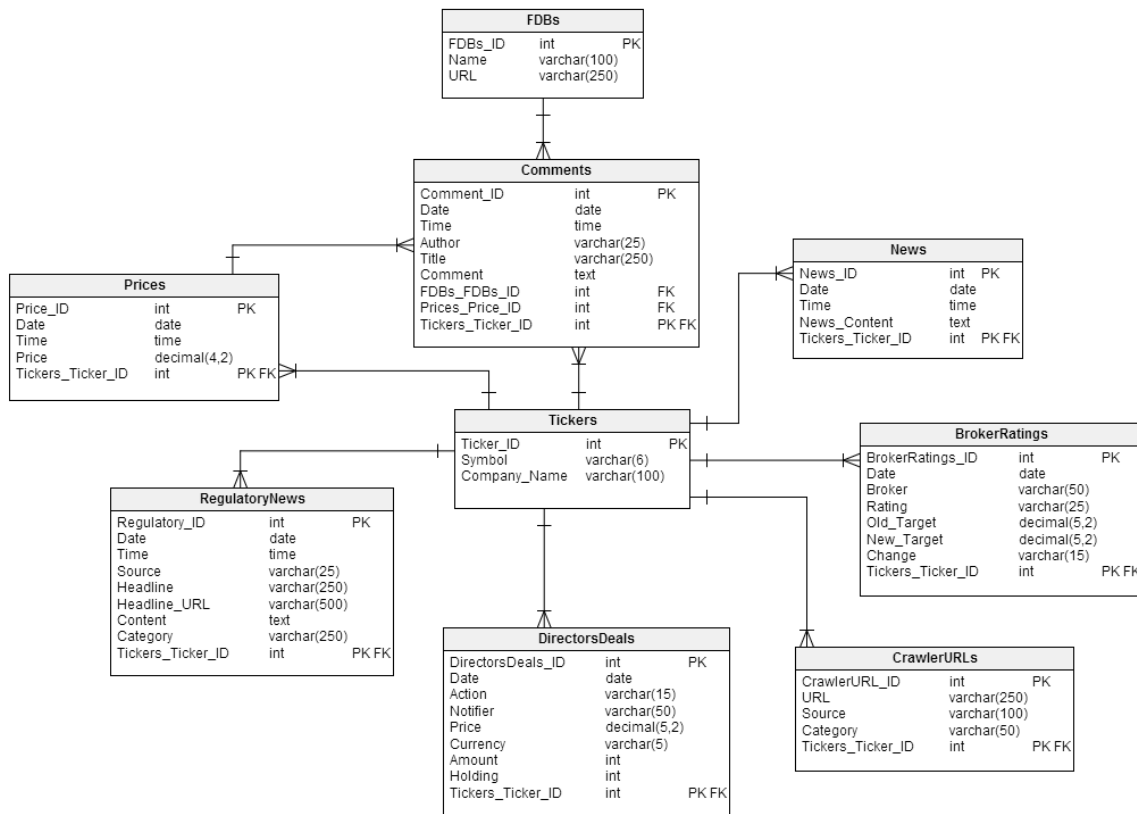


Fig. 2. FDB dataset structure.

**B. Dataset Acquisition**

Table I provides an overview of the FDB dataset (FDB-DS) in this research. These data were collected between 23<sup>rd</sup> September 2014 and 22<sup>nd</sup> December 2014.

As mentioned in Section III, A, these 941 ticker symbols were collected from two of the LSE’s indices, i.e., 100 ticker symbols from FTSE100 and 841 ticker symbols from FTSE

AIM All-Share. The comments, which belong to all these ticker symbols, made within the 12 weeks were collected from both LSE and III. As for prices, these are 12 weeks’ worth of 1-minute bar share prices belong to all the 941 ticker symbols. Director deals and broker ratings related to all the ticker symbols were also collected for potential future work. Fig. 2 depicts the FDB-DS structure.



TABLE. I. TOTAL NUMBER OF ARTEFACT RECORDS (FDB-DS)

Artefacts	Total Number of Records
Ticker Symbols	941
Comments	507,970
Prices	28,980,465
Director Deals	11,456
Broker Ratings	6,469

### C. IE Template

Pump & Dump (P&D) IE keyword template is populated by obtaining the keywords from the P&D comments demonstrated in existing research [6], [27]-[29]. The following is a sample list of the keywords and phrases that were used in this work:

- Pump dump
- Once in a lifetime
- Pump the price
- Keep ramping
- Buy now
- Good future
- Invested so heavily
- It will fly
- Sell now
- This is the chance
- Price will go up
- Buy as quickly as possible
- Get out while you can.

## IV. FORWARD ANALYSIS METHODOLOGY

This section introduces the novel forward analysis methodology. The aim of this methodology is to flag and filter the potentially illegal P&D comments using P&D keyword template with the integration of the share prices in the analysis process. This will categorise the flagged comments into different risk levels and allows relevant authorities to investigate into the flagged comments more realistically in terms of time and efforts.

The forward analysis methodology in this section will test the following hypothesis:

$H_{0a}$ : Pump and Dump activity from FDBs can be filtered using template based IE and their correlation with price movements.

$H_{1a}$ : Pump and Dump activity from FDBs cannot be filtered using template based IE and their correlation with price movements.

As shown in the architecture diagram in Fig. 1, the forward analysis component contains several functions. These functions (i.e. comments flagging, price matching, threshold calculation and threshold labelling) that are part of the forward analysis methodology which will be discussed below.

### A. Methodology

The following describes the steps taken in this methodology to flag potentially illegal comments:

#### 1) Comments Flagging:

a) Firstly, the forward analyser matches all the available keywords and phrases from the Pump and Dump IE keyword template against all the 507,970 comments which were stored in FDB dataset (FDB-DS).

b) The flagged comments which deemed potentially illegal are imported into FDB-DS as a new database table named `flaggedcomment`.

#### 2) Price and Comments Matching:

a) Once `flaggedcomment` has been populated, the forward analyser appends the price to each flagged comment by matching the ticker symbol and the exact or nearest date and time. This step is done to ensure a “base price” is set for each flagged comment. The “base price” will be used for threshold labelling in next step. Due to the extremely large 12 weeks’ worth of price data belongs to 941 companies, the process of setting a “base price” takes up to a week to complete.

#### 3) Comments Threshold Labelling:

a) After having all the “base price” set for each flagged comment in the previous step, the forward analyser labels each flagged comment with thresholds. Due to the large data set, the threshold labelling process takes up to five days to complete all threshold calculations. To determine whether a flagged comment’s base price exceeds any thresholds (i.e. various levels of spikes in prices), the forward analyser calculates all the  $\pm 2$  days’ per-minute prices against the “base price” of each flagged comment.

b) When there is a trigger, a flagged comment will be labelled accordingly. The threshold labelling rules are as follows:

- Flagged comments that have no price figure (due to empty price figures collected from ADVFN) are labelled as “N” (Null).
- If any of the  $\pm 2$  days prices calculated against the “base price” indicates a 5% price hike the comment is labelled as “Y” (Yellow).
- If any of the  $\pm 2$  days prices calculated against the “base price” indicates a 10% price hike the comment is labelled as “A” (Amber).
- If any of the  $\pm 2$  days prices calculated against the “base price” indicates a 15% price hike the comment is labelled as “R” (Red).
- Flagged comments that do not trigger any thresholds are labelled as “C”.

### B. Forward Analysis Methodology Results

By matching the keywords and phrases from P&D IE keyword template against all the 507,970 comments, a total number of 49,858 comments were flagged as potentially illegal comments (as shown in Table II). These flagged comments took up 9.82% of the total comments.

TABLE. II. TOTAL NUMBER OF FLAGGED COMMENTS

Comments	Total	Percentage
Flagged	49,858	9.82%
Non-flagged	458,112	90.18%
<b>Grand Total</b>	<b>507,970</b>	<b>100%</b>

Out of all the 49,858 flagged comments, 3,613 (7.25%) of the flagged comments triggered the “R” 15% price hike threshold, 2,555 (5.12%) flagged comments triggered the “A” 10% price hike threshold and 5,197 (10.42%) flagged comments triggered the “Y” 5% price hike threshold. 37,895 (76.01%) flagged comments labelled as “C” did not trigger any price thresholds. The total number of flagged comments that triggered the thresholds is summarised in Table III and visualised in Fig. 3.

TABLE. III. TOTAL NUMBER OF FLAGGED COMMENTS IN EACH PRICE HIKE THRESHOLD

Threshold	Total	Percentage
C (<5%)	37,895	76.01%
Y (5%)	5,197	10.42%
A (10%)	2,555	5.12%
R (15%)	3,613	7.25%
Null	598	1.2%
<b>Grand Total</b>	<b>49,858</b>	<b>100%</b>

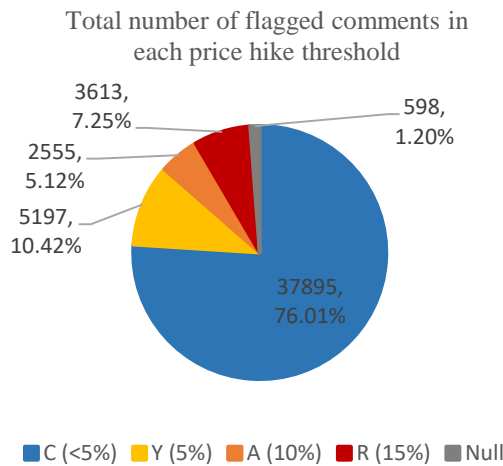


Fig. 3. Total number and percentage of each threshold.

The results show the possibility to filter comments that may be indicative of Pump and Dump activities by using template based IE and the correlation with price movements. For 12 weeks’ worth of 941 companies’ share prices data, the forward analyser took approximately seven days to completely calculate all the price thresholds and labelling the flagged comments. The length of time taken in this process heavily relied on the computer machine power and the efficiency of the programming in FDBM. In this research, the server machine used is a quad core CPU (2.50GHz Intel(R) Xeon(R) CPU E5-2680 v3). Although the forward analysis process takes a long time to process, this is due to the massive amount of data being processed altogether in this research. In real world scenario, this methodology can significantly help relevant authorities to narrow down and focus on the potentially illegal comments

with higher risks. Therefore, the hypothesis for this section is met.

### V. BACKWARD ANALYSIS METHODOLOGY

As an enhancement to the forward analysis process, the novel backward analysis process will test whether simple moving average (SMA) technique can be used to reduce false positives in the comments flagging process by highlighting abnormalities in the share prices and backward classify the flagged comments.

The backward analysis methodology in this section will test the following hypothesis:

H<sub>0b</sub>: Backward analysis can be performed by matching abnormal stock prices with the flagged comments to further classify flagged comments to reduce false positive.

H<sub>1b</sub>: Backward analysis cannot be performed by matching abnormal stock prices with the flagged comments to further classify flagged comments to reduce false positive.

The moving average is one of the technical analysis methods that is often being used by financial analysts to predict the future price patterns, learning stocks’ behaviour and trends by studying historical price data. The most basic moving average technique being used by financial analysts is SMA. Some research even used such moving average techniques to predict the rate of traffic congestions and road accidents [30]. However, it appears that there was no attempt to integrate moving average techniques in the detection process of potential FDB crimes in the past.

The backward analysis attempts to use SMA to test if it can be of helpful to detect flagged comments while reducing false positives. SMA technique is integrated and applied to the share prices before performing backward analysis. The moving average technique is used in backward analysis because it can calculate and highlight whether a price figure exceeds a certain threshold. The following section discusses the methodology to perform backward analysis.

#### A. Methodology

The following describes the steps taken to produce results for analysis:

##### 1) Moving Average Calculation

a) Firstly, decide time periods use for this experiment i.e. 1 day, 3 days and 5 days.

b) Next, calculate the Simple Moving Average (SMA) using its formula as below and record calculation results in database:

$$SMA = \frac{P_1 + P_2 + \dots + P_n}{n}$$

##### 2) Alert Labelling

a) Apply 5%, 10% and 15% thresholds calculation based on the calculated SMA figure+++ and save in database table. Table IV shows an example of the threshold calculations, assuming the SMA is \$15.4:

TABLE IV. SMA THRESHOLD CALCULATION EXAMPLE

Threshold	SMA Threshold Price
5%	\$15.4 * 1.05 = \$16.17
10%	\$15.4 * 1.10 = \$16.94
15%	\$15.4 * 1.15 = \$17.71

b) Once the SMA figures and threshold figures above SMA are obtained, check each original price against the calculated threshold figures. If an original price exceeds the calculated threshold figure, label these threshold alerts accordingly (i.e. 5%, 10% or 15%). The alert labelling rules are as follows:

- Label as “5%”: If the original price figure of a particular date and time is between 5% and 10% higher than the SMA price figure.
- Label as “10%”: If the original price figure of a particular date and time is between 10% and 15% higher than the SMA price figure.
- Label as “15%”: If the original price figure of a particular date and time is 15% and above the SMA price figure.

3) Alert Matching

a) Next, the backward analyser appends the price alerts back to the ‘flaggedcomment’ table by matching the ticker symbol and the exact or nearest date and time between both ‘price’ and ‘flaggedcomment’ tables.

B. Backwards Analysis Methodology Results

Table V shows the total number of flagged comments that matched 5% threshold from both forward and backward analysis for the 1 day, 3 days and 5 days’ time period. Out of 49,858 flagged comments there are 228 flagged comments from the 1 day time period experiment labelled with Y (5% threshold from forward analysis) which are also labelled with 5% threshold from backward analysis. Next, there are 306 flagged comments from the 3 days’ time period labelled with Y (5% threshold from forward analysis) and 5% threshold from backward analysis. Lastly, there are 274 flagged comments from the 5 days’ time period labelled with Y (5% threshold from forward analysis) and 5% threshold from backward analysis.

TABLE V. TOTAL NUMBER OF FLAGGED COMMENTS THAT MATCHED 5% THRESHOLD FROM BOTH FORWARD AND BACKWARD ANALYSIS

	5% 1D	5% 3D	5% 5D
C (<5%)	518	1039	1300
Y (5%)	228	306	274
A (10%)	89	259	183
R (15%)	154	126	84

Table VI shows the total number of flagged comments that matched 10% threshold from both forward and backward analysis for the 1 day, 3 days and 5 days’ time period. Out of 49,858 flagged comments there are 40 flagged comments from the 1 day time period experiment labelled with A (10% threshold from forward analysis) which are also labelled with 10% threshold from backward analysis. Next, followed by 49 flagged comments from the 3 days’ period labelled with A

(10% threshold from forward analysis) and 10% threshold from backward analysis. Lastly, there are 64 flagged comments from the 5 days’ period labelled with A (10% threshold from forward analysis) and 10% threshold from backward analysis.

TABLE VI. TOTAL NUMBER OF FLAGGED COMMENTS THAT MATCHED 10% THRESHOLD FROM BOTH FORWARD AND BACKWARD ANALYSIS

	10% 1D	10% 3D	10% 5D
C (<5%)	204	291	366
Y (5%)	99	62	100
A (10%)	40	49	64
R (15%)	79	85	97

Table VII shows the total number of flagged comments that matched 15% threshold from both forward and backward analysis for the 1 day, 3 days and 5 days’ period. Out of 49,858 flagged comments there are 199 flagged comments from the 1 day time period experiment labelled with R (15% threshold from forward analysis) which are also labelled with 15% threshold from backward analysis. There are 408 flagged comments from the 3 days’ time period labelled with R (15% threshold from forward analysis) and 15% threshold from backward analysis. Lastly, there are 500 flagged comments from the 5 days’ time period labelled with R (15% threshold from forward analysis) and 15% threshold from backward analysis.

TABLE VII. TOTAL NUMBER OF FLAGGED COMMENTS THAT MATCHED 15% THRESHOLD FROM BOTH FORWARD AND BACKWARD ANALYSIS

	15% 1D	15% 3D	15% 5D
C (<5%)	242	356	395
Y (5%)	74	127	146
A (10%)	42	65	94
R (15%)	199	408	500

The results in Tables V, VI and VII show it is possible to perform backward analysis by matching the abnormal stock prices backwards to the flagged comments to resolve false positives.

Take ticker symbol “BOX” as an example, there are 50 comments belong to this stock flagged as “R (15%)” threshold in the forward analysis process. Subsequently, some of these comments are flagged with SMA 15% threshold alert in the backward analysis process. This indicates that there are very high chances of potentially illegal activities going on during ± 2 days’ time of the comments made. A further look at these flagged comments can confirm a highly potential P&D crime. One comment suggests that P&D has indeed happened which pumped the price up and then dumped. Another comment shows that there is still an attempt to pump up the price after the P&D event. Author “ne14t” has a series of BOX comments showing that he/she could possibly involve in a P&D crime. As an enhancement to the forward analysis methodology, the backward analysis aims to resolve false positives and reduce the need of a lot of manpower and time to read through initially flagged comments. The time taken in both forward and backward analysis process in this research is long; however, this is only due to the significant amount of data being processed and analysed altogether. If the prototype system and both methodologies are applied in real time in real world scenarios, it can significantly reduce the time, effort and cost of

monitoring and detecting P&D crimes on FDBs. Therefore, this concluded that the hypothesis is met.

Date/time: 06/10/2014 14:42:38  
Author: bigwod  
Comment: slow build up is what i wanted had some  
fools ramp it up and it was gone now its back

Date/time: 07/10/2014 09:02:19  
Author: ne14t  
Comment: buys now showing the correct colour!

## VI. CONCLUSION AND FUTURE WORK

This paper has introduced two novel methodologies for detecting potentially illegal activities on share price based FDBs by looking not only at the comments but also the per minute share prices. IE techniques were used to collect FDB artefacts such as ticker symbol, comments and prices which made the forward analysis possible to be conducted in this research. A total of 49,858 comments were flagged when matching against the P&D IE keyword template. On average, this is 4,154 flagged comments per week or 593 flagged comments a day. More importantly, these comments belong to only 941 listed companies, not the entire stock market in the UK. Furthermore, according to the results, a large portion of these flagged comments are belong to the listed companies under FTSE AIM All-Share index, where it contains many smaller companies since it is an index that has a more flexible regulatory system, thus, allowing the smaller companies to enter LSE. In order to perform a more realistic investigation into such financial crime on all the FDBs and for all listed companies in the UK on a daily basis, the forward and backward analysis methodologies integrate share prices in the analysis process. This makes it possible for the relevant authorities to prioritise on investigating the flagged comments that have higher risks. The methodologies implemented in FDBM can significantly reduce the time and efforts needed by the relevant authorities to investigate P&D crime on FDBs in real time. As suggested by [29], regulators need to monitor share price based FDBs closely as share price based FDBs are becoming increasingly popular and the authors also find strong positive relationship between the stock prices of smaller companies and the investors' sentiments on FDBs.

The current limitations of this research are such as, not having a predefined IE keyword template for other financial crimes that can happen on the FDBs, namely Insider Information; secondly, the prototype system has not yet taken other artefact data such as broker ratings and director deals into account during the forward and backward analysis; thirdly, the prototype system has previously relied on an XML file format to obtain comments artefact data from the FDBs, thus, it should be programmed to be able to obtain comments through HTML file format, so that it can crawl comments data from FDBs that do not provide comments through XML file format.

### REFERENCES

[1] Leinweber, D.J., & Madhavan, A.N., "Three Hundred Years of Stock Market Manipulations," *Journal of Investing*, p. 7–16, 2001.

- [2] Campbell, J.A., "In and Out Scream and Shout: An Internet Conversation about Stock Price Manipulation," *Proceedings of the 34th Hawaii International Conference on System Sciences*, p. 1–10, 2001.
- [3] Riem, A., "Cybercrimes of the 21st Century: Crimes against the individual — Part 1. Computer Fraud & Security," 6, 13–17, 2001.
- [4] Cybenko, G., Giani, A., & Thompson, P., "Cognitive Hacking: A Battle for the Mind," 2002.
- [5] Delort, J. Y., Arunasalam, B., & Paris, C., "Automatic Moderation of Online Discussion Sites," *International Journal of Electronic Commerce*, 15(3), p. 9–30, 2011.
- [6] Knott, E., & Owda, M., "The detection of potentially illegal activity on financial discussion boards using information extraction," 2nd International Conference on Cybercrime, Security and Digital Forensics, London, UK, 2012.
- [7] Antweiler, W., & Frank, M. Z., "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards," *The Journal of Finance*, 59(3), p. 1259–1294, 2004.
- [8] Cook, D. O., & Lu, X., "Noise, Information, and Rumors: Internet Boards Messages Affect Stock Returns," University of Alabama, 2009.
- [9] Delort, J. Y., Arunasalam, B., & Leung, H., "The Impact of Manipulation in Internet Stock Message Boards," *International Journal of Banking and Finance*, 8(4), p. 1–18, 2011.
- [10] Bettman, J., Hallett, A., & Sault, S., "Rumortrage: Can Investors Profit on Takeover Rumors on Internet Stock Message Boards?," 2011.
- [11] Leung, H., and Ton, T., "The impact of internet stock message boards on cross-sectional returns of small-capitalization stocks," *Journal of Banking & Finance*, p. 37–55, 2015.
- [12] Lee, P. S., Owda, M., & Crockett, K., "The detection of fraud activities on the stock market through forward analysis methodology of financial discussion boards," *Future of Information and Communications Conference*, Singapore, 2018.
- [13] Cowie, J., & Lehnert, W., "Information Extraction," *Communications of the ACM*, 39(1), p. 80–91, 1996.
- [14] Seo, K., Choi, J., & Choi, Y., "Research about Extracting and Analyzing Accounting Data of Company to Detect Financial Fraud. *Intelligence and Security Informatics*, p. 200–202, 2009.
- [15] Limanto et al, "An Information Extraction Engine for Web Discussion Forums," Nanyang Technological University, Singapore. ACM 1-59593-051-5/05/0005, May 2005.
- [16] Owda, M., Lee, P. S., Crockett, K., "Financial Discussion Boards Irregularities Detection System (FDBs-IDS) using Information Extraction," *Intelligent Systems Conference 2017*, London, UK, 2017.
- [17] Masterson, D., & Kushmerick, N., "Information Extraction from Multi-Document Threads," 2003.
- [18] Soderland, S., "Learning Information Extraction Rules for Semi-structured and Free Text," 1999.
- [19] Cunningham H., "Information Extraction, Automatic," in Keith Brown, (Editor-in-Chief) *Encyclopedia of Language & Linguistics*, Second Edition, 5, p. 665-677, 2006.
- [20] Chiticariu, L., Li, Y., & Reiss, R. F., "Rule-based Information Extraction is Dead! Long Live Rule-based Information Extraction Systems!," *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 827–832, Seattle, Washington, USA, 2013.
- [21] Kaiser, C., & Bodendorf, F., "Mining consumer dialog in online forums," *Internet Research*, 22(3), p. 275-297, 2012.
- [22] Westerman, D., Spence, P. R., & Van Der Heide, B., "Social Media as Information Source: Recency of Updates and Credibility of Information," *Journal of Computer-Mediated Communication*, 19, p. 171-183, 2014.
- [23] Wolfram, M. S. A., "Modelling the Stock Market using Twitter," 2010.
- [24] Mittermayer, M., "Forecasting Intraday Stock Price Trends with Text Mining Techniques," in *Hawai'i International Conference on System Sciences*, 2014.
- [25] Siering, M., "All Pump, No Dump? The Impact of Internet Deception on Stock Markets," *ECIS 2013 Completed Research*, 115, 2013.

- [26] Alić, I., "Supporting Financial Market Surveillance: An IT Artifact Evaluation, BLED 2015 Proceedings, Paper 36, 2015.
- [27] Felton, J., & Kim, J., "Warnings from the Enron Message Board," *Journal of Investing*, 11(3), p. 29-52, 2002.
- [28] Campbell, J.A. & Cezec-Kecmanovic, D., "Communicative practices in an online financial forum during abnormal stock market behavior. *Information and Management*, 48, p. 37-52, 2011.
- [29] Sabherwal, S., Sarkar, S.K., & Zhang, Y., "Do Internet Stock Message Boards Influence Trading? Evidence from Heavily Discussed Stocks with No Fundamental News," *Journal of Business Finance & Accounting*, 38(9) & (10), p. 1209–1237, 2011.
- [30] Raiyn, J., and Toledo, T., "Real-Time Road Traffic Anomaly Detection," *Journal of Transportation Technologies*, 4(3), p. 256-266, 2014.

# Inferring of Cognitive Skill Zones in Concept Space of Knowledge Assessment

Rania Aboalela  
Kent State University  
Kent, Ohio, U.S.A

Javed Khan  
Kent State University  
Kent, Ohio, U.S.A

**Abstract**—In these research zones of the knowledge, the assessed domain is identified. Explicitly, these zones are known as Verified Skills, Derived Skills and Potential Skills. In detail, the Verified Skills Zone is the set of tested concepts in the knowledge domain, while Derived Skills Zone is the set of the prerequisite concepts to the tested concepts based on the cognitive skills relation, whereas Potential Skills Zone is the set in the domain that have never been tested or prerequisite to the tested concepts but they are related to the tested concept based on the cognitive relation skills. Identifying cognitive relations between the concepts in one domain simplifies the structure of the assessment, which helps to find the knowledge state of the assessed individual in a short time and minimum number of questions. The existence of the concepts in the assessment domain helps us to estimate the set of the concepts that are known or not known or ready to be known or not ready to be known. In addition, it provides the output of the assessment in concept centric values in addition to the quantity values. The assessment result gives binary values of the assessed domain. “1” implies knowing the concept, whereas “0” implies not knowing the concept. The output is six sets of concepts: 1) Verified Known Skills; 2) Verified Not Known Skills; 3) Derived Known Skills; 4) Derived Not Known Skills; 5) Potential Known Skills; and 6) Potential Not Known Skills. The experiment is conducted to show the binary output of the assessed domain based on the participants’ answers to the asked questions. The results also highlight the efficiency of the assessment.

**Keywords**—Cognitive skill; bloom’s taxonomy; assessment of knowledge; concept space; concepts zones

## I. INTRODUCTION

The knowledge assessments of knowledge space [1] requires large number of questions and the assessment structure is complicated. They don’t identify cognitive difficulty variations of learning. The previous research work applied the assessment of knowledge in one space such as [1], [2] are only applicable for a domain that has clear relations like Mathematical Fields. Applied science fields like Software Engineering and Medical Science need identifying internal cognitive relations to simplify the assessment structure and to give an accurate assessment result. The difficulty in assessing the knowledge in one domain studied by [3] introduced ontological relation between the concepts in the course and test questions. In addition, the research work [4] proposed a skill that can be characterized as a pair consisting of a concept and an activity. As an example of such a pair, they give “Apply the Pythagorean Theorem”. In this study, the researchers concentrate on the concepts as they appear in the text in either

phrase form or single word form and identify the link between the concepts, as the skill required to learn the concept at a certain skill level, which identifies the prerequisite relation between the concepts. Moreover, [5], [6] studied and validated the efficiency and importance using the parameter of cognitive skill level to assess the knowledge in one domain. The research work [7] introduce a model of the cognitive level relation between the concepts in the learning assessment. The cognitive skill that used is the verbs of the Revised Bloom’s Taxonomy. The revised version identified six verbs infer to six categories of skills. The original taxonomy was created in 1956 by Dr. Benjamin Bloom [8]. The Blooms’ Taxonomy arranges what the learner has to learn in a hierarchy of six levels. In 2001, the six major categories were changed from noun to verb forms and renamed [9]. In this research, the verbs of the revised Blooms’ Taxonomy used to identify the prerequisite cognitive relation between the concepts in the domain. The six verbs are inferred using skill numbers that indicate the cognitive difficulty as the following:

“1” means recall, “2” means understanding, “3” means applying, “4” means analyzing, “5” means evaluating and “6” means creating. The researchers in this research work use the five higher levels in referring to the level needed to acquire the concept at the cognitive skill levels of understanding, applying, analyzing, or creating the concepts.

The present study is structured as follows: materials and methods in Sections II and III. The complexity of assessing the knowledge in one domain is discussed in Section IV. The component of the assessment is discussed in Section V. The experimental results in Section VI. The conclusion is given in Section VII. The discussion and future work is discussed in Section VIII.

## II. COGNITIVE SKILL ZONES

### A. The Concept Zones of Verified Skills

First, Verified Skills zone is the zone of the main concepts that have to be tested and they infer to the most concepts in the assessed domain. VS is defined as where there is a direct evidence that a learner knows concept  $C_x$  at a cognitive skills level  $k$  it is considered to belong to verified set  $VS(k)$ . To illustrate the VS, let us consider a question  $Q_i$ , which can ascertain that a student knows a specific concept  $C_x$ . The verified skills satisfy the condition that:

If  $(Q_i, C_x)L_k$  &  $C_x$  is correct answer  $\rightarrow C_x \in VS(k) \forall C_x \in$  completely correct answer concepts.

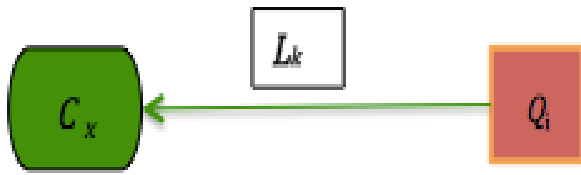


Fig. 1. Verified Skills (VS) Zone.

Where,  $Q_i \in$  Test questions,  $C_x \in$  Tested concepts,  $L_k \in$  Bloom link of level  $k$ ,  $VS(k) \in$  Verified skills at level  $K$  and  $(Q_i, C_x)L_k$  means existing link between the question  $Q_i$  and the concept  $C_x$  at level  $K$ . The link means that to answer  $Q_i$  correctly  $C_x$  must be learned at level  $k$ . Fig. 1 shows verified skill link.

**B. Concept Zones of Derived Skills at Level 2, DS (K=2)**

The Zones of the derived skills is the set of the concepts that must be understood to attain learning the VS. Derived Skill is defined as where there is indirect evidence that a student knows a concept  $C_i$  at a cognitive skills level 2 (the understand level), it is considered to be a part of DS (2). In other words, if there is indirect evidence that the concept  $C_i$  is understood by the student then it will belong to DS ( $K=2$ ). The condition of the relation is expressed as the following:

If  $C_i$  is not a member in a verified set but there exists two links such that  $(Q_i, C_x)L_k$ ,  $(C_i, C_x)L_m$  &  $C_x \in VS(k)$  that it is a member in VS and  $m=2$  and  $k \geq m$ , then  $C_i$  is a member in DS at level 2, i.e.  $C_i \in DS(2)$ ,  $Q_i \in$  Test Questions,  $C_x \in VS$ ,  $C_i \in$  another concept in the concept space. The  $(Q_i, C_x)L_k$ ,  $(Q_i, C_x)L_m$  means existing link between the Question  $Q_i$  and the Concept  $C_x$  at level  $k$  and  $m$  respectively. Fig. 2 illustrates DS relation at level 2.

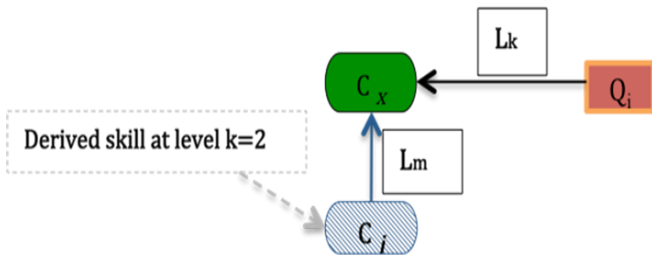


Fig. 2. Derived Skill (DS) Zone at level 2.

**C. Zones of the Support Set (SS)**

To distinguish the higher-level relations, the classification of the set into Support Set and Supported Set must be identified. The support set means the prerequisite set of the supported set. Let  $C_A$  be a node. Let  $C_B$  be another node from where there is a level  $k$  link to  $A$ . Then  $C_B$  at level  $k$  is called the support node of  $C_A$ . That means  $C_B$  is the prerequisite set of  $C_A$  concept at level  $k$ . Let  $S(C_A, k)$  be the set of all such  $C_B$  nodes in the complete concept graph  $G$ . The  $S(C_A, k)$  is the level  $k$  Support Set for  $C_A$ . i.e. all concepts in this set must be learned to have a level  $k$  skill in  $A$ . Fig. 3 illustrates the Support Set & Support Node, which is any node in the Support Set.

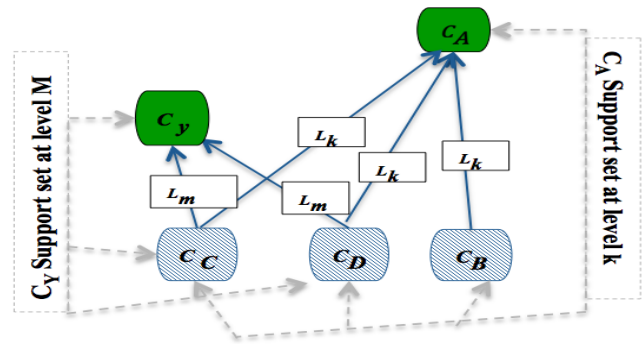


Fig. 3. Support Node (SN) & Support Set (SS).

**D. Zones of Derived Skill, DS ( $k > 2$ )**

DS ( $k > 2$ ) means that there is a direct evidence a learner knows a concept  $C_y$  at a cognitive skill level 2, and there is indirect evidence he knows it at a cognitive skill level higher than a cognitive skill level 2. In other words, by inference a learner could either apply/analyze/evaluate/create, a concept  $C_y$ . The relation condition is illustrated as the following: If  $C_y$  is known i.e. it is in  $DS(2)$  or  $VS(2)$ , and if all level  $k$  support nodes of  $C_y$  i.e.  $S(C_y, k)$  is in  $VS(2) \vee DS(2)$ , then  $C_y$  will be considered as a Derived Skill at level  $k$ . In other words. If  $C_y \in DS(2) \vee VS(2)$  and  $S(C_y, k)$  is subset of  $DS(2) \vee VS(2) \rightarrow C_y \in DS(k)$ . Fig. 4 illustrates DS ( $k > 2$ ).

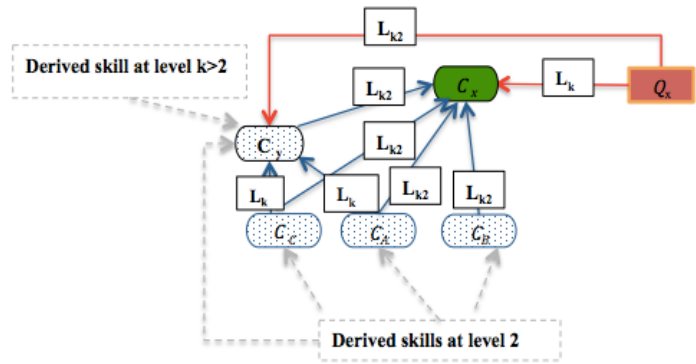


Fig. 4. Derived Skills relation at level  $>2$ .

**E. Zones of Potential Skill PS [ $k > 2$ ]**

The zones of Potential Skills ( $k > 2$ ) is defined as where there is indirect evidence that a learner knows a particular concept such as “A” at a cognitive skills level higher than 2 (apply/ analyze/ evaluate/ create), it is a part of PS ( $k > 2$ ). The concepts in the Zone PS have never been tested but their prerequisite concepts at the target skills level are tested either directly or indirectly. The relation condition is illustrated as the following: Let  $S(A, k)$  is the support set of  $A$  at level  $k$ . If every node in the  $S(A, k)$  is subset of  $VS \vee DS$  at any level (doesn't matter-because The study only want to guarantee that the set is known) i.e.  $S(A, k) \subset VS() \vee DS()$ , but there is no evidence that  $A$  is known, then  $A$  is in potential skill set  $PS(k)$  i.e.  $A \in PS(k)$ , where  $C_d, C_x \in (VS)$  and  $C_c, C_A, C_B \in (DS)$  and  $L_k \in$  Bloom's link at level  $k$ . Fig. 5 illustrates Potential Skills relation.

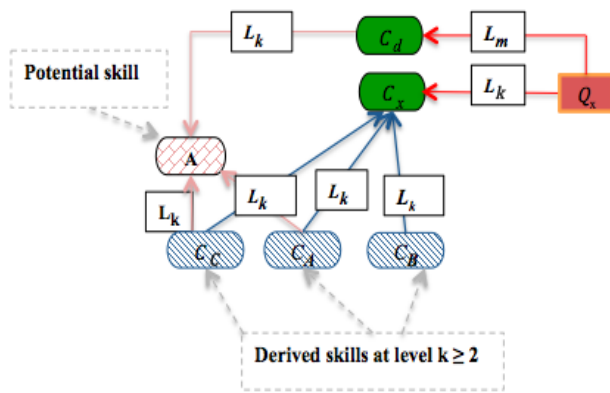


Fig. 5. Potential skills zone.

### III. CONCEPT MAPPED TESTING AND EVALUATION METHOD

To measure the student learning, a *concept mapped testing and evaluation* method is set up. A test is composed of a set of questions. The learners are required to answer the questions based on their knowledge. Grader evaluates the student knowledge based on the answers. In conventional evaluation, a grader grades the answers and assigns a quantitative score for the student. The researchers slightly modify the evaluation method where the grader instead of a numerical score, is asked to evaluate if there is evidence in the answer that the student has succeeded or failed to attain learn a concept at a certain cognitive skill level. The researchers called it *concept mapped testing & evaluation method*. Each tested concept in the assessment domain is labeled to the question based on the cognitive level as the following theory: To answer the question  $Q_i$  correctly the concept  $C_x$  must be known at skill level  $L_k$ . Fig. 6 shows the relation.

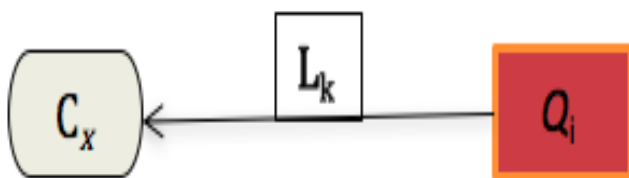


Fig. 6. The theory of connect the question to the tested concept.

### IV. THE COMPLEXITY OF ACCESSING THE KNOWLEDGE IN ONE DOMAIN

Assessing the knowledge in one domain is complex and requires large number of questions to ask about the target skill level of each concept in the domain. It is hard to capture a question for each concept in the domain at the target skill level. To solve this problem, assessment zones to classify which concepts should be directly tested or indirectly tested were obtained. Zones of the assessment were identified to simplify and decrease the number of the tested concepts. In this study, the researchers propose making the link between any two concepts a verb of the skill which needs to be learned, for example: to “apply” a concept B, the individual must know the prerequisite concept A at skill level 2, which is the understanding level. The “apply” is indicated by the number 3

in the link. Once the proposed zones are considered then the number of questions would be minimized to test only the concepts in the VS zones. Visualizing the concepts of one domain by using cognitive level mapped concepts graphs accomplished in [9]. The idea of automatically discovering and extracting the Bloom’s Taxonomy from the text in one knowledge space is studied by Nafa, Khan and their colleagues [10]-[12].

### V. ASSESSMENT COMPONENTS

#### A. The Input of the Assessment

The input is composed of:

- 1) The assessment domain graph, which is the cognitive skill level mapped concept graph of the assessed domain. The concepts mapped together with the prerequisite relation based on the cognitive skills.
- 2) The set of the questions mapped with the concepts in the concepts space.
- 3) The set of learners’ responses to the set of questions.

#### B. The Output of the Assessment

The assessment goal is to find the knowing concepts inferred by “1” and the not knowing concepts inferred by “0” in the zones of the concepts (which concepts are known and not in known in which zone). This is the binary of knowledge (learning states) of these concepts in the various zones. The learning states are six states of knowing and not knowing the concepts in the concepts zones. 1) Verified Known Skills, 2) Verified Not Known Skills, 3) Derived Known Skills, 4) Derived Not Known Skills, and 5) Potential Known Skills, 6) Potential Not Known Skills. The sets of known zones of VKS include such a concept whose question is answered correctly and given the value 1. The sets of known zones of DKS include such a concept has never been tested but it is prerequisite to the concept in VKS. Each estimated concept in DKS is given the value 1. The set of PKS zone, which they have never been tested but all their prerequisite concepts at a certain skill level are tested and given the result that they all are known. Thus, the concepts are considered ready to be known PKS. Each estimated concept in PKS is given the value 1. The set of PNS zone, which they have never been tested but all their prerequisite concepts at a certain skill level are tested and given the result that they all are not known. Thus, the concepts are considered not ready to be known PNS. Each estimated concept in PNS is given the value 1.

### VI. EXPERIMENT

An experiment is organized to prove the efficiency of identifying the relations of the cognitive skill level between the concepts to maximize the estimation of measurement the concepts from few tested concepts.

#### A. The Experiment Setup

A human subject test is organized to prove the efficiency of the methods. The test is composed of 9 questions, which are selected from the midterm questions that have been given to the learners by the instructor of the class. The class is CS 61002 Algorithms and Programming in the Computer Science department. The test was introduced online in one session. The



participants are 154 graduate students, attending the class. In this setup, the questions are specially redesigned to directly test a certain skill level of each concept belonging to the assessment domain. Nine questions are asked about 18 concepts at certain skill levels. Thus the 18 concepts are in the zone of VS. The concepts in the zones DS and PS are not tested but their evaluation of their binary states values is estimated. This experiment proves that using the proposed methods associated with the cognitive relations optimizes the knowledge assessment. The result of the evaluation of the perfect learner shows that the amount of the estimated knowledge of the assessed learner could be increased by at least 3 times over the conventional assessment which uses just numerical methods. The perfect learner's answers are used to calculate the experiment footprint and the size of footprint of each relation method. Table I and Fig. 7 show the size of footprint according to the perfect learner. Footprint is the number of fundamental tested concepts. The perfect learner is the student who gives a correct answer to all the asked questions. As evident the size of VS footprint is 18, the size of DS footprint is 31 and the size of PS footprint is 31, which means that if the learner answered the questions correctly, then it would tell he knows certain levels of each of CONCEPTS AT EACH SKIL knowledge assessment methods with the cognitive relation, one can maximize the amount of the estimation knowledge of the assessed learners.

TABLE I. THE NUMBER OF TESTED AND ESTIMATED CONCEPTS AT EACH SKILL LEVEL IN THE VARIOUS ZONES VS, DS AND PS

Skill Level	Verified	Derived	Potential
	VS	DS	PS
L2	7	12	13
L3	4	11	11
L4	2	3	2
L5	1	2	2
L6	4	3	3
Sum	18	31	31

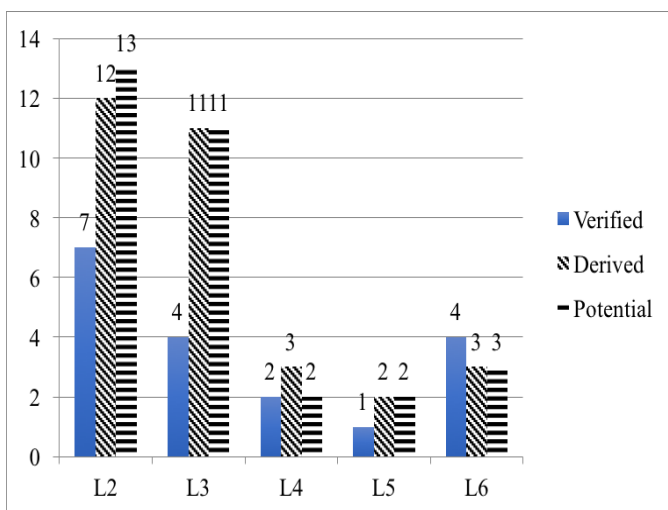


Fig. 7. The number of tested and estimated concepts at each skill level in the Various Zones VS, DS and PS.

### B. The Binary Concept States of the Human Subject Test

The experiment is conducted to find out the binary set of 154 participants. Each participant of the 154 participants is assigned to number. Two participants we rechosen to show their binary concept states. One of them is the perfect student who gets all the answer correctly. The perfect students assigned to number 1 and the second student assigned to number 23. Also, each concept in the concept space is assigned to an integer number. The researchers show the binary concept state for the three zones VS, DS and PS. Thus, six concept zones are illustrated for the two participants. Fig. 8, 9 and 10 show the binary concept states of the 31 concepts in the zone of VS, DS and PS of the perfect student respectively. Fig. 11, 12 and 13 show the binary concept states of the 31 concepts in the zone of VS, DS and PS of the of the laziest student #23 respectively.

## VII. CONCLUSION

The concept zones of the assessed domain are proposed. The efficiency of identifying cognitive relations between the concepts in the assessed domain is proved. Using the cognitive skills relation between the concepts in the assessment increases the amount of the estimated concepts, even though the number of tested concepts may be minimized and eliminated under the conditions laid down by the target cognitive skill levels. The binary concept state is assigned to the participants and the estimated binary concept states of the untested concepts are concluded.

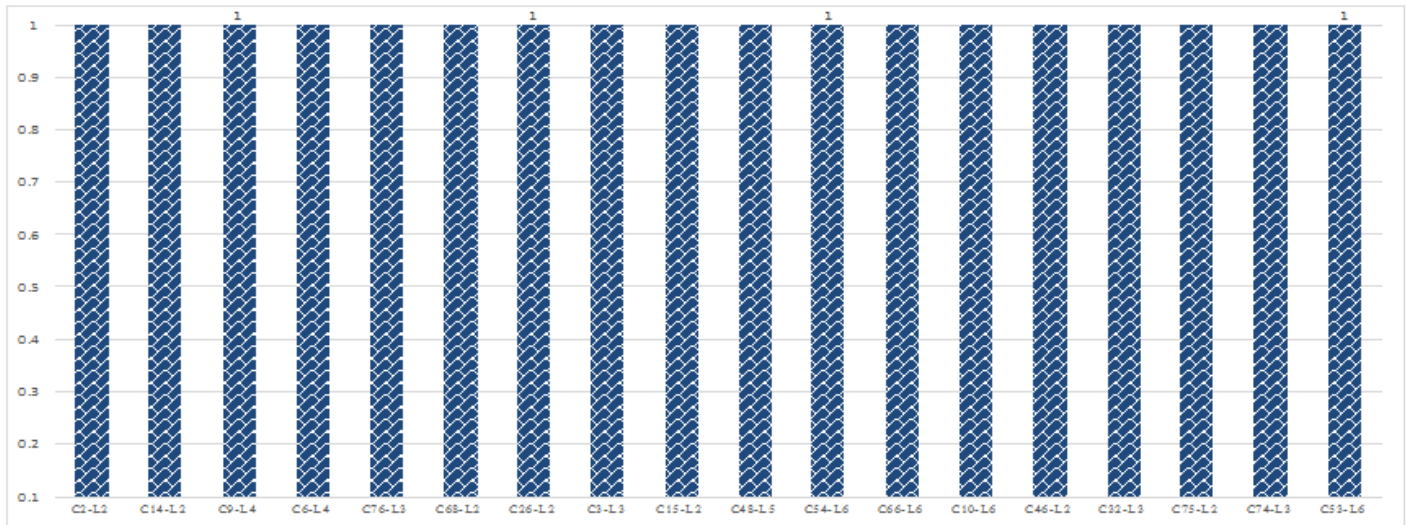
## VIII. DISCUSSION AND FUTURE WORK

There are many evidences or references that can infer the target skills of the concepts in the learning states zones. Sometimes many questions would be asked about the concepts in the domain and there can be a contradiction between the answers given. Also, errors might arise in the estimation of the learning states. In real exams, there can be other phenomena like lucky guess or careless mistakes. People have variant levels of initial knowledge. Accordingly, the probability computation should be used. This fact would be studied in the future work. The probability of the concepts states based on the cognitive relation of the concept zones would be analyzed.

### REFERENCES

- [1] J.-C. Falmagne, E. Cosyn, J.-P. Doignon, and N. Thiery, "The assessment of knowledge, in theory and in practice ," in International Conference on Integration of Knowledge Intensive Multi-Agent Systems, pp. 609-615, 2003.
- [2] J. C Falmagne and J. P. Doignon, Knowledge Spaces, Berlin, Springer, 1999.
- [3] J. Khan, M. Hardas, Y. Ma, "A Study of problem difficulty evaluation for Semantic Network Ontology based intelligent courseware sharing," IEEE/WIC/ACM International Conference on Web Intelligence, WEB Intelligence, pp. 426-429, 2005.
- [4] Heller, J., Steiner, C., Hockemeyer, C. & Albert, D. "Competence-Based Knowledge Structures for Personalised Learning.," International Journal on E-Learning: Association for the Advancement of Computing in Education (AACE), 5(1), Chesapeake, VA, pp. 75-88, 2006.
- [5] R. Aboalela & J. Khan, "Are we asking the right questions to grade our students in a knowledge-state space analysis?," In IEEE 8th International Conference on Technology for Education (T4E), Mumbai, pp. 144-147, 2016.

- [6] R. Aboalela & J. Khan, "Model of Learning Assessment to Measure Student Learning: Inferring of Concept State of Cognitive Skill Level in Concept Space," in In proceeding of the 3rd International Conference on Soft Computing & Machine Intelligence (ISCM), Dubai, United Arab Emirates, pp. 189-195. doi: 10.1109/ISCM.2016.26
- [7] B.S. Bloom, M.D. Engelhart, E.J. Furst, W.H. Hill, & D.R. Krathwohl, (Eds.) Taxonomy of Educational Objectives. The Classification of Educational Goals, Handbook I: Cognitive Domain. David McKay Company, New York, 1956.
- [8] L.W. Anderson, D.R. Krathwohl, P. W. Airasian, K.A. Cruikshank, R.E. Mayer, P.R. Pintrich, et al., A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York, U.S.A. Longman, 2001.
- [9] R. Aboalela, J. Khan, "Visualizing concept space of course content," IEEE 7th International Conference on Engineering Education (ICEED), Japan, November 2015. pp. 609-615  
DOI: 10.1109/ICEED.2015.7451512
- [10] F. Nafa, & J. Khan "Conceptualize the Domain Knowledge Space in the Light of Cognitive Skills," In Proceedings of the 7th International Conference on Computer Supported Education. SCITEPRESS-Science and Technology Publications, pp. 285-295, 2015.
- [11] F. Nafa, J. Khan, S.Othman, & A. Babour, "Discovering Bloom Taxonomic Relationships between Knowledge Units Using Semantic Graph Triangularity Mining," International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Chengdu: IEEE, pp. 224-233, 2016.
- [12] F. Nafa, J.Khan, S. Othman, & A. Babour "Mining Cognitive Skills Levels of Knowledge Units in Text Using Graph Triangularity Mining," Web Intelligence Workshops (WIW), International Conference (IEEE/WIC/ACM), IEEE, pp. 1-4, 2016.



X: Concept Number  
Y: Binary Number, "1" the concept is known. "0" The concept is unknown

Fig. 8. The VKS and VNS Zones of perfect learner.

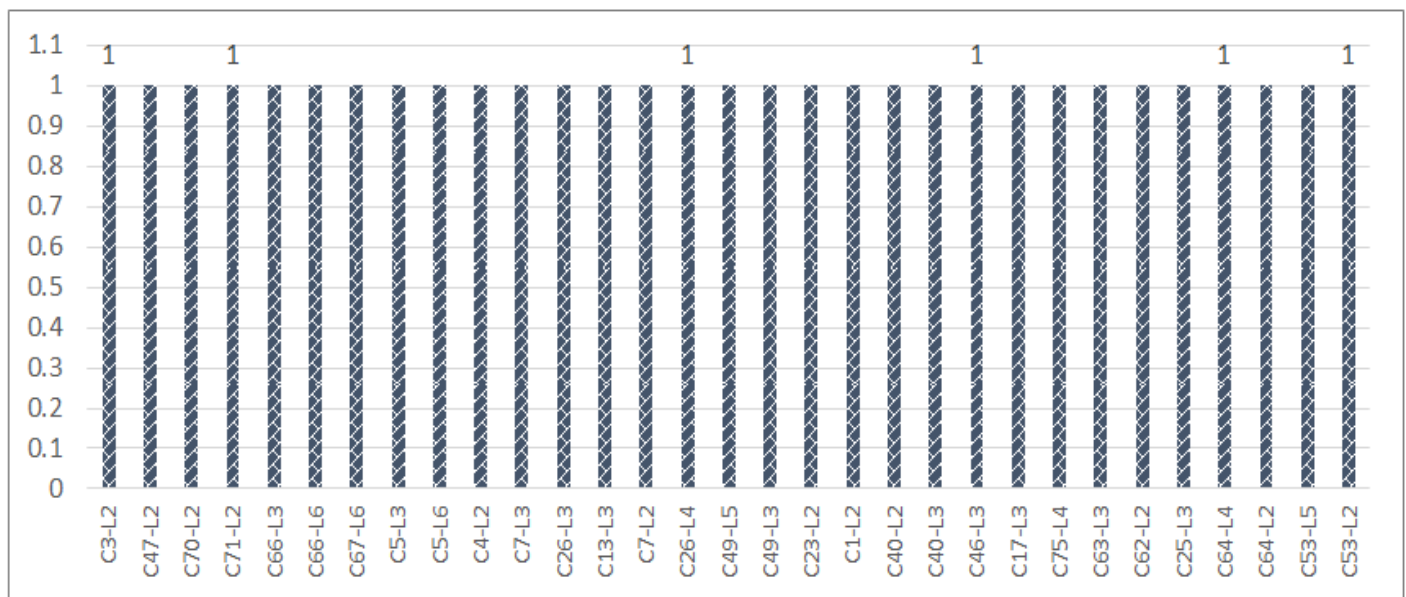


Fig. 9. The DKS and DNS Zones of perfect learner.

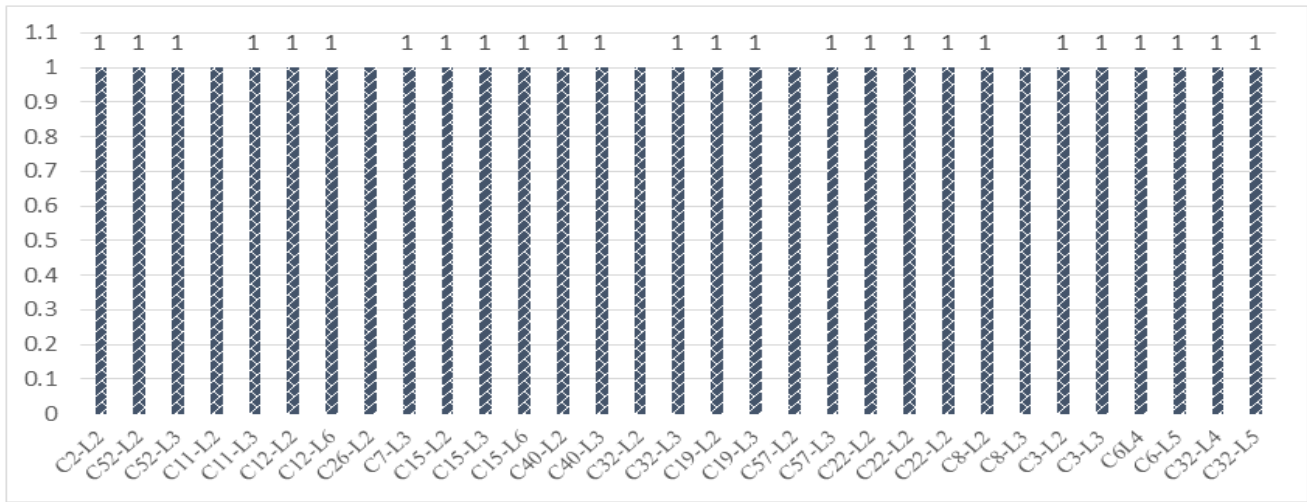
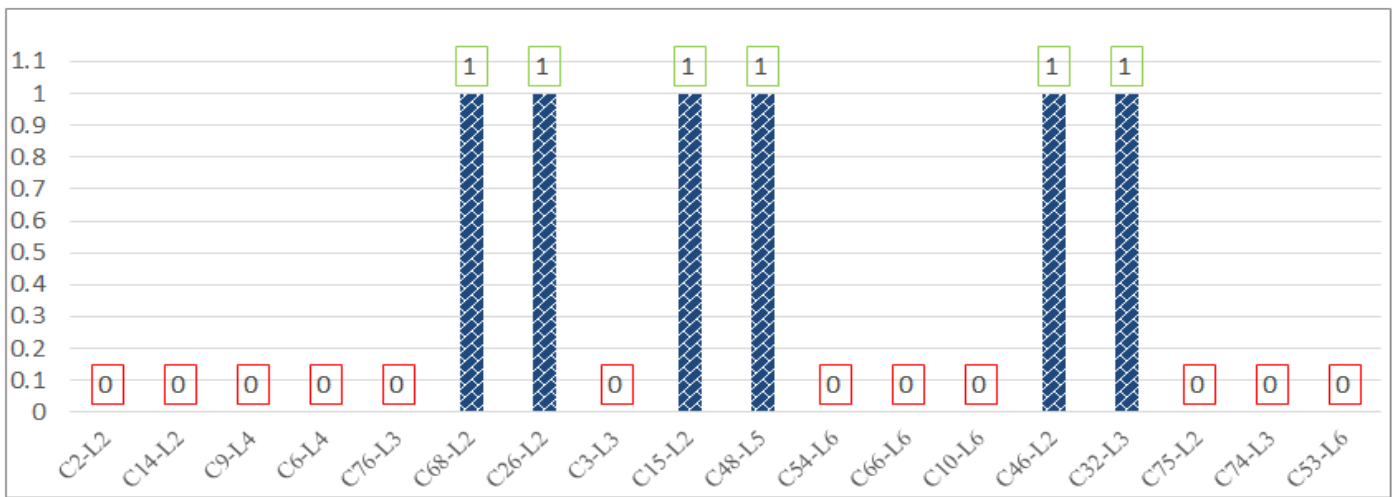
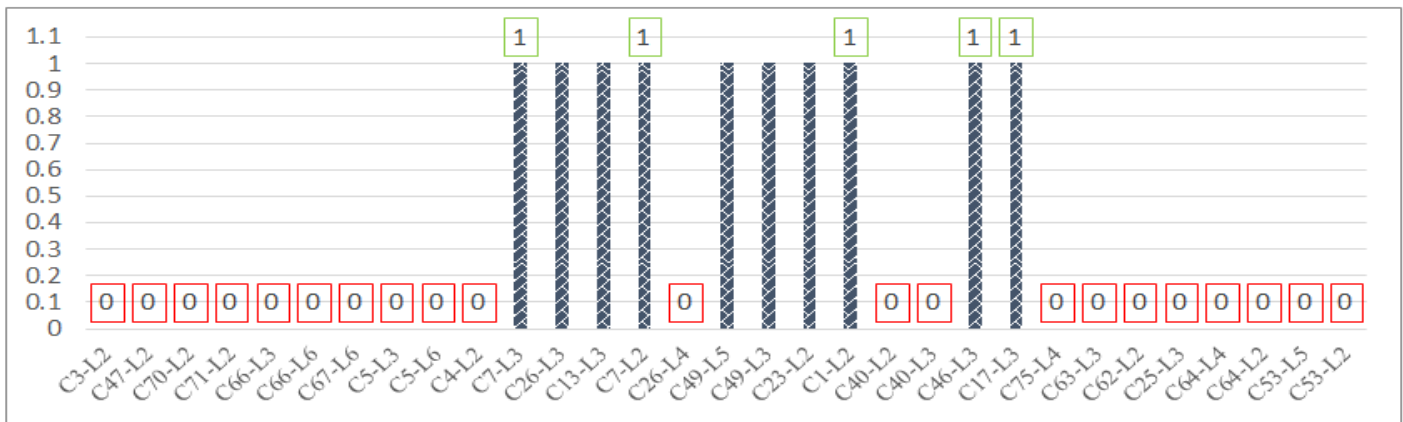


Fig. 10. The PKS and PNS Zones of perfect learner.



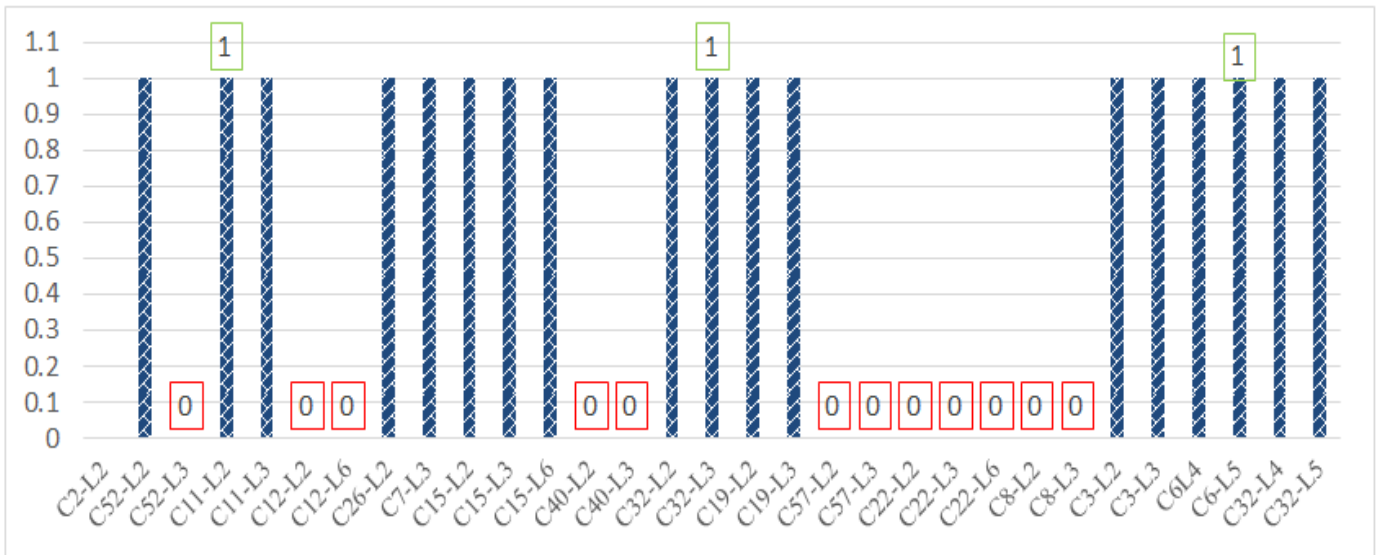
X: Concept Number  
Y: Binary Number, "1" the concept is known. "0" the concept is unknown

Fig. 11. The VKS and VNS Zones of learner # 23.



X: Concept Number  
Y: Binary Number, "1" the concept is known. "0" the concept is unknown

Fig. 12. The DKS and DNS Zones of learner # 23.



X: Concept Number  
 Y: Binary Number, "1" the concept is known. "0" the concept is unknown

Fig. 13. The PKS and PNS Zones of learner # 23.

# Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine

Apapan Pumsirirat, Liu Yan

School of Software Engineering, Tongji University  
Shanghai, China

**Abstract**—Frauds have no constant patterns. They always change their behavior; so, we need to use an unsupervised learning. Fraudsters learn about new technology that allows them to execute frauds through online transactions. Fraudsters assume the regular behavior of consumers, and fraud patterns change fast. So, fraud detection systems need to detect online transactions by using unsupervised learning, because some fraudsters commit frauds once through online mediums and then switch to other techniques. This paper aims to 1) focus on fraud cases that cannot be detected based on previous history or supervised learning, 2) create a model of deep Auto-encoder and restricted Boltzmann machine (RBM) that can reconstruct normal transactions to find anomalies from normal patterns. The proposed deep learning based on auto-encoder (AE) is an unsupervised learning algorithm that applies backpropagation by setting the inputs equal to the outputs. The RBM has two layers, the input layer (visible) and hidden layer. In this research, we use the Tensorflow library from Google to implement AE, RBM, and H2O by using deep learning. The results show the mean squared error, root mean squared error, and area under curve.

**Keywords**—Credit card; fraud detection; deep learning; unsupervised learning; auto-encoder; restricted Boltzmann machine; Tensorflow

## I. INTRODUCTION

Fraud detection in online shopping systems is the hottest topic nowadays. Fraud investigators, banking systems, and electronic payment systems such as PayPal must have an efficient and complex fraud detection system to prevent fraud activities that change rapidly. According to a CyberSource report from 2017, the present fraud loss by order channel, that is, the percentage of fraud loss in their web store was 74 percent and 49 percent in their mobile channels [1]. Based on this information, the lesson is to determine anomalies across patterns of fraud behavior that have undergone change relative to the past.

A good fraud detection system should be able to identify the fraud transaction accurately and should make the detection possible in real-time transactions. Fraud detection can be divided into two groups: anomaly detection and misuse detection [2]. Anomaly detection systems bring normal transaction to be trained and use techniques to determine novel frauds. Conversely, a misuse fraud detection system uses the labeled transaction as normal or fraud transaction to be trained in the database history. So, this misuse detection system entails

a system of supervised learning and anomaly detection system a system of unsupervised learning. What is the difference between supervised learning and unsupervised learning? The answer is that supervised learning studies labeled datasets. They use labeled datasets to train and to render it accurate by changing the parameters of the learning rate. After that, they apply parameters of learning rate to the dataset, the techniques that implement supervised learning such as multilayer-perceptron (MLP) to build the model based on the history of the database. This supervised learning has a disadvantage, since if new fraud transactions happen that do not match with the records of the database, then this transaction will be considered genuine. While, unsupervised learning acquires information from new transactions and finds anomalous patterns from new transaction. This unsupervised learning is more difficult than supervised learning, because we have to use appropriate techniques to detect anomalous behavior.

Neural networks were introduced to detect credit card frauds in the past. Now, we focus on deep learning that is a subfield of machine learning (ML). Based on deep learning in the first period, they use deep learning to know about an image's processing. For example, Facebook uses deep learning in the function to tag people and to know who the person is for subsequent reference. Further, deep learning in neural networks have many algorithms for use in fraud detection, but in this paper, we selected the AE and RBM to detect whether normal transaction of datasets qualified as novel frauds. We believe that some normal transaction in datasets that were labeled as fraud also show suspicious transaction behavior. So, in this paper we focus on unsupervised learning.

In this paper, we use three datasets in these experiments; these datasets are the German, Australian, and European datasets [4], [3], [18]. The first dataset is German, provided by Professor Dr. Hans Hofman [4]. There are twenty attributes that describe the capability, such as credit history, purpose to use credit card, credit amount, job, among others. The German dataset were 1000 instances. The second dataset is from Australia. [3] The attributes' names and values in this dataset have been changed to meaningless symbols to protect the confidentiality of the data. There were 690 instances. The last dataset was from a European cardholder from September 2013. This dataset shows the transaction that occurred in two days with 284, 807 transactions. There were 31 features in this dataset. The 28 features, such as V1, V28 is a numerical input variable result of PCA transformation. Other 3 feature that do

not bind with PCA are “Time”, “Amount”, and “Class”. This experiment will bring together three datasets to compare different receiver operating characteristics (ROC) to understand the performance of binary classifiers.

## II. RELATED WORK

In the past decade, credit card was introduced in the financial segment. Now, credit card has become a popular payment method in online shopping for goods and services. Since the introduction of credit cards, fraudsters have tried to falsely adopt normal behavior of users to make their own payments. Due to these problems, most research on credit card fraud detection has focused on pattern matching in which abnormal patterns are identified as distinct from normal transactions. Many techniques for credit card fraud detection have been presented in the last few years. We will briefly review some of those techniques below.

The K-nearest neighbor (KNN) algorithms are used to detect credit card frauds. This technique is a supervised learning technique. KNN is used for classification of credit card fraud detection by calculating its nearest point. If the new transaction is coming and the point is near the fraudulent transaction, KNN identifies this transaction as a fraud [5]. Many people confuse KNN with K-means clustering, whether they are the same techniques or not. K-means and KNN are different. K-means is an unsupervised learning technique, used for clustering. K-Means tries to determine new patterns from the data and by clustering the data into groups. Conversely, KNN is the number used to compare the nearest neighbor to classify or predict a new transaction based on previous history. The distance in KNN between two data instances can be calculated by using different method, but mostly by using the Euclidean distance. KNN is very useful.

The outlier detection is another method used to detect both supervised and unsupervised learning. Supervised outlier detection method studies and classifies the outlier using the training dataset. Conversely, unsupervised outlier detection is similar to clustering data into multiple groups based on their attributes. N. Malini and Dr. M. Pushpa mention that the outlier detection method based on unsupervised learning is preferred to detect credit card fraud over outlier supervised learning, because unsupervised learning outlier does not require prior information to label data as fraudulent. So, it needs to be trained by using normal transactions to discriminate between a legal or illegal transaction [5].

Some credit card fraud transaction datasets contain the problem of imbalance in datasets. Anusorn Charleonnann mentions that the unbalance of datasets has many characteristics that emerge during the classification. He uses RUS, a data sampling technique, by trying to relieve the problem of class unbalance by editing the class distribution of training datasets. There are two major methods of adjusting the imbalance in datasets, undersampling and oversampling. In his research, he also uses the MRN algorithm for the classification problem of credit card fraud [6].

Artificial neural network (ANN) is a flexible computing framework used to solve a comprehensive range of non-linear

problems. The main idea of ANN is mimicking the learning algorithm of the human brain. The smallest unit of ANN is called a perceptron, is represented as a node. Several perceptrons are connected as a network like the human brain. Each node has a weighed communication with several other nodes in the adjacent layer. A weight is simply a floating-point number, and it can be adjusted when the input eventually comes to train the network. Inputs are passed from input nodes through hidden layers to output nodes. Each node can learn and adjust itself to make it more accurate and appropriate.

The problem of credit card fraud detection has been analyzed with the Chebyshev Function Link Artificial Neural Network (CFANN). CFANN consists of two components, functional expansion and learning. Mukesh Kumar Mishra and Rajashree Dash, authors who used CFANN to detect credit card fraud by comparing it with MLP, and the Decision Tree [7]. MLP infers that the topology was structured into a number of layers. The first layer is called input layer, the middle layer is called the hidden layer. This layer can have more than one layer, and the last layer is called the output layer. Feed forward infers that all information flows in the same direction, the left-to-right direction, without recurrent links. Decision Tree is a structured tree that has a root node and a number of internal and leaf nodes. Their paper compares the performance of CFANN, MLP, and Decision Tree. The result of their study suggests that MLP outperforms CFANN and Decision Tree in fraud detection. Conversely, CFANN makes accurate predictions over the other two techniques [7].

Deep learning forms a state of the art technology in the present day. Most people in IT should follow this. First, ANN was introduced. After that, ML becomes a subset of ANN, and deep learning, a subfield of ML. Deep learning has been used in many fields such as image recognition in Facebook, speech recognition in Apple or Siri, and natural language processing in Google translator. Yamini Pandey used deep learning with the H2O algorithm framework to know complex patterns in the dataset. H2O is an open source for predictive data analytics on Big Data. Supervised learning is based on predictive analytics. The author used H2O based multi-layered, feed forward neural network to find credit card fraud patterns. H2O's performance based on the deep learning model shows less error in mean squared error, root mean squared error, mean absolute error, and root mean squared log error. Hence, these errors are low that enhances accuracy. The model's accuracy is also high in relation to the errors mentioned above [8]. Another concern before registering credit cards is credit cards' analysis' judgement. Ayahiko Niimi uses deep learning to judge whether a credit card should be issued to the user if they satisfy particular criteria. Transaction judgement refers to the validity of a transaction's attributes before making the decisions. To verify the transaction, the author uses the benchmark experiment based on deep learning and confirms that the result of deep learning has similar accuracy as the Gaussian kernel SVM. For the comparison, the authors use five typical algorithms and change the parameters of deep learning for five times, such as activation function and dropout parameter [9].

### III. DEEP LEARNING TECHNIQUE FOR DETECT CREDIT CARD FRAUD

Deep learning is the state of the art technology that recently attracted the IT circle's considerable attention. The deep learning principle is an ANN that has many hidden layers. Conversely, non-deep learning feed forward neural networks have only a single hidden layer. The given picture shows the comparison between non-deep learning as in Fig. 1 and deep learning with hidden layers as in Fig. 2.

Now, we know about ANN, ML, and Deep Learning (DL). If these three words are metaphorically equated with the human body, they would be comparable as follows: artificial intelligence is like the body that contains the traits of intelligence, reasoning, communication, emotions, and feeling. ML is like one system that acts in the body, especially the visual system. Finally, deep learning is comparable to the visual signaling mechanism. It consists of a number of cells, such as retina that acts as a receptor and translates light signals into nerve signals. Now, we shall compare all the three categories with the human body.

Deep learning is a generic term used for multilayer neural network. Based on deep learning, there are many algorithms to implement such as AE, deep convolutional network, support vector machine, and others. One problem in selecting the algorithm to solve the problem is that the developer should know the real problem and what each algorithm in deep learning does. The three algorithms of deep learning that do unsupervised learning are RBM, AE, and the sparse coding model. Unsupervised learning automatically extracts the meaningful features of your data, leverages the availability of unlabeled data, and adds a data-dependent regularization for training.

In this study, we use AE for credit card fraud detection. AE has the input equal to the output in the hidden layer that has more or less the kind of input units depicted in the Fig. 3.

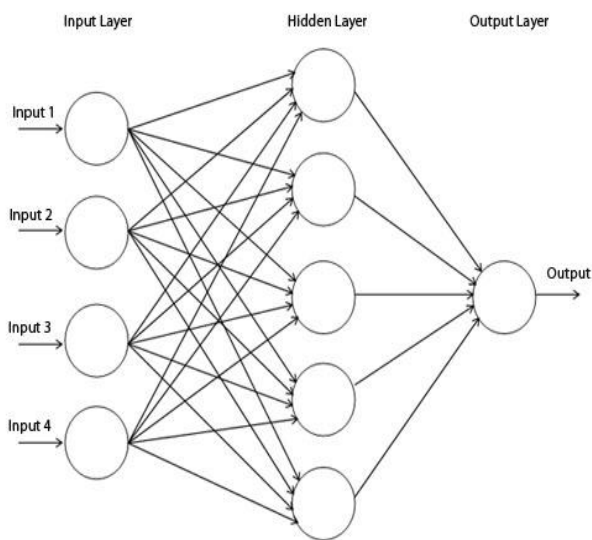


Fig. 1. Single layer hidden neural network [10].

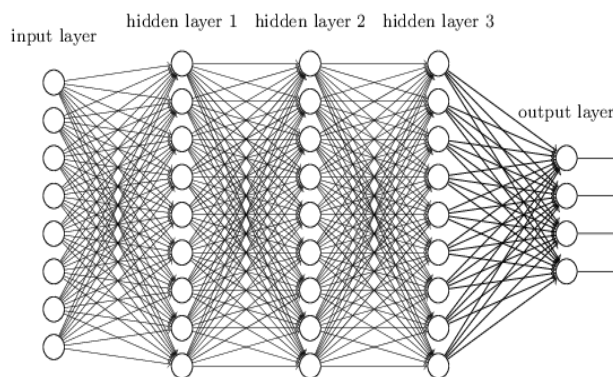


Fig. 2. Deep neural network [11].

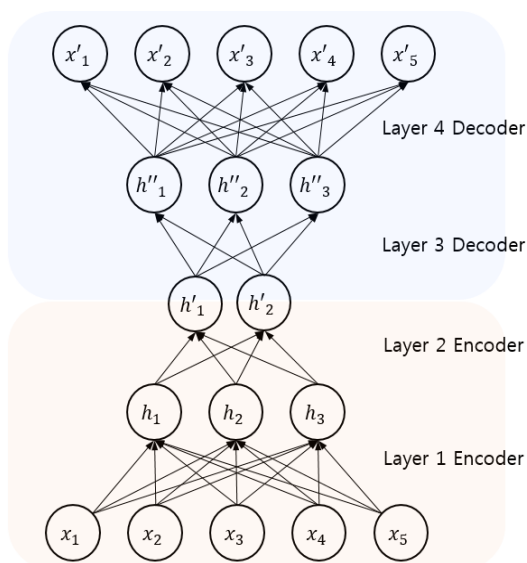


Fig. 3. Auto-encoder [12].

The equation of an encoder and a decoder are presented here:

Encoder

$$h(x) = g(a(x)) \\ = \text{sigm}(Wx) \text{ or} \\ = \text{tanh}(Wx)$$

Decoder

$$\hat{x} = o(\hat{a}(x)) \\ = \text{sigm}(W * h(x)) \text{ or} \\ = \text{tanh}(W * h(x))$$

In this study to implement AE, we use the hyperbolic tangent function or “tanh” function to encode and decode the input to the output. As a sample of a neural network, when we have already used the AE model, we should reconstruct the error by using backpropagation. Backpropagation computes the “error signal”, propagates the error backwards through network that starts at the output units by using the condition that the error forms the difference between the actual and desired output values. Based on the AE, we use parameter gradients for realizing backpropagation.

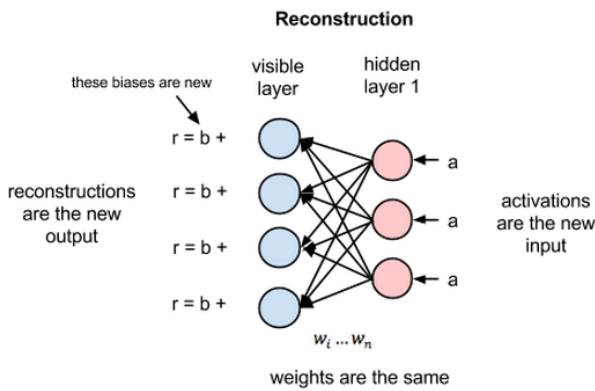


Fig. 4. Reconstruction of RBM [17].

behavior to be trained first, and then uses the new coming transaction as a validation test for the transaction. AE does not use labeled transactions to be trained, because it is unsupervised learning. RBM uses all transactions that transfer from acquiring bank as visible input and then that goes to the hidden node, and after the calculation of the activation function, the RBM reconstructs the model by transferring the new input from the activation function back to the output or visible function. As a conclusion of this in Fig. 5, if the transaction is fraudulent, the system will record this transaction as a fraud in the database and will then reject it. Next, the acquiring bank sends a SMS alert to the real consumer that the transaction has not been processed, because the system suspects the transaction as fraudulent.

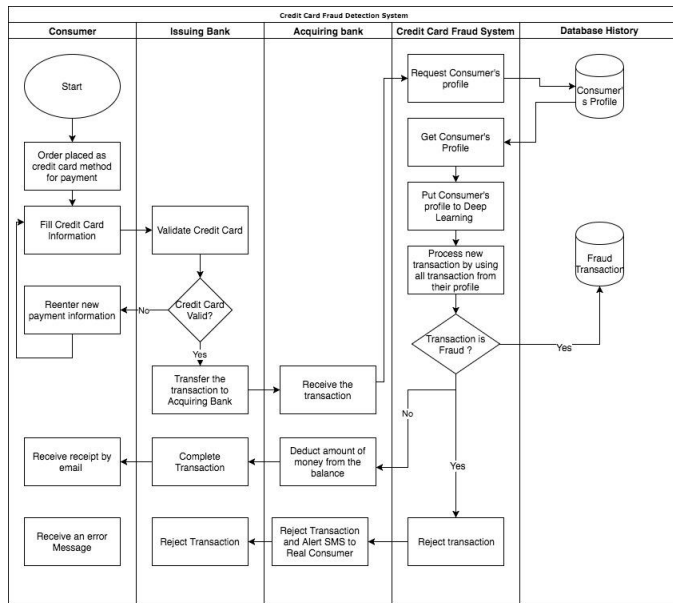


Fig. 5. Credit card fraud detection system using deep learning.

Another algorithm is RBM. There are two structures in this algorithm, visible or input layer and hidden layer. Each input node takes the input feature from the dataset to be learned. The design is different from other deep learning, because there is no output layer. The output of RBM is getting the reconstruction back to the input as shown in the picture below or Fig. 4. The point of RBM is the way in which they learn by themselves for data reconstruction; this is unsupervised learning.

Let us proceed to our design of credit card fraud detection system by using deep learning between AE and RBM in Fig. 5. First, the consumer orders the product via internet by using the credit card payment method. After that, the issuing bank sends the transaction to the acquiring bank by sending the amount of money, date and time of payment, location of internet usage, and more. Now, this is the credit card fraud detection system used to validate the behavior of credit card. As you can see, the credit card fraud system requests consumer's profile from the database to bring their behavior into the AE and RBM by using deep learning. Based on the AE, the acquiring bank transfers the input that is the amount of money, date and time, location of internet use, and other information. Then, the AE uses past

IV. COMPARATIVE FRAUD DETECTION TECHNIQUES

Before focusing on the study of AE and RBM, this paper would prefer to compare it with other techniques to show that deep learning is suitable for finding anomalous patterns against normal transactions in Table I.

TABLE I. COMPARISON OF FRAUD DETECTION TECHNIQUE

Fraud Detection Techniques	Advantage	Disadvantage
K-nearest Neighbor Algorithm	KNN method can be used to determine anomalies in the target instance and is easy to implement.	KNN method is suitable for detecting frauds with the limitations of memory.
Hidden Markov Model (HMM)	HMM can detect the fraudulent activity at the time of the transaction.	HMM cannot detect fraud with a few transactions.
Neural Network	Neural networks have learned the previous behavior and can detect real-time credit card frauds.	Neural networks have many sub-techniques. So, if they pick-up this which is not suitable for credit card fraud detection, the performance of the method will decline.
Decision Tree	Decision Tree can handle non-linear credit card transaction as well.	Decision Tree have many type of input feature, DT can be constructed using different induction algorithm like ID3, C4.5 and CART. So, the cons are how to bring up induction algorithm to detect fraud as well. DT cannot detect fraud at the real time of transaction.



Outlier Detection Method	Outlier detection detects the credit card fraud with lesser memory and computation requirements. This method works fast and well for large online datasets.	Outlier detection cannot find anomalies accurately like other methods.
Deep Learning	A key advantage of deep learning is the analysis and learning of a massive amount of unsupervised data. It can extract complex patterns [13].	Now, deep learning is widely used in image recognition. No information to explain the other domains is available. The library of deep learning does not cover all algorithms.

### V. PROPOSED METHOD

In this paper, we use Keras [15] as a high-level neural network API implemented by python. Another program that we implement in AE is H2O [16] package. We use the H2O package to find MSE, RMSE, and variable importance across each attribute of the datasets. Conversely, we used Keras in parallel processing to get AUC and confusion matrix. Both frameworks, we coded in python on Jupyterlab.

Before we could develop the program AE by using Keras API and code the program AE by using H2O, the datasets needed to be cleansed. As we know, the German credit card data set and the Australian dataset classified characteristics for each attribute. You can see the details of these attributes in [3], [4].

This is the step of cleansing data.

1) Classified the data into a number of classifications such as attribute 4 (qualitative) purpose

- A40: Car (new)
- A41: Car (used)
- A410: others

We transform it to the number of classifications, such as A40 = 1, A41 =2, ..., A410 = 10 and so on.

2) After obtaining the classification for each attribute, we transform those classifications into PCA by using XLSTAT [14].

TABLE II. AUTOENCODER MODEL USING KERAS

```

Input_Dimension = Training.shape[1]
Hiddenlayer = 16
Input_layer = Input(shape=Input_Dimension,)
Encoder1 = Dense(Hidden_layer,activation="tanh")(Input_layer)
Encoder2 = Dense(Hidden_Layer/2,activation="tanh")(Encoder1)
Encoder3 = Dense(Hidden_Layer/4,activation="tanh")(Encoder2)
Decoder1 = Dense(Hidden_Layer/4,activation="tanh")(Encoder3)
Decoder2 = Dense(Hidden_Layer/2,activation="tanh")(Decoder1)
Decoder3 = Dense(Input_Dimension,activation="tanh")(Decoder2)
AutoEncoderModel = Model(inputs=Input_layer,outputs=Decoder3)
    
```

TABLE III. AUTOENCODER MODEL USING H2O

```

Autoencoder =
h2o.estimators.deeplearning.H2OAutoEncoderEstimator(hidden=hidden_stru
cture, epochs=200, activation='tanh', autoencoder=True)
Autoencoder.train(x = Input,
                  Training_frame = data_set)
Print(Autoencoder)
    
```

In the Keras method, we designed 6 hidden layers by having 3 encoders and 3 decoders. In each hidden layer, we designed the following units:

- Input : 21 attributes or 21 Input
- Encoder1 (H1) : 16
- Encoder2 (H2) : 8
- Encoder3 (H3) : 4
- Decoder3 (H1) : 4
- Decoder2 (H2) : 8
- Decoder3 (H3) : 21
- Output : 21

As mentioned above, every hidden layer we used was the “Tanh” activation function. In Keras, there are many activation functions to implement. Based on the experiment, we used “Tanh” function, because it achieves a high level of AUC. We divide the train and test with 80 and 20 percentage of data by using normal transactions to predict fraudulent transactions.

This is an example of Python Coding in Keras as in Table II.

As you can see, in Keras API, we need to build our model by preparing the command ourselves. Conversely, in the H2O package, we use the command of AE in Table III.

Base on methodology of our research, we coded in Python and then we used Area of Under Curve to identify the success rate of the model. If the percentage of AUC is high then mean that we found unsupervised learning rate with true positive rate on our model. Conversely, some datasets that has less amount of data will get more false positive rate because they has not much data to be trained.

### VI. EVALUATE THE RESULT

These are the result of the German Dataset show in Fig. 6, 7 and 8; as we mentioned above that the Dataset was divided for training and testing in a ratio 80:20 by using the normal labeled transactions in the column “Creditability” to find anomalous patterns. These form the AUC and confusion matrix.

This form the MSE and RSME from H2O the package of the German Dataset.



Fig. 6. AUC of German Dataset by using AE.



Fig. 9. AUC of Austrian Dataset by using AE.

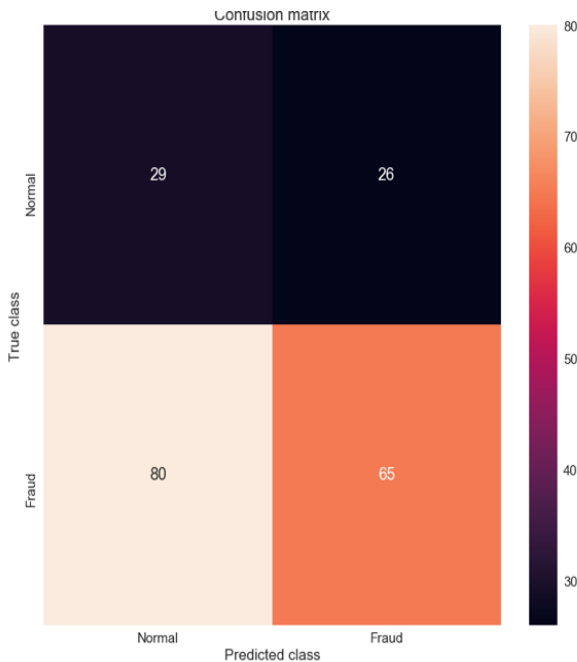


Fig. 7. Confusion Matrix of German Dataset by using AE.

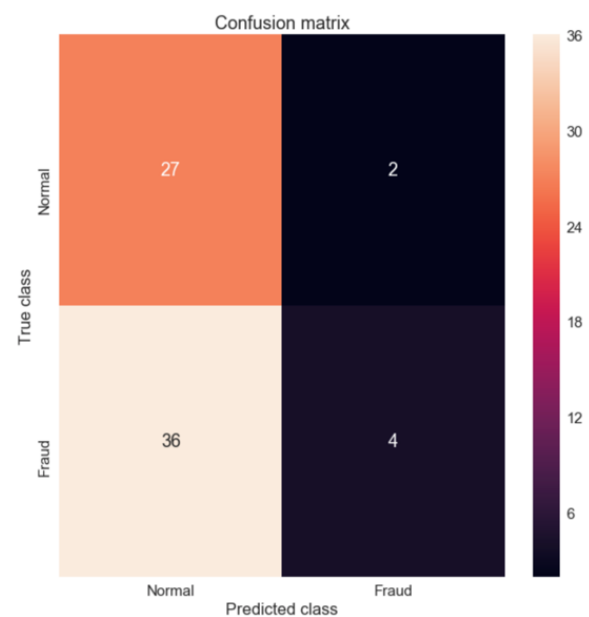


Fig. 10. Confusion Matrix of Australian Dataset by using AE.

**Model Details**

```
=====
H2OAutoEncoderEstimator : Deep Learning
Model Key: DeepLearning_model_python_1513057801244_1
```

```
ModelMetricsAutoEncoder: deeplearning
** Reported on train data. **
```

```
MSE: 0.00129114108292
RMSE: 0.0359324516687
```

Fig. 8. AE Model of deep learning report of German Dataset on H2O framework.

Let us move on to another dataset, the Australian Dataset. The AUC result is given, and the confusion matrix from Keras. The results are shown in Fig. 9, 10 and 11.

**Model Details**

```
=====
H2OAutoEncoderEstimator : Deep Learning
Model Key: DeepLearning_model_python_1513081686672_1
```

```
ModelMetricsAutoEncoder: deeplearning
** Reported on train data. **
```

```
MSE: 0.000604421565799
RMSE: 0.0245849865934
```

Fig. 11. Auto Encoder Model deep learning report of Australian Dataset based on H2O framework.

This is the Australian Dataset's MSE and RSE obtained by running the H2O package.

Here, we move on to the large dataset, the European Dataset with 284, 807 transactions. The results are shown in Fig. 12, 13 and 14.

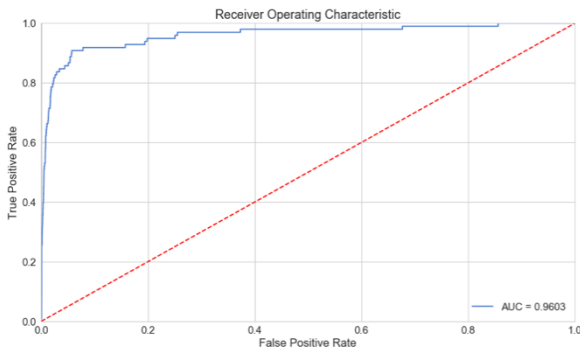


Fig. 12. AUC of European Dataset by using AE.

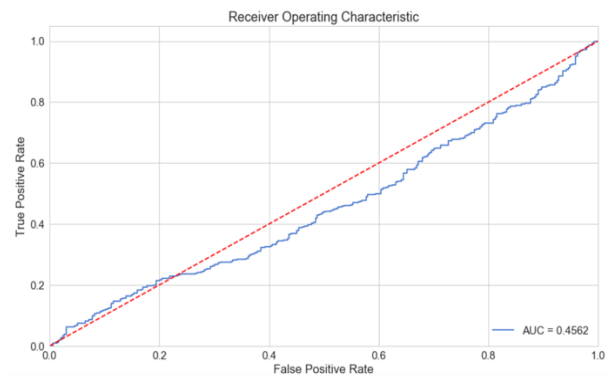


Fig. 15. AUC of German Dataset by using RBM.

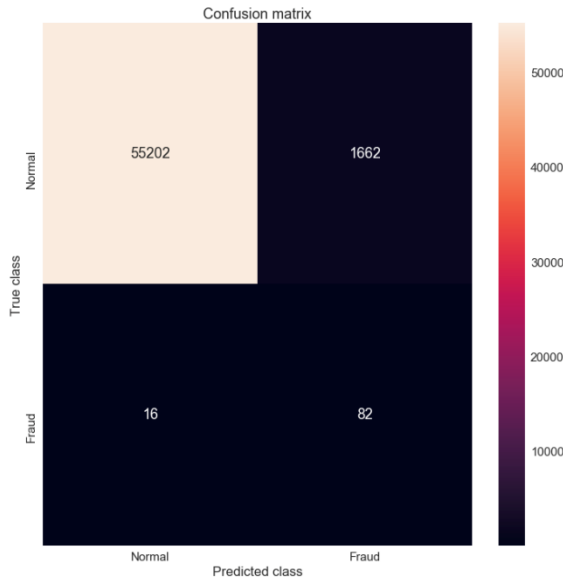


Fig. 13. Confusion Matrix of European Dataset by using AE.

```

Model Details
=====
H2OAutoEncoderEstimator : Deep Learning
Model Key: DeepLearning_model_python_1513687461083_1

ModelMetricsAutoEncoder: deeplearning
** Reported on train data. **

MSE: 1.30914842402e-05
RMSE: 0.0036182156155
    
```

Fig. 14. Auto Encoder Model deep learning report of European Dataset based on H2O framework.

As summarized by three datasets, there is lesser data in the German and Australian datasets. So, when we find anomalies in fraud detection, we obtain a lower of AUC, because we trained the systems for a small number of data and validated the test data for a lesser amount. Conversely, when we apply this AE model based on Keras with a large amount, the European Dataset, we got AUC of 0.9603. AE is suitable for large datasets.

Further RBM's results based on the three datasets are presented: we begin by explaining the German Dataset in Fig. 15. As you can see, the AUC of German Dataset is 0.4562.

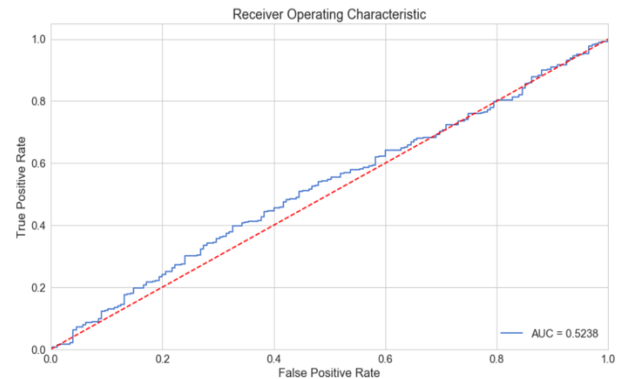


Fig. 16. AUC of Australian Dataset by using RBM.

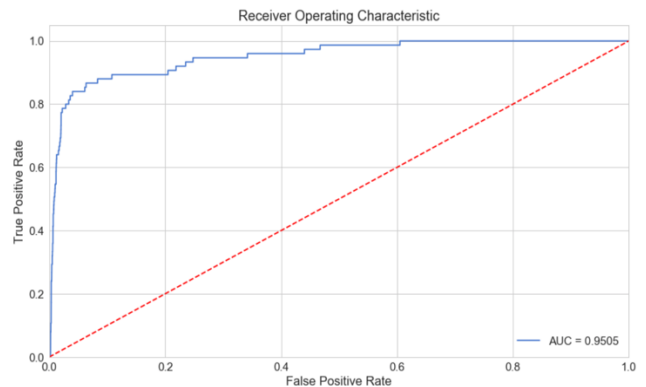


Fig. 17. AUC of European Dataset by using RBM.

The graph shows the result of the Australian Dataset by using the RBM algorithm to implement in Fig. 16. The AUC score is 0.5238.

While the biggest dataset is the European Dataset that produced an AUC value greater than the other two datasets shown above (Australian and German Dataset). The AUC score of European dataset is 0.9505 which can be seen in Fig. 17.

This is the summary of AUC's score that implemented AE and RBM of three different datasets.

From this research, we can conclude that AE and RBM produce high AUC score and accuracy for bigger datasets,

because there is a large amount of data to be trained. You can see the details of AUC's score in Table IV.

TABLE IV. COMPARISON AUC'S SCORE BETWEEN THREE DATASETS

Dataset Name	No. of transactions	AUC's score based on AE	AUC's score based on RBM
German Dataset	1000	0.4376	0.4562
Australian Dataset	690	0.5483	0.5238
European Dataset	284, 807	<b>0.9603</b>	<b>0.9505</b>

Based on two popular datasets, we can conclude that supervised learning dataset is suitable for history database for credit card fraud detection. Supervised learning such as multilayer perceptron in neural network that uses the prediction algorithm to identify whether new transactions are legal or illegal. When a credit card used, the neural network based on the fraud detection system checks for the pattern used by the fraudster and corroborates the pattern in question or checks for attributes that have been determined as illegal; if the pattern matches with genuine transaction behavior, then the transaction is considered legitimate. Conversely, unsupervised learning entails knowing about normal transactions and finding anomalous patterns, and then, responding in real-time to the system as a fraud or legal transaction.

## VII. CONCLUSION AND FUTURE WORK

Nowadays, in the global computing environment, online payments are important, because online payments use only the credential information from the credit card to fulfill an application and then deduct money. Due to this reason, it is important to find the best solution to detect the maximum number of frauds in online systems. AE and RBM are the two types of deep learning that use normal transactions to detect frauds in real-time. In this study, we focused on ways to build AE based on Keras, RBM, and H2O. To verify our proposed methods, we used benchmark experiments with other tools to confirm that AE and RBM in deep learning can accurately achieve credit card detection with a large dataset such as the European Dataset. Although, for these experiments, it will be better to use real credit card fraud transactions with a huge amount of data. We guarantee that AE and RBM can make more accurate AUC for receiver operator characteristics than that observable from the results from the European Dataset.

## REFERENCES

- [1] CyberSource. (2017, Nov. 29). *2017 North AMERCA edition, online fraud benchmark report persistence is critical* [Online]. Available: [http://www.cybersource.com/content/dam/cybersource/2017\\_Fraud\\_Benchmark\\_Report.pdf?utm\\_campaign=NA\\_17Q3\\_2017%20Fraud%20Report\\_Asset\\_1\\_All\\_Auto&utm\\_medium=email&utm\\_source=Eloqua](http://www.cybersource.com/content/dam/cybersource/2017_Fraud_Benchmark_Report.pdf?utm_campaign=NA_17Q3_2017%20Fraud%20Report_Asset_1_All_Auto&utm_medium=email&utm_source=Eloqua)
- [2] L. Seyedhossein and M. R. Hashemi, "Mining information from credit card time series for timelier fraud detection," in 5th International Symposium on Telecommunications (IST'2010), 2010 © IEEE. doi: 978-1-4244-8185-9/10/\$26.00
- [3] UCI Machine Learning Repository. (2017, Nov. 29). *Stalog (Australian credit approval) dataset* [Online]. Available: [http://archive.ics.uci.edu/ml/datasets/Statlog+\(Australian+Credit+Approval\)](http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval))
- [4] UCI Machine Learning Repository. (2017, Nov. 29). *Stalog (German credit data) dataset* [Online]. Available: <http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>
- [5] N. Malini and Dr. M. Pushpa, "Analysis on credit card fraud identification techniques based on KNN and outlier detection" in 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEEICB17).
- [6] A. Charleonnan, "Credit card fraud detection using RUS and MRN algorithm," in The 2016 Management and Innovation Technology International Conference (MITiCON-2016), 2016 © IEEE. doi: 978-1-5090-4105-3/16/\$31.00
- [7] M. K. Mishra and R. Dash, "A comparative study of chebyshev functional link artificial neural network, multi-layer perceptron and decision tree for credit card fraud detection" in 2014 13th International Conference on Information Technology, 2014 © IEEE. doi: 978-1-4799-8084-0/14 \$31.00
- [8] Y. Pandey, "Credit card fraud detection using deep learning" *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5, May–Jun. 2017.
- [9] A. Niimi, "Deep learning for credit card data analysis," in World Congress on Internet Security (WorldCIS-2015), 2015 © IEEE. doi: 978-1-908320-50/6 \$31.00
- [10] Single Hidden Layer Neural Network [Online]. Available: <http://nicolamanzini.com/single-hidden-layer-neural-network/>
- [11] Chapter 6 (2017, Aug. 4). *Deep learning* [Online]. Available: <http://neuralnetworksanddeeplearning.com/chap6.html>
- [12] Introduction Auto-encoder (2015, Dec. 21). *Auto-encoder* [Online]. Available: <https://wikidocs.net/3413>
- [13] M. M. Najafabadi et al., "Deep learning applications and challenges in big data analytics," *J. Big Data*. doi: 10.1186/s40537-014-0007-7
- [14] *XLSTAT your data analysis solution* [Online]. Available: <https://www.xlstat.com/en/>
- [15] *Keras the python deep learning library* [Online]. Available: <https://keras.io/>
- [16] *H2O API* [Online]. Available: <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html>
- [17] *A beginner's tutorial for restricted Boltzmann machine* [Online]. Available: <https://deeplearning4j.org/restrictedboltzmannmachine#define>
- [18] *Credit card fraud detection anonymized credit card transaction labeled as fraudulent or genuine* [Online]. Available: <https://www.kaggle.com/dalpozz/creditcardfraud/data>

# Voice Detection in Traditionnal Tunisian Music using Audio Features and Supervised Learning Algorithms

Wissem Ziadi, Hamid Amiri

Signal, Images, Information Technologies (LR-SITI)  
Tunisian National School of Engineering, ENIT  
Tunis, Tunisia

**Abstract**—The research presented in this paper aims to automatically detect the singing voice in traditional Tunisian music, taking into account the main characteristics of the sound of the voice in this particular music style. This means creating the possibility to automatically identify instrumental and singing sounds. Therefore different methods for the automatic classification of sounds using supervised learning algorithms were compared and evaluated. The research is divided into four successive stages. First, the extraction of features vectors from the audio tracks (through calculation of the parameters of sound perception) followed by the selection and transformation process of relevant features for singing/instrumental discrimination. Then, using learning algorithms, the instrumental and vocal classes were modeled from a manually annotated database. Finally, the evaluation of the decision-making process (indexing) was applied on the test part of the database. The musical databases used for this study consists of extracts from the national sound archives of Centre of Mediterranean and Arabic Music (CMAM) and recordings made especially for this research. The possibility to index audio data (classify/segment) into vocal and instrumental recognition allows for the retrieval of content-based information of musical databases.

**Keywords**—Tunisian voice timbre; audio features extraction; singing voice detection; sung/instrumental discrimination; supervised learning algorithms

## I. INTRODUCTION

Faced with the increasing availability of sound data broadcasted online, the importance to search these immense volumes of data by musical, sound and information-based content have become apparent for archiving and classification purposes. Several multidisciplinary researches have been carried out for this purpose, notably on the description of sound and musical contents. Among others the QUAERO<sup>1</sup> project, a collaborative industrial research and innovation program addressing automatic analysis and enrichment of digital, multimedia and multilingual content.

This project gathers 32 French and German partners. The CUIDADO<sup>2</sup> project from IRCAM is another major development scheme for the description of audio-visual content (the MPEG-7 standardization process).

In Tunisia, the Telemeta<sup>3</sup> platform [1] is being developed in the national sound archives of the Centre of Mediterranean and Arab Music<sup>4</sup> (CMAM). The project is conducted under the direction of the Ethno-Musicology Research Centre in France. Telemeta is a collaborative platform for CNRS sound archives to analyse, identify and index digital sound resources.

In a piece of music, the singing voice is the main element. She carries the message. In order to automatically detect the sung voice in an audio stream, multiple methods have been implemented so far, but never on a corpus of traditional Tunisian music. This research explores different methods using supervised learning algorithms to extract the vocals in this particular music: K-nearest neighbours, the support vector machine and the Gaussian mixture model. The method chosen to locate the sung voice in an audio stream is inspired by the systems developed by Tong Zhang [2] and Peeters Geoffroy [3], [4]. This is a statistical method based on two phases: a learning phase and a classification phase. This process of supervised classification is based on a succession of four stages: First the extraction of relevant audio features [5], [6]. Followed by the selection and transformation of these feature vectors to minimize redundancy and reduce the dimensions. Then a modelling procedure of the instrumental and vocal classes is needed using learning algorithms [5], [7], [8]; finally, ending with an indexing phase. The training phase was carried out on a training database. A second database was used for the test phase (indexing). The extracts of Tunisian music used during this research were obtained from the funds of National Phonetics collected and saved by the National Archive of the Centre for Arab and Mediterranean Music. Another database was composed of live recordings and recordings made specifically for this study at Ixir studio. This was done in order to compare the outcome of different recording techniques and acoustic elements. The whole corpus was manually annotated indicating the sung and instrumental parts in the audio tracks.

This paper includes, in the first section, the musical context and a presentation of the audio features used for sung/instrumental discrimination. Then, in the second section, it presents a few methods for dimension reduction and a brief description of the supervised learning algorithms aiming to model the two instrumental and singing classes. The final

<sup>1</sup> www.quaero.org

<sup>2</sup> anasynth.ircam.fr/home/english/projects/cuidado

<sup>3</sup> telemeta.org

<sup>4</sup> phonotheque.cmam.tn/

section is reserved to the experimental results including corpus, post-treatments and evaluation.

## II. MUSICAL CONTEXT AND AUDIO FEATURES

There are different styles of traditional Tunisian music, including Soufi, Mezoued, Stambeli, Salhi combining spirituality, poetry, festivity and religion. The best known style though is Mâlouf. It is influenced by Arabic poetry brought to Tunisia by muslim-Andalusian immigrants in the 13<sup>th</sup> and 14<sup>th</sup> century. It is organized by quartertones following a classical Arab mode called ‘maqâm’ and carries Berber and Turkish elements in its rhythm. It features instruments like the violin, various percussion instruments, the ‘Ud’, and flutes, but it is essentially carried by the sung voice. The acoustic characteristics and timbre of the singing voice are genre-specific and distinguished by its harmonic sound and its vibration [9]. Styles differ greatly from region to region, each presenting a range of acoustic characteristics specific to each environment, instrumental technique and peculiarity of the vocals.

The first stage of the research tried to identify the typical aspects that characterize the sung voice in traditional Tunisian music. The analysis of the voice was carried out through the visualization of the items extracted from the sound signal (using VAMP plugin with Sonic Visualizer). Each detail is capable of describing the precise behaviour of an analysed signal. A multitude of features have been proposed in the literature, both about the field of speech processing as well as the classification of sounds, in our case singing vs. instrumental discrimination. The aspects best capable of distinguishing the singing voice from the rest of a musical sound stream of the Tunisian sung repertoire were identified. The most relevant features are: auto-correlation, Zero Crossing Rate (ZCR), Spectral Centroid, Mel -frequency Cepstral Coefficients (MFCC) and Harmonic Pitch Class Profiles (HPCP). Each feature describes one or more acoustic characteristics of sound and will be used by learning algorithms to establish a model for both instrumental and vocal classes [5], [7], [10].

### 1) Auto-correlation

Auto-correlation is used to compare the time lag of a signal with a delayed copy of itself. A periodic signal is perfectly correlated with itself if the delay time is the same as the duration of the signal. Autocorrelation is a relevant parameter for describing the mechanism of human listening, hence its relevance in the process of sound differentiation.

### 2) Zero Crossing Rate (ZCR)

This feature consists in locating the number of times the signal changes sign during a given time interval (in seconds), from negative to positive and vice versa. ZCR is the most relevant feature in voice / noise classification in the process of speech recognition and music information retrieval. The smaller the ZCR, the closer the sample is to the human voice.

### 3) Spectral centroid

This feature indicates the center of gravity of an audio signal. It is calculated as the weighted average of the frequencies present in the signal, determined using a Fourier transformation with their magnitudes as weight. The spectral

centroid is used to estimate the brightness of a sound. It is a key aspect in describing the musical timbre.

### 4) Mel-Frequency Cepstral Coefficients (MFCC)

The MFCC is the most used feature in existing methods of automatic speech recognition and sound indexing. Bridle and Brown were the first to use it in 1974. The Mel-Frequency Cepstral Coefficients is able to simulate part of the speech production and perception. More exactly, the MFCC is a logarithmic representation of the loudness and pitch of a sound.

### 5) Harmonic Pitch Class Profiles (HPCP)

HPCP is a set of qualities commonly used for the recognition and identification of string instruments in an audio signal. This feature presents a sequence of vectors that describe the distribution of pitches in a single octave, specifying their tones and intensities according to a distribution over 12 temperate ranged classes. With HPCP you can determine the key of a song, search by similarity or, as in our case, index and classify sound signals.

## III. SINGING VOICE DETECTION

### A. Dimension Reduction

In this experimental study, audio features were extracted with a sliding window called “Hamming window” [11] with a size of 50 ms, the hop size set to 25 ms. Temporal modelling was applied using the mean and variance values over a 2s window with a hop size of 1s. After the extraction of features from an audio signal, the selection and transformation of the most relevant features was performed to describe the two classes (vocal and instrumental). A very large amount of features can cause confusion of the class models during the course of the experiment. Therefore the algorithms that select the features should be able to detect a minimal set of relevant (informative and meaningful) features in relation to the class models, avoiding a redundancy of data. The most popular filter type selection algorithm is the inertia ratio maximization using feature space projection (IRMFSP) algorithm [12]. A features space transformer was then applied. It aims to reduce the size of the features after the features selection phase. Several types of transformations are presented in the literature, such as:

- Box-Cox transformation which reduces the size of the features space while preserving the variance.
- The linear discriminant analysis (LDA) transformation reduces the space of audio features by maximizing class separation. LDA is based on predicting the belonging of a feature to a class depending on its characteristics measured using predictive variables.
- The principal component analysis method (PCA) extracts the principal component by a linear transformation, computed using singular value decomposition algorithms [13]. A new set of features are ordered according to their importance. This procedure consists of five steps: First, subtracting the mean from each audio feature vectors (mean is zero). Then, calculating the covariance matrix. After that, computing the eigenvalues and eigenvectors of the

covariance matrix. Then, order eigenvalue in descending order (the number of eigenvectors is equal to the dimension of audio features vectors). Finally, derive the new audio feature vectors (multiply the transpose of the audio feature vector to the left of the original data set. Final Features Vectors = RFE \* RDM. Where RFE is the matrix with the eigenvectors (columns transposed) and RDM is the matrix mean-adjusted data transposed.

### B. Supervised Learning Algorithms

The supervised learning phase of this system aims to model the two instrumental and vocal classes and subsequently permitting automatic classification of an unknown sound. The models are based on the information provided by the audio features vectors and manual annotations of the database, indicating the instrumental and vocal parts of each audio track. The most efficient methods chosen for the learning phase are:

#### 1) K-Nearest Neighbors (KNN)

This method is considered the simplest and most popular of the supervised learning algorithms for the classification and automatic indexing of sounds [12], [14]. In our case, the data devoted to learning are formed by a set of vectors of audio features. Each sample contains a class label (instrumental or vocal) and is recorded in memory during the learning phase. During the indexing phase, the test samples are classified by assigning the class tags using the closest learning sample.

By definition, the implementation of this method results in two fundamental questions: the choice of neighbourhood (the value of K) and what distance to take into account. The results obtained by the method depend on these two criteria. We can arrive at totally different results depending on the choices made.

The most used method to determine the similarity between the samples is the Euclidean distance.

#### 2) Support vector machine (SVM)

This method makes it possible to use the samples closest to the separation boundary, assuming that it provides the most useful information for the classification. These are called support vectors. SVMs were developed in the 1990s based on Vladimir Vapnik's theoretical considerations on the development of a statistical theory of learning: the Vapnik-Chervonenkis theory [15], [16].

This theory seeks to maximize the margin between the separation boundary and the closest samples. The problem then lies in finding the optimal decision surface for the separation between classes, and subsequently predicting which class a test sample belongs to.

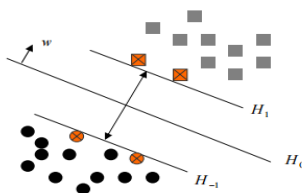


Fig. 1. SVM model.

In our case the issue is typical: the classification into two classes: instrumental and vocal, in which the samples (or feature vectors) are linearly separable. The separation boundary is used to classify a frame to one of two classes.

It is named in Fig. 1 the optimal hyper plan  $H_0$ .

We note the function  $f$  that has an input vector  $x \in \mathcal{R}^n$  and matches an output  $y$ :  $f(x) = y$ . In our case (question of two-class discrimination)  $y \in \{-1,1\}$ . The input vector  $x = (x_1 \dots x_n)$  and a weight vector  $w = (w_1 \dots w_n)$ . Which gives:

$$f(x) = w \cdot x + b \quad (1)$$

$H_0$  is the region of the vector  $x$  that checks the equation  $f(x) = 0$ .  $H_1$  and  $H_2$  are two parallel hyper plans to  $H_0$  and which are defined respectively by  $f(x) = 1$  and  $f(x) = -1$ .

The distance between the two hyper plans  $H_1$  and  $H_2$  is  $\frac{2}{\|w\|}$

The intention is to maximize the margin. The decision hyper plan  $H_0$  depends directly on the vectors closest to the two hyper plans  $H_1$  and  $H_2$  who we call support vectors.

#### 3) Gaussian mixture model GMM

For its ability to approximate the global distribution of the features collection for each class [7], [8], the Gaussian Mixture Model was used. A GMM is used to model the distribution of data in the features space at D dimensions. This space is obtained from the weighted sum of N probability density function (pdf).

The probability to observe the feature vector  $x$  knowing its GMM is defined by the parameters  $\lambda = \{\mu, \Sigma\}$  resulting in:

$$p(\vec{x} | \lambda) = \sum_{i=1}^N p_i b_i(\vec{x}) \quad (2)$$

Where  $\vec{x}$  is a D dimension of feature vector and  $p_i$ ,  $i = 1 \dots N$ , is the weight associated with each GMM.  $b_i(\vec{x})$ ,  $i = 1 \dots N$ , is the probability density, which can be written as follows:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} (\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i)\right\} \quad (3)$$

In the above equation  $\vec{\mu}_i$  and  $\Sigma_i$  represent the estimated means vector and the covariance matrix. The mixture weight satisfies the constraint  $\sum_{i=1}^N p_i = 1$ . The parameters of the GMM are the mean vector, the covariance matrix and the mixture weight. They can be represented as follows:

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}; i = 1, \dots, N. \quad (4)$$

These parameters will be initialized by the classical K-means algorithm and then iteratively adjusted by the Expectation-Maximization algorithm [17]. The model of the Gaussian mixture of the vocal class was noted as «  $\lambda_c$  » and that of the instrumental class as «  $\lambda_{In}$  ». These two models will be trained by the learning algorithm from the manual annotations of the database, distinguishing the vocal parts from the instrumental parts.

During the test phase, the classifier takes as input the features vectors of  $T_{x-frame}$   $X = (x_1, x_2, \dots, x_{T_x})$ , extracts from an audio track frame in the test data base, and outputs the log likelihood of each frame  $\log p(x_t | v_c)$  and  $\log p(x_t | v_{In})$ ;  $1 < t < T_x$ , for the vocal and Instrumental

GMM. Each frame was assigned to the corresponding class according to the logarithmic probabilities. Depending on the choice of analysis interval, there are many variants and combinations that had to be taken into account during the decision making process [8]. In this study, decisions were made based on a fixed segment:

If  $\log p(x_t|v_c) > \log p(x_t|v_{in})$ , the frame is classified as vocal. Consequently if  $\log p(x_t|v_c) \leq \log p(x_t|v_{in})$ , the frame is classified as instrumental.

#### IV. EXPERIMENTAL STUDY

##### A. Corpus

The research corpus is composed of two databases; the CMAM database and the Ixir database. All tracks in both databases are 16-bit coded with a sampling frequency of 44.1 kHz. Both databases were annotated manually, dividing them into two classes, vocal and instrumental. The learning process was done on 2/3 of the database, called the training database. The test phase was conducted on the 1/3 of the database left, called the test database.

- The CMAM database: Was extracted from the Funds of National Phonetics, collected and saved by the Centre of Mediterranean and Arab Music (CMAM). In this centre, Tunisian phonographic collections (collected since the beginning of the 20th century) are catalogued and indexed. Access to the music records is obtained through a Telemeta platform. Our research database is composed of 403 tracks in wav format and corresponds to different Tunisian music styles and epochs. The duration of the extracts is between 2 and 60 seconds. All tracks combined represent about 3:50 hours (210 min) of music. Table I gives a description of the distribution.

TABLE I. DESCRIPTION OF THE DISTRIBUTION OF THE CMAM DATABASE

	Vocal	Instrumental	Total
Training	155	139	294
Test	56	53	109
Total	211	192	403

- The Ixir database: has been partly recorded in the studios of “el Xir Labs – studio” at the Centre for Music and Sound Studies in Tunis and partly live specifically for this research. We did so in order to compare the outcome of different recording techniques and acoustic environments. This database consists of 186 tracks. We opted for a variety of styles from the Tunisian repertoire. The covers are performed in voice and lute (‘ud’) by Wissem Ziadi. The audio tracks are between 2 and 60 seconds long, which makes the set about 1h30 (90 min) of music. Table II gives a description of the distribution of the Ixir database:

TABLE II. DESCRIPTION OF THE DISTRIBUTION OF THE IXIR DATABASE

	Vocal	Instrumental	Total
Training	62	62	124
Test	31	31	62
Total	93	93	186

##### B. Tests and Evaluation

The tests were done on a standard PC; Intel core i4, 2.3 Gh, 4 GB of RAM. Many simulations have been performed to evaluate the performance of the classification methods. All these simulations were made in python. The goal of this system is to assign each unknown analysis track from the test database to either vocal or instrumental.

###### 1) Post-treatment

Due to the use of a very short decision window (50 ms), much less information is used for the calculation of each feature. This necessarily implies a greater variability of the estimated results. To reduce the noise and to refine the results after assigning the class probabilities (abrupt or accidental values), several types of filters are proposed in the literature [18], [19]. A median filter and a smoother using a hidden Markov model were applied. The median filter is a nonlinear filter. Therefore a median is a value  $m$  which serves to partition a set of values in two equal parts; on one side the smaller ones, on the other the bigger ones. In our case, the set of values is a probability distribution. The median can only be the value for which the density function is 0.5. The median filter applies to a distribution of  $N$  odd values to find the median that divides the samples into two equal groups. The application of the median filter brought a remarkable improvement in the performance of the classification system, but it remained blind to the nature of the classes. Therefore the Markov hidden model process was applied. The hidden Markov model smoothing process presented in this experiment is inspired by the famous Rabiner tutorial [20] and implemented in practice by Ramona [16]. Here, the purpose of the Markov model is to present the transitions between a set of states (the acoustic classes: vocal and instrumental).

###### 2) Résultats

###### a) Cross validation method

For each classification method, recall, precision and F-score were measured. The recall is the fraction of vocal frames (or instrumental) existing on all found frames. It is a measure of sensitivity. Precision is the fraction of the vocal frames found on all the vocal frames of the database. It is a measure of confidence.

$$F\text{-score} = 2 (\text{Recall} * \text{Precision}) / \text{Recall} + \text{Precision}.$$

This is the measure of competence.

For the evaluation of this system the cross validation method was used.

Table III first shows that the GMM model, applied to both databases, is the most reliable model for instrumental / vocal discrimination. The precision rate the GMM method has given ranges from 89.6% with the CMAM database up to 95.8% with the Ixir database. The Ixir database gave the best results with all three classification methods. This could be explained by the difference in the acoustic nature (recording techniques) of the records between the CMAM and Ixir databases, the complexity of the models and the large number of parameters related to the number of samples.



TABLE III. TEN-FOLD CROSS VALIDATION VOCAL CLASSIFICATION RATE OF KNN, SVM AND GMM METHODS FOR THE TWO DATABASES

	Database	
	CMAM	Ixir
<b>KNN</b>	81.4%	91.5%
<b>SVM</b>	74.4%	88.3%
<b>GMM</b>	89.6%	95.8%

The different values of the recall, precision and F-score factor should be analysed. These values depend directly on the distribution of the entire database. Running the CMAM database with the SVM learning model for example, F-score=74.4%. We got 94.4% as recall for the vocal class. While the precision rate is only 61.4%. This seems very unsatisfactory as a significant part of the vocal frames are classified as instrumental. But, in reality, this is due to the imbalance of the data composition between both databases since they don't have the same percentage of sung frames compared to instrumental frames. The CMAM database consists of 22620 frames belonging to the vocal class and 75875 frames belonging to the instrumental class. While the Ixir DB is composed of 7512 sung frames compared to 4375 instrumental frames.

*b) Cross database validation*

In this part of the test, a cross database validation method was applied. With this method the generality of the established classification systems is tested and checked if the system learned the general and specific acoustic characteristics of the sound signal. The procedure consists of starting the supervised learning process with the first database and later launching the test phase with the second and vice versa. The results of this test are given in the tables below.

Tables IV, V and VI show the classification results. They are declining and do not exceed 67.5% with the Gaussian mixture model. This is a fairly logical result due to the differences between the acoustic characteristics of both CMAM and Ixir records. The extracts retrieved from the CMAM database are recordings registered since the beginning of the 20th century. Hence the quality of the recording and post processing of the sound varies greatly from one era to another. While the Ixir database is registered under technically optimal conditions for this study. These differences imply a mutation of the extracted features from the audio frames and their distributions and therefor generate confusion in the classification process.

TABLE IV. CROSS-DATABASE VALIDATION (CMAM AND IXIR DATABASE) - KNN ETHOD

		Test database	
		CMAM	Ixir
<b>Training Database</b>	<b>CMAM</b>		61.2%
	<b>Ixir</b>	60.5%	

TABLE V. CROSS-DATABASE VALIDATION (CMAM AND IXIR DATABASE) - SVM ETHOD

		Test database	
		CMAM	Ixir
<b>Training Database</b>	<b>CMAM</b>		56.5%
	<b>Ixir</b>	59.3%	

TABLE VI. CROSS-DATABASE VALIDATION (CMAM AND IXIR DATABASE) - GMM ETHOD

		Test database	
		CMAM	Ixir
<b>Training Database</b>	<b>CMAM</b>		65.7%
	<b>Ixir</b>	67.5%	

V. CONCLUSION

This article presented the search to find the technical means and the appropriate tools to automatize the detection of the voice in traditional Tunisian music. An analytic spectrum of an audio signal to identify relevant audio features for vocal/instrumental discrimination was established. Then selections and transformations algorithms to minimize redundancy and reduce the dimensions of the features vectors space extracted from audio tracks were implemented. The KNN, SVM and GMM methods were used to model the singing and instrumental classes in an approach that is based on firstly a learning and then a testing phase. The experimental results show that the supervised learning algorithm based on the Gaussian mixture models GMM have the best precision rate, resulting in 95.8% accuracy with the Ixir database. The Ixir database gave the best results with all three classification methods.

During this study some problems were encountered that distorted the results, like the complexity of the statistic models, the difference in the acoustic nature of the audio records and the large amounts of parameters in the classification process, related to the number of samples.

Despite these complications, the results were encouraging and open new perspectives in terms of sound analysis and supervised classification of Tunisian music through learning algorithms. However, the complexity and large amount of perimeters used during this study, makes it still a very time consuming process.

For further study, we would like to propose experimenting with the extraction of other audio features such as Linear Predictive Coefficients (LPC) or Perceptual Linear Prediction (PLP) [7] which give a spectral representation of the spoken voice and which is widely used in the field of speech-processing.

Other learning algorithms for supervised classification such as naive Bayesian or Artificial Neural Network classification have also given excellent results in other studies [20] and might be interesting to try with Tunisian music.

It would also be interesting to extend this study on all acoustic categories of Tunisian music; an analysis of the different timbres of the instruments used in the Tunisian tradition. This study would lead to the detection of any Tunisian instrument to classify and index a musical database (National Phonetics) in different acoustic categories to facilitate access and exploration of the musical and sound heritage.

REFERENCES

[1] Thomas Fillon, Josephine Simonnt, Marie-France Mifune, Stéphanie Khoury, Guillaume Pellerin, Maxime Le Coz, Estelle Amy de la Bretèque, David Doukhan and Dominique Fourer. Telemeta: An open source web framework for ethnomusicological audio archives

- management and automatic analyses. Conference Paper, Journal of New Music Research, 2014.
- [2] T. Zhang, "System and method for automatic singer identification". IEEE International Conference on Multimedia and expo, HPL, 2003.
- [3] G. Peeters, "A generic system for audio indexing: Application to speech/music segmentation and music genre recognition". Proc. of the 10 international conference on digital audio effects (DAFx-07), Bordeaux, France, 2007.
- [4] P. Herrera, G. Peeters and S. Dubnov, "Automatic classification of musical instrument sound". Journal of new music research, 2010.
- [5] P. Herrera, G. Peeters and S. Dubnov. Automatic classification of musical instrument sound. Journal of new music research. Vol : 32, 2010.
- [6] G.Peeters. Descripteurs audio: de la simple représentation aux modèles de connaissances. Geste sonore et paramètres. L'analyse musicale à l'heure des outils multimédia, Jan 2015, Paris, France, 2015.
- [7] T. Ratanpara and N. Patel. Singer identification using perceptual features and cepstral coefficients of an audio signal from Indian video songs. Ratanpara and Patel EURASIP Journal on Audio, Speech, and Music Processing, 2015.
- [8] W.Tsai and H.Wang. Automatic singer recognition of popular music recording via estimation and modeling of solo vocal signals. 2-MUSI Signal Processing for music, 2005.
- [9] R. Miller and J. Franco. Analyse spectrographique de la voix chantée. NATS Journal, 1995.
- [10] Inderjeet Singh, Shashidhar Koolagudi. Classification of Punjabi Folk Musical Instruments Based on Acoustic Features. Proceedings of the International Conference on Data Engineering and Communication Technology, pp.445-454, 2017
- [11] Shruti and Bharti Chhabra. An Approach for Singer Identification Technique Using Artificial Neural Network. International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, 2015.
- [12] G. Peeters, "Automatic classification of large musical instrument data base using hierarchical classifiers with inertia ratio maximization". Audio Engineering Society, Convention Paper, USA, 2003.
- [13] P. Huang, S.D. Chen, P. Smaragdis and M. Hasegawa-Johnso. "Singing-Voice separation from monaural recordings using robust principal component analysis". IEEE, 2012.
- [14] M. Kalamani<sup>1</sup>, Dr.S.Valarmathy, S.Anitha. Automatic Speech Recognition using ELM and KNN Classifiers. International Journal of Innovative Research in Computer and Communication Engineering. Vol. 3, Issue 4, 2015.
- [15] Lhoucine Bahatti , Omar Bouattane , My Elhoussine Echhibat , Mohamed Hicham Zaggaf . An Efficient Audio Classification Approach Based on Support Vector Machines. International Journal of Advanced Computer Science and Applications pages 205 -211, 2016.
- [16] Mathieu Ramona. Classification automatique de flux radiophoniques par Machines à Vecteurs de Support. Thésée à l'Ecole Télécom ParisTech , Spécialité : Signal et Images, 2010.
- [17] Frédéric Santos. L'algorithme EM : une courte présentation. CNRS, 2015.
- [18] Guillaume Noyel. Filtrage réduction de dimension, classification et segmentation morphologique hyper-spectrale. Thèse à l'Ecole des Mine de Paris, spécialité : morphologie mathématique, 2008.
- [19] Lawrence Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proc. IEEE, 77(2): pages 257–286. (1989).
- [20] Vikramjit Mitra, Chia-Jiu Wang. Content based audio classification: a neural network approach. Soft computing, Methodology and Application, volume 12, issue 7, pp 639–646. 2008

# Predicting Fork Visibility Performance on Programming Language Interoperability in Open Source Projects

Bee Bee Chua

University of Technology, Sydney  
Australia

**Abstract**—Despite a variety of programming languages adopted in open source (OS) projects, fork variation on some languages has been minimal and slow to be adopted, and there is little research as to why this is so. We therefore employed a K-nearest neighbours (KNN) technique to predict the fork visibility performance of a productive language from a pool of programming languages adopted in projects. In total, 38 showcase OS projects from 2012 to 2016 were downloaded from the GitHub website and categorized into different levels of programming language adoption clusters. Among 33 languages, JavaScript is one of the popular languages that adopted by community. It has been predicted the language chosen when fork visibility is high can increase project longevity as a highly visible language is likely to occur more often in projects with a significant number of interoperable programming languages and high language fork count. Conversely, a low fork language reduces longevity in projects with an insignificant number of interoperable programming languages and low fork count. Our results reveal the survival of a productive language is in response to high language visibility (large fork number) and high interoperability of multiple programming languages.

**Keywords**—Open Source Programming Languages; K-nearest neighbors (KNN) Algorithm; interoperability; survivability

## I. INTRODUCTION

Programming languages constantly evolve to meet the demand of the software development industry. However variation of programming languages adopted in open source (OS) projects must comply with other programming languages so that developers can fork (copy) language files into their own local development environment. To ensure interoperability, programming languages must be expressive, generic and compliant, otherwise developers will not be interested in downloading or forking new OS libraries, as the frameworks are not compatible with their environment. There are different ways to define programming language success, with programming language interoperability performance being a major contributor to success. Despite this, unfortunately, most languages are not interoperable.

To understand when and why developers would fork a programming language file, language needs and motivation are two important factors. Some developers may fork a language because it is a new language that compiles with the original language, while other developers may fork a language

because it is a subset of the original language, with features added, removed or amended.

In spite of these motivating reasons to inspire developers to fork languages, many programming languages are experiencing a ‘fork crisis’, that is, they have low or minimal fork counts. This may be due to social factors [1]-[3] and environmental reasons [4]-[6], or the languages may lack expressiveness, be too generic or have compliance with the original or other languages. Interestingly, many OS project owners tried to increase programming language interoperability by adopting different programming languages; however this does not seem to increase forking.

Our motivation for this paper is firstly to make an intelligent recommendation system for developers and project owners to adopt programming languages that are compliant with other language interoperability. Secondly, to understand how a productive language fork may be affected by low programming language interoperability and low compliance with many programming languages’ interoperability.

This paper is organised into the following sections: Section 2, literature around language forking prediction, the problem and research questions; Section 3 research methodology on KNN algorithm, data quantisation methods and a case study of OS projects; Section 4 results, Section 5 outcomes of the four scenarios tested; and lastly, justification and conclusions.

## II. PROGRAMMING LANGUAGE FORKING

### A. Language Forking Prediction Problem and Research Questions

We investigated whether it was possible to predict with reasonable accuracy the fork visibility performance of any programming language with respect to interoperability compliance. In addition, we sought to determine the probability of new projects adopting a productive language where fork visibility performance is impacted by low versus high programming language interoperability.

Two research questions were developed to address these aims:

1) How can we predict, with reasonable accuracy, a programming language fork visibility performance in projects

that is in compliance to other languages interoperability?

2) For a new project, how can we predict productive programming language fork visibility performance based on the level of programming language interoperability?

In this paper, we define a ‘more’ interoperable programming language project as a language that has more healthy forks in the majority of programming languages, and a ‘less’ interoperable programming language project as one with fewer healthy forks in each language.

### III. METHODS AND DATASET PREPARATION

#### A. K Nearest Neighbour (KNN) Algorithm

The KNN algorithm is based on representation of statistics and distributions in training data. While the method was first discovered in 1961 by a group of American researchers who showed it works effectively on actual instances of training data [7], it remains unpublished. It has since been applied to machine learning and data mining, and more recently has successfully been applied in education research to predict student learning success and failure rates [8]-[12]. The KNN method is effective at predicting different types of data, is simple and versatile, and handles noisy or incomplete data, when in many situations a classification is required [13]-[17].

The baseline KNN predicts the fork performance of a given project by first calculating the actual project (project being predicted) similarity to all instances in the training set and finds the K most similar ones. The similarity is calculated with a simple Euclidean distance between the features of the test subject and corresponding features of each instance in the training set [12].

In this study, KNN was used to predict fork visibility performance of languages that were adopted as interoperable language in projects to differing degrees (‘more’ or ‘less’). Firstly the algorithm applied Euclidean distance formula (see Fig. 1) to calculate the distance of a productive language fork for less adopted interoperable language projects. X refers to the number of language repositories created in the project and Y refers to the number of programming languages adopted in the project. X1 is the actual number of language repositories from 38 project showcases and X2 is the predicted number of new project language repositories. Similarly, Y1 and Y2 are the actual and predicted numbers of programming language from the 38 project showcases and the new project.

We classified the outcome of that algorithm into two categories: 1) JavaScript in a project with low fork visibility; and 2) JavaScript in a project with high fork visibility. Next, we used K=3 to predict the language project on JavaScript fork visibility outcomes.

$$D(x,y)=\sqrt{(x1 - y1)^2 + (x2 - y2)^2}$$

Fig. 1. KNN equation.

#### B. Case Study: Showcase Projects

Of the 40 OS show case projects available on GitHub on from January 2012 through August 2016 (www.github.com), 38 projects have complete information such as the type of programming languages and the fork count. We rejected 2 projects because of some programming languages were not

stated (unknown). As our goal was to predict the language fork visibility performance, defined as success or survival of different programming languages in a project, the 38 projects were classified into types of projects and by different levels of programming languages (Fig. 2).

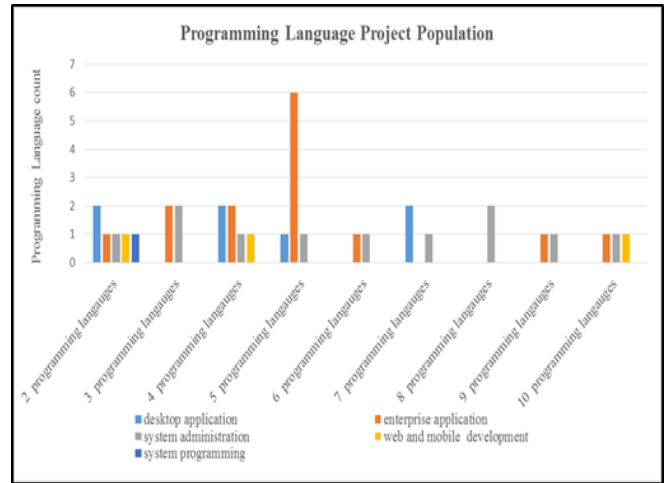


Fig. 2. Programming language project population.

The types of projects ranged from desktop application, enterprise application, systems administration, systems programming and website development.

Next, we categorized productive programming languages by types of programming language tier level according to the TIOBE programming community index, which ranks various programming languages [18], [19]. Fig. 3 shows that projects adopted from 2 to 9 programming languages, and JavaScript was the most popular.

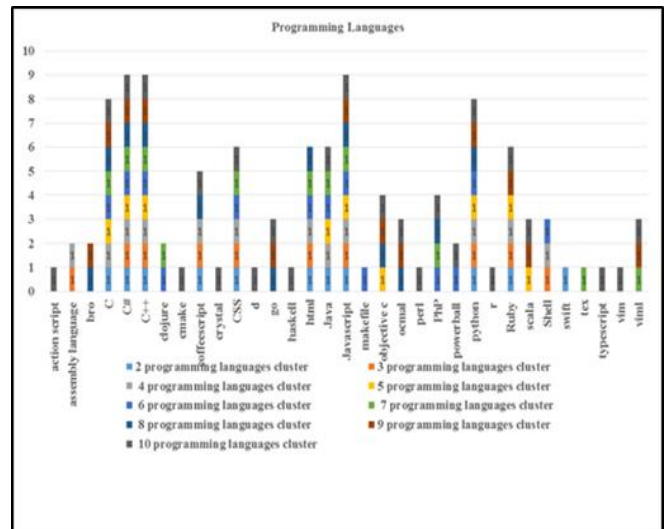


Fig. 3. Tier levels of programming languages.

#### C. Programming Language Fork Visibility Performance and Data Quantisation

Prior to applying the KNN algorithm (see below), we first identified the features of programming language fork visibility performance that responded to programming language interoperability. These included individual programming

language type, the number of individual programming languages adopted per project, the individual language repository number, and individual language fork frequency, when available (published on the project webpage). Then, due to the large quantity of fork counts, the data underwent quantisation, with each feature weighted as per Table 1.

TABLE I. PROGRAMMING LANGUAGE FORK PERFORMANCE FEATURE

Feature	Range	Weight	
		Min	Max
Number of adopted programming languages	1–10	0.1:1	1.0:10
Adopted language repository file number	1–10	01:1	1.0:10
Specific language fork number	300–200,000	0.01:1–500	0.2:200,000

Quantisation produced a total number of 2652 data features. An example of each project data that converted to data quantisation as follows to each field as: number of adopted programming languages, adopted language repository file number, from specific programming language fork number 1 to number 33.

0.1,0.2,0.01,0.1,0.01,0.3, 0.01,0.0001,0,0,0,0,0,0,0.1, 0.0001,0,0, 0,0,0.0001,0,0,0.01,0.0001,0.01,0.1,0,0,0,0.1,0

D. Averaging Programming Language Number and Programming Language Fork Count

To confirm the programming language fork visibility performance, we set a threshold on programming language number and fork count size, with minimum and maximum values. To support the threshold, we derived an equation to determine the threshold outcome based on two further equations: 1) Average Programming Language Number (APLN); and 2) Average Programming Language Interoperability (APLI), for the APLN and APLI, the formulas were:

$$APLN = \frac{\text{Total number of project language number}}{\text{Total number of project in the case study}} \quad (1)$$

$$APLI = \frac{\text{Total number of project language number}}{\text{Total number of project in the case study}} \quad (2)$$

Next, we compared each APLN against the APLI in the project. If the APLN score was greater than the APLI score then the project was defined as having adopted high programming language interoperability. Conversely, if the APLN was less than the APLI score then the project was defined as having adopted low programming language interoperability.

IV. RESULTS

Fig. 4 shows a simple example illustrating KNN with two features (programming language fork size count as the x axis and programming language number as y axis) to find the JavaScript visibility performance.

The justification on JavaScript as it produces many libraries and frameworks on OS projects that are compliant for cross-platform integration. Moreover, the JavaScript language community is large because it is familiar to developers who learned it during training and qualification. In the context of

this paper, we were interested to find out the predicted outcomes for JavaScript fork performance on low and high programming language interoperability for a new project.

We generated four scenarios to predict their outcomes using the KNN algorithm. The first scenario was a project that was likely to receive low fork count in JavaScript, which adopts low average programming language interoperability (APLI). The second scenario was a project likely to experience high JavaScript fork in the adopted low APLI. The third scenario was a project with low JavaScript fork in a high APLI, and the fourth scenario was high JavaScript fork in a high APLI.

Scenario 1: JavaScript low fork visibility performance with low adopted programming language interoperability

The first scenario was a new project that adopted very low programming language interoperability, including JavaScript. Fig. 4 shows the new project (orange circle) distance is close to projects A, C and G. By majority voting, project C was predicted as the nearest to the new project, that is, the new project JavaScript language fork was predicted to be low if the adoption of programming languages interoperability was low.

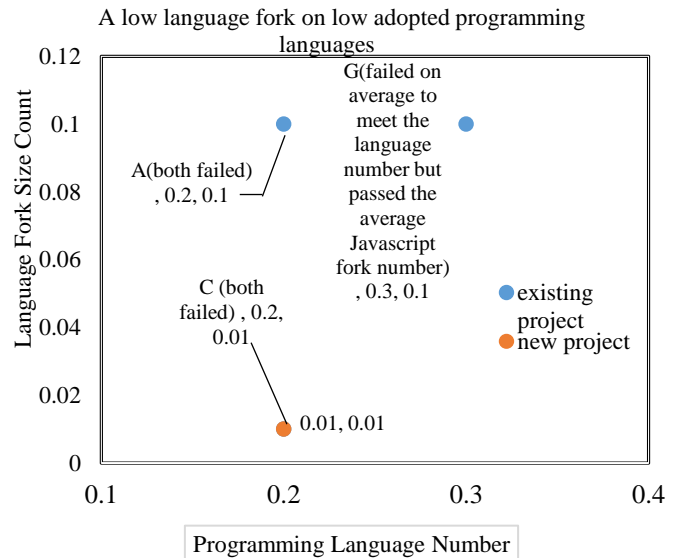


Fig. 4. Scenario 1: JavaScript low fork visibility performance in a low APLI.

Project A was a website development that had adopted JavaScript and Ruby and, based on their fork population; it was very close to the JavaScript fork size on the new project. Project C, on the other hand, was an enterprise application and adopted only 2 programming languages – JavaScript and CSS. The project failed to receive high fork attention because CSS is used for formatting structured content on HTML documents. As a result, it is less interesting to developers as a problem-solving technique. For Project G, despite having JavaScript, Python and HTML as marked up languages adopted, they face survival problems being unable to find developers to fork the language file, possibly because Python is less compliant with JavaScript [6], thus lessening JavaScript forking.

Overall, these project languages failed to pass the average adopted programming language interoperability levels and average JavaScript fork count size. By majority voting – where  $K=3$  – a new project was predicted to fail in a low adopted programming languages and low JavaScript fork environment.

**Scenario 2: JavaScript high fork visibility performance with low adopted programming language interoperability**

The second scenario outlined JavaScript high fork visibility performance in a low APLI, which was the reverse of the first scenario. Fig. 5 shows the new project (orange circle) is close to projects A1, Q and D1. We applied  $K=3$  which resulted in a tied vote, with a different outcome on the three projects. Project A1 had a sufficient APLI number but failed to generate a high JavaScript fork. Project Q failed on the APLI but passed on the average number of JavaScript forks. In contrast, Project D1 satisfied both conditions, passed APLI and average number of JavaScript forks. However as the data set was small no one single outcome can predict whether a new project would be likely to be near to an existing project. We further examined each project cause, finding that JavaScript language files added new features that attracted developer attention.

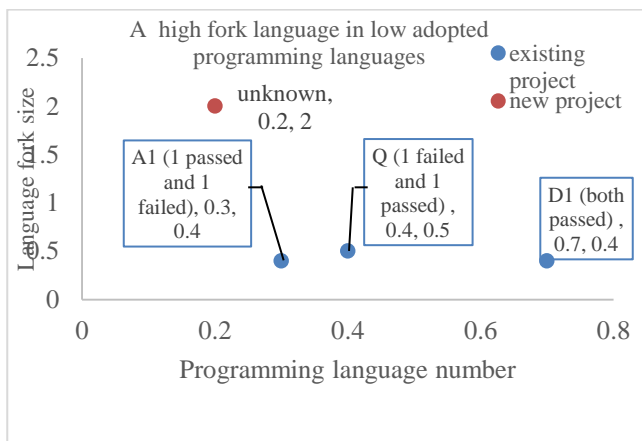


Fig. 5. Scenario 2: JavaScript high fork visibility performance in a low APLI.

**Scenario 3: JavaScript low fork visibility performance with highly adopted programming language interoperability**

The third scenario was a new project with high APLI and low fork count on JavaScript. Based on the majority voting, the three projects predicted to the nearest distance of the new project were J1, L1 and K1 (Fig. 6). Successfully all passed both the average programming language interoperability number and the average JavaScript fork number. The results showed that low language fork can arise in a project with some languages adopted with weak compliance to JavaScript. In Scenario 3, non-JavaScript language files focused on back-end development; as such they were of core project value. Consequently, it has a high impact on JavaScript developers' fork behaviour to download and fork less the JavaScript files.

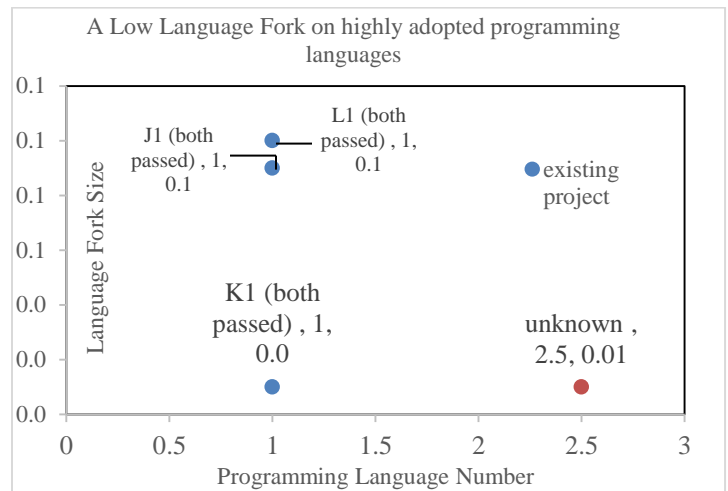


Fig. 6. Scenario 3: JavaScript low fork visibility performance in a high APLI.

**Scenario 4: JavaScript high fork visibility with highly adopted programming language interoperability**

The fourth scenario was a new project that adopted a variety of programming languages; the JavaScript language is one of the most well-known languages that contain a high fork count. Fig. 7 shows the distance of a new project status (orange circle) and existing projects D1, Q and L1. The three existing projects passed the average adopted programming language interoperability number and the average JavaScript fork count. We applied  $K=3$  to detect the possible outcome for the new project. The result shows by majority voting in this case all 3 projects have the same outcome and they are predicted the nearest projects to the new project.

These projects seemed to perform better because they were compliant with other programming languages, such as Ruby, PhP, Python C and C++. As JavaScript shows a high connectivity with Ruby and PHP [20], JavaScript can fetch a high fork count from developers. From the project development perspective, the topic domain or field interest to developers, and the selective programming languages, contribute to the high fork frequency. From our observation on the three projects' fork aggressiveness, the languages adopted in these projects are compatible to cross platforms.

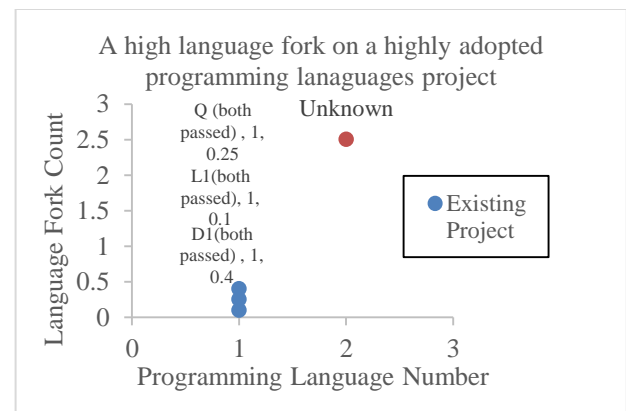


Fig. 7. Scenario 4: JavaScript high fork visibility performance in a high APLI.

## V. JUSTIFICATION

1) Positioning a productive language in a pool of compliant language interoperability:

Our previous work [21] introduced a technique to detect the chance of programming languages used in Apache, Mozilla and Ubuntu surviving from a forking perspective. The current work from the evidence, the productive language, JavaScript, showed less difficulty to survive when placed in a pool with low APLI. In addition, a low survival of JavaScript could be expected in conjunction with high APLI because JavaScript is less compliant with other languages' interoperability, except Ruby and PHP [19], [20].

2) Programming languages fork visibility performance: A new perspective on survivability and longevity:

The scenario-based evidence presented here provides a new perspective on developers' fork behaviour – particularly on programming language interoperability number adopted in a project and how they might influence each other, especially the productive languages. However, there is no certainty on which programming language can survive longer in terms of emerging technology, except it must be compliant with other language interoperability. Due to an increased change in emerging technology – such as mobile application and cloud development – more projects will increasingly add more programming languages for interoperability. As such, the more a language is compliant, the more likely a language will increase fork visibility, and in turn increase language survivability.

Previous work [22]-[29] showed OS variables' impact on developers' fork behaviour is generally related to project topics and domains, developers' language preference, and programming language popularity. However, our findings show another possible cause, that is, poor fork visibility in a low APLI is less compliant. A language with low fork visibility is likely to decline its longevity and survivability whereas a language with high fork visibility is likely to increase its longevity and survivability, which will serve to keep the project viable and accessible by developers. Understanding programming language fork success or failure draws a new perspective out of the literature, highlighting that fork success is highly dependent on specific language domination [1] and/or a productive language [20].

3) The relationship between visibility and vulnerability in the context of open source programming languages:

The term 'visibility' is described in the context of meteorology as transparency of air, in the dark, etc. In a disruptive or sustainable technology, visibility is a metric used to determine factors such as project vulnerabilities, consumer confidence, or purchasing pattern or behaviour. In the context of OS programming languages, visibility exposes the vulnerability of a language, which can become less significant as a result to sustain frameworks and libraries, front-end, back-end, etc. As such, new programming languages with better implementation performance are likely to dominate and replace existing language source codes.

## VI. CONCLUSION

This research focused on applying an algorithm to a case study of four scenarios. The preliminary findings require further validation in a larger dataset to examine programming language strength, in terms of compliance, compatibility and connectivity. This paper introduced a new perspective to OS programming language survivability research, particularly the fork visible performance that different programming languages exhibit and their interoperability performance across different ecosystems and environments.

### REFERENCES

- [1] Nyman, L. 2013. Freedom and forking in open source software. Proceedings of the Nordic Academy of Management Conference ,Reykjavik, Iceland
- [2] Tsay, J. Dabbish, L. and Herbsleb, J. 2014. Influence of Social and Technical Factors for Evaluating Contribution in Github, Proceeding of ICSE'14, Hyderabad, India, ACM.
- [3] Khondhu, J. Capiluppi, A. , Stol, K.J. 2013. Is It All Lost? A Study of Inactive Open Source Projects. In Proceedings of the 9th International Conference on Open Source Systems
- [4] Crowston, K. Howison, J. and Annabi, H. 2006. Information Systems Success in Free and Open Source Development : Theory and Measures " Software Process and Practice, Vol. 11, No 2, Pp. 123-148
- [5] V. Midha and P. Palvia, "Factors affecting the success of open source software," The Journals of Systems and Software, vol. 85, no. 4, pp. 895–905, 2012.
- [6] S. Comino and F. M. Manenti, "Government policies supporting open source software for the mass market," Journal Review of Industrial Organization, vol. 26, no 2, pp. 217–240, 2005
- [7] Johns, M. V. (1961) An empirical Bayes approach to non-parametric two-way classification. In Solomon, H., editor, Studies in item analysis and prediction. Palo Alto, CA: Stanford University Press
- [8] Tanner, T. and Toivonen, H. (2010). Predicting and preventing student failure – using the k-nearest neighbour method to predict student performance in an online course environment. International Journal of Learning Technology 5(4):56–377. <https://www.cs.helsinki.fi/u/htoivone/pubs/ijlt2010.pdf>
- [9] Ishii, N., Hoki, Y., Okada, Y. and Bao, Y. (2009). Nearest neighbor classification by relearning. In: Proceedings of the 10 International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'09), pp. 42–49.
- [10] Kotsiantis, S., Pierrakeas, C. and Pintelas, P. (2003). Preventing student dropout in distance learning systems using machine learning techniques. In: Proceedings of 7<sup>th</sup> International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Lecture Notes in Artificial Intelligence, Springer-Verlag. 2774:267–274.
- [11] Minaei-Bidgoli, B., Kashy, D.A., Kortmeyer, G. and Punch, W.F. (2003). Predicting student performance: an application of data mining methods with an educational Web-based system. In: Proceedings of the 33<sup>rd</sup> Annual Frontiers in Education, 1:T2A–18.
- [12] Shih, B. and Lee, W. (2001). The application of nearest neighbour algorithm on creating an adaptive on-line learning system. In: 31st Annual Frontiers in Education Conference, 1:T3F–10–13.
- [13] Manning, C., Raghavan, P. and Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press: Cambridge, UK.
- [14] Dumais, S., Platt, J., Heckerman, D. and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In: Proceedings of the International Conference on Information and Knowledge Management., pp. 148–155.
- [15] Dumais, S., Platt, J., Heckerman, D. and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In: Proceedings of the International Conference on Information and Knowledge Management, pp. 148–155.

- [16] Han, Y. and Lam, W. (2006). Exploring query matrix for support pattern based classification learning. *Advances in Machine Learning and Cybernetics, Lecture Notes in Computer Science* 3930:209–218.
- [17] Zou, Y., An, A. and Huang, X. (2005). Evaluation and automatic selection of methods for handling missing data. In: *Proceedings of the IEEE International Conference on Granular Computing*, 2:728–733.
- [18] Cover, T. and Hart, P. (1967). Nearest neighbour pattern classification", *IEEE Transactions on Information Theory* 13(1):21–27.
- [19] TIOBE. TIOBE programming community index definition. 2016, <http://www.tiobe.com/index.php/content/paperinfo/tpci/tpcidefinition.html>
- [20] Bissyande, T.F., Thung, F., Lo, D., Jiang, L.X. and Réveillère, L. (2013). Popularity, interoperability, and impact of programming languages in 100,000 open source projects. In: *Proceedings of COMPSAC '13: 2013 IEEE 37th Annual Computer Software and Applications Conference*, 22–26 July, 2013, Kyoto, Japan. pp. 303–312. Research Collection School of Information Systems.
- [21] Chua, B. 2015, 'Detecting Sustainable Programming Languages through Forking on Open Source Projects for Survivability', *IEEE, The 26th IEEE International Symposium on Software Reliability Engineering (ISSRE) 2015 in conjunction with a WOSAR workshop*, IEEE, Gaithersburg, USA. pp. 120-124.
- [22] Samoladas, I. Angelis, L. and Stamelos, I. 2010. Survival duration on the duration of open source projects. *Journal of Software and Information Technology*, Vol.52, No.1, Pp 902-922.
- [23] Chen, S. 2010. Determinants of Survival of Open Source Software: An Empirical Study. *Academy of Information and Management Sciences Journal*, Vol.13, No.2, Pp119-128.
- [24] Wang, J. 2012. Survival factors for Free Open Source Software projects: A multi-stage perspective," *European Management Journal*, Vol.30, No.1, Pp352-371.
- [25] Wu, J. and Tang, Q. 2007. Analysis of Survival of Open Source Projects: a Social Network Perspective. In *proceedings of Australian Conference of Information Systems (ACIS)*.
- [26] Angelis, L. Sentas, P. 2005. Duration Analysis of Software Projects. In *Proceedings of the 10th Panhellenic Conference on Informatics*, 258-269.
- [27] Raja, U. and Tretter, M. J. 2012. Defining and Evaluating a Measure of Open Source Project Survivability. *Journal of IEEE ,Transactions on Software Engineering*, Vol.38, No.1,Pp163-174.
- [28] Chengalur-Smith, I., Sidorova, A. and Daniel, S. Sustainability of Free/Libre Open Source Projects: A Longitudinal Study. *Journal of Association For Information Systems (JAIS)*, Vol.11, No. 11/12.Pp657-683.
- [29] Oskar, J., Gruszka, B., Jaroszewicz, S., Bukowski, L. and wierzbicki, A. 2014. GitHub Projects. Quality Analysis of Open-Source Software. In the proceeding of 6h International Conference, SocInfo. Barcelona, Spain, *Lecture Notes in Computer Science Volume 8851*.



# Cadastral and Tea Production Management System with Wireless Sensor Network, GIS based System and IoT Technology

Kohei Arai

Information Science Department  
Graduate School of Science and Engineering, Saga University  
Saga City, Japan

**Abstract**—Cadastral and tea production management system utilizing wireless sensor network of Internet of Things (IoT) technology is proposed. To improve efficiency of tea productions, cadastral management and tea production processes must be managed by Geographical Information System (GIS) based system. Through experiments with sensor acquired data, it is found that the required information can be estimated and represented efficiently. Thus, the system works for improvement of the tea production management and quality control.

**Keywords**—Internet of Things; geographical information system; tea production; quality control

## I. INTRODUCTION

Vegetation monitoring is attempted with red and photographic cameras [1]. Growth rate monitoring is also attempted with spectral observation [2].

Total nitrogen content corresponds to amid acid which is highly correlated to Theanine: 2-Amino-4-(ethyl carbamoyl) butyric acid for tealeaves so that total nitrogen is highly correlated to tea taste. Meanwhile fiber content in tealeaves has a negative correlation to tea taste. Near Infrared: NIR camera data shows a good correlation to total nitrogen and fiber contents in tealeaves so that tealeaves quality can be monitored with network NIR cameras. It is also possible to estimate total nitrogen and fiber contents in leaves with remote sensing satellite data, in particular, Visible and Near Infrared: VNIR radiometer data. Moreover, Vegetation Cover: VC, Normalized Difference Vegetation Index: NDVI, Bi-Directional Reflectance Distribution Function: BRDF of tealeaves have a good correlation to growth index of tealeaves so that it is possible to monitor expected harvest amount and quality of tealeaves with network cameras together with remote sensing satellite data. BRDF monitoring is well known as a method for vegetation growth [3], [4]. On the other hand, degree of polarization of vegetation is attempted to use for vegetation monitoring [5], in particular, Leaf Area Index: LAI together with new tealeaves growth monitoring with BRDF measurements [6].

It is obvious that nitrogen rich tealeaves tastes good while fiber rich tealeaves tastes bad. Theanine: 2-Amino-4-(ethyl carbamoyl) butyric acid that is highly correlated to nitrogen contents in new tealeaves are changed to catechin [7],[8],[9] due to sun light. In accordance with sunlight, new tealeaves

growth up so that there is a most appropriate time for harvest in order to maximize amount and taste of new tealeaves simultaneously.

Optical properties of tealeaves and methods for estimation of tealeaves quality and harvest amount estimation accuracy are well reported [10]-[17]. The method proposed here is to determine tealeaves harvest timing by using NIR camera images together with meteorological data. Also, the methods for estimation of vitality of tea trees (vigor) and tealeaf quality assessment are proposed together with Kyushu small satellite based tea farm area monitoring [18]-[23]. Multi-layer observation for agricultural (tea and rice) field monitoring system is proposed by the author [24], [25].

In this paper, cadastral and tea production management system utilizing wireless sensor network of Internet of Things: IoT technology is proposed. In order to improve efficiency of tea productions, cadastral management and tea production processes have to be managed by Geographical Information System: GIS based system. Through experiments at the Saga Prefectural Tea Institute situated in Kyushu, Japan with sensor acquired data, it is found that the required information can be estimated and represented efficiently. Also, it is found that the system works for improvement of the tea production management and quality control. The growth rate indicates fiber content in tealeaves while total nitrogen content in tealeaves is highly correlated with taste of tea. The vegetation monitoring should have both capabilities, fiber and nitrogen contents in tealeaves. Thus the proposed system allows to monitor the age and the taste of tealeaves with GIS system after observation.

The following section describes the proposed cadastral management system based on GIS representation. Then the proposed tea production control system with wireless sensor network is described together with some experiments conducted with the proposed system. Finally, conclusion is described with some discussions.

## II. PROPOSED TEA PRODUCTION MANAGEMENT SYSTEM

### A. System Configuration

Fig. 1 shows a portion of the proposed tea production management system. In the tea farm areas situated in Ureshino-city, Kyushu, Japan, ground based visible to near infrared

cameras are equipped for monitoring of tealeaf growing processes (fiber content in tealeaf) and tealeaf quality (Nitrogen content in tealeaf). Visible Pan-Tilt-Zoom: PTZ network camera and NIR filter (IR840) attached network camera is equipped on the pole. PTZ cameras are controlled by mobile phone as well with “mobile2PC” or Internet terminal with “LogMeIn” of VNC services [7] through wireless LAN connected to Internet. Acquired camera data are used for estimation of total nitrogen and fiber contents as well as BRDF for monitoring growth index.

The cameras are connected to the Internet through the network card of W05K that is provided by AU/KDDI. Through <http://119.107.81.166:8080>, the acquired image data are

accessible so that it is easy to access the data from Internet terminals. Panasonic BB-HCM371 cameras are used for the experiments. Solar panel of G-500 (12V, 500mA, 8.5W) with battery of SG-1000 is used together with Xpower75 (60W) of inverter. On the other hand, weather station data can be accessible from the URL of <http://katy.jp/mapstation/> of data server provider through wireless LAN connection from the weather station to the Internet terminal. Acquired imagery data can be transferred to the central station together with meteorological data. Therefore, tealeaf growing process and tea quality can be monitored in the central station. Also, the acquired images and information can be monitored with mobile terminals.

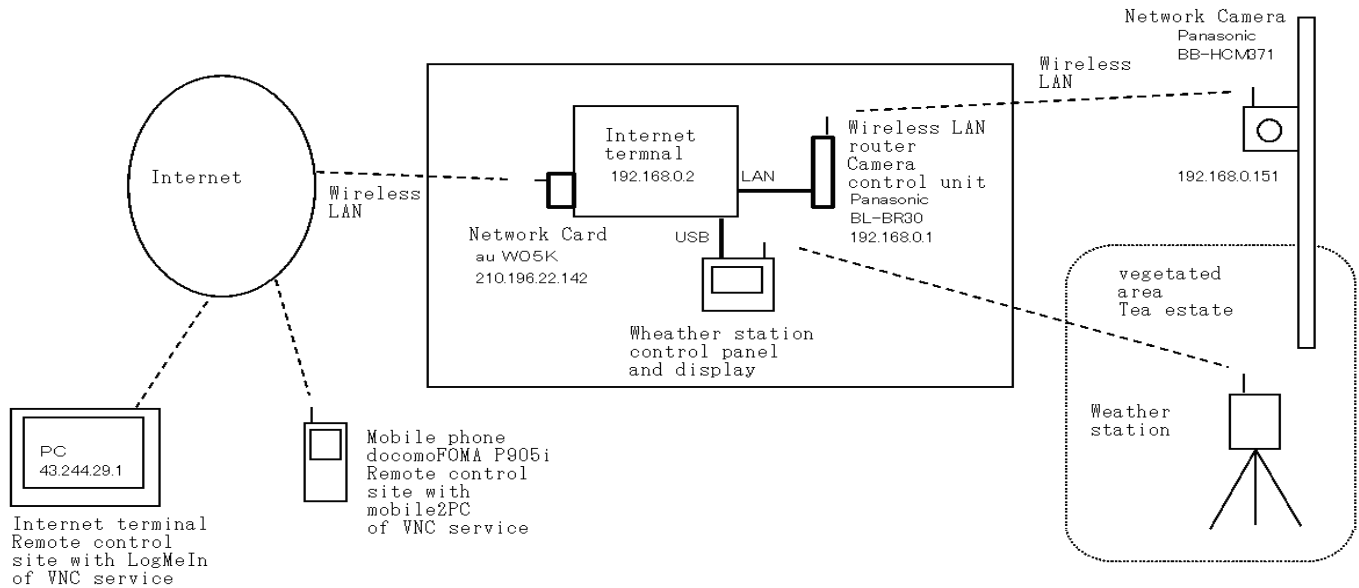


Fig. 1. Proposed wireless sensor network system configuration.

### B. Cadastral Management and GIS Representation of all the Required Data and Information for Tealeaf Growing and Quality Monitoring of Tea Farm Areas

Cadastral management can be performed with the following procedure: First, entire topographic map of the Saga prefecture is represented with the map scale of 1/100,000 in GIS as shown in Fig. 2(a). In the proposed system, SuperMap DeskPro of GIS<sup>1</sup> system is used. The GIS system has the following functions, Data Editing, Thematic Map, Topology, Attribute creation, Layout, Spatial analysis, Grid analysis, Network analysis, 3D analysis. Red circles in Fig. 2(a) indicate the tea farm areas as points. Then the tea farm areas are represented with the map scale of 1/25,000 as shown in Fig. 2(b). In this stage, polygons are used for tea farm areas. Thus area of the specific tea farm can be calculated.

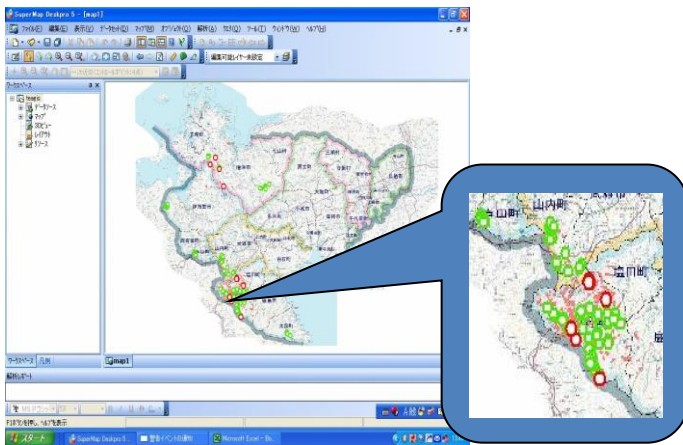
Also, all kinds of parameters are displayed in the GIS map such as owner's name, latitude, longitude, elevation, tea tree species, ID No., observation date and time, etc. Smart snapshot, topology process, registration process, control point location input, attribute table relation, hyperlink function,

spatial analysis including area calculation, buffering, overlaying analysis and geometric calculation are also available in this stage. Much large scale of map can be available up to 1/15,000 at this stage as shown in Fig. 2(c). Through these process, 1/1,500 map scale representation is available as shown in Fig. 2(d).

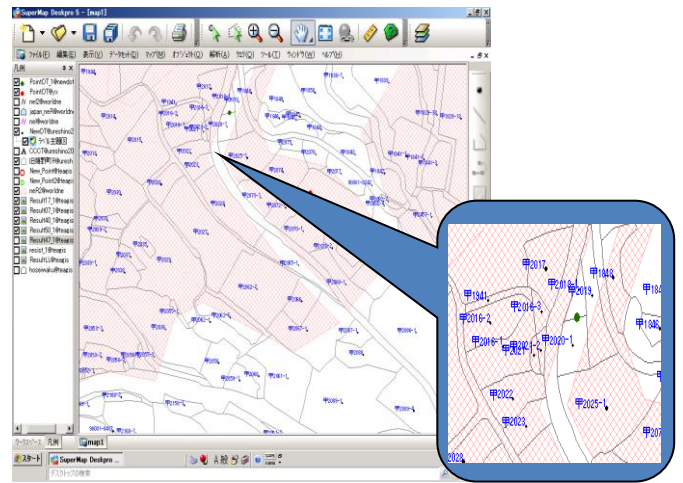
The corresponding areas of satellite imagery data to the topographic map areas can also be displayed in the GIS map as shown in Fig. 2(e). In this case, pan-sharpened image between Advanced Land Observation Satellite: ALOS/PRISM (Panchromatic band of visible sensor) and Terra (the first Earth Observing Satellite System: EOS satellite) /Advanced Sensor for Thermal Emission and Reflectance: ASTER/Visible and Near Infrared Radiometer: VNIR is displayed (2 m of spatial resolution of multispectral imagery data can be created by the pan-sharpened process<sup>2</sup>). These data are linked to not only satellite imagery data but also the ground based NIR camera data. Therefore, the corresponding linked camera data to the clicked GIS map location is displayed as shown in Fig. 2(f).

<sup>1</sup> [https://supermap.jp/products/supermap/deskpro/d\\_viewer.html](https://supermap.jp/products/supermap/deskpro/d_viewer.html)

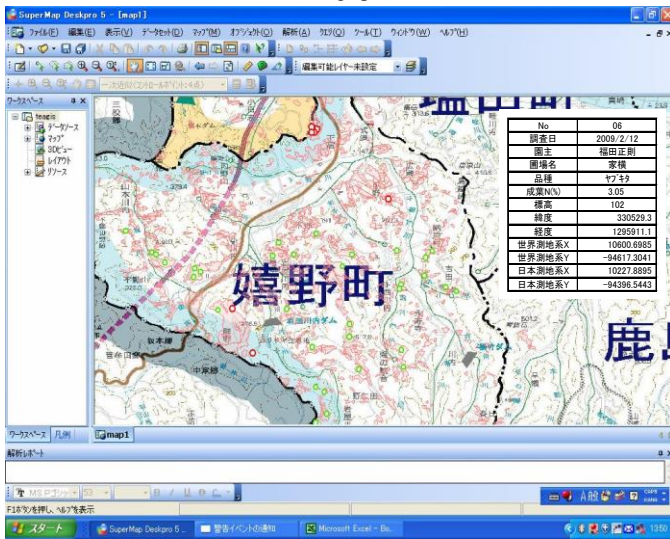
<sup>2</sup> RGB coordinate system of VNIR is converted to HIS system. Then Intensity is replaced by PRISM data. After that, HIS system is converted to RGB system.



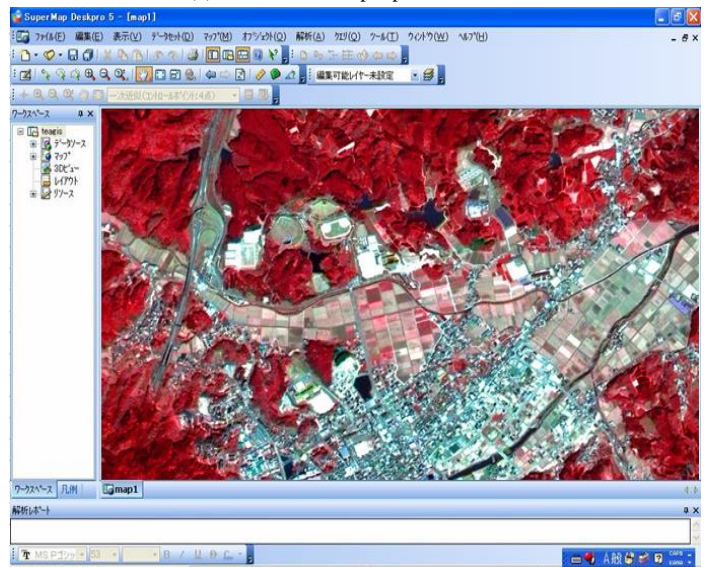
(a) Entire Saga prefecture.



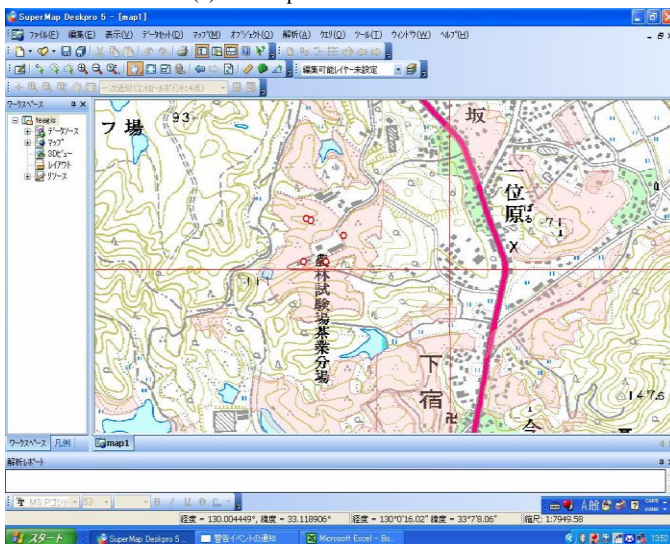
(d) 1/1,500 scale map representation.



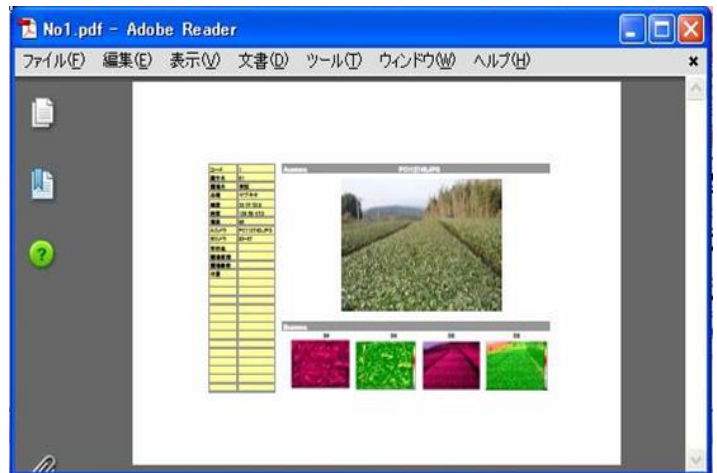
(b) Close-up for each tea farm area.



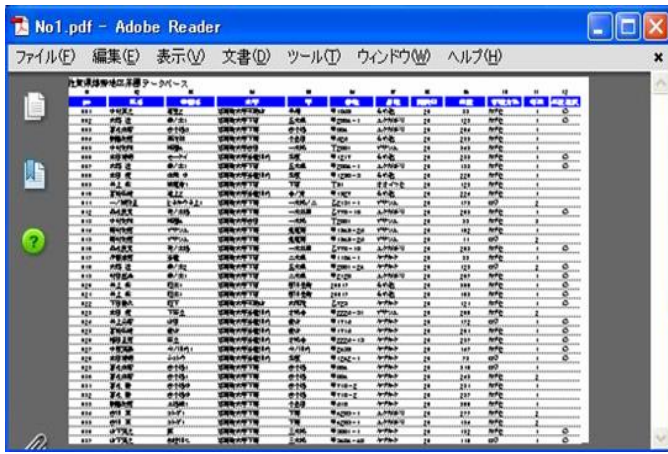
(e) Pan-sharpened image between ALOS/AVNIR-2 and Terra/ASTER/VNI.



(c) 1/15,000 scale map representation.



(f) Linked camera imagery data.



(g) All the required data and information.

Fig. 2. Cadastral Management and GIS representation of all the required data and information for tealeaf growing and quality monitoring of tea farm areas.

Thus all the required data and information for tealeaf production (tealeaf growing and quality) are displayed by clicking the location of tea farm areas on the GIS map as shown in Fig. 2(g). As shown in Fig. 2(g), the proposed GIS system allows monitoring of each tea farm field. Therefore, fertilizer, water resources, pesticide can be controlled by the data shown by GIS system by field by field. Furthermore, it can be done to arrange almost same quality of harvested tealeaves for tea production by taking a look at the quality of tealeaves in concern with the GIS system. The aforementioned function, data handling and analysis can be done in the GIS system as well.

### III. EXPERIMENTS WITH THE PROPOSED SYSTEM

#### A. Optical Property of Tealeaf

The most specific optical property of tealeaf is spectral reflectance. Fig. 3 shows examples of typical spectral reflectance.

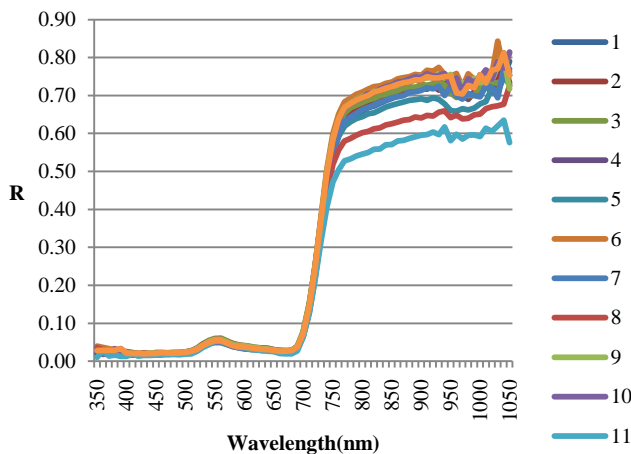


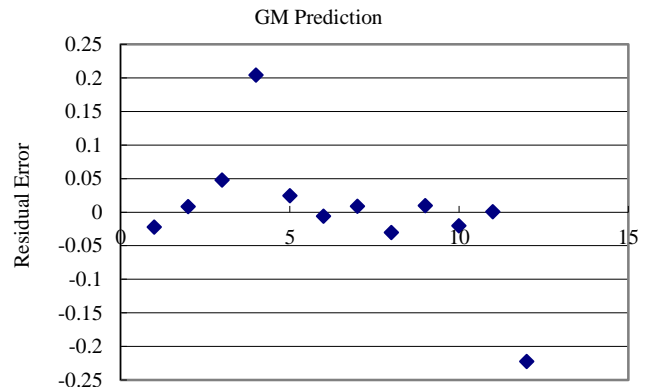
Fig. 3. Examples of spectral reflectance of tealeaves.

In the figure, spectral reflectance of 10 different tea farm areas situated in Ureshino-city, Saga, Japan are shown in Fig.3. The most specific feature of the spectral reflectance is high reflectance in the near-infrared spectral region followed by relatively high reflectance at the 550nm, green color of spectral range. Chloroplast under the cuticle of tealeaf shows high reflectance at the near-infrared wavelength region (935nm) depending on the interval between cells. The difference among the reflectance at near-infrared wavelength region indicates their quality, in particular, total nitrogen content in the tealeaves which corresponds to amino acid (Theanine) content. Therefore, tealeaf quality can be estimated with the measured reflectance at near-infrared wavelength region. The measured spectral reflectance at near-infrared region is transmitted to the cadastral and tea production management center with wireless sensor network through internet. Therefore, quality of the different tea farm areas can be monitored and estimated. Using high quality of tealeaves, high quality of bland tea (High quality tea qualified by satellites) can be produced.

#### B. Other Property of Tealeaf

Traditionally, GM(Green Meter) value is an indicator of the quality of tealeaf. GM can be measured with the commercially available GM meter (a kind of touch sensor) with relatively low cost. Therefore, most of tea farmers use the GM meter. Essentially, GM indicate the reflectance of tealeaf at the green wavelength. As shown in Fig. 3, there is strong relation between reflectance at near-infrared and green wavelength regions. As the result, it is possible to estimate GM value by using the measure reflectance at the near-infrared. Also, Total Nitrogen (TN) content in tealeaf is highly related to the reflectance at the near-infrared wavelength region. Fig. 4(a) and (b) shows residual errors of GM and TM predictions, respectively. Also, Fig. 4(c) and (d) shows the relation between estimated and actual TN as well as the regressive analysis result of water content in the tealeaves with the reflectance at near-infrared wavelength region, respectively. All these regressive analysis show a good correlation in terms of residual errors for GM and TN as well as estimations of water content and TN in tealeaves.

Therefore, it may be said that GM, TN and water content in tealeaf can be estimated with the measured reflectance at the near-infrared wavelength region.



(a) GM.

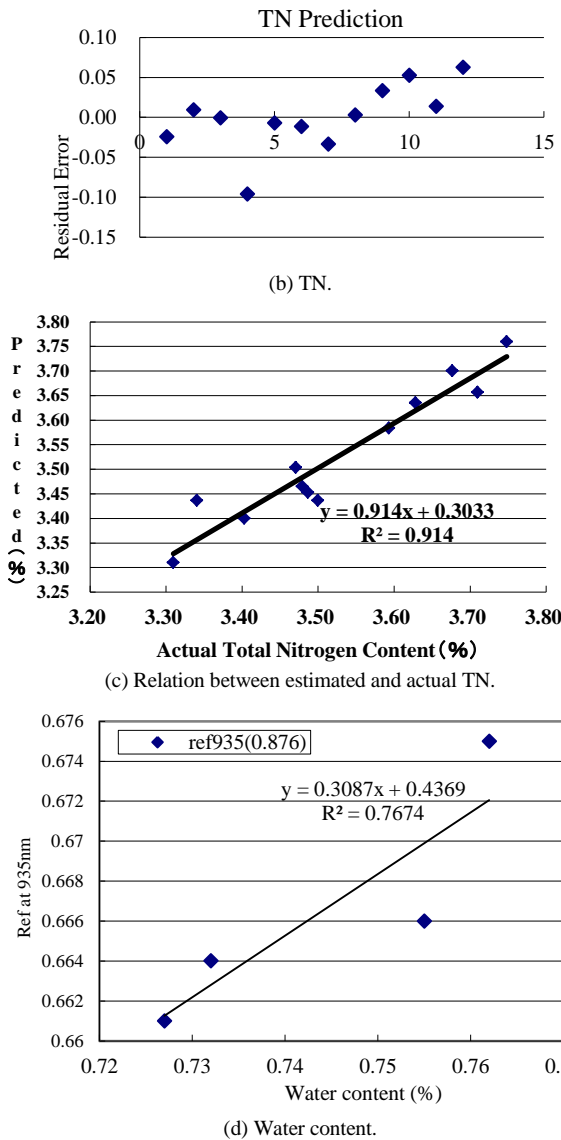


Fig. 4. Regressive analysis results of GM and TN predictions using the measured reflectance at near-infrared wavelength region.

Another important factor of tealeaf quality is fiber content (F-NIR). Fiber rich tealeaf implies elder tealeaf while fiber poor tealeaf means young tealeaf. Obviously, younger tealeaf is better than that of elder tealeaf. Fiber content in tealeaf can be estimated with reflectance at near-infrared wavelength region.

Fig. 5 shows the relation between reflectance at 935 nm and TN as well as fiber content in tealeaves. The R square value (square of correlation between both) is not good enough, around 0.74. Therefore, another index which shows much high correlation has to be found. Normalized Difference Vegetation Index: NDVI is well-known index for vegetation. If the spectral reflectance is measured with spectrometers, then reflectance at arbitrary wavelength can be used for estimation of TN and fiber content. The best index for estimation of TN and fiber content is determined through regressive analysis with the measured spectral reflectance.

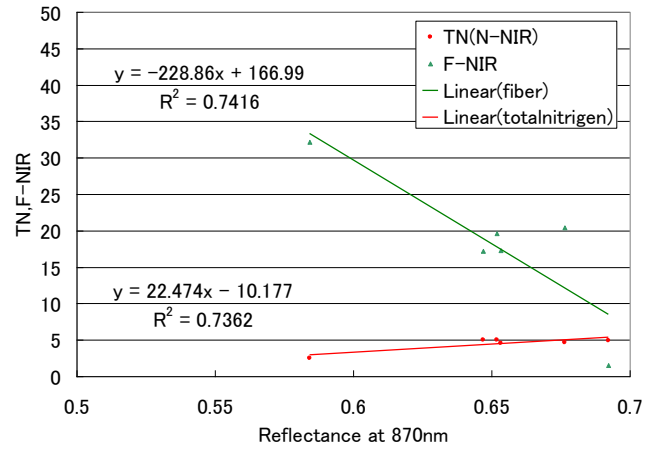


Fig. 5. Relation between the reflectance at 870nm and TN as well as fiber content in tealeaves.

Fig. 6 shows the regressive analysis results for TN and fiber content estimations. The results show that the TN and fiber content estimations with Arai's index is better than those estimations with the conventional wavelength of reflectance and NDVI.

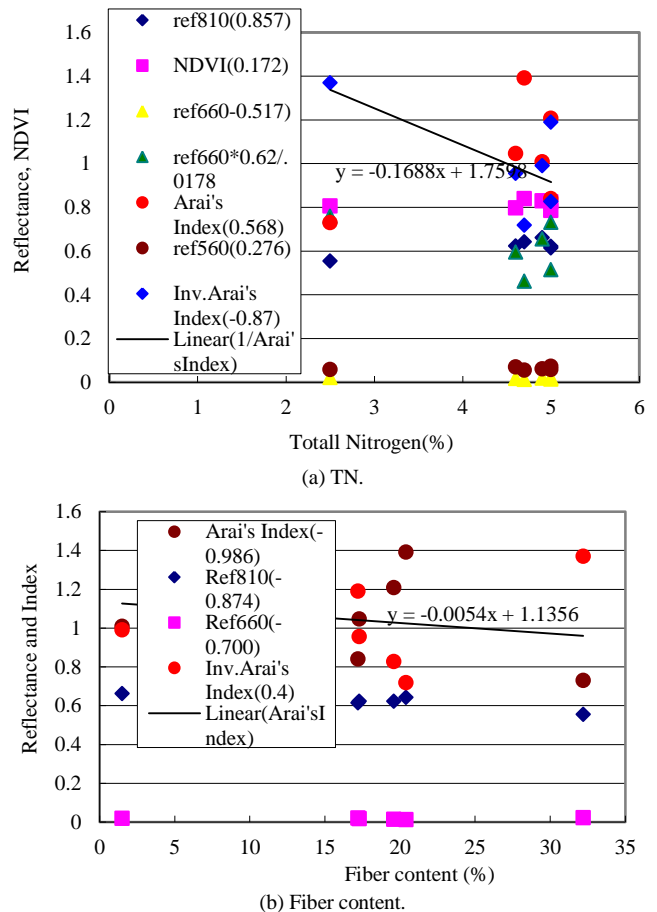


Fig. 6. Best index for estimation of TN and fiber content in tealeaf.

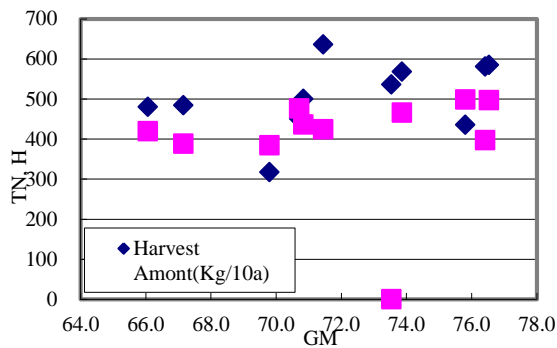


Fig. 7. Relation between GM value and TN as well as harvest amount in unit of Kg/10a.

### C. Harvest Amount Estimation

Other important factor for tea farm area evaluation is harvest amount which is highly correlated to GM value. Fig. 7 shows the relation between GM value and TN as well as harvest amount in unit of Kg/10a. As shown in Fig. 4, GM value is highly correlated to the reflectance at near-infrared wavelength region. Therefore, harvest amount can be estimated with the measured reflectance at near-infrared wavelength region.

The harvest amount of tealeaf is also highly correlated to the tea tree age. In general, expected tea quality and harvest amount proportional to their age as shown in Fig. 8. Also, vitality of tea trees are getting down with tea tree age of around 50.

### D. Trend Analysis of Total Nitrogen Content in Tealeaf

TN is changing for time being. During from October to March next year, tea tree keeps their vigor, or vitality. From the begging of April to the begging of May, tea tree has new tealeaves. Then the new tealeaves are harvested in the middle of May. After that, new tealeaves are born again from the middle of May to July. Then the second new tealeaves are harvested in July. From August to September, new tealeaves are born and growth up followed by third harvesting in October. This is the typical annual cycle of tea trees. Tealeaf quality and harvest amount is varied due to weather conditions, fertilizer, water supply, insect damage, and so on. That is why the total nitrogen content in tealeaves is changeable as shown in Fig. 9.

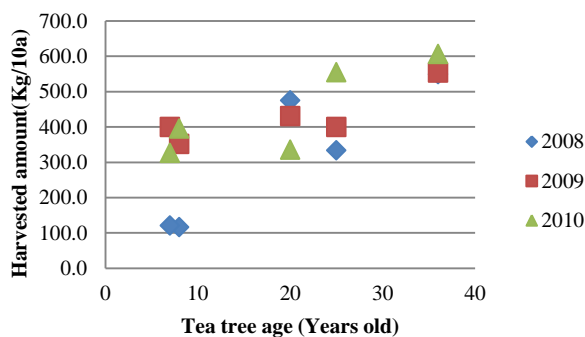


Fig. 8. Expected tea quality and harvest amount proportional to their age.

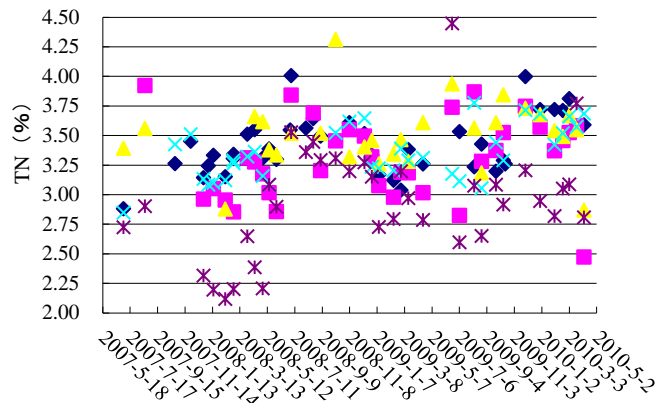


Fig. 9. Seasonal and annual changes of TN of tea farm areas situated at the 10 different farm areas in Ureshino, Saga, Japan during from 2007 to 2010.

Such the trend analysis can be done with the proposed Cadastral and Tea Production Management System with Wireless Sensor Network.

## IV. CONCLUSION

Cadastral and tea production management system utilizing wireless sensor network of Internet of Things (IoT) technology is proposed. In order to improve efficiency of tea productions, cadastral management and tea production processes have to be managed by Geographical Information System: GIS based system. Through experiments with sensor acquired data, it is found that the required information (quality: Total Nitrogen, Fiber Content, Water Content in tealeaves and harvest amount) can be estimated and represented efficiently. Thus the system works for improvement of the tea production management and quality control. For instance, it is possible to reduce tea farmer's labor cost about half (the number of look around monitoring). Fertilizer, pesticide and water resources can be reduced approximately 20%.

Further investigation is required for remote sensing satellite big data analysis utilizing Artificial Intelligence. Time series analysis for estimation of tealeaf quality and harvest amount can be done with AI much accurately.

## ACKNOWLEDGMENT

Author would like to thank Dr. Hideo Miyazaki of the Saga Prefectural Tea Research Institute for his cooperation through this research work.

## REFERENCES

- [1] J.T.Compton, Red and photographic infrared linear combinations for monitoring vegetation, *Journal of Remote Sensing of Environment*, 8, 127-150, 1979.
- [2] C.Wiegand, M.Shibayama, and Y.Yamagata, Spectral observation for estimating the growth and yield of rice, *Journal of Crop Science*, 58, 4, 673-683, 1989.
- [3] S.Tsuchida, I.Sato, and S.Okada, BRDF measurement system for spatially unstable land surface-The measurement using spectro-radiometer and digital camera- *Journal of Remote Sensing*, 19, 4, 49-59, 1999.
- [4] K.Arai, *Lecture Note on Remote Sensing*, Morikita-shuppan Co., Ltd., 2000.
- [5] K.Arai and Y.Nishimura, Degree of polarization model for leaves and discrimination between pea and rice types of leaves for estimation of leaf area index, Abstract, COSPAR 2008, A3.10010-08#991, 2008.

- [6] K.Arai and Long Lili, BRDF model for new tealeaves and new tealeaves monitoring through BRDF monitoring with web cameras, Abstract, COSPAR 2008, A3.10008-08#992, 2008.
- [7] Greivenkamp, John E., *Field Guide to Geometrical Optics*. SPIE Field Guides vol. FG01. SPIE. ISBN 0-8194-5294-7, 2004.
- [8] Seto R H. Nakamura, F. Nanjo, Y. Hara, *Bioscience, Biotechnology, and Biochemistry*, Vol.61 issue9 1434-1439 1997.
- [9] Sano M, Suzuki M ,Miyase T, Yoshino K, Maeda-Yamamoto, M.,*J.Agric.Food Chem.*, 47 (5), 1906-1910 1999.
- [10] Kohei Arai, Method for estimation of growth index of tealeaves based on Bi-Directional reflectance function: BRDF measurements with ground based network cameras, *International Journal of Applied Science*, 2, 2, 52-62, 2011.
- [11] Kohei Arai, Wireless sensor network for tea estate monitoring in complementally usage with Earth observation satellite imagery data based on Geographic Information System(GIS), *International Journal of Ubiquitous Computing*, 1, 2, 12-21, 2011.
- [12] Kohei Arai, Method for estimation of total nitrogen and fiber contents in tealeaves with ground based network cameras, *International Journal of Applied Science*, 2, 2, 21-30, 2011.
- [13] Kohei Arai, Monte Carlo ray tracing simulation for bi-directional reflectance distribution function and growth index of tealeaves estimation, *International Journal of Research and Reviews on Computer Science*, 2, 6, 1313-1318, 2011.
- [14] K.Arai, Monte Carlo ray tracing simulation for bi-directional reflectance distribution function and growth index of tealeaves estimations, *International Journal of Research and Review on Computer Science*, 2, 6, 1313-1318, 2012.
- [15] K.Arai, Fractal model based tea tree and tealeaves model for estimation of well opened tealeaf ratio which is useful to determine tealeaf harvesting timing, *International Journal of Research and Review on Computer Science*, 3, 3, 1628-1632, 2012.
- [16] Kohei Arai, Method for tealeaves quality estimation through measurements of degree of polarization, leaf area index, photosynthesis available radiance and normalized difference vegetation index for characterization of tealeaves, *International Journal of Advanced Research in Artificial Intelligence*, 2, 11, 17-24, 2013.
- [17] K.Arai, Optimum band and band combination for retrieving total nitrogen, water, and fiber in tealeaves through remote sensing based on regressive analysis, *International Journal of Advanced Research in Artificial Intelligence*, 3, 3, 20-24, 2014.
- [18] Kohei Arai, Hideo Miyazaki, Masayuki Akaishi, Tea tree vitality evaluation method and appropriate harvesting timing determination method based on visible and near infrared camera data, *Journal of Japan Society of Photogrammetry and Remote Sensing*, 51, 1, 38-45, 2012.
- [19] K.Arai, Optimum band and band combination for retrieving total nitrogen, water, and fiber in tealeaves through remote sensing based on regressive analysis, *International Journal of Advanced Research in Artificial Intelligence*, 3, 3, 20-24, 2014.
- [20] Kohei Arai, Kyushu small satellite for remote sensing (QSAT/EOS) and value added tealeaves "Eisei-no-megumi Ureshino-cha", *Journal of Society for Instrument Control Engineering of Japan*, 53, 11, 988-996, 2014
- [21] Kohei Arai, Yoshihiko Sasaki, Shihomi Kasuya, Hideto Matsuura, Appropriate tealeaf harvest timing determination based on NIR images of tealeaves, *International Journal of Information Technology and Computer Science*, 7, 7, 1-7, 2015
- [22] Kohei Arai, Yoshihiko Sasaki, Shihomi Kasuya, Hideo Matsuura, Appropriate harvest timing determination referring fiber content in tealeaves derived from ground based NIR camera images, *International Journal of Advanced Research on Artificial Intelligence*, 4, 8, 26-33, 2015.
- [23] K.Arai, Method for Vigor Diagnosis of Tea Trees Based on Nitrogen Content in Tealeaves Relating to NDVI, *International Journal of Advanced Research on Artificial Intelligence*, 5, 10, 24-30, 2016.
- [24] Kohei Arai, Bigdata Platform for agricultural field monitoring and environmental monitoring, *Proceedings of the 4th LISAT Symposium (Invited Speech)*, p.37, 2017.
- [25] Kohei Arai, Multi-Layer Observation for Agricultural (Tea and Rice) Field Monitoring, *Proceedings of the Seminar at Bogor Agriculture University, Keynote Speech*, 2016.

#### AUTHORS PROFILE

**Kohei Arai.** He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post-Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a counselor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission-A of ICSU/COSPAR since 2008. He received Science and Engineering Award of the year 2014 from the minister of the ministry of Science Education of Japan and also received the Best Paper Award of the year 2012 of IJACSA from Science and Information Organization: SAI. In 2016, he also received Vikram Sarabhai Medal of ICSU/COSPAR and also received 37 awards. He wrote 37 books and published 570 journal papers as well as 370 conference papers. He is Editor-in-Chief of International Journal of Advanced Computer Science and Applications as well as International Journal of Intelligent Systems and Applications. <http://teagis.ip.is.saga-u.ac.jp/>

# LOD Explorer: Presenting the Web of Data

Karwan Jacksi

Computer Science Department  
University of Zakho  
Zakho, Iraq

Subhi R. M. Zeebaree

Computer and Communications  
Engineering Department  
Nawroz University  
Duhok, Iraq

Nazife Dimililer

Information Technology Department  
Eastern Mediterranean University  
Gazimagusa, N. Cyprus

**Abstract**—The quantity of data published on the Web according to principles of Linked Data is increasing intensely. However, this data is still largely limited to be used up by domain professionals and users who understand Linked Data technologies. Therefore, it is essential to develop tools to enhance intuitive perceptions of Linked Data for lay users. The features of Linked Data point to various challenges for an easy-to-use data presentation. In this paper, Semantic Web and Linked Data technologies are overviewed, challenges to the presentation of Linked Data is stated, and LOD Explorer is presented with the aim of delivering a simple application to discover triplestore resources. Furthermore, to hide the technical challenges behind Linked Data and provide both specialist and non-specialist users, an interactive and effective way to explore RDF resources.

**Keywords**—Semantic web; linked open data; linked data browsers; exploratory search systems; RDF; SPARQL

## I. INTRODUCTION

Day after day, the amount of uploaded data to the Web grows, due to the simple uploading process offered by World Wide Web (www) [1]. Thus, the Web has transformed into a giant semi-structured collection of data, which makes information retrieval a challenging task. Search engines are typically used for information retrieval from the Web, but finding highly relevant retrievals, efficient search skills are necessary.

Marchionini has categorized the search approaches into two groups: lookup and exploratory search [2]. In the lookup search approach, also called keyword-based search, database systems are used to find information using keywords. This is the widely used approach in the existing Web, aka Syntactic Web, where the data sources are mainly text formats and the search elements are known [3].

Exploratory search is a special information seeking method, where the goal of users is not essentially identified through the search process [4]. In this approach, learning and investigation are more important for a user than retrievals of facts and replies to queries. The user compares, investigates and learns new ideas and concepts for the retrieved information [5], [6].

The information retrievals in the Syntactic Web is limited to keywords. Thus, search engines use the user query keywords to retrieve information, where the quality of retrieved results is rather poor. To develop the issue, the contents of the Syntactic Web are enriched with annotations forming the Semantic Web [1], [7].

The Semantic Web is an extension and next generation of the WWW through standards by the W3C<sup>1</sup>. The data of the Semantic Web has well-defined meanings, can be understood and processed by machines, and allows machines and people to work in collaboration [8]. The Semantic Web combines the technologies of RDF<sup>2</sup>, OWL<sup>3</sup>, and XML<sup>4</sup> to enable the replacement of the Syntactic Web so as to provide search engines capability to understand the meaning of data [7].

The Semantic Web cannot be completed by only annotating the data on the Web, but the data has to be linked with each other so as the Web of data can be formed and be discovered by machines and people [9]. Linked data makes it possible to discover related data of a term once only a subset is given. Hence, the terms Semantic Web and Linked Data have been coined by Berners-Lee and defined the Linked Data as “Semantic Web done right” [10].

The Linked Data (LD) term points out to a set of steps to distribute and connect structured data on the Web. These steps were introduced by Berners-Lee in his impressions about Web architecture design issues and soon turned out to be the principles of LD [11].

In the hypertext Web, HTML documents are connected with each other using untyped hyperlinks, whereas LD depends on the documents having RDF formats to create typed links that connect things globally forming the Web of Data [12], [13]. Once the LD is presented under an open license, it's called Linked Open Data (LOD).

The rest of the research is organized as follows: DBpedia dataset is described in Section 2. In Section 3, related works are addressed, and LOD Explorer is presented in Section 4. Evaluation of the application elaborated in Section 5, and results of the evaluation is detailed in Section 6. Conclusions and future work are given in Section 7.

## II. DBpedia DATASET

DBpedia is a leading project for publishing LD started by individuals at the Free University of Berlin and Leipzig University in cooperation with OpenLink Software. The project was first published in public as a Linked Open Data dataset in 2007 with the intention of becoming a large,

<sup>1</sup> World Wide Web Consortium: [www.w3.org](http://www.w3.org)

<sup>2</sup> Resource Description Framework: [www.w3.org/RDF](http://www.w3.org/RDF)

<sup>3</sup> Web Ontology Language: [www.w3.org/OWL](http://www.w3.org/OWL)

<sup>4</sup> Extensible Markup Language: [www.w3.org/XML](http://www.w3.org/XML)



multilingual, semantic knowledge graph for an open data infrastructure. It is now the center of Linked Open Data cloud<sup>5</sup>. The data of the dataset is created from the extracted information of the Wikipedia using the DBpedia Information Extraction Framework. The latest release (2016-10) of DBpedia consists of 13 billion pieces of information (RDF triples) where 1.7 billion pieces were English edition extractions of Wikipedia, 6.6 billion from other language editions and 4.8 from Wikipedia Commons and Wikidata. The English edition of the DBpedia dataset defines 6.6 million entities out of which 4.9 million have abstracts and 1.7 million have depictions. Altogether, 5.5 million resources are classified in a reliable ontology, containing 1.5 million persons, 840 thousand places, 496 thousand works such as films and music albums, 286 thousand organizations, 306 thousand species, 58 thousand plants and 6 thousand diseases. In addition to 6.6 million entities, the overall count of DBpedia for the English version is 18 million resources which include 1.7 million of SKOS concepts (categories), 7.7 million redirect pages, 269 thousand disambiguation pages and 1.7 million intermediate nodes [14].

Entities of DBpedia have different varieties of information, they normally have types, links, categories, labels, links of LD, and literal descriptions related to them. Within the DBpedia dataset, there are relations to identical entities for other languages (for instance [ar.dbpedia.org](http://ar.dbpedia.org)), and there are associations to corresponding entities reside in other datasets as in case of YAGO dataset. Additionally, there are specific domain classes and properties such as the Person typed entity `dbpedia:Carl_XVI_Gustaf_of_Sweden` has `dbo:spouse` which donates to the entity `dbpedia:Queen_Silvia_of_Sweden`.

From the time when DBpedia dataset was publicly published, various services and tools have been developed around it. DBpedia Spotlight<sup>6</sup>, which is a tool for robotically annotating mentions of DBpedia resources in text [15]. DBpedia Lookup<sup>7</sup>, a web service which allows to look up for DBpedia entities by related keywords. The DBpedia mappings wiki<sup>8</sup>, an exertion to improve the DBpedia information by obtaining mappings between the dataset ontology and Wikipedia Infoboxes. The DBpedia Extraction Framework<sup>9</sup> uses the mappings to standardize information extracted from Wikipedia before creating structured information in RDF. Besides DBpedia tools, further independent tools and services have been developed which use DBpedia as their dataset. In the following section, a few of such tools and services are employed.

### III. RELATED WORKS

In recent times, Linked Data (LD) usage on the Web has remarkably enlarged. However, for the lay-users, it is still challenging to be used. Dealing with LD to be used and visualized has been known as problems from the time when the foundation of the Semantic Web [16]. The growing

development of LD applications resulted in providing a set of approaches to let users interact and grasp the LD notion. Some of approaches present LD as outline and table modes as in Tabulator<sup>10</sup> and Explorer<sup>11</sup>, others present LD as graphs as in Graphity<sup>12</sup> and RelFinder<sup>13</sup>, whereas a combination of both features can be found in other systems such as LODmilla<sup>14</sup>. Authors of [17] present a rich and state of the art survey of LD exploration systems.

SWOC<sup>15</sup> uses semantic connections in the DBpedia dataset to let humans explore its resources [3]. Besides of using the semantic properties of DBpedia, the system uses Web search engines and social tagging systems as external resources making a hybrid approach to present DBpedia nodes. The system made up of two main modules: back-end, where the calculations of pairs between DBpedia resources are performed to produce similarities for the initial node, and a flash-based front-end presenting the results of the back-end.

At the front-end, DBpedia lookup service<sup>16</sup> is utilized to select an initial. The selected node, which should be of the ICT area, is presented on the webpage surrounding with most ten similar resources computed in the back-end. At the right side, a windowpane is available to present basic information about the selected resource.

LED<sup>17</sup> utilizes DBpedia dataset to provide users related resources to a query [18]. It uses DBpedia lookup service to return a resource of RDF dataset. Later, the system forms a cloud of tags that are semantically related to the selected resource. New tags from the formed cloud can be added to the main query resulting in a new query of the combined resources in a new tab. A pop-up pane for each resource is available while hovering on a tag presenting a description of the tag.

Aemoo<sup>18</sup> uses Encyclopedic Knowledge Pattern (EKP)<sup>19</sup> to explore the data of DBpedia [19]. When the system gets a query, it uses DBpedia first to process the query, then Wikipedia, Twitter, and Google News are used as external sources to assemble and combine the data from. The combination of data is achieved by principles of cognitively sound approaches by using knowledge patterns, the structure of hypertext links, and utilizing technologies of the semantic web. To present the retrieved data, EKP filters are used so that only related data is presented. A further utility called *curiosity* is offered by the system so that to show the filtered information by the EKP.

LodLive<sup>20</sup> explores RDF resources and visualizes them as dynamic graphs [20]. Resources in this system can be

<sup>5</sup> [lod-cloud.net/versions/2017-02-20/lod.svg](http://lod-cloud.net/versions/2017-02-20/lod.svg)

<sup>6</sup> [demo.dbpedia-spotlight.org/](http://demo.dbpedia-spotlight.org/)

<sup>7</sup> [github.com/dbpedia/lookup](http://github.com/dbpedia/lookup)

<sup>8</sup> [mappings.dbpedia.org](http://mappings.dbpedia.org)

<sup>9</sup> [github.com/dbpedia/extraction-framework](http://github.com/dbpedia/extraction-framework)

<sup>10</sup> [w3.org/2005/ajar/tab](http://w3.org/2005/ajar/tab)

<sup>11</sup> [tecweb.inf.puc-rio.br/explorator](http://tecweb.inf.puc-rio.br/explorator)

<sup>12</sup> [graphity.org](http://graphity.org)

<sup>13</sup> [visualdataweb.org/refinder.php](http://visualdataweb.org/refinder.php)

<sup>14</sup> [lodmilla.sztaki.hu](http://lodmilla.sztaki.hu)

<sup>15</sup> [sisinflab.poliba.it/semantic-wonder-cloud](http://sisinflab.poliba.it/semantic-wonder-cloud)

<sup>16</sup> [wiki.dbpedia.org/Lookup](http://wiki.dbpedia.org/Lookup)

<sup>17</sup> [sisinflab.poliba.it/led](http://sisinflab.poliba.it/led)

<sup>18</sup> [wit.istc.cnr.it/aemoo](http://wit.istc.cnr.it/aemoo)

<sup>19</sup> [ontologydesignpatterns.org/ekp](http://ontologydesignpatterns.org/ekp)

<sup>20</sup> [en.lodlive.it](http://en.lodlive.it)

connected from different endpoints. By using the Sesame Framework, RDF data can be parsed even when they are not in a SPARQL endpoint. This can be achieved by remotely creating graphs in order to store the requested resources temporarily for making queries. The system can also be used as a tool for the ontology definitions in its early stages so as to check the validity of an RDF schema and select a solution among several ones visually. The application is built using JavaScript and presents the calls from endpoints in HTML5 web pages. The retrievals of JSON format of JSONP (JSON with Padding) calls from endpoints are parsed to HTML documents without the need of a server-side programming.

LODmilla<sup>21</sup> is a LOD browser and editor that combines the features of both textual and graph-based LD browsers [21]. The system provides the abilities to connect to several LD datasets and browse the LD resources. Editing the resources is one of the main features of the application. The system consists of two main parts, a frontend side and a server side. The frontend is constructed using JavaScript while Java has been used for the server side. A dedicated server has been set for the system so as to enable search functions and support caching and fast loading of RDF triples. Two techniques can be used when loading RDF triples: a SPARQL-based query and actionable URIs. Using the Jena toolkit at the server side, several serializations can be obtained from parsing RDF data including JSON. Hence, multiple datasets can be used in parallel regardless of configuring the details of datasets at the frontend. The editing functions of the system give users abilities to add or remove resources or to make new connections between resources of a dataset.

LD Viewer is an adaptable framework of several tools to present a user-friendliness exploration of LD datasets [22]. The main target of the project is to provide a unified and powerful featured interface that can easily be accepted by several LD datasets. The retrieved information from the RDF datasets is presented in a tabular form of properties. Forward and reverse exploration of properties for each of the retrieved resource are offered, furthermore, a pagination feature for reverse properties of a large amount of values is available. Based on the nature of triples, each triple in the property table has action(s) which can be clicked. For instance, annotations to DBpedia dataset can be accomplished if the action is applicable for such triple. The application is implemented by JavaScript and largely by using AngularJS framework, and components of JASSA library<sup>22</sup> (JavaScript Suite for Sparql Access). Configuring the application with an LD dataset does not need to understand the core of the application.

#### IV. LOD EXPLORER

Thus far, the size of the LD growing intensely, subsequently, a lot of LD projects are available to be used and millions of triples have been put away in triple datasets. But from the opposing point of view, it is challenging to find exploring tools truly based on RDF standards and capable to

validate the efficiency of these standards. LOD Explorer<sup>23</sup> has been developed with the aim of:

- RDF datasets exploration employing a dynamic visual graph
- using different RDF datasets to be used and connected with each other
- expanding the norm and standardization space of LD
- providing an easy application to be used by everybody for LD Exploration
- presenting data properties of LD resources
- searching within the resources to find it's connections
- fetch and display an image of the resource
- providing flexibility for adding plugins.

The fundamental idea of the LOD Explorer is to deliver an easy approach to discover, understand, and learn the published resources along with the W3C standards for Semantic Web.

The novelty of the proposed approach is the capability to straightaway explore a SPARQL endpoint utilizing the greatness of JavaScript and its libraries without a necessity of a server-side module.

LOD Explorer uses the technologies of JSONP calls to the constructed endpoints fetching JSON formatted data to be parsed by JavaScript and presents the LOD resources in an HTML5 web page. The resources are presented as graph nodes while their properties as textual information with the aim of mixing the best of both worlds. Hence, this way, the significance of using SPARQL endpoints can be proved and promote using triplestores to develop federated queries.

LOD Explorer processes RDF data in advance and organizes them to be presented. The system presents all existing materials in RDF datasets without hiding any of its portions. For instance, property types are used to group In/Out properties.

The exploration process can be started by querying the endpoint for a particular resource either by using a resource name or a resource URI. A couple of resource examples are provided as well where one can start from. Afterward, exploring the resource is easy as can be through an attractive information presentation and following the related incoming and outgoing connections. New resources can be added to the graph and each of the newly opened resources will automatically connect to the ones already opened if and only if there is a semantic connection between them.

The system is constructed using the following technologies:

- Pure JavaScript
- jQuery libraries
- jsPlumb toolkit<sup>24</sup> to draw nodes of graph
- an HTML5 page

<sup>21</sup> lodmilla.sztaki.hu

<sup>22</sup> \_aksw.org/Projects/Jassa.html

<sup>23</sup> lodexplorer.uoz.edu.krd

<sup>24</sup> jsplumbtoolkit.com

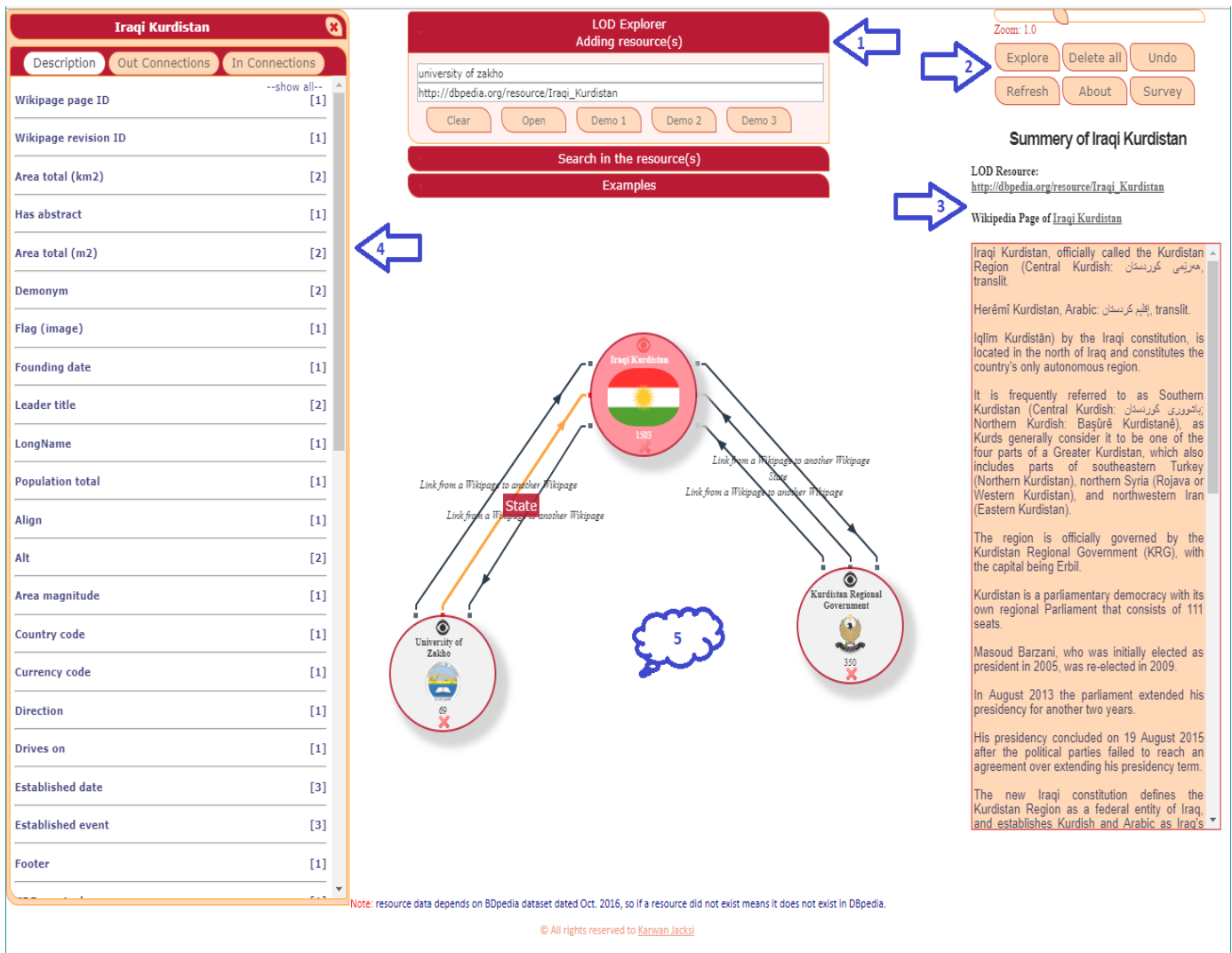


Fig. 1. System interface.

The user interface of the application consists of the following parts as in Fig. 1: System interface.

- 1) Search panel (center)
- 2) Toolbox (top right)
- 3) Description box (right)
- 4) Details Panel (left)
- 5) and the ground

The ground of the application is where the resources are presented in the form of graph nodes. The search panel is the main part of the system where resources can be found from LOD datasets and presented on the background. The resources are opened from this panel using either the resource name or the resource URI. When using resources names, an autocomplete search is offered by the system so as to select one of the offered resources. While when using URI of a resource, the available open button has to be clicked. Hence, the resources are opened this way and are drawn on the ground as graph nodes. Moreover, the nodes can shrink and enlarge by zooming them in and out, and they can be moved around anywhere on the ground using the mouse.

When a resource is opened, a search function from the search panel is activated so that to search inside the opened resource and find related information to the resource, as in Search in the resources. When multiple resources are open, the search function searches inside all of the opened resources. Results of search within resources are given in the form of active autocomplete combining suggestions of all of the related information to the opened resources right below its input box. The selected suggestion from the results opens the details panel.

The details panel contains all the details of the opened resource. This panel can be opened by either clicking on the eye button as in Resource as a graph node, or through the search within resource results. The panel consists of three main parts: 1) the description tab, 2) the out connections tab and 3) the in connections tab. The description tab contains detailed properties about the resource itself that are of the type literals. In and out connections are defined by the direction property and are presented in groups as labels having elements with targeted URIs. The panel is labeled with the resource label so

to realize the opened resource, and it can be closed to give more space to the background.

During this process, some presumptions are set to the nodes to enhance the visual appearance. For instance, a searching icon is set to let the user wait for the process to get completed. The node image is taken from the value of resource property `dbpedia:thumbnail` and `foaf:depiction`. And if that value is not available, the values of `rdf:type` property are used to show predefined icons such as no endpoint, person, group, work...etc. The values of `foaf:name`, `rdfs:label`, `skos:prefLabel` or `dc:title` properties are used for the node label.

Newly opened resources are inserted to the page without affecting the existing ones, this is helpful to let the surfer realize the new resource and to provide a least disruptive technique. After inserting new resources, the search within resources' array gets enriched with new information from the new resource. Any opened resource can be deleted as well as individual, this can be done using the cross sign (X) from the node. As a result, all LD related to the deleted node is removed from the search array.

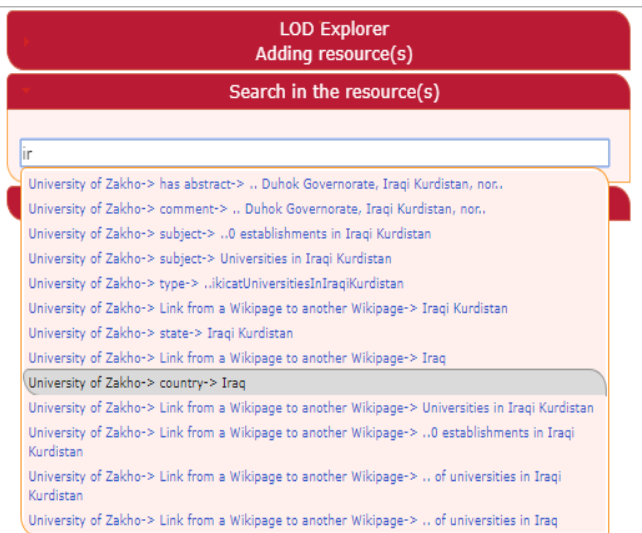


Fig. 2. Search in the resources.



Fig. 3. Resource as a graph node.

The right-hand up buttons are working as follow: the Explore button is used to expand the exploration process by inserting a predefined number of connections from the configuration file (currently set to 5). So, when this action is clicked, the system inserts 5 new nodes related to the selected node and present them to the page having direct connections to the selected node. The aim of this method is to help users get a larger vision of LOD exploration and to give them a better idea of how the system works.

A Delete all button, as from its name, it deletes all the opened resources and removes them from the search in the resources array. An Undo utility has been employed as well so as to go back to the last actions the user made sequentially.

## V. EVALUATION

To assess the proposed system, a user survey is conducted. The survey is based on System Usability Scale (SUS)<sup>25</sup>, which is an effective tool for evaluating the usability of a product and signifies a self-reported survey metric. The SUS scores score can range from 0 to 100, the highest score the highest level of efficiency, productivity, and satisfaction to the application [23].

The users of the survey have to work on the system first prior starting the evaluation. Therefore, the system has been uploaded to an online host for that purpose. With this survey, realizing whether the users in general like the system and how intuitive they're experiencing it are the targets.

The survey consists of two main parts: the first part includes questions to build a simple user profile. Only questions about users' affiliation, academic rank and degree, and discipline are asked. The second part of the survey is the standard SUS questions, which consists of 10 questions with 5 response options to show an average user satisfaction or dissatisfaction. At the end of the questionnaire, a suggestion field is also added. The SUS questions are listed below:

- Q1. I think that I would like to use this system frequently.
- Q2. I found the system unnecessarily complex.
- Q3. I thought the system was easy to use.
- Q4. I think that I would need the support of a technical person to be able to use this system.
- Q5. I found the various functions in this system were well integrated.
- Q6. I thought there was too much inconsistency in this system.
- Q7. I would imagine that most people would learn to use this system very quickly.
- Q8. I found the system very cumbersome/awkward to use.
- Q9. I felt very confident using the system.
- Q10. I needed to learn a lot of things before I could get going with this system.

<sup>25</sup> en.wikipedia.org/wiki/System\_usability\_scale

The response format is: strongly agree (SA), agree (A), neutral (N), disagree (DA), and strongly disagree (SDA).

## VI. EVALUATION RESULTS

The survey is sent to 80 individuals, out of which 62 were responded. Around 19% were Ph.D. degree holders, 8% were Ph.D. students, and 50% were Masters. The academic rank of the participants was as follows: 2% Profs, 11% Assist Profs, 10 Lecturers, 42% Assistant Lecturers and 36% with no academic title. Discipline was an important factor in the survey so as to know the feedback from the more specialized participants. 74% of the participants were from Computer Science specialists, and the rest were from Chemistry, Biology, History, Economics, Environmental Science, Law, Civil, Mechanical and Petroleum Engineering.

The initial survey shows participants overall like the application. Responses to Q1 were 39%SA, 42%A, and 16%N, which indicate the users like to use the system. Around 15% found the system is unnecessarily complex, while 84% (44%SA, 40%A) through the application is easy to use. 15% of the participants need assistance to use the application, and they're mostly from unspecialized people. 78% (23%SA, 55%A) went for Q5, and 8% thought there is inconsistency in the application. For the question: most people would learn to use this application very quickly the responses were: 26%SA, 47%A, and 19%N. Feedback for Q8 was 37%SDA, 39%DA, and 15%N. 87% felt very confident to use the application, while 13% needed to learn many things before using the application.

Most comments to the system were to compliment the efforts taken building this application while one of them was interesting since it was talking about the found resources are not up-to-date and this is of course not a fault of the system since it depends on the DBpedia dataset version 2016-10.

The suggestions part of the survey was an important plan to improve the system. Nine suggestions for the system have been recorded, some of which were well valued. Somebody suggested disabling the search within resources function when there are no resources on the ground to search within it; this has been implemented and added to the system. Someone else advised adding auto-correction feature to the search process, while other one said to include more datasets and provide an ability for users to select a shape from a list of shapes for nodes such as squares or hexagons.

The scores of SUS have been converted to a new number of all items by normalizing the scales to a range from (0-4). For positive formulated questions (or odd questions), the normalization is as follow: for the highest score, 4 is given to strongly agree and 0 to strongly disagree. But, for negative expressed questions (even questions), the range is given as 0 to strongly agree and 4 to strongly disagree. Later, the numbers are multiplied by 2.5 to transform the original scores from 0-40 to 0-100.

Based on studies, a score of a SUS survey that is below 68 is considered as below average, and above that benchmark considers above average. The SUS scores for the proposed application are 76.01 which exceed by far the benchmark of 68. However, further improvements can be made to deliver

even higher levels of usability and satisfaction. The evaluation results for each question can be seen from Average Scores to SUS Questions.

TABLE I. AVERAGE SCORES TO SUS QUESTIONS

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Over-all
3.1	2.7	3.2	2.8	3.0	2.9	2.9	3.1	3.4	2.9	76.0
5	9	7	4	2	2	5	3	2	2	1

## VII. CONCLUSION AND FUTURE WORK

The amount of publishing data consistent with the standards of Linked Data is growing dramatically. But, consumption is still limited for professionals who understand the technologies of Linked Data. Thus, a tool for intuitive presentation of Linked Data is crucial. LOD Explorer, an interactive and easy-to-use tool for exploring RDF resources, is presented. The application is made using pure JavaScript and jQuery libraries without the need for a server-side software. An evaluation of the application is employed using the known user survey System Scalability Scale (SUS) tool, and the evaluation results were by far acceptable.

The future plans for the tool are to enrich it with several further functions such as adding more RDF datasets, giving users an opportunity to select a desired shape for the nodes, adding pathfinding feature so as to find the exact relationship between two or more resources.

### REFERENCES

- [1] G. Madhu, D. A. Govardhan, and D. T. Rajinikanth, "Intelligent Semantic Web Search Engines: A Brief Survey," ArXiv Prepr. ArXiv11020831, 2011.
- [2] G. Marchionini, "Exploratory Search: From Finding to Understanding," Commun ACM, vol. 49, no. 4, pp. 41-46, Apr. 2006.
- [3] R. Mirizzi, A. Ragone, T. D. Noia, and E. D. Sciascio, "Semantic Wonder Cloud: Exploratory Search in DBpedia," in Current Trends in Web Engineering, F. Daniel and F. M. Facca, Eds. Springer Berlin Heidelberg, 2010, pp. 138-149.
- [4] T. Jiang, "Exploratory Search: A Critical Analysis of the Theoretical Foundations, System Features, and Research Trends," in Library and Information Sciences, Springer, 2014, pp. 79-103.
- [5] G. Marchionini and B. Shneiderman, "Finding Facts vs. Browsing Knowledge in Hypertext Systems," Computer, vol. 21, no. 1, pp. 70-80, Jan. 1988.
- [6] K. Jacksi, N. Dimililer, and S. R. Zeebaree, "a survey of exploratory search systems based on lod resources," Proc. 5th Int. Conf. Comput. Inform. ICOCI 2015, pp. 501-509, 2015.
- [7] J. A. R and M. Kurian, "A Survey on Tools essential for Semantic web Research," Int. J. Comput. Appl., vol. 62, no. 9, pp. 26-29, Jan. 2013.
- [8] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," Sci. Am., vol. 284, no. 5, pp. 28-37, 2001.
- [9] T. Berners-Lee, "Linked Data - Design Issues," Linked Data, 27-Jul-2006. [Online]. Available: <http://www.w3.org/DesignIssues/LinkedData.html>. [Accessed: 20-Dec-2017].
- [10] K. Krieger and D. Rosner, "Linked Data in E-Learning: A Survey," Semantic Web 0, pp. 1-9, 2011.
- [11] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, 1st edition. Morgan & Claypool., 2011.
- [12] G. Klyne, J. J. Carroll, and B. McBride, "Resource Description Framework (RDF): Concepts and Abstract Syntax," Resource Description Framework (RDF): Concepts and Abstract Syntax, 10-Feb-2004. [Online]. Available: <http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/>. [Accessed: 20-Dec-2017].

- [13] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, pp. 1–22, 2009.
- [14] "DBpedia version 2016-10 | DBpedia." [Online]. Available: <http://wiki.dbpedia.org/datasets/dbpedia-version-2016-10>. [Accessed: 23-Dec-2017].
- [15] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia spotlight: shedding light on the web of documents," presented at the Proceedings of the 7th international conference on semantic systems, 2011, pp. 1–8.
- [16] V. Geroimenko and C. Chen, *Visualizing the semantic web: XML-based internet and information visualization*. Springer Science & Business Media, 2006.
- [17] K. Jacksi, N. Dimililer, and S. R. Zeebaree, "State of the Art Exploration Systems for Linked Data: A Review," *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 7, no. 11, pp. 155–164, 2016.
- [18] R. Mirizzia, A. R. T. Di Noiaa, and E. Di Sciascioa, "Lookup, Explore, Discover: how DBpedia can improve your Web search," 2010.
- [19] A. Musetti et al., "Aemoo: Exploratory search based on knowledge patterns over the semantic web," *Semantic Web Chall.*, 2012.
- [20] D. V. Camarda, S. Mazzini, and A. Antonuccio, "Lodlive, exploring the web of data," presented at the Proceedings of the 8th International Conference on Semantic Systems, 2012, pp. 197–200.
- [21] A. Micsik, Z. Tóth, and S. Turbucz, "LODMilla: Shared Visualization of Linked Open Data," presented at the Theory and Practice of Digital Libraries--TPDL 2013 Selected Workshops, 2014, pp. 89–100.
- [22] D. Lukovnikov, C. Stadler, and J. Lehmann, "LD viewer-linked data presentation framework," presented at the Proceedings of the 10th International Conference on Semantic Systems, 2014, pp. 124–131.
- [23] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: Adding an adjective rating scale," *J. Usability Stud.*, vol. 4, no. 3, pp. 114–123, 2009.

# Agent-Based System for Efficient $kNN$ Query Processing with Comprehensive Privacy Protection

Mohamad Shady Alrahhah<sup>1</sup>, Maher Khemakhem<sup>2</sup>, Kamal Jambi<sup>3</sup>  
King Abdulaziz University (KAU)  
Jeddah, Saudi Arabia

**Abstract**—Recently, location based services (LBSs) have become increasingly popular due to advances in mobile devices and their positioning capabilities. In an LBS, the user sends a range of queries regarding his  $k$ -nearest neighbors ( $kNN$ s) that have common points of interests (POIs) based on his real geographic location. During the query sending, processing, and responding phases, private information may be collected by an attacker, either by tracking the real locations or by analyzing the sent queries. This compromises the privacy of the user and risks his/her safety in certain cases. Thus, the objective of this paper is to ensure comprehensive privacy protection, while also guaranteeing the efficiency of  $kNN$  query processing. Therefore, we propose an agent-based system for dealing with these issues. The system is managed by three software agents ( $selector_{DL}$ ,  $fragmentor_Q$ , and  $predictor$ ). The  $selector_{DL}$  agent executes a Wise Dummy Selection Location (WDSL) algorithm to ensure the location privacy. The mission of the  $selector_{DL}$  agent is integrated with the mission of the  $fragmentor_Q$  agent, which is to ensure the query privacy based on Left-Right Fragmentation (LRF) algorithm. To guarantee the efficiency of  $kNN$  processing, the  $predictor$  agent executes a prediction phase depending on a Cell Based Indexing (CBI) technique. Compared to similar privacy protection approaches, the proposed WDSL and LRF approaches showed higher resistance against location homogeneity attacks and query sampling attacks. In addition, the proposed CBI indexing technique obtains more accurate answers to  $kNN$  queries than the previous indexing techniques.

**Keywords**—Agents; attacks; dummies; fragmentation; indexing; privacy protection; resistance

## I. INTRODUCTION

Location Based Services (LBSs) are services that are customized according to the location of the user. In recent years, LBSs have received substantial attention, especially since GPS-enabled devices (such as smart phones) became popular. One of the most important advantages of LBS-enabled applications is their ability to search for the nearest Point of Interests (POIs). Searching for the nearest POIs requires construction of a query on the LBS user side. Table I summarizes the units of the constructed query.

TABLE I. GENERAL FORM OF THE LBS QUERY

Symbol	$\langle X, Y \rangle$	POI	R	ID
Description	Coordinates of the real location	Queried interests	Queried range	The identity of the LBS user

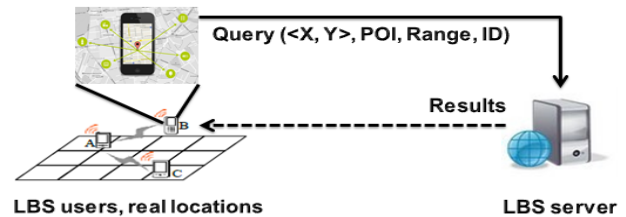


Fig. 1. Classical scenario of using LBS applications.

As a general example of LBS usage, Fig. 1 illustrates the classical scenario of using LBS-enabled applications based on the query units that are listed in Table I.

In Fig. 1, the LBS user constructs a query regarding a desired POI and sends it to the LBS server. Then, the LBS server processes the query and sends back the results. However, this classical scenario involves risk since the LBS user is forced to construct the query based on his/her real geographic location. This risk is directly related to the privacy issue of the LBS user. The reason behind this risk is that an attacker can track the real location of the LBS user [1] or intercept the sent query for analysis purposes [2]. In both cases, the attacker can collect sensitive or personal information about the LBS user, such as customs, habits, religion, or political leanings. Then, this personal information can be misused to conduct attacks in real life, such as mugging, extortion or stealing. According to [3], these two methods of personal data collection can lead to branches of two kinds of privacy: location privacy and query privacy. Therefore, if we want to achieve full privacy protection, we need to protect these two kinds of privacy. However, achieving comprehensive privacy protection requires protecting the query privacy (in addition to the location privacy) at the sending, processing, and responding levels. Comprehensive LBS privacy protection has not been addressed previously to the best of our knowledge.

The queried POI, are either static POIs (such as the nearest hotels, hospitals, or sports clubs in a defined range) or moving POIs (such as the nearest taxis that will enter a defined range). When an LBS user searches for a moving POI, it is referred to as a range query or  $k$ -nearest neighbor ( $kNN$ ) query [4]-[7]. In manipulating  $kNN$  queries, two major issues arise: The first is related to ensuring the privacy protection of the  $kNN$  queries, which in turn ensures the privacy of the LBS user. The second is related to guaranteeing the accuracy of the retrieved results (i.e., the retrieved locations of the queried moving POI) [8], [9]. Fig. 2 illustrates the uncertainty problem, which is considered a real-time problem.

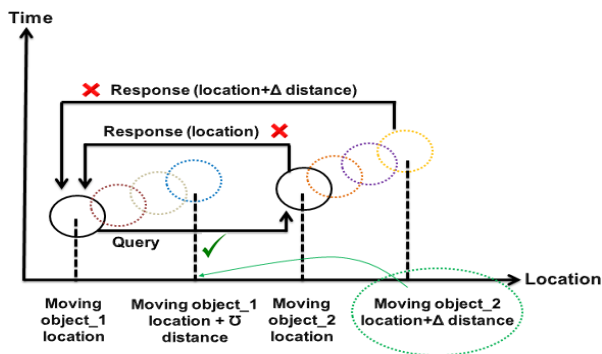


Fig. 2. Real-time uncertainty problem for k-NN queries.

According to Fig. 2, the first Moving Object (MO), as a query issuer, searches for a second MO. Because of the continuous updating of the locations of both the 1st MO and the 2nd MO in the real-time temporal and spatial domains, the query issuer will receive an unmatched value that is related to the exact location of the queried MO. The correct response to the query is  $(\text{location} + \Delta)$ , which must be delivered to the query issuer location  $(\text{location} + U)$ .

**Motivation.** Many efforts have been proposed to protect the privacy of continuous  $kNN$  queries and overcome the real-time uncertainty problem. One of the most important proposed approaches is the use of dummies. In the context of LBS privacy protection, a dummy is a query that is constructed based on a fabricated location or fabricated properties. If the LBS user surrounds his/her real location by some fabricated (or false) locations, location privacy protection will be achieved [10]-[12]. If the LBS user tampers with the properties of the query itself (changing the queried interest or POI, for example), query privacy protection will be asserted [13]. In both cases, the current query (real query) is mixed with a number of false queries (dummies) so that the attacker cannot recognize the real query among the dummies. This process (i.e., mixing process) aims at achieving  $k$ -anonymity in which the attacker cannot identify the real query among  $k-1$  dummies. However, achieving full privacy protection (i.e., location privacy and query privacy) by using dummies has not been addressed. Moreover, generating weak dummies allows an attacker to filter these dummies, thereby determining the accurate location of the LBS user. Beyond generating weak dummies, some inference attacks, such as location homogeneity attack [14] (which targets location privacy) and query analysis attack, such as query sampling attack [15] (which targets query privacy), can be applied by an attacker to circumvent the privacy protection methods. In both inference attacks and query analysis attacks, the attacker does not need to know the accurate location of the LBS user to infer the personal data. This, in turn, means that achieving robust privacy protection is a pressing need. Regarding the manipulation of  $kNN$  queries, many techniques have been proposed, such as  $R^*$ -tree [16],  $D$ -tree [17], and Grid-partition [18]. However, these techniques rely on Euclidean space to manipulate the  $kNN$  queries, whereas, in many real-life applications, the objects' movements are constrained in a road network. Moreover, these techniques cannot be applied in road networks because the network distance (i.e., the shortest path distance) cannot be computed using the boundary of the

minimum bounding rectangle (MBR) or grid cell. This, in turn, leads to a poor manipulation of the real-time uncertainty problem for  $kNN$  queries. Therefore, an efficient technique for manipulating  $kNN$  queries is a top requirement.

In this paper, based on agent software technology, we propose an agent-based system architecture for privacy protection of LBS users. Three main missions are assigned to three software agents, which are integrated with one another to ensure comprehensive privacy protection of  $kNN$  queries and overcome the real-time uncertainty problem. The main contributions of this work are as follows:

- To protect the location privacy of LBS users, we introduce a novel Wise Dummy Selection Location (WSDL) algorithm. The objective of our WSDL algorithm is to select strong dummy locations that cannot be distinguished from the real location of the LBS user. The power of the proposed WSDL algorithm comes from taking into consideration two main factors: 1) selecting the dummy locations based on the historical query probability of each cell; and 2) selecting dummy locations that are far away from one another based on the products of the distances among the selected dummies. This, in turn, gives the WSDL algorithm strong resistance against location homogeneity attack.
- To protect the query privacy, we introduce a novel Left-Right Fragmentation (LRF)-based algorithm. Our LRF-based algorithm extracts the sensitive units of the constructed query, encrypts them, and randomizes them to ensure resistance to query sampling attacks.
- To enhance the real-time uncertainty problem, we introduce a novel indexing technique called Cell-Based Indexing (CBI). Our indexing technique performs efficient motion modeling with a prediction phase to ensure that the exact locations of the queried MOs are retrieved.

The rest of this paper is structured as follows: Section II discusses related work. The threat model is provided in Section III. Our proposed agent-based architecture is provided in Section IV. Section V discusses the security analysis. In Section VI, we present the metrics that are used. Section VII presents our experimental results and the conducted evaluations. Finally, we conclude the paper in Section VIII.

## II. RELATED WORK

This section reviews some of the related work on privacy protection approaches in the LBS research field. In addition, we discuss some of the related work on techniques that are used to manipulate  $kNN$  queries.

### A. LBS Privacy Protection Approaches

Many efforts have been made to classify the privacy protection approaches in the domain of LBS, such as [3], [19], [20]. There are two major categories of LBS privacy protection approaches: server-based approaches and user-based approaches. In this subsection, we review some existing approaches from the user-based category that aim at protecting location privacy or query privacy.



The authors of work [10] proposed a dummy data array (DDA) algorithm for generating dummy locations to protect the location privacy of LBS users. For a given region, which is divided into a grid of cells, the key idea of the DDA algorithm is to calculate both the vertices and the edges of each cell in the grid. Then, the DDA algorithm randomly selects some of the cells as dummy locations. To select strong dummy locations and achieve  $k$ -anonymity, the DDA algorithm selects  $k$  cells of equal area. Similarly, [11] uses dummies to protect the location privacy of LBS users, but with a different dummy generation method. The authors proposed two algorithms. The first is called CirDummy, which generates dummies based on a virtual circle that contains the real location of the LBS user. The second is called GridDummy, which generates dummies based on a virtual grid that covers the real location of the LBS user. In [12], a dummy generation method called the Destination Exchange (Dest-Ex) method was proposed. In this method, historical motion trajectories are used to generate the dummies. To ensure that the generated dummies are strong, the Dest-Ex method chooses the historical trajectories that intersect with the current trajectory of the LBS user. Therefore, the attacker is confused when trying to determine the correct LBS user, who has several motion trajectories with different destinations. However, the main objective of all of these previous approaches was location privacy protection. To achieve query privacy protection, the authors of [13] proposed an approach called DUMMY-Q. The DUMMY-Q approach depends on the strategy of generating dummies, but the strategy is applied to the query, rather than the location. Therefore, dummy queries of different attributes from the same location are generated to hide the real query. To make the generated dummies stronger, two aspects are taken into consideration: 1) the query context; and 2) the motion model.

Encryption techniques have been employed to protect the privacy of LBS users. The authors of [21] proposed the idea of using buddies to protect both location privacy and query privacy against the LBS server (a malicious party). This approach depends on notifying the friends (buddies) of an LBS user who are located in the vicinity, thereby avoiding the revelation of any personal data to the LBS server. This approach assumes that each user shares a secret with each of his buddies and uses symmetric encryption techniques. Another approach was proposed based on using Private Information Retrieval (PIR) [22] to achieve full privacy protection. The key idea of the PIR technique depends on the quadratic residuosity assumption, which states that it is computationally hard to find the quadratic residues in modulo arithmetic of a large composite number for the product of two large primes. Therefore, the LBS server can process and answer the query without knowing any sensitive information about the query.

### B. Techniques of $kNN$ Query Manipulation

The Global Positioning System (GPS), which is integrated with the mobile devices of the LBS users, allows the users to obtain their locations from the satellite and send them to the LBS server. During movement, the locations of the LBS users are continuously updated on the LBS server side. This results in inaccurate retrieved locations when the LBS user asks for the  $kNN$  MOs as POIs. Therefore, the final goal of any

techniques that is used for manipulating the  $kNN$  queries is to retrieve approximate locations of the MOs as responses to the  $kNN$  queries.

Many techniques have been proposed for manipulating the  $kNN$  queries. In [16], a traditional method called P\*-tree was proposed for supporting range queries. The P\*-tree technique efficiently manipulates range queries with static POIs, but not moving POIs. Another technique was provided in [17], which is called D-tree. The key idea of D-tree is to index the data regions based on the divisions among them so that a binary D-tree index is constructed. For a given  $kNN$  query, two main phases are used to find and retrieve the queried POIs: region partitioning and location-dependency query processing based on paging the D-tree index. The authors of [18] developed the D-tree technique, proposing a Grid-partitioning technique. The authors used the Voronoi Diagram to partition the service area into disjoint Voronoi cells (VCs), with each corresponding to one object. An object  $a$ , is guaranteed to be the nearest neighbor to any client that is located inside the same VC. In [23], a new  $kNN$  query processing technique was proposed by Jang et al. based on the density of the POIs. A PIR protocol was used to search for the POIs within a clocking region, so that the clocking region was expanded to overlap other regions based on the  $k$ -d overlap index. However, in all the previous techniques, the index is constructed for large regions, thereby ignoring the cells that are included in the divided regions.

### III. THREAT MODEL

In this section, we define the threat model, which specifies the attacker and his/her objective. In addition, we determine the ways that are used by the attacker to collect personal information about the victim, in addition to inference and analysis attacks.

#### A. Attacker and His/Her Objective

The objective of the attacker is to obtain privacy information about a particular LBS user, including location, POI and queried range. To achieve his/her objective, the attacker can track the location of the LBS user or analyze the sent query, as shown in Fig. 3 below.

In the context of the threat model, we define two terms: passive attack and active attack. In a passive attack, any LBS user can act as an attacker. In an active attack, the LBS server (or its maintainer) is an attacker and all the information (related to the trajectories of the LBS user's motion) that is stored in the LBS server is accessible. Since an active attack is stronger than a passive attack, we only address active attack.

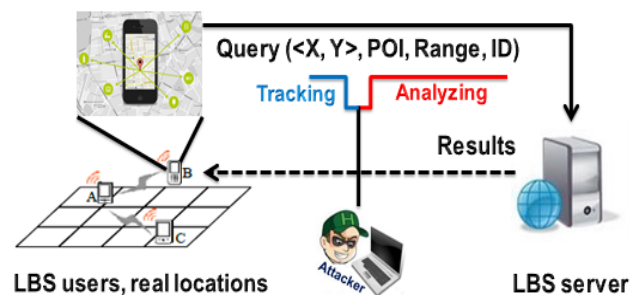


Fig. 3. Attacking the privacy of LBS users.

B. Inference Attacks and Query Analysis Attacks

The LBS server (an attacker) can apply inference attacks, such as location homogeneity attack, and query analysis attacks, such as query sampling attack.

In a location homogeneity attack, the attacker analyzes the locations of all LBS users. If their positions are almost identical, then the position information of each member is revealed. For instance, if the users are located in a place that represents a landmark such as a hospital, the attacker can infer that those users (including the victim) have problems related to their health, without needing to accurately identify their locations. Fig. 4 illustrates a location homogeneity attack.



Fig. 4. Location homogeneity attack: (H) hospital or medical area, (S) sport cub or athletic area, (R) restaurant or rest area.

In a query sampling attack, the attacker employs the uneven location distribution of the LBS users for his own malicious purposes. This attack targets isolated users in sparse regions, as illustrated in Fig. 5. Therefore, it relies on the traffic statistics of the environment where the users are located. In detail, the attacker tries to calculate a probability distribution function of the user location over a given area. If the distribution is not uniform, then the attacker can determine the areas where the user is located with a high probability. Once the location of the victim is determined, the attacker focuses on analyzing the sent queries.

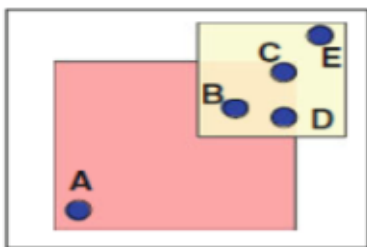


Fig. 5. Query sampling attack.

IV. OUR PROPOSED PRIVACY PROTECTION ARCHITECTURE

In this section, we provide our agent-based privacy protection architecture, followed by the roles of the agents. The details of the architecture are represented by a sequence diagram.

The framework of the proposed architecture consists of an untrusted LBS server (a malicious party) and a group of mobile devices, which are connected via a network. The system is managed by three agents ( $selector_{DL}$ ,  $fragmentor_Q$ , and  $predictor$ ), as shown in Fig. 6.

Table II lists the agents and identifies the main mission of each one, its type, and where it is installed.

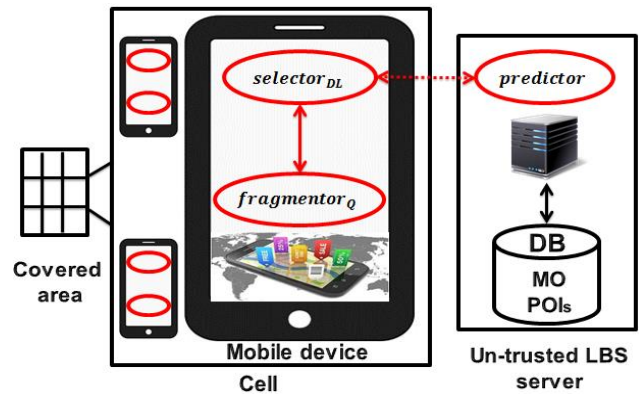


Fig. 6. Our agent-based architecture.

TABLE II. AGENTS

Agent Name	Type	Main Mission	Location
$Selector_{DL}$	Stationary	Location privacy protection	Each mobile device
$Fragmentor_Q$	Mobile	Query privacy protection	Each mobile device
Predictor	Stationary	Uncertainty real-time problem solution	LBS server

A. Roles of the Agents

**$Selector_{DL}$ :** This stationary agent executes the Wise Dummy Selection Location (WSDL) approach. It targets the location privacy protection against the untrusted LBS server, which can apply location homogeneity inference attack, as described below.

1) Wise Dummy Selection Location (WSDL) approach

The final objective of the WSDL approach is to generate strong dummy locations to protect the location privacy of the LBS user. In the dummy generation process, suitable locations are selected that cannot be distinguished from the real location of the LBS user. Consider a region (G) divided into a grid of cells. Each cell has a probability of being queried, which is based on past queries. This is referred to as the query probability. For a given LBS user in a cell within G, randomly selecting cells to be the dummy locations, as proposed in the DDA approach [10], for an example, it is a poor strategy. In contrast, selecting the cells (to be a dummy locations) that have the same query probabilities as the cell where the LBS user is located is an efficient solution. Fig. 7 illustrates this solution, where G is divided according to the coordinates (X, Y).

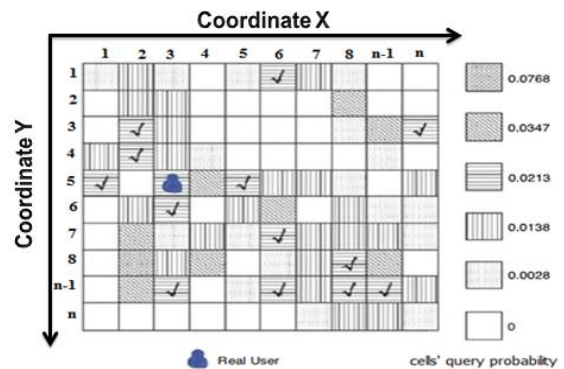


Fig. 7. Dummy locations selection in the WSDL approach.

In Fig. 7, if the LBS user who is located in the cell that is identified by row number five and column number three (i.e., the coordinates C[5, 3]) wants to protect his/her location privacy by achieving 4-anonymity level (i.e.,  $k=4$ ), he/she can select three of the cells that are marked by the  $\surd$  symbol. Since the query probability of any of the three selected cells equals the query probability of the original cell, the attacker cannot determine the real location of the LBS user among the  $k-1$  dummy locations.

In a formal way, for a given region (G) that is divided into  $(n \times n)$  cells, let (qp) refers to the query probability of a cell. Then,  $\sum_{i=1}^n qp_i = 1$ . Each of the  $k$  locations (i.e., cells) that are contained in a query, which include one real location and  $(k - 1)$  dummies, has a conditional probability of being the real location. Let  $\hat{p}_i$  ( $i = 1, 2, \dots, k$ ) denote the probability that the  $i^{\text{th}}$  location is the real location. Then,  $\hat{p}_i = \frac{qp_i}{\sum_{j=1}^k qp_j}$ .

The entropy (E) of identifying the real location out of the dummy set is defined as:

$$E = - \sum_{i=1}^k \hat{p}_i \times \log_2 \times \hat{p}_i \quad (1)$$

The first factor that is taken into consideration is the maximization of the entropy value in the dummy selection process.

$$\text{Max} (- \sum_{i=1}^k \hat{p}_i \times \log_2 \times \hat{p}_i) \quad (2)$$

### 2) Danger of location homogeneity inference attack

If the LBS user selects cells C[5, 1], C[4, 2], and C[6, 3], as shown in Fig. 8, some personal information can be inferred by the attacker without the need to determine the real location of the LBS user. This occurs because the three selected dummy locations belong to a medical area (which includes hospitals as a POIs, for example), then the attacker can infer that the LBS user has a health problem. Therefore, it is better to select the following three cells, for example, C[3, n], C[n-1, n-1], and C[1, 6].

To defend against location homogeneity attacks, a second factor is taken into consideration in the process of dummy location selection: “the selected dummy locations must be far away from one another”. In this context, the question arises as to how to determine the furthest dummy location from the real location of the LBS user and spreads away from the other dummy locations. This can be accomplished by calculating the distance between the real location of the LBS user and each dummy location based on the product distance rather than the normal sum distance. Fig. 8 illustrates the strategy of wise dummy location selection.

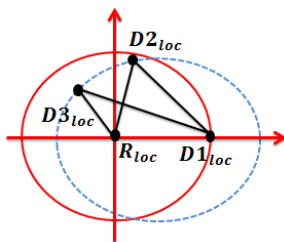


Fig. 8. Wise dummy location selection.

In Fig. 8,  $R_{loc}$  represents the real location of the LBS user, and  $D1_{loc}$ ,  $D2_{loc}$ , and  $D3_{loc}$  represent the dummy locations, where the query probability of each dummy location equals the query probability of the real location. Let the distance between two points  $PT_i$  and  $PT_j$  be given by  $\sum_{i \neq j} \text{dis}(PT_i, PT_j)$ .  $D1_{loc}$  is the first dummy location that can be directly selected since it is the furthest location from  $R_{loc}$ . If we want to achieve 3-anonymity level, we can choose  $D2_{loc}$  or  $D3_{loc}$ . If we consider the sums of distances between pairs of dummy locations, we can choose either of them ( $D2_{loc}$  or  $D3_{loc}$ ) because  $(|D2_{loc}R_{loc}| + |D2_{loc}D1_{loc}| = (|D3_{loc}R_{loc}| + |D3_{loc}D1_{loc}|)$ . However, to achieve higher resistance  $D2_{loc}$  is preferred over  $D3_{loc}$  since it spreads dummy locations farther. Therefore, instead of using the sum of distances between pairs of dummy locations, we can use their product. Note that  $(|D2_{loc}R_{loc}| \times |D2_{loc}D1_{loc}| > (|D3_{loc}R_{loc}| \times |D3_{loc}D1_{loc}|)$ . This leads to the choice of  $D2_{loc}$  as the second dummy location.

Mathematically, the two previous factors form two objectives in a Multi-Objective Optimization Problem (MOP). Let  $DL = [D1_{loc}, D2_{loc}, D3_{loc}, \dots, Dk_{loc}]$  denote the set of real and dummy locations. The MOP is defined as:

$$FD_{loc} = \arg \max \{- \sum_{i=1}^k \hat{p}_i \times \log_2 \times \hat{p}_i, \prod_{i \neq j} \text{dis}(D_{i_{loc}}, D_{j_{loc}})\} \quad (3)$$

Where,  $FD_{loc}$  represents the final selected dummy locations.

The first objective of the MOP was previously optimized in formula 2 because, from all the given dummy locations (i.e., all cells that form the region G), we select a set of dummy locations based on similarity of query probability. This set is called the set of candidate dummy locations ( $CD_{loc}$ ), which yields the maximum entropy value. Out of the candidate dummy locations, we optimize the second objective of the MOP as follows, which determines the final selected dummy locations:

$$FD_{loc} = \arg \max \{\prod_{i \neq j} \text{dis}(D_{i_{loc}}, D_{j_{loc}})\} \quad (4)$$

In steps, we first sort the cells according to their query probabilities. Second, we select  $4k$  cells from outside the queried range (R) of the real query ( $k$  cells from each direction around the real location of the LBS user  $R_{loc}$ ). All  $4k$  selected cells have the same query probability as the cell of the real location of the LBS user. The  $4k$  selected cells form the candidate set of dummy locations. Third, out of the candidate set, we randomly select the furthest  $(k - 1)$  cells as the actual and final dummy locations. Algorithm 1 provides details of the WSDL approach.

#### Algorithm 1: Wise Dummy Selection Location (WSDL)

**Input:**  $qp$  (query probability of each cell),  $R_{loc}$  (the real location of the LBS user),  $k$  (anonymity level).

**Output:**  $FD_{loc}$ .

- 1: sort cells based on their query probabilities;
- 2: **for** (direction=1; direction <4; direction ++)
- 3:      $CD_{loc} = FD_{loc} = \emptyset$ ;
- 4:     select  $k$  cells from each direction around  $R_{loc}$ ;
- 5:     Count  $\leftarrow 0$ ;

```

6:   while (count candidate  $\neq$  k)
7:     if ( $qp(C_i) = qp(R_{loc})$ ) then
8:        $CD_{loc} \leftarrow CD_{loc} \cup C_i$ ;
9:       Count  $\leftarrow$  count + 1;
10:    end if
11:  end while
12:  for ( $i = 1; i \leq \text{length}(CD_{loc}); i++$ )
13:    Dis-Array-core[i]  $\leftarrow$  calculate distance ( $C_i, R_{loc}$ );
14:    core candidate  $\leftarrow$  max (Dis-Array-core);
15:    for ( $j = 1; j \leq \text{length}(CD_{loc}); j++$ )
16:       $dis_1 = \text{dis}(\text{core candidate}, \text{candidate}_j)$ ;
17:       $dis_2 = \text{dis}(R_{loc}, \text{candidate}_j)$ ;
18:      Dis-Array[j]  $\leftarrow dis_1 \times dis_2$ ;
19:    end for
20:    Selected-Dummies [direction]  $\leftarrow$ 
21:      {Top (Sort (Dis-Array),  $\frac{k-1}{4}$ )  $\cup$  core candidate};
22:  end for
23:   $FD_{loc} \leftarrow \bigcup_{direction=4} Selected - Dummies [direction]$ ;
24:  output  $FD_{loc}$ 

```

After generating the final  $(k - 1)$  dummy locations, the  $selector_{DL}$  agent delivers them (as a set of coordinates) to the  $fragmentor_Q$  agent to start its mission, as described below.

**Fragmentor<sub>Q</sub>:** The final goal of this mobile agent is to protect the privacy of the issued query during the sending and processing phases. To complete this mission, the  $fragmentor_Q$  agent constructs  $k$  queries ( $k - 1$  queries based on the  $k - 1$  dummy locations that are received from the  $selector_{DL}$  agents, plus the query based on the real location of the LBS user). Then, it executes a fragmentation approach called Left-Right-Fragmentation (LRF) to protect the privacy of each constructed query. After that, it migrates to the LBS server, carrying the protected queries, which are manipulated and answered there with the help the  $predictor$  agent. After the queries are answered on the LBS server side, the  $fragmentor_Q$  migrates back to the home machine (i.e., the mobile device of the LBS user) to deliver the results.

### 3) Left-Right-Fragmentation (LRF) approach

The  $fragmentor_Q$  agent receives the set of actual dummy locations that were generated by the  $selector_{DL}$  agent. Each dummy location has its own coordinates  $(X, Y)$ . Let  $FD_{loc-coor}$  denote the set of the coordinates of the generated dummy locations, where:

$$FD_{loc-coor} = \{\langle X_1, Y_1 \rangle, \langle X_2, Y_2 \rangle, \langle X_3, Y_3 \rangle, \dots, \langle X_{k-1}, Y_{k-1} \rangle\} \quad (5)$$

For each coordinate  $\langle X_i, Y_j \rangle \in FD_{loc-coor}$  ( $i, j = 1, 2, \dots, k - 1$ ), a query is built according to the format that is specified in Table I, which consists of the following units: coordinates of the LBS user  $\langle X, Y \rangle$ , queried interest POI, queried range  $R$ , and identity of the LBS user ID. Each constructed query is referred to as an original query.

The key idea of the fragmentation technique is to extract the sensitive data from the query, encrypt them, and then randomize them, as shown in Fig. 9.

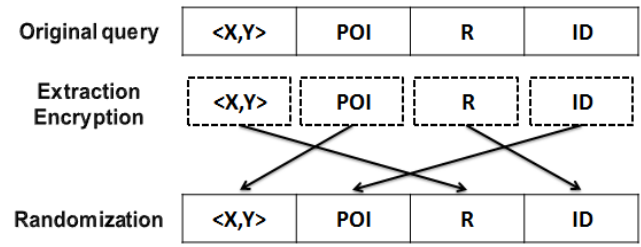


Fig. 9. Fragmentation technique.

In the context of fragmentation, we address the sensitive units of the query and the sensitive associations among the units because the attacker focuses on either one unit or the associations among two or more units to infer personal information. For instance, if the LBS user always queries the nearest hospitals as POIs, then the attacker can infer that the LBS user has a health problem. Meanwhile, if the attacker associates the ID of the LBS user with the queried POIs, then he/she can accurately identify the LBS user who has a health problem. Therefore, protecting the sensitive association is more important than protecting the sensitive units.

In this paper, the sensitive units of a given query are  $(\langle X, Y \rangle, POI, \text{ and } R)$ . For the LBS user ID, it is not considered a sensitive unit because the attacker cannot gain any private information from the ID unit alone. Moreover, even if the attacker associates the ID unit with any of the other units, he/she will fail to gather private information due to the encryption and randomization processes. Thus, if the attacker applies a query analysis attack, he/she will obtain, for instance, the following information: "the LBS user whose ID is (Bob-1) issues a query from an unknown location that asks for nameless POIs that are located in non-existent range  $R$ ". This statement does not reveal any private information.

In detail, the sensitive units are protected by public key infrastructure (PKI) and the sensitive associations are protected by a randomization phase. This forms our proposed Left-Right-Fragmentation (LRF) algorithm.

Formally, for a given set of queries  $Q = (q_1, q_2, \dots, q_i)$ , where ( $i = 1, 2, \dots, k - 1$ ), encoding a query ( $q_i$ ) consists of splitting it into two main parts: the left part ( $P_{q_i}^l$ ) and the right part ( $P_{q_i}^r$ ). Both parts are necessary for reconstructing the original query  $q_i$ .

$$q_i = P_{q_i}^l \circ P_{q_i}^r \quad (6)$$

In the first step of the randomization phase, we place the ID unit in the middle since it is not considered a sensitive data, as shown in Fig. 10.

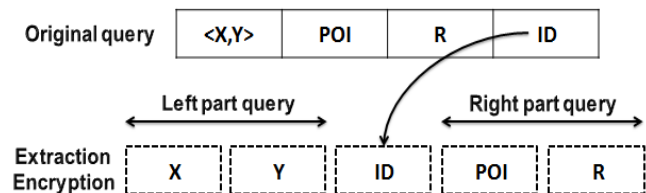


Fig. 10. First step in the randomization phase of the LRF-based algorithm.

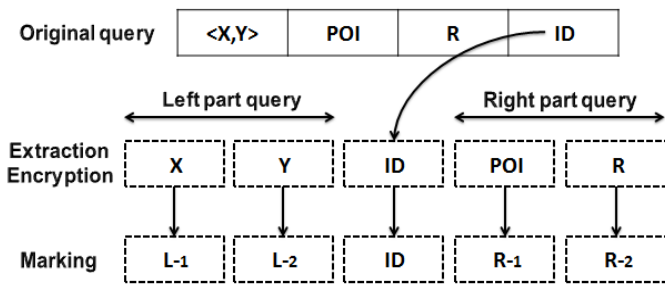


Fig. 11. Marking step in the randomization phase of the LRF-based algorithm.

The left part of the query includes one unit, which consists of two sub-units (X, Y). As for the right part query, it includes two units: (POI and R). In the second step of the randomization phase, we mark each unit or sub-unit by a (letter-number) pair that indicates the correct order in the original query, as shown in Fig. 11.

In Fig. 11, for instance, R-1 indicates that the encrypted unit (POI) must be placed directly to the right of the ID unit.

Since we have five different sites for ordering the units of the original query, there are  $(1 \times 2 \times 3 \times 4 \times 5 = 120)$  probable sites for randomizing the original units. Thus, the *fragmentor<sub>Q</sub>* agent can periodically change the randomization strategy, which prevents the attacker from discovering the correct order of the original query's units. Fig. 12 illustrates one possible choice and the reconstruction process.

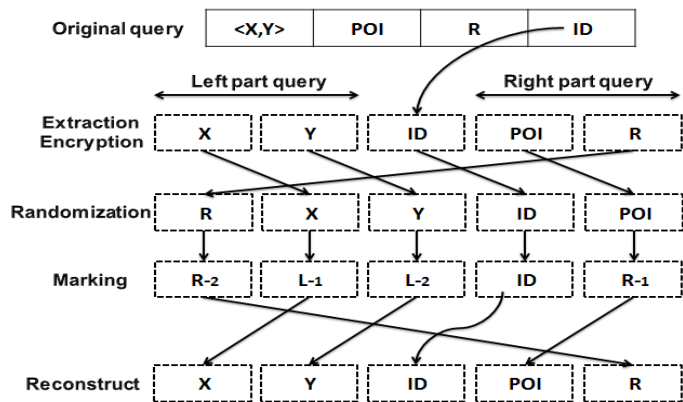


Fig. 12. Left-Right-Fragmentation (LRF) algorithm and reconstruction.

Reconstructing the original query is necessary for the stationary predictor agent to perform its mission (i.e., manipulating and answering the queries). The reconstruction process is carried out in four steps:

- 1) Putting the ID unit in the middle.
- 2) Decrypting the units of the query based on the shared encryption key between the *fragmentor<sub>Q</sub>* agent and the *predictor* agent.
- 3) Performing the marking step.
- 4) Moving the ID unit to the end.

Algorithm 2 illustrates the details of the left-right fragmentation approach.

**Algorithm 2:** Left-Right Fragmentation (LRF)

```

Input: kNN ( $\langle X, Y \rangle, POI, R, ID$ ) query.
Output: protected kNN ( $\langle X, Y \rangle, POI, R, ID$ ) query.
1: units{ } = extract ( $X, Y, POI, R$ );obtaining the sensitive data.
2: units{ } = encrypt (units )using 3DES algorithm;
3: new-units{ }=null;
// randomization
4: count =0; rand-array [5] ={-1};
5: while (count <5)
6:     random-value = rand(4);
7:     if (! contains (rand-array, random-value))
8:         rand-array[count] = random-value;
9:         Count ++;
10:    end if
11: end while
12: for (i=0; i<5;i++)
13:   new-units {i}= units{ rand-array[i] };
14: Return new-units;
    
```

By the LRF-based algorithm, all queries that are constructed on the LBS user side are protected before being sent to the LBS server. Then, all the queries are packaged and carried together by the *fragmentor<sub>Q</sub>* agent to the LBS server, which in turn means that the queries are protected during the sending phase. Because the LRF-based algorithm mixes the real query with  $k - 1$  dummy queries (which are constructed based on the  $k - 1$  dummy locations and selected by the *selector<sub>DL</sub>* agent) and the mission of manipulating the queries is assigned to the *predictor* agent, the queries are protected during the processing phase. The task of protecting the queries during the responding phase is included in the role of the *predictor* agent.

**Predictor:** This stationary agent receives the  $k$  queries that were constructed and carried by the *fragmentor<sub>Q</sub>* mobile agent. Then, it manipulates each query individually. After answering the received queries, the results are delivered to the *fragmentor<sub>Q</sub>* mobile agent, which, in turn, migrates back to the mobile device of the LBS user (i.e., the home machine). The process of manipulation requires the reconstruction of the  $k$  protected queries. This is performed according to the four steps that are listed above, where the same shared encryption key as was used to encrypt the units of the queries is used for decryption. After reconstructing the queries, the *predictor* agent manipulates each query according to an indexing technique, as described below.

4) *Cell-Based Indexing (CBI) technique*

In the *kNN* queries, the LBS user asks for the nearest  $k$  moving POIs that are located within a specified range  $R$  of the LBS user. Because of the continuous updatings of the locations of the moving POIs, the locations of the queried moving POIs are updated during the sending and processing of the queries. In addition, the location of the query issuer is also updated since it is considered an MO. Therefore, we need to retrieve the new exact locations of the queried moving POIs, and these new locations must be delivered to the new exact location of the query issuer. To achieve this, we model the motion of the moving POIs first. Then, the *predictor* agent indexes the

moving POIs and, based on their indices, predicts their new locations.

The given region (G), which is divided into  $(n \times n)$  cells of equal size, is modeled as an undirected graph  $GR(H,A)$ , where H represents the headers, and A represents the arms. The numbers that are associated with the arms denote weights (W), which represent the physical distances between two headers, as shown in Fig. 13.

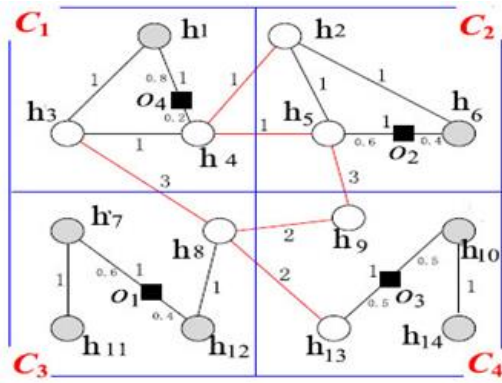


Fig. 13. Modeling the motion of the POIs.

In the context of the model, the terms path, boundary path, and MO are defined as follows:

**Definition 1.** For a given undirected graph  $GR(H,A)$ , a path P between a start header  $h_s$  and end header  $h_e$ , which is illustrated as a black line in Fig. 13, is expressed by the following formula:

$$P(h_s, h_e) = \{ h_s, h_{p1}, h_{p2}, \dots, h_{pm}, h_e \} \quad (7)$$

where,  $h_{pm}$  represents a sub path in the case in which there exist many headers from the start to the end.

**Definition 2.** For a given path P, a path is called a boundary path, if its start header  $h_{ps} \in C_i$  and its end header  $h_{pe} \in C_j$ , where C represents a cell and  $i \neq j$  (i.e., passing from one cell to another). Boundary paths are shown as red lines in Fig. 13.

**Definition 3.** For a given path P, an MO that is located on a path at time t is expressed by the following triple:

$$mo^t = \langle cl, p, d \rangle \quad (8)$$

where cl denotes the current location of the MO, p denotes the path that is linked to the MO, and d denotes the direction of the MO from the start header to the end header.

Based on the previous three definitions, four neighboring cells are shown in Fig. 13. The MOs are illustrated as black boxes.  $C_1$  contains three boundary paths ( $\{h_4, h_2\}$ ,  $\{h_4, h_5\}$ ,  $\{h_3, h_8\}$ ), which are weighted as 1, 1, and 3, respectively. The moving object  $mo_1$  resides in  $C_3$  on the path between  $h_7$  and  $h_{12}$ , and moves in the direction of  $h_{12}$ , with a distance of 0.4.

Based on the model that was presented above, the predictor agent creates and manages an index at the cell level. This index includes two parts: an index part and a data part, as shown in Fig. 14.

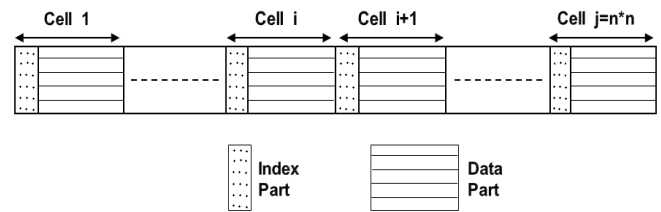


Fig. 14. General structure of CBI.

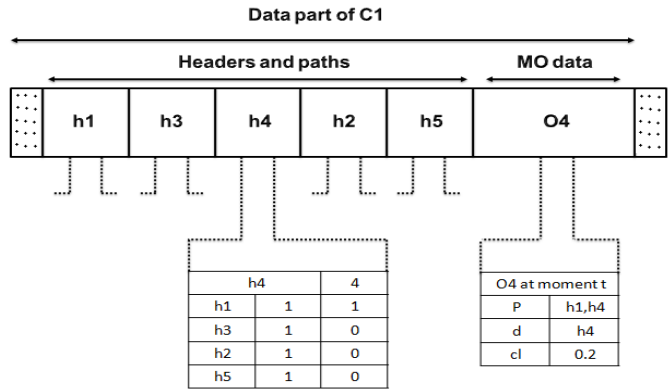


Fig. 15. Data part structure of  $C_1$ .

The data part holds detailed information about both the cell and the MOs that are located within the cell. This information mainly includes headers, paths, numbers of MOs on the paths, and data about the MOs, as illustrated in Fig. 15.

Two tables are shown in Fig. 15. The first record in the left table indicates that there are 4 headers that are linked to the header  $h_4$  and contained in  $C_1$ , which form two paths ( $\{h_4, h_1\}$ ,  $\{h_4, h_3\}$ ) and two boundary paths ( $\{h_4, h_2\}$ ,  $\{h_4, h_5\}$ ). The rest of the records carry information about the physical distances (or weights) of the formed paths/boundary paths and the number of MOs on each. For example, the second record states that  $w(h_4, h_1) = 1$ , and this path has one moving object. The right table states that moving object  $O_4$  moves on path  $p = \{h_4, h_1\}$  towards  $h_4$ , and its current location is a distance of 0.2 from  $h_4$ .

Based on the information that is related to the MO ( $O_4$  in the right table of Fig. 15), the predictor agent can calculate the speed of the MO based on its two previous consecutive locations and moments, as follows:

$$speed = \frac{distance}{time} = \frac{|cl \text{ at the moment } t_2 - cl \text{ at the moment } t_1|}{t_2 - t_1} \quad (9)$$

After calculating the speed, the predictor agent can estimate the future location of the MO by calculating the  $\Delta$  distance (illustrated in Fig. 2 in the introduction section) and adding it to the current location of the MO, taking into consideration the direction of the MO.

The index part of a given cell contains the cell identifier ( $C_{i,j}$ ); the area of the cell, which is represented by the width of the cell ( $wth^2$ ); and the number of MOs that are located in the cell. In addition, it includes the same previous information about the eight cells that surround the given cell, as shown in Fig. 16.

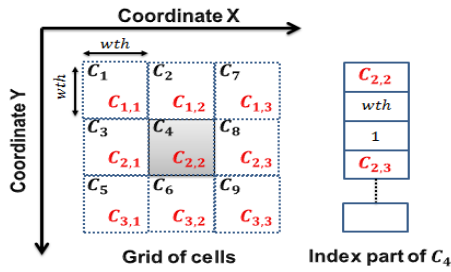


Fig. 16. Index part structure of  $C_4$ .

The index part will be the input of a bloom filter [24]. The benefit of the bloom filter is that it can give a direct answer regarding the existence or non-existence an element within a set. We exploit this to determine whether there is an MOs within the cells that are covered by the range  $R$ , which is specified in the  $kNN$  query. If no MOs are found in a cell, it is not necessary to search inside the cell. Thus, we can move to the next cell. This greatly speeds up both the response time and the processing time of  $kNN$  queries since time is not wasted on examining empty cells.

In detail, for a given  $kNN$  query with a range  $R$ , we first determine which cells are covered by  $R$ . Then, the index part of each cell is used to determine which contain the queried MOs using the bloom filter. For only the cells that have MOs, the actual search is performed on the data part of each limited cell with a prediction phase; to retrieve the future locations of the queried MOs. Algorithm 3 illustrates the steps of processing a  $kNN$  query based on the proposed CBI technique.

After retrieving the results (i.e., the predicted locations of the queried MOs), the *predictor* agent encrypts the results and delivers them to the *fragmentor<sub>Q</sub>* agent. The *fragmentor<sub>Q</sub>* mobile agent migrates back to the home machine to deliver the results to the LBS user. The process of encrypting the results ensures the privacy protection of the queries during the responding phase. Algorithm 4 describes the itinerary of the *fragmentor<sub>Q</sub>* mobile agent.

**Algorithm 3:** CBI based  $kNN$  query processing

```

Input: cells,  $R_{loc}$  real location, range  $R$ .
Output: POIs [] //moving objects.
1: covered-cells[]=null;
2: for (i=1; i<=count(cells); i++)
3:   distance  $R_{loc,Cell_{loc}} = \sqrt{(Rx_{loc} - Cellx_{loc})^2 + (Ry_{loc} - Celly_{loc})^2}$ ;
4:   if (distance  $R_{loc,Cell_{loc}} \leq R$ )
5:     add (cell[i], covered-cells);
6:   end for
7: foreach cell in covered cells
8:   if (bloom (index-part of cell))
9:     fetch (data-part of cell)
10:    foreach path in data-part
11:      if (path contains MO)
12:        future-cell=prediction (MO);
13:        add(future-cell, POIs);
14:      end if
15:    end foreach
16:   end if
17: end foreach

```

18: return POIs;

**Algorithm 4:** Trip of *fragmentor<sub>Q</sub>* mobile agent

```

Input:  $kNN$  query.
Output: report results.
1: agent = new fragmentorQ (); (create an agent)
2: itinerary = new itinerary ();
3: itinerary.Adddistenation ("LBS server", "execute encryption method");
4: itinerary.Adddistenation ("LBS mobile device", "execute report results method");
5: output: report results;

```

Since the query issuer (i.e., the LBS user) is an MO, his/her location changes during the sending and processing the query. Therefore, the results must be delivered to the query issuer according to his/her new location. Because the *fragmentor<sub>Q</sub>* is a mobile agent that is created in the mobile device of the LBS user, which represents the home machine, it must return back to the same home machine without any additional predictions on the location of the query issuer, as shown in Algorithm 4. Therefore, the future locations of the queried MOs are calculated in the prediction phase, while the future location of the query issuer is naturally obtained due to the returning step in the itinerary. In other words, it is not necessary to compute the  $\bar{U}$  distance. As a result, the two parts of the real-time uncertainty problem are solved, as shown in Fig. 17.

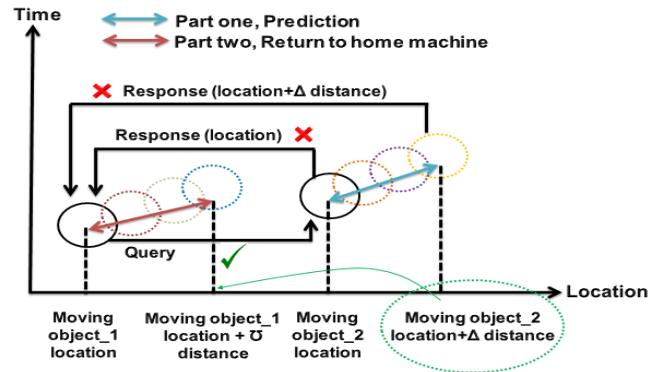


Fig. 17. Solved uncertainty real-time problem.

**B. Details of Our Proposed Architecture**

We use sequence diagrams to illustrate the general scenario of our proposed agent-based architecture. Fig. 18 shows the steps for processing a  $kNN$  query with comprehensive privacy protection.

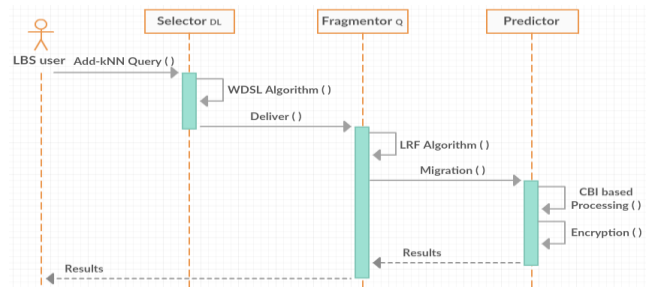


Fig. 18. Sequence diagram of processing a  $kNN$  query.

## V. SECURITY ANALYSIS

In this section, we discuss two main security issues. The first is related to the agents themselves and the second is related to the WSDL and LRF-based algorithms, which were proposed for privacy protection.

### A. Security of Agents

The main obstacle to the widespread deployment of the mobile agent technology is the security issue, in particular, the problem of protecting a mobile agent from malicious hosts that may completely block the agent or modify its carried data. Since the scope in this paper is privacy protection, security is out of scope. Therefore, we assume that all the agents are secure. Specifically, the approach that was proposed in [25] is followed. The integrating of privacy protection with security agents will be considered in future work.

### B. Security against Inference Attacks and Query Analysis Attacks

In this subsection, we prove that our proposed architecture is robust by discussing the resistance of the WSDL algorithm and the LRF-based algorithm against location homogeneity attack and query sampling attack, respectively. Since we consider active attack (as discussed in Section III), Table III lists the capabilities of the LBS server (the attacker).

TABLE III. CAPABILITIES OF THE ATTACKER (LBS SERVER)

Cap- No	Description
1	Can eavesdrop on the wireless channel.
2	Can monitor the current queries of LBS users.
3	Can obtain all the stored information that is related to the LBS users.
4	Can obtain the historical location data of the LBS users.
5	Knows the query privacy protection method (LRF-based algorithm).
6	Knows the location privacy protection method (WDSL algorithm).

We follow the definition-theorem-proof style in discussing resistance against inference attacks.

**Definition 1.** An algorithm is query sampling attack resistant if the units of the sent query cannot be obtained and correctly reordered.

**Theorem 1.** The proposed LRF-based algorithm is query sampling attack resistant.

**Proof 1.** Obtaining the units of a query requires eavesdropping on the wireless channel. Since a cryptographic technique (PKI) is used to protect the sensitive data (the units of the query), the attacker cannot obtain the units. Moreover, even if the attacker were to successfully break the encryption phase, he/she would need to form the query in a correct order due to the randomization phase. Furthermore, if the attacker tries to reverse the LRF-based algorithm, he/she will fail because of the periodic changing of the query's units in the LRF-based algorithm, which confuses the attacker and forces him/her to randomly guess the correct order of the units to form the original query. This means that the query sampling attack fails.

**Definition 2.** An algorithm is location homogeneity attack resistant if the probability of successfully guessing the real location of an LBS user is very low.

**Theorem 2.** The proposed WSDL algorithm is location homogeneity attack resistant.

**Proof 2:** We assume that the attacker completely breaks the LRF-based algorithm, thereby obtaining the location of the LBS user. In addition, the information that the attacker holds is the query probability of each individual cell (qp) and all the submitted ( $k$ ) locations  $l_1, l_2, \dots, l_k$  (i.e., the mixture of real and dummy locations). Let  $PS_{(event)}$  refer to the probability of the attacker successfully guessing whether (event) is true. The WDSL algorithm is resistant to location homogeneity attack if the following two conditions are satisfied:

- 1)  $PS_{(l_i)} = PS_{(l_j)} \quad \forall (0 < i \neq j \leq k)$  (10)
- 2)  $dis(l_i, l_j)$  is long.

First, since the dummy locations are selected based on the query probabilities (qp) of the cells being similar to the query probability of the LBS user's cell (i.e., his/her real location), the attacker can obtain no benefit from employing the query probabilities to determine the real location of the LBS user. Second, since we have  $k$  submitted locations, the probability of successful guessing the real location is  $\left(\frac{1}{k}\right)$ . The previous probability value is the same for all  $k$  submitted locations because no benefit is obtained from knowing the query probabilities of the locations. This means that the first condition is satisfied. Third, since the dummy locations are selected based on the product of the distances rather than the sum of the distances, the second condition is satisfied. Moreover, even if the attacker tries to reverse the WDSL algorithm, he/she will fail to determine the real locations of the dummies. That is because of the random selection of the final and actual dummy locations, which leads to uncertainty in the dummy selection results. Therefore, the attacker can only randomly guess the real location of the LBS user. As a result, the location homogeneity attack fails.

## VI. METRICS

In this section, we provide the metrics that are used for evaluation purposes. In this paper, two kinds of metrics are employed: privacy metrics and performance metrics.

### A. Privacy Metrics

We use two privacy metrics: the entropy  $E$  and a metric that is derived from the entropy. To evaluate the location privacy, we employ  $E$  to quantify the privacy. It is better to achieve a higher  $E$  value.  $E$  is defined by formula 1.

Suppose an LBS user sends a query to dummy locations to protect his/her privacy. The highest entropy value that can be achieved is  $\log_2(k)$ , which is achieved when all the submitted locations have the same probability of being treated as the real location of the query issuer (LBS user). Therefore, if the LBS user achieves an entropy value that is less than  $\log_2(k)$ , the extent to which the privacy was breached by the attacker (LBS server) will be  $(\log_2(k) - E)$ . As time progresses, the attacker achieves a small success with each sent query. The sum of these small successes represents the degree of danger that the



privacy will be compromised, which represents the second privacy metric.

More formally, let  $\mathcal{T} = (\tau_1, \tau_2, \tau_3, \dots, \tau_n)$  refer to the moments at which the LBS user issues queries, where each query is protected by  $k - 1$  dummy locations. The degree of danger  $D_{\text{danger}}$  is defined as:

$$D_{\text{danger}} = \sum_{i=1}^n (\log_2(k) - E(\tau_n)), \text{ where } \tau_n \in \mathcal{T} \quad (11)$$

When an encryption technique is used to protect the privacy, no privacy metric is used to quantify the privacy. This is clearly stated in the survey in [3]. Therefore, we rely on a performance metric for evaluating the query privacy protection.

### B. Performance Metrics

Since we used encryption in the proposed LRF-based algorithm to protect the query privacy, we introduce the computation time  $T_{\text{comp}}$  for evaluation. Here, the computation time refers the time that is spent on both sides (the LBS mobile device's side and the LBS server side). On the LBS mobile device's side, the computation includes the time spent constructing a query based on a dummy location and passing sensitive units, which is equal to the sum of the durations of the extraction, encryption, randomization, and marking phases.

$$T_{\text{Q-cons}} = T_{\text{ext}} + T_{\text{enc}} + T_{\text{rand}} + T_{\text{mark}} \quad (12)$$

On the LBS server side, the computation time is the time that is spent preparing the query (i.e., reconstructing the query), which is equal to the sum of the durations of the decryption, marking, unit ordering, and processing phases.

$$T_{\text{Q-recons}} = T_{\text{dec}} + T_{\text{mark}} + T_{\text{ordering}} + T_{\text{processing}} \quad (13)$$

Thus, the computation time is defined as:

$$T_{\text{comp}} = T_{\text{Q-cons}} + T_{\text{Q-recons}} \quad (14)$$

To evaluate the proposed indexing technique, we use two times as performance metrics: access latency and tuning time. Access latency ( $T_{\text{AL}}$ ) refers the elapsed time between the moment when a query is issued and the moment when it is satisfied. Therefore, it depends on  $T_{\text{comp}}$  as follows:

$$T_{\text{AL}} = T_{\text{comp}} + T_{\text{sending}} + T_{\text{receiving}} \quad (15)$$

The tuning time  $T_{\text{TU}}$  is the time that the mobile LBS user stays active to receive the requested data.

## VII. EXPERIMENTAL RESULTS AND EVALUATIONS

### A. Simulation Setup

In this paper, Matlab software is used to implement the proposed algorithms, with the help of Java Agent DEvelopment Framework (JADE). The performance evaluation is simulated on a Genuine Intel(R) 2.4 GHz PC with 4.00 G RAM, running Microsoft Windows 7 Ultimate. Table IV lists the parameter settings. A data base is constructed for the moving POIs, where timestamps are attached to each POI and each query. The query probability is generated randomly with the help of the Google Maps API.

TABLE IV. PARAMETER SETTINGS

Parameter	Setting
Number of cells ( $n \times n$ )	160 × 160
Number of headers (H)	21,103
Number of arms (A)	21,246
Number of users	10,000
Number of moving POIs	500

For comparison, we selected three dummy-based approaches for location privacy protection: DDA [10], CirDummy [11], and Dest-Ex [12]. The Buddies [21] and PIR [22] approaches are selected for query privacy protection. As  $kN$  query processing techniques, we selected D-tree [17] and density [23].

### B. Evaluations of Resistance Against Attacks

There is a direct correlation between the  $k -$  anonymity level and the resistance against attacks because a higher  $k -$  anonymity level provides higher resistance. Increasing the  $k -$  anonymity level requires increasing the number of generated dummies. Therefore, based on the entropy value, we first measure the privacy protection level against the  $k -$  anonymity level, assuming that the defenses of the fragmentation technique have been broken. Then, we calculate the number of LBS users that reach dangerous states based on the  $D_{\text{danger}}$  privacy metric.

Fig. 19 below shows a snapshot at a time progress of 120 minutes. Among the approaches, the DDA approach performs the worst because DDA fills the array of dummies by selecting locations in a random way based on the principle that "the dummy locations must be equal in area". Thus, the entropy of the dummy locations mainly relies on the current query probabilities of the grid of cells. The CirDummy approach slightly outperforms the DDA approach. That is because the selected dummy locations are limited by a virtual circle. Since the variation of the query probabilities is not large within the circle, which covers only a few cells (a small region), the corresponding entropy value is only slightly higher. The Dest-Ex approach outperforms both DDA and CirDummy. The main factor that contributes to the enhancement of the entropy values is the direction, which may be changed to include more cells with the same query probability. Compared to the previous approaches, the WDSL approach performs the best. The underlying reason is that the dummy locations are selected based on having similar query probabilities to the real location. This guarantees much higher entropy values and higher corresponding privacy levels.

Regarding to the evaluation based on the  $D_{\text{danger}}$  privacy metric, we evaluated the situations of LBS users under location homogeneity attack. In this context, a threshold is defined ( $\text{thr} = 0.75$ ) at which the LBS user is considered vulnerable to attack by the LBS server. The level of anonymity is fixed to ( $k = 6$ ) (i.e., at any moment, the sent query is protected by five queries, which are built based on dummy query locations). Twenty LBS users are randomly selected from each of the compared approaches and a snapshot at ( $t = 120$ ) is taken, as shown in Fig. 20.

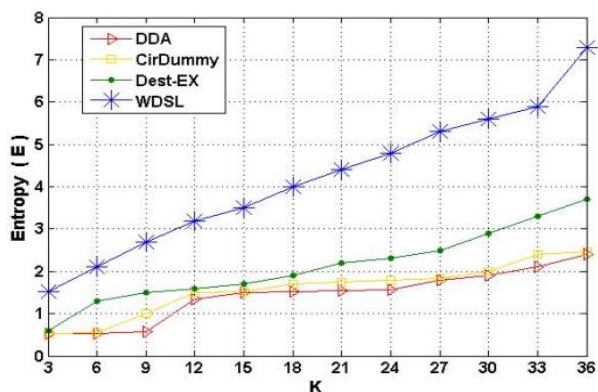


Fig. 19. Entropy vs. k, t = 120.

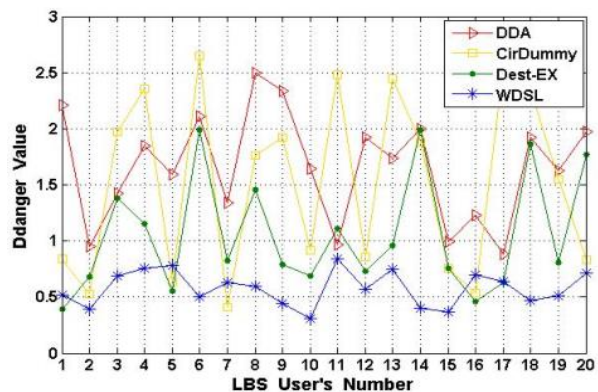


Fig. 20.  $D_{danger}$  values for 20 LBS users, k = 6, t = 120.

TABLE V. COMPARISON OF VULNERABILITY STATUSES OF LBS USERS

Settings: t = 120, k = 6, thr = 0.75.		
Approach	Term	Percentage of encroachment
WDSL	Number of users that exceed the threshold	3
Dest-Ex	Number of users that exceed the threshold	10
CirDummy	Number of users that exceed the threshold	16
DDA	Number of users that exceed the threshold	20

Table V shows that all LBS users in the DDA approach exceeded the threshold. That is because all the selected dummy locations are close to one another since they are formed by the vertices and the edges of the grid. More than three-quarters and half of the LBS users exceeded the threshold in CirDummy and Dest-Ex, respectively. Compared to DDA, the CirDummy has higher resistance against location homogeneity attack since the radius of the circle may be enlarged to include some dummy locations that are far away from the real location of the LBS user. Dest-Ex achieved a higher resistance than CirDummy because the directions can be changed to include dummy locations that are further away from the real location. The proposed WDSL approach performs the best since it has the minimum number of LBS users that exceeded the threshold, and, consequently, the highest resistance against the location homogeneity attack. That is because the dummy locations are selected based on the product of their distances.

Under different threshold values, snapshots, and numbers of LBS users, Table VI supports the results in Table V.

TABLE VI. PERCENTAGE OF ENCROACHMENT OF THE PREDEFINED THRESHOLDS

Try NO	NO of LBS users	t	thr	Percentage of encroachment			
				WDSL	Dest	Cir	DDA
1	40	130	0.7	0.11	0.5	0.62	100
2	60	140	0.65	0.12	0.53	0.78	100
3	80	150	0.6	0.2	0.4	0.61	100
4	100	160	0.55	0.18	0.41	0.55	100
5	120	170	0.5	0.13	0.34	0.53	100

### C. Evaluations of Computation Costs

We use the  $T_{comp}$  performance metric to evaluate the efficiency of the proposed LRF-based approach against the buddy and PIR approaches. Two aspects are considered in the evaluation: the impact of increasing the k value that is associated with a query and the impact of increasing the number of sent queries.

In general, the computation time increases as k increases. Fig. 21 shows a snapshot at t = 120, where we randomly selected an LBS user who sends a privacy-protected query at different levels of k. The PIR-based approach performs the worst since it performs many computations to protect the privacy of the query. Despite the times spent in the various phases (i.e., the extraction, encryption, randomization, and marking phases), our proposed LRF fragmentation technique performs the best. The reason behind this is the efficient employment of the bloom filter to enhance the processing time of the query. Specifically, the help that is provided by the predictor agent through the proposed CBI technique efficiently contributes to the shortening of the query processing time. In depth, the process of encapsulating the index part by the bloom filter has a positive impact on the search time, as it avoids searching in the empty cells.

The results that are shown in Fig. 21 are supported by those in Fig. 22, in which the number of protected queries increased. Again, the bloom filter is the underlying feature that accelerates the answering of the queries, which is not utilized by the other approaches.

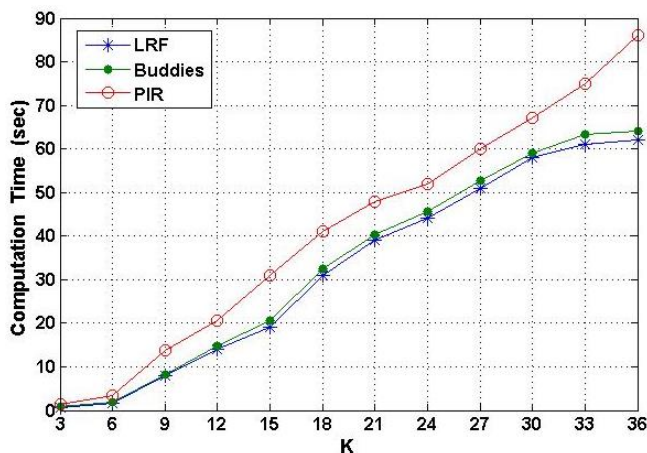


Fig. 21.  $T_{comp}$  vs. k, t = 120.

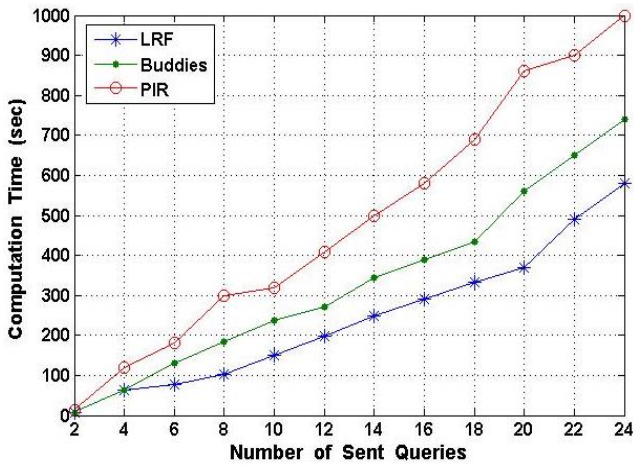


Fig. 22.  $T_{comp}$  vs. Number of sent queries,  $k = 6, t = 120$ .

#### D. Access Latency ( $T_{AL}$ ) and Tuning Time ( $T_{TU}$ ) Evaluations

In the previous subsection, the sending time of the query (from the mobile device of the LBS user to the LBS server) and the receiving time of the query's answer (from the LBS server to the mobile device of the LBS user) are completely ignored. When evaluating  $T_{AL}$  and  $T_{TU}$ , the two previous times must be taken into account. We assume that the sending time of the query is the same for all of the compared techniques (i.e., D-tree, Density, and the proposed CBI). Fig. 23 shows the access times for different numbers of sent queries.

As shown in Fig. 23, the proposed CBI technique outperforms the Density and D-tree techniques. The main factor in this is the migration of the *fragmentor<sub>Q</sub>* mobile agent back to the home machine, to deliver the answers to the sent queries. Meanwhile, in both the Density and D-tree techniques, a significant amount of time is needed to search for the queries' issuer (since it is considered an MO) to deliver the answers. In other words, the receiving time of the queries' answers is longer for these two methods than for the proposed CBI technique. The access latency time reflects the efficiency of the proposed CBI technique in solving the second part of the real-time uncertainty problem (see Fig. 17).

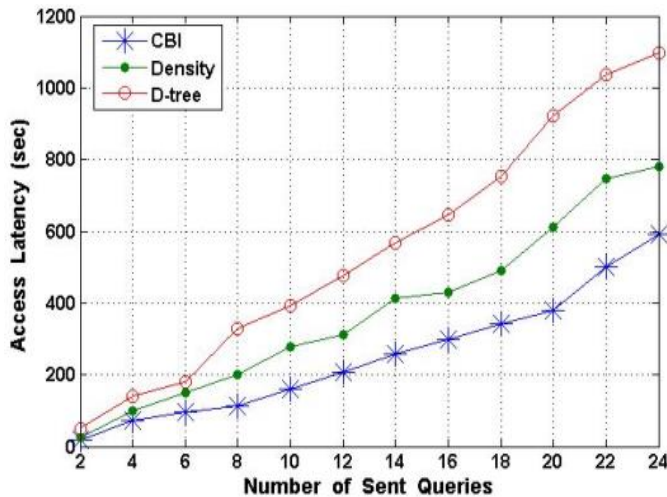


Fig. 23.  $T_{AL}$  vs. Number of sent queries,  $k = 6, t = 120$ .

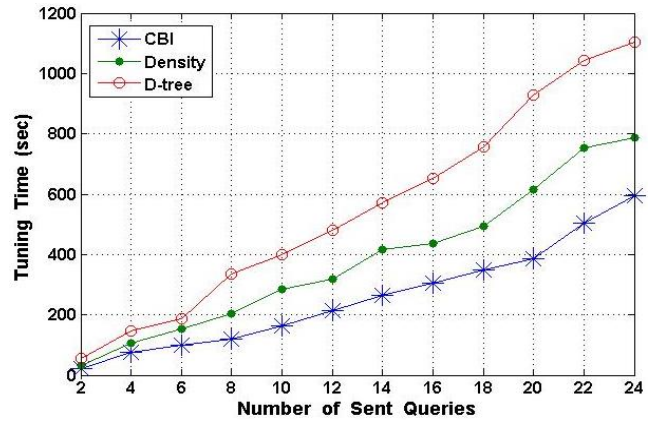


Fig. 24.  $T_{TU}$  vs. Number of sent queries,  $k = 6, t = 120$ .

The results that are shown in Fig. 24 support those that are illustrated in Fig. 23, but with higher tuning time values since the LBS user spends additional time preparing the queries and exploring the received answers. However, the proposed CBI technique provides the minimum tuning time values. Since the tuning time refers the time that mobile device of the LBS user stays active, the proposed CBI technique reduces the battery consumption of the mobile device. Short battery life is a main drawback of user-based privacy protection approaches.

#### E. Evaluations of the Prediction Phase of the CB Technique

The migration of the *fragmentor<sub>Q</sub>* mobile agent back to the home machine contributes to solving the second part of the real-time uncertainty problem, and the prediction phase in the proposed CBI technique contributes to solving the first part of the real-time uncertainty problem (see Fig. 17 above). In this context, we evaluate the number of retrieved moving POIs and the precisions of the locations of the retrieved moving POIs.

Fig. 25 shows the number of retrieved moving POIs when the LBS user searches for the nearest 6 taxis that are located within a 0.5 km range around different real locations of the LBS user. For instance, in response to the query ( $< 70,70 >$ , taxis, 0.5, Bob), 3, 6, and 11 moving taxis were retrieved by the D-tree, Density, and CBI techniques, respectively. The Density technique outperforms the D-tree technique since it uses the overlap among the cells to build the index. The proposed CBI technique outperforms the Density technique due to two factors: First, the index is built on the level of cells, which accurately covers all the cells that are included in the 0.5 km range. Second, in the prediction phase, because of the motion of the queried POIs, many additional POIs may enter the cells (covered by the given range) from the surrounding cells. Therefore, the prediction phase can include the POIs that entered the range in the answer to the query.

Since the number of the retrieved POIs does not accurately reflect the efficiency of the proposed indexing technique, we evaluate the precision of the retrieved locations of the moving POIs. Here, the precision term is the degree of matching between the current location (i.e., the exact future location of the moving POI) and the predicted one. From Fig. 25, we select the nine retrieved taxis that are related to the query ( $< 90,90 >$ , taxis, 0.5, Bob) for evaluation.

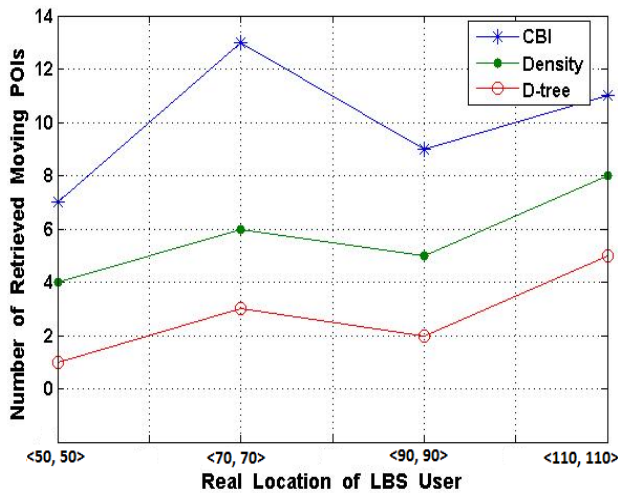


Fig. 25. Number of retrieved POIs,  $k = 6$ ,  $R = 0.5$  km,  $t = 120$ .

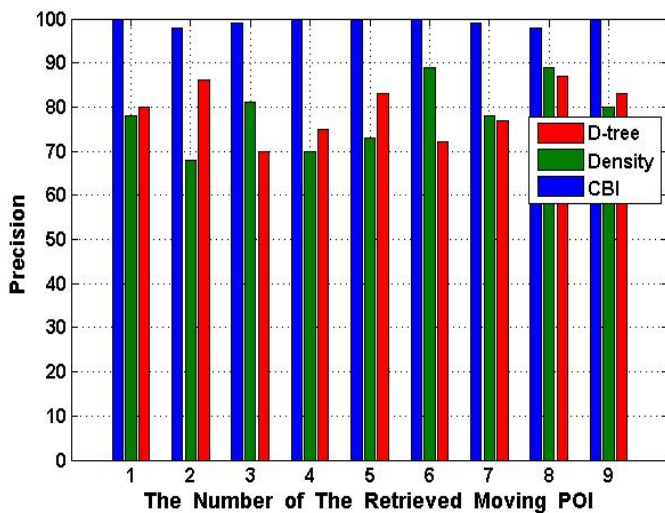


Fig. 26. Locations precision of the retrieved moving POIs.

As shown in Fig. 26, the precision of the retrieved locations is variable for both the D-tree and Density techniques. In contrast, the stability of the precision of the locations that were retrieved by the proposed CBI technique high: it varies between 100% and 98%. This is due to the prediction phase.

### VIII. CONCLUSION

With the impressive development of both wireless networks and mobile devices, Location Based Services (LBSs) have become popular. LBSs enable network users to perform range queries or  $k$ -Nearest Neighbor ( $kNN$ ) queries. However, it is extremely important to ensure comprehensive privacy protection, in addition to guaranteeing the efficiency of  $kNN$  query processing. We propose a Wise Dummy Selection Location (WDSL) approach for ensuring the location privacy of  $kNN$  queries. To ensure a high protection level of location privacy, the WDSL approach selects dummy locations that satisfy two conditions: (1) the query probabilities of the selected dummy locations are the same as that of the real location of the LBS user and (2) the selected dummy locations are distributed over a wide region to ensure resistance against location homogeneity inference attack. Resistance against

query sampling attack, which targets the query privacy, is considered. Extracting, encrypting, and randomizing the sensitive units of the sent query based on a Left-Right Fragmentation (LRF) technique results in robust defense against the query sampling attack and ensures the query privacy. The integration of the WDSL approach and the LRF technique ensures the  $kNN$  query privacy during the sending, processing, and responding phases. To manipulate the  $kNN$  query efficiently, an index is built based on an efficient motion model at the level of cells, in which the moving POIs are moving. The index consists of two parts: a data part and an index part. The data part is supported by a prediction phase, which estimates the future locations of the queried moving POIs. The index part is encapsulated by a bloom filter to speed up the response to the  $kNN$  query. In terms of resistance against inference attacks and query analysis attacks, computational cost, and number and accuracy of retrieved moving POI locations, the proposed system outperforms similar approaches and techniques.

In future work, we intend to ensure the integrating of agents security and privacy. In addition, we intend to develop defenses against other inference attacks, such as map matching attacks and semantic location attacks.

### REFERENCES

- [1] Dardari, Davide, Pau Closas, and Petar M. Djurić. "Indoor tracking: Theory, methods, and technologies." *IEEE Transactions on Vehicular Technology* 64.4 (2015): 1263-1278.
- [2] Wernke, Marius, et al. "A classification of location privacy attacks and approaches." *Personal and Ubiquitous Computing* 18.1 (2014): 163-175.
- [3] Shin, Kang G., et al. "Privacy protection for users of location-based services." *IEEE Wireless Communications* 19.1 (2012).
- [4] Yi, X., Paulet, R., Bertino, E., & Varadarajan, V. (2014, March). Practical  $k$  nearest neighbor queries with location privacy. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on* (pp. 640-651). IEEE.
- [5] Ni, Weiwei, Mingzhu Gu, and Xiao Chen. "Location privacy-preserving  $k$  nearest neighbor query under user's preference." *Knowledge-Based Systems* 103 (2016): 19-27.
- [6] Yi, Xun, et al. "Practical approximate  $k$  nearest neighbor queries with location and query privacy." *IEEE Transactions on Knowledge and Data Engineering* 28.6 (2016): 1546-1559.
- [7] Ma, Tinghui, et al. "Protection of location privacy for moving  $kNN$  queries in social networks." *Applied Soft Computing* 64.2 (2017): 485-158.
- [8] Dai, Jian, Zhi-Ming Ding, and Jia-Jie Xu. "Context-Based Moving Object Trajectory Uncertainty Reduction and Ranking in Road Network." *Journal of Computer Science and Technology* 31.1 (2016): 167-184.
- [9] Zhang, Xu, et al. "A novel location privacy preservation method for moving object." *International Journal of Security and Its Applications* 9.2 (2015): 1-12.
- [10] Alrahhah, Mohamad Shady, et al. "AES-Route Server Model for Location based Services in Road Networks." *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS* 8.8 (2017): 361-368.
- [11] Lu, Hua, Christian S. Jensen, and Man Lung Yiu. "Pad: privacy-area aware, dummy-based location privacy in mobile services." In *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pp. 16-23. ACM, 2008.
- [12] Hara, Takahiro, et al. "Dummy-Based User Location Anonymization Under Real-World Constraints." *IEEE Access* 4 (2016): 673-687.
- [13] Pingley, Aniket, Nan Zhang, Xinwen Fu, Hyeong-Ah Choi, Suresh Subramaniam, and Wei Zhao. "Protection of query privacy for

- continuous location based services." In Infocom, 2011 Proceedings IEEE, pp. 1710-1718. IEEE, 2011.
- [14] Pan, Xiao, et al. "Protecting personalized privacy against sensitivity homogeneity attacks over road networks in mobile services." *Frontiers of Computer Science* 10.2 (2016): 370-386.
- [15] Wang, Yong, et al. "A fast privacy-preserving framework for continuous location-based queries in road networks." *Journal of Network and Computer Applications* 53 (2015): 57-73.
- [16] Hambrusch, Susanne, Chuan-Ming Liu, Walid G. Aref, and Sunil Prabhakar. "Query processing in broadcasted spatial index trees." In *International Symposium on Spatial and Temporal Databases*, pp. 502-521. Springer, Berlin, Heidelberg, 2001.
- [17] Xu, Jianliang, Baibua Zheng, W-C. Lee, and Dik Lun Lee. "Energy efficient index for querying location-dependent data in mobile broadcast environments." In *Data Engineering, 2003. Proceedings. 19th International Conference on*, pp. 239-250. IEEE, 2003.
- [18] Zheng, Baihua, et al. "Grid-partition index: a hybrid method for nearest-neighbor queries in wireless location-based services." *The VLDB Journal—The International Journal on Very Large Data Bases* 15.1 (2006): 21-39.
- [19] A. Beresford and F. Stajano, —Location Privacy in Pervasive Computing, IEEE Pervasive Computing, vol. 2, no. 1, 2003, pp. 46–55.
- [20] Zhang, Xu, and Hae Young Bae. "Location Positioning and Privacy Preservation Methods in Location-based Service." *International Journal of Security and Its Applications* 9.4 (2015): 41-52.
- [21] Mascetti, Sergio, et al. "Privacy in geo-social networks: proximity notification with untrusted service providers and curious buddies." *The VLDB Journal—The International Journal on Very Large Data Bases* 20.4 (2011): 541-566.
- [22] Ghinita, Gabriel, Panos Kalnis, Ali Khoshgozaran, Cyrus Shahabi, and Kian-Lee Tan. "Private queries in location based services: anonymizers are not necessary." In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 121-132. ACM, 2008.
- [23] Jang, Mi Young, and Jae Woo Chang. "A new k-nn query processing algorithm enhancing privacy protection in location-based services." In *Computer and Information Technology (CIT), 2011 IEEE 11th International Conference on*, pp. 421-428. IEEE, 2011.
- [24] Solomon, Brad, and Carl Kingsford. "Improved Search of Large Transcriptomic Sequencing Databases Using Split Sequence Bloom Trees." In *International Conference on Research in Computational Molecular Biology*, pp. 257-271. Springer, Cham, 2017.
- [25] Madkour, Mohamed A., et al. "Securing mobile-agent-based systems against malicious hosts." *World Applied Sciences Journal* 29.2 (2014): 287-2

# A Multiclass Deep Convolutional Neural Network Classifier for Detection of Common Rice Plant Anomalies

Rommel R. Atole

Institute of Information Technology  
Partido State University  
San Juan Bautista, Goa, Camarines Sur, 4422 Philippines

Daechul Park

Dept. of Computer, Comm. and Unmanned Technology  
Hannam University  
70 Hannam-ro, Daeduk-gu, Daejeon, Korea

**Abstract**—This study examines the use of deep convolutional neural network in the classification of rice plants according to health status based on images of its leaves. A three-class classifier was implemented representing normal, unhealthy, and snail-infested plants via transfer learning from an AlexNet deep network. The network achieved an accuracy of 91.23%, using stochastic gradient descent with mini batch size of thirty (30) and initial learning rate of 0.0001. Six hundred (600) images of rice plants representing the classes were used in the training. The training and testing dataset-images were captured from rice fields around the district and validated by technicians in the field of agriculture.

**Keywords**—Deep neural network; convolutional neural network; rice; transfer learning; AlexNet

## I. INTRODUCTION

### A. Background

Advances in computer processing power have revolutionized not only the scope of its applications but as well as its capability to process large amounts of data. What were once constrained to a few layers of neurons, neural networks can now span several layers each comprising thousands of computational neurons, largely due to significant improvements in computing hardware. With this structure, neural networks have evolved into more powerful computational tools closely mimicking human intelligence.

One of the hottest applications of machine learning nowadays is on computer vision and object recognition in general and in plant health monitoring in particular [1]-[3]. Neural networks and deep learning currently provide the best solutions to many problems in image recognition, speech recognition, and natural language processing [4]. Many researchers around the world continue to exploit this computational power in almost every problem domain.

In the case of rice farming, the concept supporting the application of DIP (Digital Image Processing) in the detection of rice plant diseases, is bolstered by the observation that most of these diseases are manifested in the appearance of the leaves

and on the general visual features of the plant. It is therefore not hard to imagine that these visual patterns can be taken collectively to form multivariate basis of identity and traits unique to each type of disease. For instance, Brown Spot, a fungal disease, is characterized by presence of lesions that are initially small, circular, and dark brown to purple-brown [5], while Bacterial Blight is identified with wilting and yellowing of leaves, which, among older plants, turn yellow to grayish white with black dots due to the growth of various saprophytic fungi [6].

### B. Machine Learning and Transfer Learning with AlexNet

Transfer learning is an approach in Deep Learning where a large, deep neural network previously trained on other datasets, is adopted and used in another application. Although designing and training a deep network from scratch remains an interesting alternative, adoption of pre-trained deep networks is an appealing prospect for a number of reasons:

- 1) “Reinventing the wheel” and training networks from scratch takes time and demands high computing power. A convolution network the size and topology of Alexnet finished training in 5 to 6 days;
- 2) Pre-trained networks have been trained over millions of images to classify thousands of object and classes; As such, the weights and biases connecting its neurons have been optimally calculated; and
- 3) Many successful applications of pre-trained networks can be found in the literature in applications such as speech recognition and object detection.

AlexNet is a deep, convolutional neural network originally designed to classify 1.2 million high-resolution images in the ImageNet Large Scale Visual Resolution Challenge (ILSVRC) in 2010 into 1000 different classes. It has 650,000 neurons in a total of eight (8) hidden layers of neurons.

During training, Alexnet used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers the network employed “dropout” method that proved to be very effective. In the ILSVRC-2012 competition a variant of Alexnet achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry [7].

---

This work is partially supported by the LINC+, Hannam University (Project No. 201701540001) and by the Partido State University. The authors would like to thank Mr. Jerry Mercado, the Municipal Agricultural Officer of Goa, Camarines Sur, and the Crop Protection Office of DA-ROV.

Instead of sigmoid activation functions, Alexnet used Rectified Linear Units (ReLU), and a softmax function at the output of the last fully connected layer.

### C. The Rice Plant and Its Infestations/Diseases

Rice, also known as *Oryza Sativa*, is one of the most important plants with over half the world population depending on it for food. It is primarily grown in Asia and in the Philippines for instance, rice is a major staple food for millions of Filipinos. As the Department of Agriculture [8] puts it, “an average Filipino diet is based on rice. It provides half of our calorie requirements and one-third of our protein intake. Rice accounts for 20% of food expenditures for average households, which increases to 30% for households belonging to the bottom third of our society.”

However, despite numerous programs being orchestrated by the government, the ever growing multiplicity of diseases that affect rice productivity remain a serious issue. Data from the International Rice Research Institute (IRRI) Knowledge Bank show that rice farmers lose an estimated average of 37% of their rice crop to pests and diseases every year [9].

If the onset of a rice disease is instantly detected, its spread can be prevented by administering timely interventions. But before anything can be done, the immediate detection of the early signs or stages and symptoms of any disease is paramount. As intimated in [9], “in addition to good crop management, timely and accurate diagnosis can significantly reduce losses.”

Generally, a rice disease is an abnormal condition that injures the rice plant and diminishes its ability to produce food. These diseases are readily recognized by their symptoms primarily by visual features on the leaves of the rice plant. There are a lot of disease types such as Bakanae, Rice Blast, Bacterial Blight, Sheath Blight, Brown Spot, Bacterial Leaf Streak, False, Smut, Tungro, Leaf Scald, and Stem Rot [9].

### D. Original Contribution

A number of ideas have been proposed on the use of image processing techniques in the identification and detection of plant diseases such as those in [10]-[16].

The use of Support Vector Machines (SVM) as a classification algorithm was demonstrated in the work of Singh, et. al. [15] to identify Leaf Blast in rice plants. The authors claim 82% classification rate.

In [16], the image processing algorithm developed was enhanced with an interface for digitally illiterate users, especially farmers to efficiently and effectively retrieve information. This work therefore takes into account some principles of Human-Computer-Interaction (HCI), which is a significant step forward considering that most farmers are alien to the digital world.

Phadikar et al. [12] is also a study featuring use of Support Vector Machines. However, the proposed system has two (2) stages: first, detection of disease is accomplished through histogram characterization; and second, either a Bayes' or SVM algorithm is applied. The system gives 79.5% and 68.1%

recognition rates for the Bayes' and SVM classifiers, respectively.

Aside from those mentioned above, while some researchers have also ventured into the possibility of using shallow, fully-connected networks in identifying rice diseases/infestations, the use of deep convolutional networks for this application still remains to be examined. This is the main contribution of this paper.

### E. Organization of the Paper

The rest of the paper is organized as follows: Section II discusses the methodologies followed, particularly the image data sets used and the leaning scheme applied. Section III presents the results, followed by the Conclusion in Section IV.

## II. METHODS

### A. Fine-Tuning AlexNet

We first customized AlexNet in order to accommodate our multi-class classification problem. The objective of our proposed system is to classify an input image of rice plant of no a priori class into whether: (a) it is infested with golden apple snails; (b) it is afflicted with diseases; and (c) it is normal and healthy. And since AlexNet was designed to handle 1000 classes, its output layer also has to be retrofitted to handle our 3-class system.

The learning algorithm used was the same with the pre-trained network, Stochastic Gradient Descent, while adopting the mini-batch size of thirty (30), and base learning rate of 0.0001.

### B. Rice Image Datasets

A total of two hundred twenty-seven (227) rice images were captured in ricefields around the district particularly in Goa (Digdigon, Buyo, Matacla, Abucayan, Halawig-gogon, Catagbacan, and Belen), San Jose (Bilog, Pugay, and Dolo), Tigaon (San Rafael, Vinagre, and San Antonio), and Sagnay (Huyon-huyon, and Nato). The images were resized into 227 x 227 resolution in order to fit the input layer of AlexNet.

These images were manually augmented by operations such as cropping and rotations, as well as by downloading public images from [17]. A priori classification was conducted with the assistance of technicians at the Municipal Agriculturist's Office of Goa, Camarines Sur, and the Crop Protection Office of Department of Agriculture, Regional Office V.

Of the total images images, 70% or six-hundred (600) were used for training and the remaining 30% or two hundred fifty-seven (257) were used for testing. There were also unstructured interviews conducted with farmers during the image capture activities.

Training images were labeled a priori classes and fed into the AlexNet network for fine-tuning. After 10 epochs, the accuracy was calculated using testing images.

The simulation data were stored and the activations are characterized visually as feature maps.

### III. RESULTS

Our unstructured interviews with farmers revealed that the three (3) most common rice plant infestations/diseases in the district are: golden apple snail, tungro, and black bug. Fig. 1 shows sample images of rice plants afflicted with these three (3) most common diseases/pests in the district.

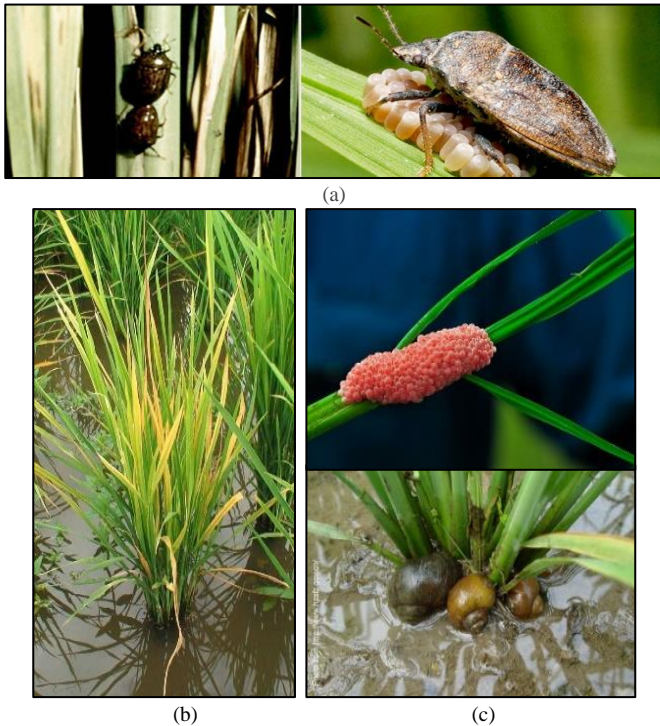


Fig. 1. Images of Rice Plants with Infestations and/or Diseases: (a) Black Bug, (b) Tungro, and (c) Golden Apple Snails. Images were downloaded from [9].

Fig. 2 is a sample image of rice infested with golden apple snail used in the test. The corresponding activations of all channels for convolution layers 1 and 2 are in Fig. 3 and 4, respectively. These activations are characterized as feature maps for visual analyses. Feature maps are results of the convolution layers which after the convolution filters are applied, which explain the similarity with edges.



Fig. 2. Sample test image of rice with golden apple snail infestation.

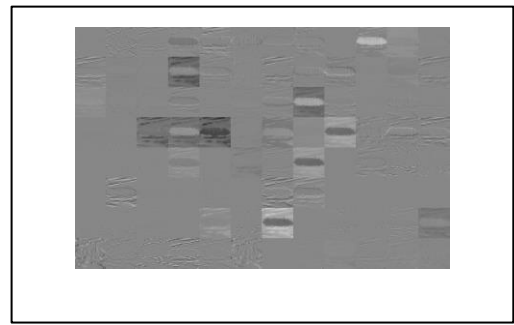


Fig. 3. Activations / Featuremaps in All Channels of Convolution Layer 1 for the Test Image in Fig. 2.

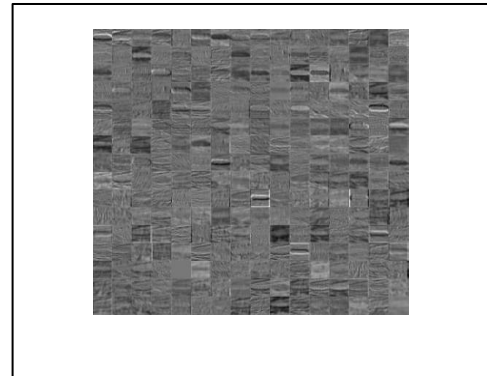


Fig. 4. Activations / Featuremaps in All Channels of Convolution Layer 2 for the Test Image in Fig. 2.

It can be noticed from the figures above that in convolution layer 1, there are 96 feature maps, while in layer 2, there are 256 thumbnail feature maps resulting from the 96 filters in layer 1, and 256 filters in layer 2 of AlexNet network, respectively.

Table I below summarizes the training and learning results, where stochastic gradient descent was used, with a base learning rate of 0.0001 and batch size of 30. The speed of convergence indicated by the figures under the column “time elapsed” is noticeable as is always the case in transfer learning. In matter of minutes the algorithm converged into an acceptable accuracy.

The column ‘mini-batch accuracy’ refers to the accuracy of the algorithm measured against the sample set of images in the batch. Recall that in stochastic gradient descent method, the algorithm estimates the gradient vector using a number of random sample of images taken from the training set known as mini-batch size. In the case of our study, batch size is 30 images.

TABLE I. SUMMARY OF LEARNING DATA

Epoch	Iteration	Time Elapsed	Mini-batch Loss	Mini-batch Accuracy
1	1	1.04	1.7476	30%
3	50	27.87	0.5537	76.67%
5	100	57.98	0.3009	93.33%
8	150	88.49	0.1430	96.67%
10	200	118.25	0.0888	96.67%



The next figure, Fig. 5 shows a sample output of the classifier for twenty (20) test images. All images in the figure were correctly classified, the last image being labeled “Kuhol Detected”. “Kuhol” is the Filipino term for Golden Apple Snail.

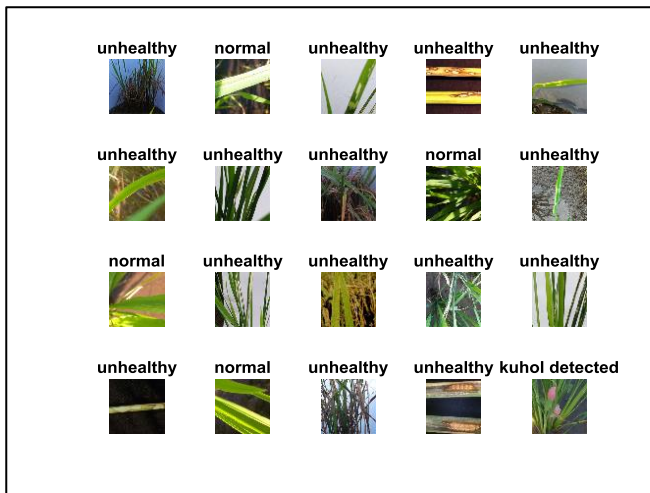


Fig. 5. Classification results of 20 test images fed into the network. The last image at the lower right corner is labeled “kuhol detected”. “Kuhol” is the Filipino word for golden apple snail.

#### IV. CONCLUSION

Leveraging on the architecture of AlexNet, we developed in this paper, a deep convolutional neural network applying the pre-trained weights and biases for classifying rice plants' images into three (3) classes: normal, unhealthy, or golden apple snail infested. The dataset of images was comprised of images captured in several ricefields around the district as well as public images from the internet. The images were resized and augmented and divided into training-testing set of 70%-30% ratio.

Stochastic gradient descent was the learning algorithm applied under a base learning rate of 0.0001 and batch size of thirty (30), which produced the result of 91.23% accuracy.

#### V. FUTURE WORK

The researchers are interested to expand this work to include other classes not yet covered in the present study, such as Brown Spot and Leaf Blythe, and other rice abnormalities. The reason these were not yet included is the absence of sufficient image sets. Furthermore, the use of multispectral high altitude images are also being considered.

#### ACKNOWLEDGMENT

This work is partially supported by the LINC+, Hannam University (Project No. 201701540001) and by the Partido State University. The authors would like to thank Mr. Jerry

Mercado, the Municipal Agricultural Officer of Goa, Camarines Sur, and the Crop Protection Office of DA-ROV.

#### REFERENCES

- [1] Fuentes, S. Yoon, S. C. Kim and D. S. Park, "A Robust Deep-Learning-Based Detector for Real-Time Tomato Plant Diseases and Pests Recognition," *Sensors*, vol. 2017, no. 9, 6 September 2017.
- [2] S. P. Mohanty, D. Hughes and M. Salathe, "Using Deep Learning for Image-Based Plant Disease Detection," 22 September 2016. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpls.2016.01419/full>. [Accessed 12 November 2017].
- [3] Z. Qin and M. Zhang, "Detection of Rice Sheath Blight for In-Season Disease Management Using Multispectral Remote Sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 7, no. 2, pp. 115-118, 2005.
- [4] M. A. Nielsen, *Neural Networks and Deep Learning*, Determination Press, 2015.
- [5] International Rice Research Institute, "Brown Spot," 29 February 2016. [Online]. Available: <http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/diseases/item/brown-spot>.
- [6] International Rice Research Institute, "Bacterial Blight," 29 February 2016. [Online]. Available: <http://www.knowledgebank.irri.org/decision-tools/rice-doctor/rice-doctor-fact-sheets/item/bacterial-blight>.
- [7] Krizhevsky, Alex, S. Ilya and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems*, 2012.
- [8] Department of Agriculture, "RICE PROGRAM," February 2016. [Online]. Available: <http://davao.da.gov.ph/index.php/programs/rice-program>.
- [9] International Rice Research Institute, "How to Manage Pests and Diseases," 29 February 2016. [Online]. Available: <http://www.knowledgebank.irri.org/step-by-step-production/growth/pests-and-diseases>.
- [10] V. Singh and A. Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques," *Information Processing in Agriculture*, vol. 4, no. 1, pp. 41-49, 2017.
- [11] S. D. Khirade and A. B. Patil, "Plant Disease Detection Using Image Processing," in *International Conference on Computing Communication Control and Automation*, Pune, 2015.
- [12] S. Phadikar, J. Sil and A. K. Das, "Classification of Rice Leaf Diseases Based on Morphological Changes," *International Journal of Information and Electronics Engineering*, pp. 460-463, 2012.
- [13] S. K. Tichkule and D. H. Gawali, "Plant diseases detection using image processing techniques," in *Online International Conference on Green Engineering and Technologies (IC-GET)*, Coimbatore, 2016.
- [14] V. Singh, A. K. Misra and V. Misra, "Detection of unhealthy region of plant leaves using image processing and genetic algorithm," in *International Conference on Advances in Computer Engineering and Applications*, Ghaziabad, 2015.
- [15] A. K. Singh, A. Rubiya and B. S. Raja, "Classification of Rice Disease Using Digital Image Processing and SVM Classifier," *International Journal of Electrical and Electronics Engineers*, pp. 294 - 299, 2015.
- [16] N. Mittal, B. Agarwal, A. Gupta and H. Madhur, "Icon Based Information Retrieval and Disease Identification in Agriculture," *International Journal of Advanced Studies in Computer Science & Engineering*, pp. 26-31, 2014.
- [17] The University of Georgia-WSFNR; The University of Georgia-CAES; USDA Identification Technology Program; "IPM Images," 2017. [Online]. Available: <https://www.ipmimages.org/>. [Accessed 10 December 2017].

# Improve Mobile Agent Performance by using Knowledge-Based Content

Tarig Mohamed Ahmed

Assoc. Prof., Department of MIS, Prince Sattam Bin Abdulaziz University AL-Kharj, Saudi Arabia  
Faculty of Mathematical Science, University of Khartoum, Khartoum, Sudan

**Abstract**—Mobile agent technology is one of the mobile computing areas. This technology could be used in several types of applications, such as cloud computing, e-commerce, databases, distributed systems management, network management, etc. The purpose of this paper is to propose a new model for increasing the mobile agent systems performance. The performance is considered as one of the important factors that makes the system reliable. This paper suggests a knowledge-based content to be used to improve the mobile agent systems performance. In the beginning, this work started by conducting intensive survey about related models and mechanisms to investigate the gaps in the performance. A comparative discussion has been conducted between some researches issued and the proposed model. The proposed model has been described in full details based on the components. A scenario-based approach has been used to implement the proposed model by using .Net framework and C# language. The model has been tested and evaluated based on different scenarios. As findings, the overall performance has been improved by 83% when the knowledge-based content is used. In addition, the system performance will improve automatically by the time because the content of the knowledge is increased. The proposed model is suitable to be used in any type of mobile agent applications. The originality of the model is based on conducted survey and own knowledge.

**Keywords**—Mobile agent; mobility; performance; intelligent system

## I. INTRODUCTION

Mobile agent technology is one of the mobile computing areas. This technology could be used in several types of application such as cloud computing, e-commerce, databases, distributed systems management, network management.etc. The mobile agent systems have many benefits if we compared with the client/server model such as saving network bandwidth, reducing network latency and reducing network consuming cost. The mobile agent system works based on concept of remote programming. The mobile agent travels to several nodes to accomplish tasks on behalf of users. The agent is one of the system components. It is represented as object which consists of two parts: Data Stat and Code. The data state represents the information domain of the agent. The code represents the statements that will be executed in the hosts (Service Providers). By using an itinerary table and mobility mechanism, the mobile agent travels among network hosts. The mobile agent home creates the agents according to users' requests. In addition, it dispatches and receives the agents with results. The host or service provider represents the node that is visited by mobile agents. The hosts can receive

multiple mobile agents simultaneously. The mobile agent system uses a communication mechanism that allows the agents to communicate together and with other system components. The mobility feature is a key feature of the mobile agent which allows the agents move from host to another. Vigna and Fuggetta et al. [12] defined the agent component as two parts: execution and resources units. The execution unit represents a computation algorithm. The resource unit represents as information domain that will be used by the execution units. There are two types of mobility: strong and weak mobility. Strong mobility allows the mobile agent that can carry a code and an execution state during the journey. Weak mobility allows the mobile agent to carry only a code with some initial values.

The mobile agent performance is one of the key issues [5] that make the system reliable and successful. The performance depends on many factors, one of them is the agent journey duration time. The duration time depends on two items: number visited hosts and the services execution time. In this paper a new model is proposed to improve the mobile agent performance. The model is based on reducing the number of visited hosts by the mobile agent. The main idea of the model is to use a knowledge-based content. The knowledge-based content represents the mobile agent experiences from pervious journeys. Before the mobile agent starts a new journey, it should consult the knowledge-based content in order to check if there is a previous knowledge could be used to reduce the number of visited hosts. This idea works only if the mobile agent wants to make a selection or searching for benefits among hosts for example, buying books or tickets or any goods.

The reset of the paper was organized in five sections: Section 2 explores some models and mechanisms the mobile agent performance with some discussions. Section 3 presents the proposed model with full details of all components. The model implementation has been mentioned with full discussion of the results in Section 4. This work was concluded in Section 5 with some recommendations as future work to enhance the model.

## II. RELATED WORK

In this section, some researches issued have presented and discussed. These researches were agreed on how to find a mobile agent model with high performance by using different mechanisms such as: network protocols, using parallel

processing, ranking services provider, etc. the following section presents some of them.

Salamat et al. proposed extended hierarchical query retrieval (EHQR) approach to enhance the mobile agent performance. The main idea behind this approach was to send many agents simultaneously in order to reduce the time taken for tasks. To evaluate EHQR, two experiments had been conducted by using query online and offline [1]. By using SNMP (simple network management protocol), Rantes et al. developed a model for evaluating mobile agent performance factors. After conducting many experiments, the results mentioned that the mobile agent performance depends on the network management and some parameters related to a network topology, network latency [2]. Holt et al. also used SNMP to analysis mobile agent performance factors. The model evaluated the mobile agent in two environments: Local Area Network (LAN) and Wide Area Network (WAN). The study mentioned that the bounded size of the mobile agent is optimal in a large network domain and the adopting of clustering strategy controls the mobile agent size [3]. Devadas et al. proposed a knowledge based component in a mobile agent system that can help the agents to communicate together. By this way the mobile agent performance will be increased [4]. Tarig proposed a new mechanism for increasing the mobile agent performance by reducing the mobile agent size during the agent journey. The mechanism called Free Area Mechanism (FAM). The mechanism was implemented using .Net framework and many experiments had been conducted to test the performance [5]. Sasirekha et al. proposed a new mechanism to improve the mobile agent routing algorithm. The algorithm called cluster-chain mobile agent routing (CCMAR). It used a wireless sensor network (WSN) into a few clusters and runs. Two phases were used to implement the algorithm. First, the nodes in the chained cluster aggregate the data in the cluster. Second, the mobile agents collect the data that aggregated in the cluster [6]. Aloui et al. proposed a solution for Multiple agents Itinerary Planning (MIP). The solution was based on agent's location and their size to make balance in consuming network energy. After conducting many experiments, the results mentioned that the performance was increased [7].

The accurate prediction for the resources is very important for the mobile agent to improve the performance. Chaudhar et al. proposed to use Cognitive Agent in Mobile Ad hoc Network. The cognitive agent makes the mobile agent thinking like human to take the right decision regarding to the resources. By this way, the mobile agent can determines the best traffic plan to achieve its tasks [8]. Channappagoudar et al. had conducted a study related to the resource allocation protocol. The study used static and mobile agents to evaluate the performance. The main idea of the proposed protocol was to allow the static agents to collect resource information about nodes in the network and providing the mobile agent by this information. By this way, the mobile agents will increase their performance [9]. Prapulla et al. proposed a model for multi mobile agents to reduce the energy consummation and latency. The model based on two types of mobile agents: Link Agent and Data Agent. The link agent aimed to monitoring the network resources and status. The data agent aimed to transfer

data among nodes. The idea of this model helps mobile agents to prepare their itinerary tables. By clustering the network nodes, the model was implemented and the efficiency was discussed [10]. Based on opinion-based, Zuo et al. proposed a model for increasing the mobile agent system performance. The model ranks the reputation of network nodes by aggregating information. The node reputation ranking was based on set of categories such as services quality. By this way, the mobile agents were owned valuable information before starting their journeys and the overall performance will be increased. The model was implemented and evaluated by using Algets technology [11].

Baek et al. [13] suggested planning algorithms tried to search a minimum number of agents, and the total consuming time of route by setting lime of time execution. There are two important planning factors affecting the performance of the agent system in the network environment that are the mobile agent's itinerary and the number of the mobile agent. The experiment of this research proves that if the size of a mobile agent is begin increased while retrieval operations are performed, the bandwidth varies from link to link. In this case, the agent will consume more time. Cook [14] has mentioned that building software system composed of mobile agents introduces interesting new concerns for software engineering research. He described some assumptions behind mobile agent systems and software engineering. One of them: Code is cheaper to move than data and this assumption implies that the size of the mobile agent is important and it should be reduced.

As mentioned above, all these models or mechanism were aimed to improve the mobile agent systems performance. If we compare them with our proposed mechanism, we find the dynamic and increment improvement of performance by using the knowledge-based. This fact is more suitable with nature of the information systems because the information related to the service providers are not fixed and rapidly changed. Also, the proposed mechanism allows the mobile agent systems automatically adapting with information available in the knowledge based-content to improve the performance. In addition, the performance will depend on the mobile agents' experiences.

### III. PROPOSED MODEL

#### A. Model Concepts and Components

This research aims to propose a new model to increase the mobile agent systems performance. The main idea behind this model is to use a knowledge-based content. The knowledge-based content helps the mobile agent system to reduce number of visited hosts. It can provide the mobile agents by valuable information related the services located in hosts. The knowledge-based content is incremental database that consists of information that collected by the mobile agents. When the mobile agent completes its journey, the mobile agent home extracts the information about hosts available in the mobile agent and stores it in the knowledge-based content. This process is repeated every time when the mobile agents back to their home. In addition, before the mobile agents start their journey, they should optimize their itinerary tables based on information available in the knowledge-based content. By this way the number of visited hosts will be decreased and the

overall performance automatically is going to be improved. In addition, the knowledge-based content is dynamically updated and rapidly increased by knowledge the mobile agents' experiences. Fig. 1 presents the main components of the mechanism.

As presented above, Fig. 1 depicts the mobile agent model components: Mobile Agent, Mobile Agent Home, Knowledge-based content and Hosts. Each one has specific role in the model as following:

Mobile Agent is an object, which contains tasks to be performed on behalf of users. The mobile agent visits hosts according to its itinerary table. The journey duration depends of many factors and one of them is number of visited hosts. The number of hosts is specified based on the mobile agent's tasks.

Mobile Agent Home is a place where the mobile agents start their journey. After finishing their journey, the mobile agents return home with results. Mobile Agent Home is responsible to create the mobile agent and extract their results after completing their journey. The results will be saved in the knowledge-based content.

Knowledge-Based content is a knowledge container which can store services information provided by hosts. Before mobile agents start their journey, they visit the knowledge-based content to obtain information related to their duties. In addition, after mobile agents finishing their journey, they

come again to the knowledge-based content and update the information. By this way, the knowledge-based content is incremented and that make the model more efficient and intelligent. The knowledge format in the knowledge base depends on kind of services that are provided by the hosts.

Hosts represent the service providers. It can receive and serve multiple mobile agents simultaneously. The hosts could introduce the services for example, selling books, selling electronic devices, air tickets etc.

As mentioned above in Fig. 1, the model works in specific logic steps such as:

- 1) The mobile agent is created in the mobile agent home based on user's request. The mobile agent visits the knowledge-based content to search for related knowledge to its tasks. If any, the mobile agent can benefit from it by reducing the number of hosts in the itinerary table (Key issue to increase the performance).
- 2) The mobile agent returns to the home to start its journey by using updated itinerary table.
- 3) The mobile agent starts its journey by visiting hosts (Service Providers) to perform the tasks.
- 4) After completing the journey, the mobile agent returns home with results.
- 5) The knowledge that is collected by the mobile agent will be stored in the knowledge-based content.

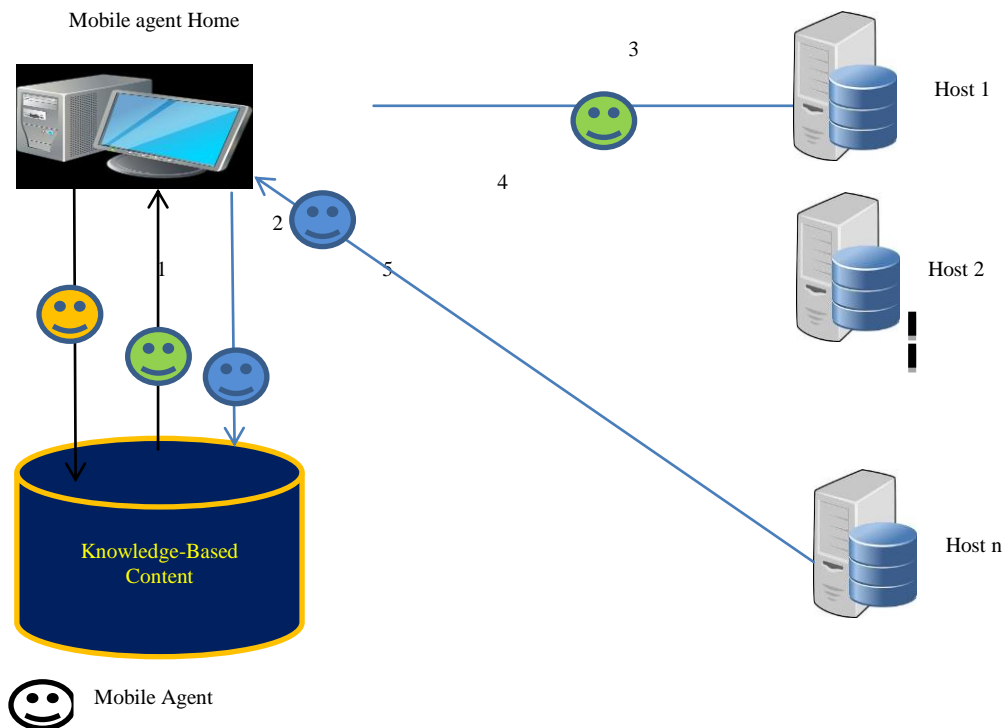


Fig. 1. Model components.

### B. Model Implementation

By using .Net framework and C# language, the model has been implemented and tested. Each component has a separate program as following:

Mobile agent is represented as object form, a class called "Agent". The class consists of several variables represent the object state such as Agent Name, Start journey time, finish journey time, itinerary table, Tasks, etc. The class also contains some methods to access those variables. Fig. 2 presents the pseudo code for the Mobile Agent Class. This class also was assigned as a Serializable to enable the object to be converted as a network stream.

The mobile agent home was implemented as two separate programs: sender and receiver mobile agents. The sender

program was represented as class called "Agent\_Sender". It consists of some variables and methods. It aims to create and dispatch the mobile agent object. For example, assigning the tasks, itinerary table and sending the mobile agent after converting it as a network stream. Fig. 3 presents the pseudo code.

The mobile agent home (Receiver program) was implemented as class. It aims to receive the mobile agents after completing their journey. It consists of two main methods: Receiving the mobile agent and extracting results from the mobile agents. The results will be stored in the knowledge-based content to be used in the future journeys. Fig. 4 presents the class of receiver agents. In addition, the program works as server listener to receive multiple mobile agents simultaneously.

```
[Serializable]
public class Agent
{
    public string agent_name;
    public DateTime start;
    public DateTime end;
    public String[] Tasks = new string[100];
    public int[]itinerary = new int[100];
    public int Current_location = -1;
    // Access methods
}
```

Fig. 2. Mobile agent class.

```
class Agent_Sender
{
    public Socket agent_socket;

    public Agent_Sender(Socket s)
    {
        this.agent_socket = s;
    }

    public void send_agent(Agent agent, int p)
    {
        int port = p;
        IPAddress ip = IPAddress.Parse("127.0.0.1");
        TcpClient client = new TcpClient();
        client.Connect(ip, port);
        NetworkStream stream_data = client.GetStream();
        BinaryFormatter bf = new BinaryFormatter();
        MemoryStream ms = new MemoryStream();
        //...
        client.Close();
    }

    // setting Tasks
    // Setting Nodes Information
    send_agent(Agent agent, int p)
}
}
```

Fig. 3. Mobile agent home (sender).

```
class Receiver_Agent
{
    public Socket agent_socket;
    public Receiver_Agent(Socket s)
    {
        this.agent_socket = s;
    }
    public void Extract_result(Agent a,int duration,string[] v)
    {
        OleDbConnection con = new
        OleDbConnection("provider=Microsoft.ace.Oledb.12.0; data
        source=d:\\AgentDB.accdb;Persist Security Info=False");
        con.Open();
        // Add result to database (Knowledge base
        con.Close();
    }
    public void create_agent()
    {
        byte[] agent_data = new byte[50000];
        int z = agent_socket.Receive(agent_data, agent_data.Length, 0);
        BinaryFormatter bf = new BinaryFormatter();
        MemoryStream ms = new MemoryStream(agent_data);
        Agent a = new Agent();
        ms.Seek(0, 0);
        a = (Agent)bf.Deserialize(ms);
        // create the mobile agent
    }
    static void Main(string[] args)
    {
        System.Console.WriteLine("***** MOBILE AGENT HOME *****");
        string ipaddress = "127.0.0.1";
        int port = System.Convert.ToInt32("8081");
        IPAddress ip = IPAddress.Parse(ipaddress)
        System.Net.Sockets.TcpListener listener = new
        System.Net.Sockets.TcpListener(ip, port);
        listener.Start();
    }
}
```

Fig. 4. Mobile agent home (receiver).

```
namespace host1
{
    class Program
    {
        public Socket agent_socket;
        public Program(Socket s)
        {
            this.agent_socket = s;
        }
        public static void send_agent(Agent ag, int p)
        {
            int port = p;
            IPAddress ip = IPAddress.Parse("127.0.0.1");
            TcpClient client = new TcpClient();
            client.Connect(ip, port);
            NetworkStream stream_data = client.GetStream();
            BinaryFormatter bf = new BinaryFormatter();
            MemoryStream ms = new MemoryStream();
            bf.Serialize(ms, ag);
            //...
        }
        public void create_agent()
        {
            System.Console.WriteLine("new agent under processing");
            byte[] agent_data = new byte[50000];
            int z = agent_socket.Receive(agent_data, agent_data.Length, 0);
            //...
        }
        static void Main(string[] args)
        {
            System.Console.WriteLine("**** The Service Provider Name: HOST No 1 ****");
            string ipaddress = "127.0.0.1";
            int port = System.Convert.ToInt32("8082");
            IPAddress ip = IPAddress.Parse(ipaddress);
            System.Net.Sockets.TcpListener listener = new System.Net.Sockets.TcpListener(ip,
port);
            listener.Start();
            System.Console.WriteLine();
            System.Console.WriteLine("##### Now Host 1 Running ##### ");
            try
            {
                while (true)
                {
                    Socket line = listener.AcceptSocket();
                    //...
                }
            }
        }
    }
}
```

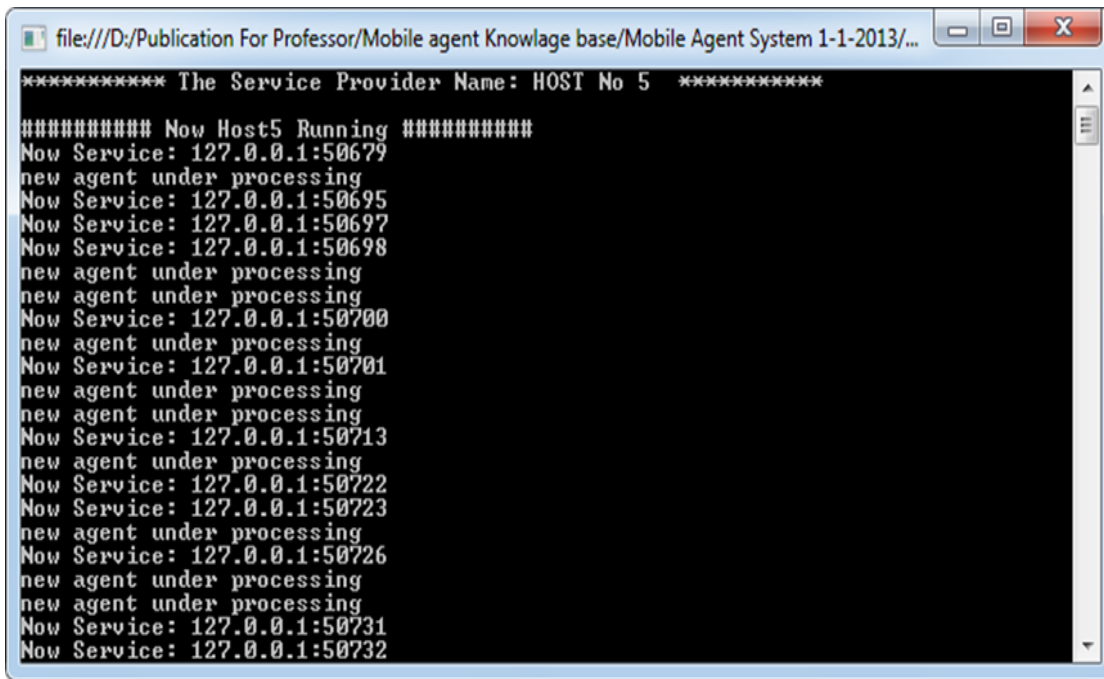
Fig. 5. Host (service provider).

The host (Service Provider) was represented as a program. Each node has its one running program with unique IP address. The node can receive multiple mobile agents and it serves them according to their tasks. After completing their jobs, the node dispatches the mobile agent according to the itinerary table. Fig. 5 presents the host program. This program is run each time when a new host is started.

By using Microsoft Access 2007, the knowledge base was implemented as a database. The database consists of two tables. The first table was used to store results from mobile agent journeys. The second table represents the visited hosts.

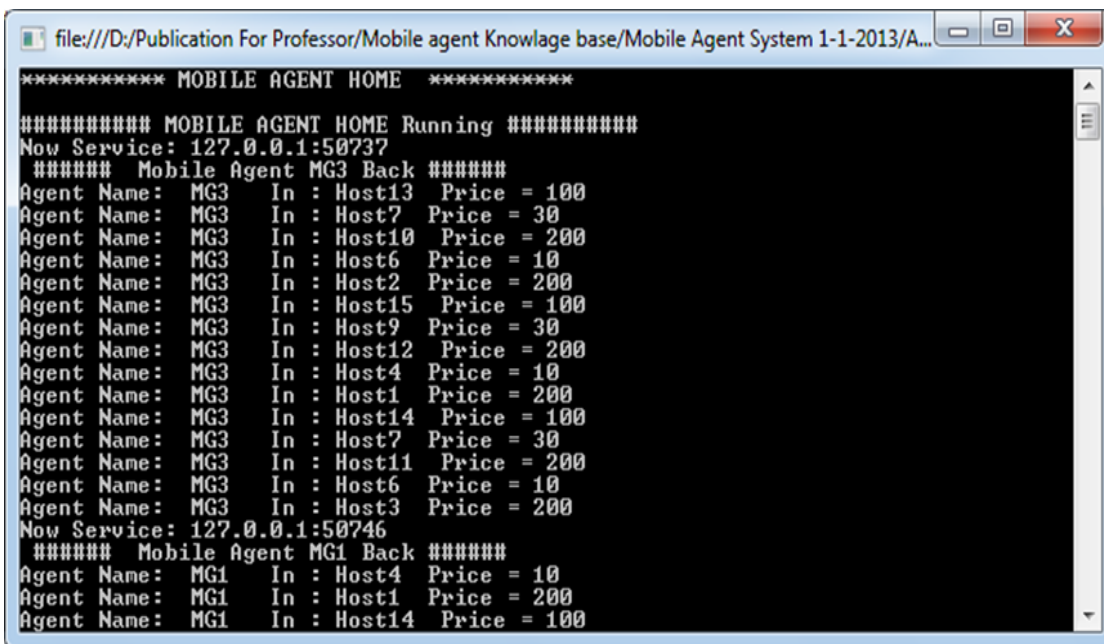
### C. Model Test and Evaluation

A scenario-based approach has been used to test and evaluate the proposed model. In the scenario, the mobile agent buys books on behalf of users based on best price. The user specifies the required book information. By using this information, the mobile agent searches among the nodes (Hosts), the required book and its price. The host represents a bookstore. After completing its journey, the mobile agent saves all collected information in the knowledge-based content. As sample, Fig. 6 presents multiple mobile agents visit Host 5.



```
file:///D:/Publication For Professor/Mobile agent Knowledge base/Mobile Agent System 1-1-2013/...
***** The Service Provider Name: HOST No 5 *****
##### Now Host5 Running #####
Now Service: 127.0.0.1:50679
new agent under processing
Now Service: 127.0.0.1:50695
Now Service: 127.0.0.1:50697
Now Service: 127.0.0.1:50698
new agent under processing
new agent under processing
Now Service: 127.0.0.1:50700
new agent under processing
Now Service: 127.0.0.1:50701
new agent under processing
new agent under processing
Now Service: 127.0.0.1:50713
new agent under processing
Now Service: 127.0.0.1:50722
Now Service: 127.0.0.1:50723
new agent under processing
Now Service: 127.0.0.1:50726
new agent under processing
new agent under processing
Now Service: 127.0.0.1:50731
Now Service: 127.0.0.1:50732
```

Fig. 6. Mobile agents under processing.



```
file:///D:/Publication For Professor/Mobile agent Knowledge base/Mobile Agent System 1-1-2013/A...
***** MOBILE AGENT HOME *****
##### MOBILE AGENT HOME Running #####
Now Service: 127.0.0.1:50737
##### Mobile Agent MG3 Back #####
Agent Name: MG3 In : Host13 Price = 100
Agent Name: MG3 In : Host7 Price = 30
Agent Name: MG3 In : Host10 Price = 200
Agent Name: MG3 In : Host6 Price = 10
Agent Name: MG3 In : Host2 Price = 200
Agent Name: MG3 In : Host15 Price = 100
Agent Name: MG3 In : Host9 Price = 30
Agent Name: MG3 In : Host12 Price = 200
Agent Name: MG3 In : Host4 Price = 10
Agent Name: MG3 In : Host1 Price = 200
Agent Name: MG3 In : Host14 Price = 100
Agent Name: MG3 In : Host7 Price = 30
Agent Name: MG3 In : Host11 Price = 200
Agent Name: MG3 In : Host6 Price = 10
Agent Name: MG3 In : Host3 Price = 200
Now Service: 127.0.0.1:50746
##### Mobile Agent MG1 Back #####
Agent Name: MG1 In : Host4 Price = 10
Agent Name: MG1 In : Host1 Price = 200
Agent Name: MG1 In : Host14 Price = 100
```

Fig. 7. Mobile agents return home.

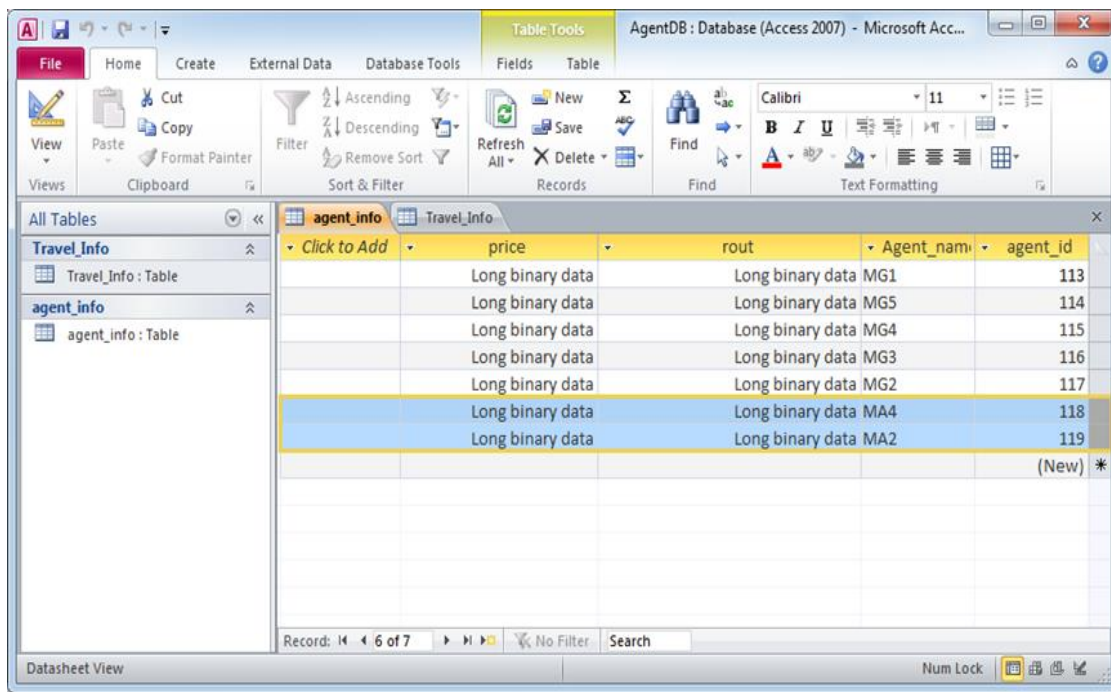


Fig. 8. Mobile agents in home.

In Fig. 7, the mobile agents return home with results from hosts.

The mobile agent home saves all information collected by the mobile agents in knowledge-based content. Fig. 8 presents the information as a table of object. Each object represents collected information of one journey as prices and visited hosts (route).

To evaluate the model, the durations of journeys have been computed with listing of visited hosts. In Fig. 9, all these

information are presented. As mentioned, the duration of Mobile Agents 4, 2(MG4, MG2). In the first journey the durations were 292 msec and 332 for MG4 and MG2 msec respectively. The two mobile agents had sent again for the same tasks. In the second journey, the durations were 56 msec and 51 for MG4 and MG2 msec, respectively. The time of journey was reduced by 83% average. According to this fact, the model has improved the performance of the mobile agent system by using the knowledge-based content.

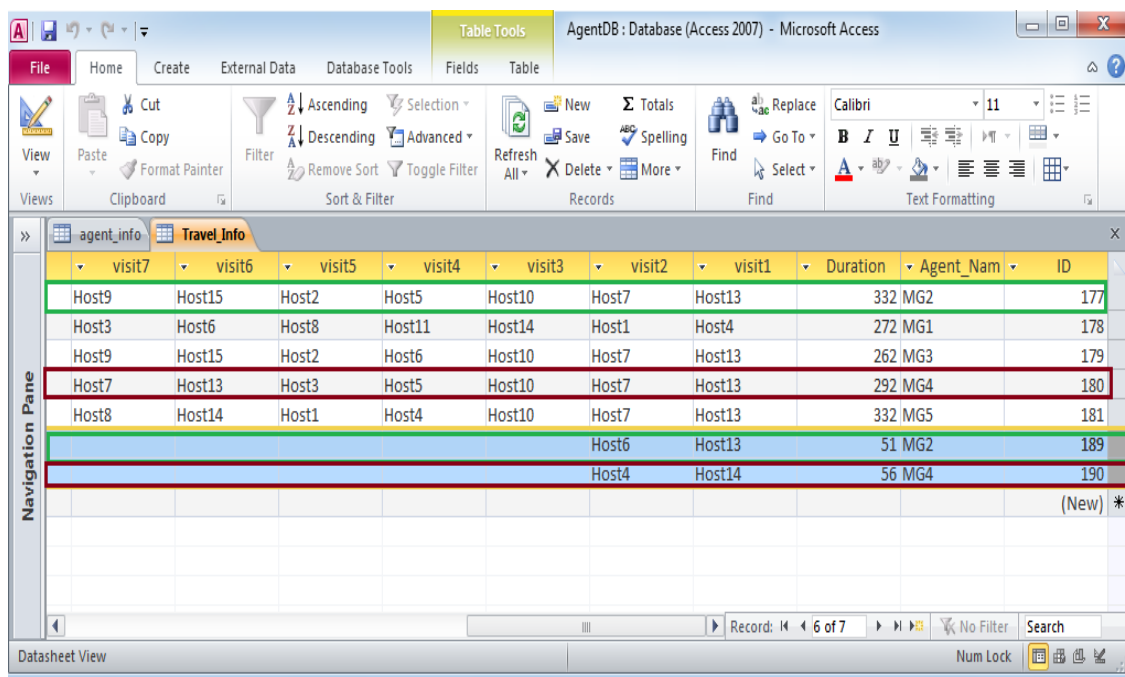


Fig. 9. Model performance measurements.



#### IV. CONCLUSION

The purpose of this paper is to propose a new model for increasing the mobile agent systems performance. The performance is considered as one of important factors that makes the system reliable. This paper suggests a knowledge-based content to be used to improve the mobile agent systems performance. In the beginning, this work started by conducting intensive survey about related models and mechanisms to investigate the gaps in the performance. A comparative discussion has been done between some researches issued and the proposed model. The proposed model has been described in full details based on the components. A scenario-based approach has been used to implement the proposed model by using .Net framework and C# language. The model has been tested and evaluated based on different scenarios. As findings, the overall performance has been improved by 83% when the knowledge-based content is used. In addition, the system performance will improve automatically by the time because the content of the knowledge is increased. The proposed model is suitable to be used in any type of mobile agent applications. The originality of the model is based on conducted survey and own knowledge.

As future work of this work, it will be a good idea if some effort be done in the knowledge-based content by setting protocol for knowledge formats. By this way the knowledge-based content will be used by heterogeneous mobile agent systems. In addition, this idea will help to share the knowledge between different systems.

#### REFERENCES

- [1] Ali Selamat, Hafiz Selamat, Analysis on the performance of mobile agents for query retrieval, Information Sciences, Volume 172, Issues 3-4, 9 June 2005, Pages 281-307, ISSN 0020-0255.
- [2] rantes, J. A., Westphall, C. B., Custódio, R. F., & de Chaves, S. A. (2010). Analytical model to evaluate the performance of mobile agents in a generic network topology. *Journal of Network and Systems Management*, 18(4), 357-373. doi:http://dx.doi.org/10.1007/s10922-010-9173-x
- [3] Holt, A., Huang, C., & Monk, J. (2007). Performance analysis of mobile agents. *IET Communications*, 1(3), 532-538.
- [4] Devadas, T. Joshva, and R. Ganesan. "INTELLIGENT AGENT-BASED KNOWLEDGE MANAGEMENT." *International Journal of Advanced Research in Computer Science* 3.2 (2012).
- [5] Tarig Mohamed Ahmed: "Increasing Mobile Agent Performance by Using Free Areas Mechanism", in *Journal of Object Technology*, vol. 6, no. 4, May-June 2007, pp. 125-140
- [6] Sasirekha, Selvakumar, and Sankaranarayanan Swamynathan. "Cluster-chain mobile agent routing algorithm for efficient data aggregation in wireless sensor network." *Journal of Communications and Networks* 19.4 (2017): 392-401.
- [7] Aloui, I., Kazar, O., Kahloul, L., & Servigne, S. (2015). A new itinerary planning approach among multiple mobile agents in wireless sensor networks (WSN) to reduce energy consumption. *International Journal of Communication Networks and Information Security*, 7(2), 116-122.
- [8] Chaudhari, Shilpa Shashikant, and Rajashekhar C. Biradar. "Traffic and mobility aware resource prediction using cognitive agent in mobile ad hoc networks." *Journal of Network and Computer Applications* 72 (2016): 87-103.
- [9] Channappagoudar, Mallikarjun B., and Pallapa Venkataram. "Performance evaluation of mobile agent based resource management protocol for MANETs." *Ad Hoc Networks* 36 (2016): 308-320.
- [10] Prapulla, S. B., et al. "Multi Mobile Agent itinerary planning using Farthest Node First Nearest Node Next (FNFNNN) technique." *Computation System and Information Technology for Sustainable Solutions (CSITSS)*, International Conference on. IEEE, 2016.
- [11] Zuo, Yanjun, and Jigang Liu. "A reputation-based model for mobile agent migration for information search and retrieval." *International Journal of Information Management* 37.5 (2017): 357-366.
- [12] Fuggetta, Alfonso, Gian Pietro Picco, and Giovanni Vigna. "Understanding code mobility." *IEEE Transactions on software engineering* 24.5 (1998): 342-361.
- [13] J. Cook, Software Engineering Concerns for Mobile Agent Systems, paper, proceeding of Workshop on Software Engineering and Mobility, Ontario, Canada, May 2001.
- [14] L. Ismail and D. Hagimont. A performance evaluation of the mobile agent paradigm. In *Proceedings of the Conference on Object-Oriented Programming, Systems, Languages, and Applications*, pages 306-313, 1999.

# Kit-Build Concept Map with Confidence Tagging in Practical Uses for Assessing the Understanding of Learners

Jaruwat Pailai, Warunya Wunnasri, Yusuke Hayashi, Tsukasa Hirashima  
Graduate School of Engineering  
Hiroshima University  
Hiroshima, Japan

**Abstract**—An answer of a learner can be interpreted as a learning evidence for demonstrating the understanding of the learner, while a confidence on the answer represents the belief of the learner as the degree of understanding. In this paper, we propose Kit-Build concept map with confidence tagging. Kit-Build concept map (KB map in short) is a digital tool for supporting a concept map strategy where learners can create the learning evidence, and the instructor can access the correctness and confidence information of learners. The practical uses were conducted for demonstrating the valuable of correctness and confidence information in the lecture class. The correctness information was visualized in the control classes, while the correctness and confidence information were visualized in the experiment classes. The observed evidence illustrates that the different information was used for selecting and ordering the supplementary content when the system visualized the different information. The normalized learning gains and effect size demonstrate the different learning achievements between control- and experiment- classes. The results suggest that the confidence information of learner affects the instructor behaviors, which is the positive changing behavior for improving the understanding of their learners. The results of questionnaire suggest that the KB map with confidence tagging is an accepted mechanism for representing the learner's understanding and their confidence. The instructors also accepted that the confidence information of learners is valuable information for recognizing the learning situation.

**Keywords**—Kit-Build concept map; confidence tagging; effect of confidence information; behavior changing of instructor

## I. INTRODUCTION

Knowledge is invisible, but it is possible to observe its effect. The knowledge is often defined as a belief, which is true and justified, while the certainty is an essential component of knowledge [1]. More discussion about the concept of knowledge was described by Hunt who mentioned that a knowledge measurement requires a measuring of correctness and sureness. The quality of knowledge can be represented by the certainty of the answer such as a learner who is sure on the correct answer and a learner who is not sure on the incorrect answer. Besides, the confidence is essential to influence real-life behavior, many decision-making, and learning processes [2]-[8]. Several researchers mentioned the confidence in the various situations. For instance, the confidence can encourage a deeper understanding of the material [9], and also increase

reflection and justification of the answers [10]. Consequently, the answer of learners represents their understanding, and the confidence on their answer indicates the degree of understanding. The answers and its confidence are possible to be utilized as the learning evidence of formative assessment for identifying the information of current learning situation, which the instructor can use to design and provide the feedback as the evidence-based feedback. The value of information will be indicated by an instructor when s/he used the information to provide the feedback for improving learners' understanding in their class.

The Kit-Build concept map (KB map in short) is a digital tool for supporting a concept map strategy, which can identify the correctness of learners-build concept map based on the instructor-build concept map automatically [11]. The learning goal is represented in the form of a concept map for indicating the expectation of the instructor, which the instructor-build map is called a goal map. Learners can construct the learner maps as the learning evidence by connecting the provided components of concept map (as we called "Kit") to form each proposition. The kit is the list of concepts and linking words from a decomposing the goal map. The assessing process of the KB map is to identify the correctness information by using the propositional level exact matching for generating the diagnosis results. The correctness information is available in the diagnosis results, which can be divided into individual-diagnosis results for informing the performance of learners individually and group-diagnosis results for informing an overview of the class. These abilities can help the instructor to reduce the gathering and assessing time instantly. The instructor can use the correctness information in various scenarios such as an intra-class feedback and an inter-class feedback [12], [13]. For instance, the group-diagnosis results identified the incorrect propositions that represent the misunderstanding of the class in only one map. The instructor can find the overview of class easily and prepare to provide the feedback shortly. Accordingly, the ability of the KB map suites to support the instructor for implementing the formative assessment in a classroom situation [14].

In this paper, we propose KB map with confidence tagging for eliciting learning evidence of learners and informing the correctness and confidence information to the instructor. The confidence tagging is integrated into the structuring task of the KB map, which learners can construct the map to represent

their understanding and identify their confidence on each unit of meaning. A completed proposition, which is able to tag the confidence, comprises one connected linking word between two concepts. The confidence of an answer is simplified in the form of confidence- and unconfidence-value, which the learner can assign to every complete proposition. Thus, the system can elicit learning evidence that includes the understanding of learners and the degree of the understanding in the gathering process. The confidence information of learners is utilized in the diagnosis results of the KB map for visualizing the degree of learner's understanding. Therefore, we present the practical uses of the KB map with confidence tagging in the classroom situations when the instructors implement the formative assessment in the lecture classes for illustrating the encouragement of correctness and confidence information in their instruction. Five paired classes were conducted in the practical uses, which each paired class was conducted by the same instructor, the same lecture topic, and two different classes. Only the correctness information was provided to the instructors of five control classes as a control group, while both correctness and confidence information were provided to the instructors of five experimental classes as an experiment group.

The investigation procedure focuses on the different behavior of the same instructor when s/he received the different information on the diagnosis results. From this procedure, we assume that the confidence information of learners effects on the supplementary content ordering of the instructor. The actual ordering of supplementary lecture was used as observed evidence to indicate how the instructor used the correctness and confidence information. Moreover, the normalized learning gains of class and the effect size demonstrate the different learning achievement between both groups, which can illustrate that the correctness and confidence based feedback of the experiment group can contribute the improvement of learning achievements better than the correctness based feedback of the control group in several classes. The learners of the experiment group have an ability to discriminate and interpret their understanding between correctness and confidence better than the learners of the control group significantly. Analysis of change of proposition type presents that the unconfident propositions are easier to be changed than the confident proposition. Finally, the questionnaire presents that the KB map with confidence tagging is an accepted mechanism. The learners accepted the mechanism for presenting their understanding as propositions and for tagging their confidence to each proposition. The instructors accepted that the confidence information of learners was the valuable information to identify learning situation and identify the degree of learners' understanding.

This paper is structured as follows: Section II demonstrates the utilizing of correctness and confidence information for classifying an answer of a learner. The formative assessment in a lecture class for improving the learning achievements, and the KB map for assessing the understanding of learners are also described in this section. Section III presents KB map with confidence tagging, the practical uses of the KB map in a lecture class, and the description of procedure. The results section, outlined in Section IV presents the observed evidence of the instructors and the learning achievements of learners.

Section V is the discussion about the effect of confidence information of learners on the instructors' behavior. Lastly, Section VI is the conclusion of this study.

## II. BACKGROUND

### A. An Assessment by using Correctness and Confidence

The confidence was used to ensure the performance of learning outcomes as the quality of knowledge or the actual performance [15] as one of assessment criteria. Confidence based learning promotes a fusion of correctness and confidence to identify the answer of learners in four quadrants. There is a definition of correctness and confidence for referencing following:

- *Correctness* is the justification of an answer, which consists of a correct answer and an incorrect answer.
- *Correct- or incorrect-* answer is justified by the criteria.
- *Confidence* is the certainty of an answer, which can be simplified the values as confidence and unconfidence.
- *Confidence- or unconfidence-* of the answer is stated by learners on their answer.

The two-dimensional assessment process was used to classify the answer into four quadrants based on the correctness and confidence simultaneously. The four quadrants of two-dimensional assessment following:

- A correct answer with confidence.
- A correct answer with unconfidence.
- An incorrect answer with confidence.
- An incorrect answer with unconfidence.

Several researchers have already proposed the scoring method based on the correctness and confidence for promoting the critical awareness and self-assessment [16]-[19], for instance, Certainty-based Marking (CBM), Confidence-based Scoring (CBS), and Certainty-based Assessment (CBA). The correct answer that learner has a confidence can get the score more than the correct answer with unconfidence, while the learner can get some score on the incorrect answer when s/he has no confidence on the answer. Zero scores or penalty score is given to the incorrect answer with confidence. The task to identify the confidence of learners on their answer is provided to learners in various strategies such as the answering of descriptive question, True/False question, or the multiple-choice question. The different values of confidence were applied to the scoring method. For instance, the two different values of sureness consist of sure and not sure, or the three different levels of certainty consist of low, middle, and high.

### B. Formative Assessment

A formative assessment provides an opportunity to improve learning achievements, which is different from evaluation in the form of a summative assessment. The key questions of formative assessment following: "Where are learners going?", "Where are learners now?" and "How to close the gap?" [20]. The information through the assessment can encourage the instructor for giving the feedback to improve the learners'

understanding in a timely manner, which is the most efficient feedback [21]. The interaction based on formative information is the formative assessment key feature [22]. The gathering and assessing the learning evidence for providing the feedback are the processes of completing the formative assessment, and are also creating an opportunity for improving learning achievements concurrently. Thus, the formative assessment approach is used to monitor the learning of learners for providing ongoing feedback, which is a key for helping the learners to achieve a learning goal. The learning goal indicates the answer of “where learners are going?” question and can be used as criteria for examining the learner’s knowledge. Subsequently, the learning evidence is assessed by the criteria to identify the correctness for responding “where are learners now?” question. In other words, the correctness of the evidence can inform the learning gap based on learning goal and the evidence of learners. The remaining requirement is “how to close the gap?” question, which can be solved by feedback. The instructor’s feedback is provided in a lecture class as a group feedback for improving the understanding of learners when the instructor duels with a large number of learners. Moreover, the individual feedback is possible to provide in the proper situation such as a focused class with a small number of learners. For instance, the instructors can give the feedback as the supplementary lecture based on the overview of their class, while the different feedback can be provided to some learners individually according to each learner’s misunderstanding after finished class. Thus, the implementation of formative assessment is a completion of the formative assessment cycle following gathering and assessing learning evidence and providing the feedback. It can create a chance to improve learning achievements in every cycle.

### C. Kit-Build Concept Map

Concept maps are graphical tools that are used to represent and organize knowledge [23]. A proposition is constructed by connecting two concepts via a relation with linking word for representing a unit of meaning. The propositions are a core component of measuring a map score. In education areas, concept maps strategy is utilized to represent and assess knowledge of learners in classes as the learning evidence. An instructor can gain the information of a classroom situation then give the feedback based on the information in various situations such as the using individual or group discussion can contribute self-awareness of learners [24], or the using of concept maps as a formative strategy to find discrepancies based on the criteria map before instructor gives the feedback to learners [25]. The concept map strategy is simple to use, effective, and satisfy on problem-solving in classroom situations [26], [27]. Accordingly, the concept map is an effective strategy in a classroom situation that affects achievements and interests of learners. Although the traditional lecturing contributed learning achievements and meaningful learning in the classroom situation, the concept map can significantly improve learning achievements of learners when compared with the lecturing and is also more effective than the traditional lecturing in encouraging meaningful learning [28]-[30].

The KB map is a digital tool for supporting a concept map strategy, which includes a construction tool where users can

construct concept maps and an automatic concept map assessment where the system can report diagnosis results [11]. The different type of concept map is available in different tasks and different meaning, which are connected to each other to form a reasonable relationship respectively. The primary map is a goal map (Fig. 1) that an instructor builds a traditional concept map as criteria for indicating a learning goal of the class. The learning evidence is constructed by the learners from connecting provided components. The provided components are the decomposed components of the goal map as a “kit” (Fig. 2) in the form of a list of concepts and linking words. A learner can connect a linking word between two concepts to form a proposition as a unit of meaning, and then all propositions become a learner map (Fig. 3) for representing their understanding as an answer. Subsequently, the correctness of each learner map is indicated by the propositional level exact matching automatically. The correctness information is reported in the form of an individual-overlay map, an individual-difference map, and a similarity score. The individual-overlay map contains the similarity score and modified visualization of learner map where the displaying of the line connection of the correct proposition is different from the incorrect proposition. The individual-difference map is a visualizing only the mistake of learners in the form of three types of error link, and the linking word of correct propositions are disappeared in this map.

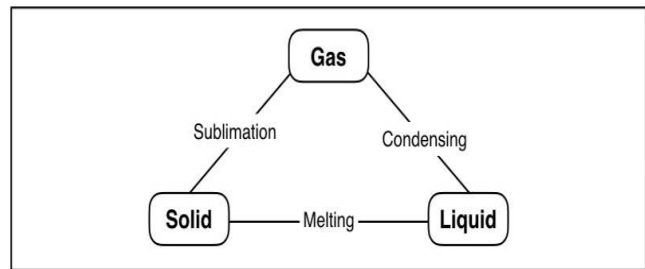


Fig. 1. An example of a goal map.

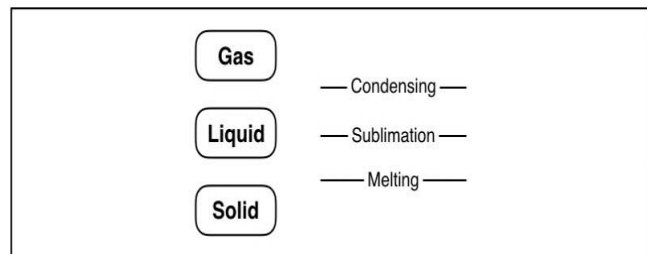


Fig. 2. An example of a kit.

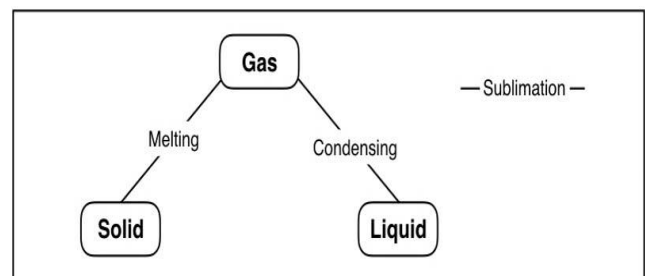


Fig. 3. An example of a learner map.

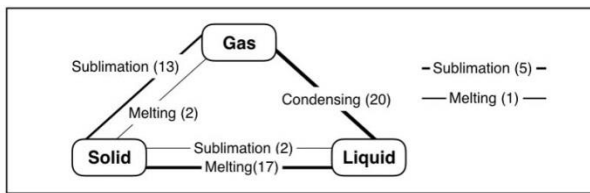


Fig. 4. An example of a group map.

The three types of error link consist of excessive links, leaving links, and lacking links. The link that is used to connect two concepts in learner map but at least one concept which is different from the goal map is called excessive link. The link that is not connected to any concept is called leaving link. The lacking links are used to call the link that is in the goal map but does not exist in the learner map, which the lacking link is the correcting of the excessive link or the leaving link. These are the individual-diagnosis results of the KB map.

Moreover, the advantage of the KB map is group-diagnosis results [31], [32]. A significant component of the KB map is the “kit” which is provided to all learners for constructing learner maps. Thus, all of the learner maps are constructed based on the same components, and overlaying all of the learner maps can be formed as the group-diagnosis results. The group map (Fig. 4) presents the overview of learners’ understanding by visualizing the difference of line weight and tagged the number of learners according to the constructors of each proposition.

In the group-goal difference map, the concepts will be located as same as the concepts in the goal map and only relations of mismatch propositions are visualized. The group-difference map visualizes three types of error link as same as the individual-goal difference map. The excessive link is represented in the form of the solid line which the link is connected with two concepts. It can identify the relations that learners confused or misunderstood, and the tagged number presents the number of learners who constructed the link. The leaving link is also represented in the form of the solid line which the link is not connected with any concept, and indicates that the learners do not understand the linking word. The tagged number means the number of learners who do not use the link to connect with any concept. The dashed line represents the lacking link which is an error correction for displaying the correcting information of excessive- and leaving- links. The tagged number of lacking link is the total number of excessive link and leaving link, which related to the weight of line. The more tagged number in each relation will represent with a thicker line. For instance, an example of a group-goal difference map is shown in Fig. 5. “Melting (3)” dashed line is the lacking link while “Melting (2)” solid line is the excessive link, and “Melting (1)” solid line is the leaving link.

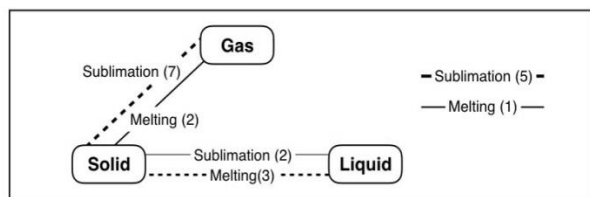


Fig. 5. An example of a group-goal difference map.

In addition, a filtering function of the Kit-Build analyzer can provide more efficient investigation by adjusting the intensity of three types of error link. The filtering function of group-diagnosis results is more explicit with the line weight, which a filtering tool can help the instructor filter out a few error links and keep the big number of error links. A thickness line and a number in parenthesis refer to the number of learners who connect those links. The link of each proposition is available for clicking to discover the learners who are the constructor of the link. Hence, the learner maps will be evaluated through the propositional level exact matching methodology that is the procedure for reporting individual-diagnosis results. The system can provide the additional procedure for reporting the group-diagnosis results at the same time.

Providing the components of the concept map is a kind of “closed-end” approach which is a realizing the automatic diagnosis of the concept map built by a learner [33]. The learner maps of the KB map are composed of the same components with the goal map. Thus, it is possible to detect the difference between them in the form of the diagnosis results. The learners are able to make a map in the limitation of providing parts, which is different from the traditional concept maps where learners can create concept map components by themselves. Therefore, the learners deal with only recall and understanding level in Bloom’s taxonomy [34]. The KB map can utilize in the aspect of confirming the understanding between the instructor and learners in classroom situations with the benefit of the automatic assessment for implementing formative assessment. In addition, the related studies presented the contribution of the KB map on learning effect [35]-[38]. The contribution of the KB map framework has been researched in reading comprehension topic where a direct interaction between the digital tool and learners has been examined. The results show that the KB map can help the learners to retain and recall the information for the longer period of time. The provided components illustrate the effective towards memory as same as the traditional concept map when the learning materials have the clear structure. The arrangement of the KB map on formative assessment also illustrates that an instructor used the suggestion of the diagnosis results for improving learning achievements [12]-[14].

For identifying the degree of learner’s understanding, the confidence identification of learners is the necessary task to indicate their confidence on each unit of meaning. The confidence tagging is utilized to facilitate the gathering of confidence information. The learners can indicate the degree of their understanding on the learner maps, and the diagnosis results also can inform the degree of the learners’ understanding to the instructor. Thus, the KB map with confidence tagging was developed to gather and assess the learning evidence for visualizing both learner’s understanding and the degree of learners’ understanding.

### III. METHODOLOGY

#### A. Kit-Build Concept Map with Confidence Tagging

For gathering learning evidence and identifying the degree of learner’s understanding, the KB map with confidence

tagging was developed for eliciting learning evidence, and associating the correctness and confidence information. In this study, the KB map is reinforced by uniting with the confidence tagging, which is a mechanism for representing learner's understanding on lecture content, and identifying learner's confidence on each proposition of a learner map. The confidence tagging is integrated into the structuring task where the learner constructs a learner map, and a tagging tool (Fig. 6) appears when two concepts and a linking word are connected as a completed proposition. Learners are required to identify their confidence by selecting "sure" or "not sure" on each completed proposition. It is also expected that the tagging task promotes learners to reconsider about their proposition again. The confidence values include "sure" for stating the certainty on the proposition, and "not sure" for indicating unconfidence on the proposition and the system allows the learners to change the values freely. If the learners disconnected the link of the completed proposition, the confidence tagging tool of the link would be disappeared, and the confidence value is reset then. The learners have to identify the confidence value again even they constructed the same proposition after disconnecting. Accordingly, the structuring task of learners can gather the answer of learners and confidence on their answer. Through this task, the system is able to gather the correctness and confidence information of each proposition in all learner maps, and then, the results of the diagnosis about the correctness and confidence are visualized at the same time.

Fig. 7 shows an example of individual-overlay map and Fig. 8 shows an example of a group-difference map, where the correctness and confidence information are reported to the instructor. An additional visualization is a confidence badge. The badge is added into the linking word to indicate the confidence of learners on the link. For instance, a dark tone badge on the dashed line illustrates the incorrect answer with confidence in the individual-overlay map (Fig. 7) of individual-diagnosis results, while a light tone badge on the solid line represents the correct answer with unconfidence.

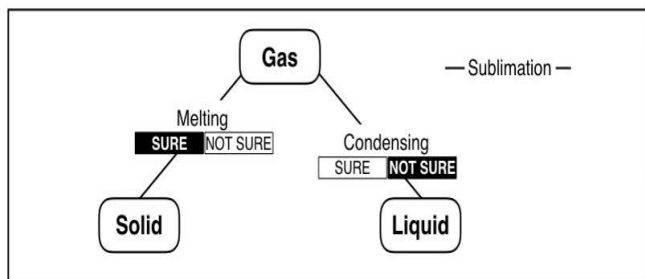


Fig. 6. An example of a learner map with confidence tagging.

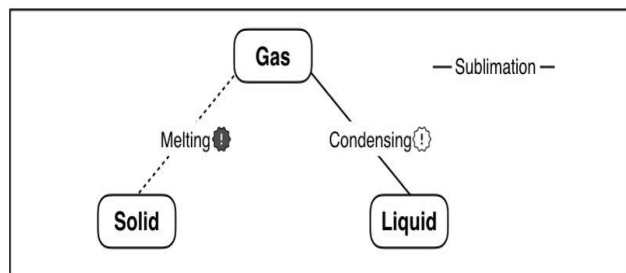


Fig. 7. An example of an individual-goal overlay map.

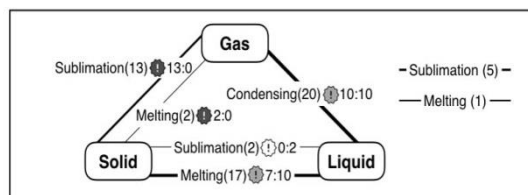


Fig. 8. An example of a group map with confidence information.

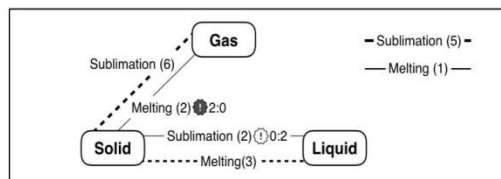


Fig. 9. An example of a group-goal difference map with confidence information.

On the other hand, the mismatch propositions are visualized in group-goal difference map (Fig. 9) of group-diagnosis results where the excessive link indicates the incorrect answer and the lacking link represents the correcting information. A dark tone badge on the solid line illustrates the excessive link with confidence, while a light tone badge on the solid line represents the excessive link with unconfidence. The group-diagnosis results has more details about the confidence information, which the color tone of the badge is varied according to the number of learners who have confidence against unconfidence on the same proposition. For instance, the darkest tone badge has appeared on the link that all of the constructors pressed on "sure" value. A middle tone badge has appeared on the link that the number of "sure" and "not sure" values are equal. The lightest tone badge appeared on the link that no one "sure" on the link. Another indicator is a tagged number of confidence information on the right-hand side of the badge. The colon is punctuation mark for separating the number of learners. The number of learners who pressed on "sure" is displayed on the left-hand side of the mark, while the right-hand side number displays the number of learners who press on "not sure." Fig. 9 shows an example of a group-difference map, where the correctness and confidence information are visualized.

### B. Practical Uses of Kit-Build Concept Map in Lecture Class

The practical uses of the KB map with confidence tagging are an implementation of formative assessment in lecture class for investigating the encouragement of the correctness and confidence information. The instructors can recognize the current learning situation for selecting and ordering the content of supplementary lecture through the analyzer of the KB map with confidence tagging. The participants are three instructors from three different schools, and learners from three different elementary schools who study in the fourth-, fifth-, and sixth-grade. The instructor of fourth grade conducted one practical use, the instructor of fifth grade conducted two practical uses, and the instructor of sixth grade also conducted two practical uses. Ten basic science classes of five paired class are separated into five control classes and five experiment classes. The arrangement of the KB map on formative assessment was used in the practical uses of this study following [14]: the first step is the general scenario of the lecture class, the instructors

created lecture contents and then constructed a goal map for indicating a learning goal of the class. The next step is to give the lecture to learners in a class period. During the lecture, the instructor checks the learner's understanding by requesting learners to construct learner maps and identify their confidence. Then, the diagnosis results are provided to the instructor immediately for informing about current understanding of learners. These steps are gathering and assessing the evidence of learners. The fifth step is to provide intra-class feedback during the class period, which requires an instant practical information for capturing an overall understanding of class. This requirement is responded by the group-diagnosis results that include the group map which can inform the common understanding, and the group-goal difference map which can inform the common misunderstanding of class in one map. Even the inter-class feedback of the sixth step was ignored in the practical uses of this study; we have an additional short discussion session with the instructors after finished classes for summarizing the classroom situation. Fig. 10 illustrates the arrangement of the KB map on formative assessment in a classroom situation.

The supplementary lecture is a feedback of the instructors in the lecture class, which a supplementary content should correspond with the misunderstanding of learners. Even the diagnosis results can identify the understanding and the misunderstanding of learners, the instructor still remains to be the most influential of the class who select the content of the supplementary lecture to raise the understanding of learners as a fulfilling the gaps. The valuable of correctness and confidence information investigation focusses on the behavior of instructors in selecting and ordering the supplementary lecture when the instructor received the different the diagnosis results. The correctness information is also available in the control group, while both the correctness and confidence information are available only in the experiment group. The excessive links of the group-goal difference map present the correctness information, indicate an overview of the incorrect answers, and represent the misunderstanding of learners. The number of excessive links was generally used to order the content of the supplementary lecture. The location of each excessive link was also used for ordering the excessive links that have an equal amount of the constructors (unordering of correctness information). Hence, an assumption of the control group is that the instructor selects the excessive links to provide the supplementary lecture following the correctness information and the location of visualization. The group-diagnosis results arrange the location of concepts and lacking links at the same location with the goal map's location. An alignment of each excessive link location is central between two connected concepts. The Z-pattern layout is the route of the instructor's eye traveling when they used the location for selecting the proposition in unordering of correctness information. The direction to select the content follows the shape of the letter Z as left to right, top to bottom of visualization screen. It can be used with a hierarchy of concept map that the components are ordered the most important from top to bottom. It can help the instructor to remember the selected- and unselected- excessive links even in the unstructured concept maps. We call this way to provide supplementary instruction as "basic strategy" in this paper.

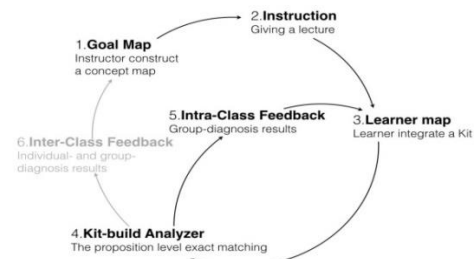


Fig. 10. The arrangement of the KB map on formative assessment.

On the other hand, because the correctness and confidence information are provided in the experiment group, it is assumed that the ordering of supplementary content is different from the ordering of the basic strategy. The difference between the basic strategy and the actual ordering in the practical uses in the experiment group demonstrate the effect of confidence information.

### C. Description of Procedure

The KB map with confidence tagging was utilized in ten science classes. All of the learners were requested to construct the learner map and tagging the confidence two times in each class. The first constructing was requested at the middle of class after the instructor lectured the content, and the second constructing was requested after the instructor gave the supplementary lecture at before the end of class. On the other hand, the different diagnosis results were provided to the instructors for investigating the behavior. A paired class consists of a control class where only the correctness information was visualized and an experiment class where the correctness and confidence information were visualized. Three instructors from three different elementary schools are the participants of the practical uses. An instructor A is the lecturer of fourth-grade who conducted one paired class. An instructor B is the lecturer of fifth-grade that conducted two paired classes, and an instructor C is the sixth-grade lecturer who conducted two paired classes. The instructor lectures the same content in both control- and experiment- classes of each paired class. Fig. 11 displays the practical flow of the paired class to distinguish the different diagnosis results between control- and experiment- group. The correctness information was visualized in both classrooms. The confidence information was blinded as the diagnosis results without confidence in the control classes, while the confidence information was visualized as the diagnosis results with confidence in the experiment classes.

Accordingly, there are no different activities in the learner role, while different information visualizing is the different factor of the instructor role. The different behavior of the same instructor should be observed in each paired class, which is the basic assumption to indicate the relation between the instructor's behavior and the confidence information. The same content of lecturing was conducted with the same instructor, but the supplementary lecturing may be different based on the provided information. The instructor will use the confidence information of learners when s/he accepted the information as the valuable information. In contrast, the behavior of the instructor in the experiment class has a possibility to behave as same as in the control class, even the confidence information was visualized.

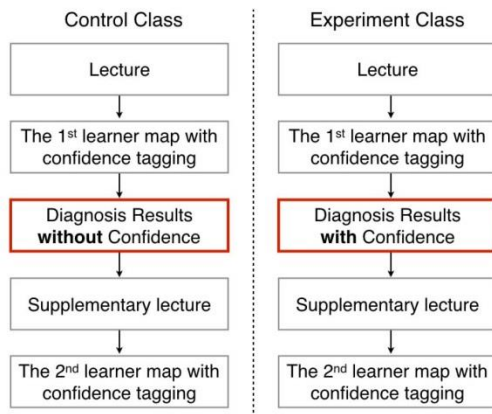


Fig. 11. The practical flows of each paired class.

The primary investigation is about how is the different behavior of the instructors when the system provided the confidence information of their learners. From the assumption, the instructor will use the confidence information for selecting and ordering supplementary content. The gathered evidence of the instructor's behavior consisted of the order of supplementary content in each class, the discussion session at the end of class, and an information evaluation session of the instructor's questionnaire. "What is an effect of the different behavior of the instructor?" is analyzed to be three values which contain a normalized learning gain, a discrimination value, and a hit rate. The normalized learning gain of each group was referred to describe the effectiveness of the different behavior of the instructor. The discrimination value illustrates the recognition of the different understanding based on correctness and confidence information. The discrimination value presents how learners have the confidence on the correct proposition and have no confidence on the incorrect proposition. The hit rate focuses only on the correct proposition that learners have confidence. Lastly, the questionnaire was conducted to assess the satisfaction of the KB map with confidence tagging in the aspect of both the learners and the instructors when it was utilized in the classroom situation.

#### IV. RESULTS

##### A. Different Behavior of the Same Instructor

The investigation of the control group is a comparison of excessive links ordering between basic strategy and the actual ordering of each control class, which the assumption is a perfect similarity between the basic strategy and the actual ordering of the class. Fig. 12 shows the goal map of the first paired class. Fig. 13 shows a part of diagnosis results of the control class in the first paired class where the instructor used the filtering function to screen out some excessive links that have the number of the constructor less than three. An observed evidence is the ordering of supplementary content based on the diagnosis results of the class. The first selected excessive link was "composed of 25%," and supplementary content mentioned to "Water" and "Air" which the action indicates that the most number of excessive links was selected for providing the feedback. The second selected excessive link was "composed of 45%". These selected excessive links can be ordered by using the correctness information, while the

remaining excessive links have the same tagged number as in ordering of correctness information. The supplementary lecture mentioned to "Water" again with the explanation of "composed of 5%" and the content of "Organic." Thus, the third selected excessive link was "composed of 5%" on the left-hand side. Then the "composed of 5%" was mentioned with the content of "Organic" again with "Inorganic" content. Hence, the fourth selected excessive link was "composed of 5%" on the right-hand side. The third- and fourth- selected excessive links demonstrate that the location visualization can help the instructor to select the excessive links in unordering of correctness information. Accordingly, the actual ordering of the instructor is the same ordering of basic strategy. The similarity value between basic strategy and actual ordering of the class is 100%. The perfect similarity value illustrates that the instructor used the correctness information and location visualization for ordering feedback, and there are no other factors in this ordering process.

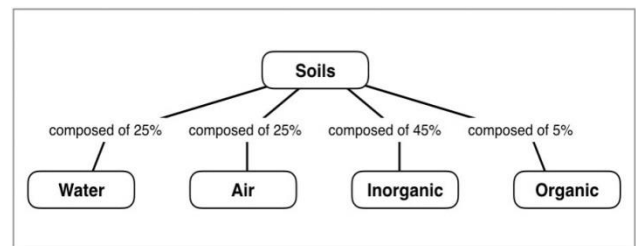


Fig. 12. The goal map of the first paired class.

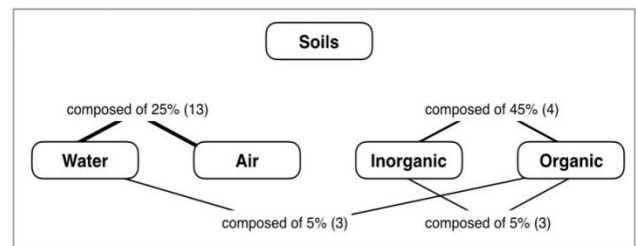


Fig. 13. The group-goal difference map of the control class in the first paired class.

Table I displays the similarity values between the basic strategy and actual ordering of five paired classes. In the control group, all of five control classes can get the perfect similarity value that represents that the instructors used the basic strategy for ordering the supplementary content where the correctness information was provided.

TABLE I. THE PERCENTAGE OF SIMILARITY BETWEEN BASIC STRATEGY AND ACTUAL ORDERING OF FIVE PAIRED CLASSES

Lecturer	Grade of learners	Paired class	Percentage of similarity	
			Control class	Experiment class
Instructor A	4	1 <sup>st</sup> paired	100.00	100.00
Instructor B	5	2 <sup>nd</sup> paired	100.00	60.00
	5	3 <sup>rd</sup> paired	100.00	14.29
Instructor C	6	4 <sup>th</sup> paired	100.00	100.00
	6	5 <sup>th</sup> paired	100.00	16.67



On the other hand, the different order of supplementary content was found in the experiment group where the system provided the correctness and confidence information to the instructor. Imperfect similarity values were found in three of five experiment classes, which indicate the different behavior of the instructors in selecting and ordering the supplementary content.

### B. How the Instructors used the Information of the Diagnosis Results

The different behavior of the same instructor was found when the system provided the different information, and the confidence information has the possibility to encourage the different behavior of the instructor. This section summarizes how the instructors used the diagnosis results from the short discussion sessions with the instructors after finished classes and the evaluation session from the questionnaire of the instructors. The summary mentions to the importance of each information in the diagnosis results, which consist of correctness, confidence information, and location visualization. The instructors commented that the correctness is only one learning evidence in the control group and they focused on the correctness information from the diagnosis results firstly, while the location visualization can help them to point out selected- and remain- excessive links. On the other hand, two learning evidences are provided in the experiment group. The correctness information is still the most important information, and confidence information becomes valuable information as the second priority, then the last priority is visualization location. The result of questionnaire also presents the order of information, which the instructors tried to pay attention to the incorrect proposition first and then looked for its tagged number of confidence information. The incorrect with confidence is the most crucial type of proposition that the all of the instructors want to provide the feedback for this proposition type before the others. Besides, even the strategy of ordering between the control- and experiment- group is different because the different behavior of the instructor on the different diagnosis results, the ordering of the first- and fourth- classes of both groups are the same order with basic strategy as shown in Table I.

Fig. 14 shows an example of the group-goal difference map layout that visualize the group-goal difference map in blinded concept label and linking words for investigating the ordering of the experiment group where the system provides both correctness and confidence information to the instructor. The correctness information is visualized in the form of the number of excessive links for indicating the misunderstanding of learners. The most number of the excessive link is displayed as “Link O (7)” for informing seven learners who connected “Concept A” and “Concept C” with the “Link O.” Thus, the first selected excessive link was selected by using only the correctness information. However, only the correctness information cannot suggest the next selected excessive link because there are six candidates that are possible to be the second selected excessive link. The confidence information is visualized for informing how many learners have the confidence and unconfidence on each excessive link. The tagged number of confidence information on six candidates

suggests that three of three confidences on two excessive links, and two of three confidences on four remaining excessive links. Subsequently, the supplementary lecture mentions to “Link N” with the error explanation, which is according to the “Concept A” and the “Concept D,” and then still keep an attention on the “Link N” again but the error explanation is according to the “Concept C” and “Concept D”. The order of supplementary content demonstrates that the confidence information was used for selecting these selected excessive links. The second selected excessive link is the upper “Link N (3) 3:0”, and the third selected excessive link is the lower “Link N (3) 3:0”. Hence, the order also demonstrates the location visualization was used for ordering when the correctness and confidence information have an equal amount.

Table II displays the used information of ordering process which can represent the amount of time that the instructor used each information. The instructor tended to incorporate the confidence information with the correctness information and location visualization. Thus, we define “CCL” strategy as the ordering supplementary content based on correctness, confidence information, and location visualization respectively. Moreover, there is the possibility, that the instructor used different strategy but both strategies can produce the same order of supplementary content. For instance, the ordering of selected excessive links in the first experiment class was ordered by using five times of correctness and two times of confidence based on CCL strategy. The same ordering can be produced from the basic strategy.

### C. Normalized Learning Gain and Effect Size

The same instructor and the same lecture content are lecturing in each paired class, while the different feedbacks produced the different intervention between the control- and experiment classes. The investigation of normalized learning gains and effect size are presented in this section, and an assumption is the different behavior based on different used strategy affects learning achievements. That means the confidence information effects to the behavior of the instructor, and then the different feedback also effects to the understanding of learners. The normalized learning gain ( $g$ ) is used to represent the effectiveness of the educational intervention [39]. The first learner map was constructed after the instructor gave the lecture (Formative map) and the second learner map was constructed after the instructor gave the supplementary lecture (Final map), which correspond to the arrangement of the KB map on the formative assessment. The learner map scores and the normalized learning gain of each learner can be calculated following:

$$\text{Map score} = \frac{\text{Correct propositions in learner map}}{\text{The number of propositions in the goal map}} \quad (1)$$

$$g = \frac{\text{Final map score} - \text{Formative map score}}{1 - \text{Formative map score}} \quad (2)$$

Correspondingly, the gain of averages ( $\langle g \rangle$ ) was used to indicate the normalized learning gain of class that can be classified into three regions of  $g$  for substantial using following “Low” when ( $\langle g \rangle$ ) less than 0.3, “Medium” when ( $\langle g \rangle$ ) from 0.3 to 0.7, and “High” when ( $\langle g \rangle$ ) more than 0.7 [39].

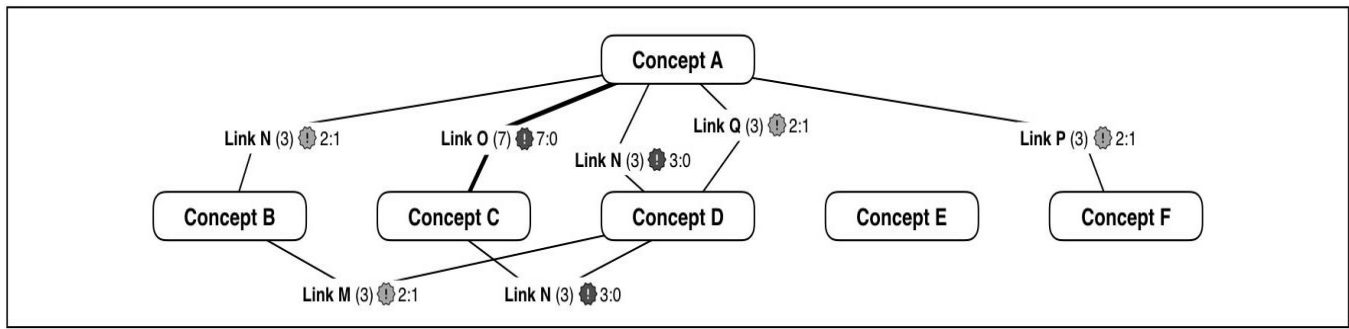


Fig. 14. The layout of the group-goal difference map of the experiment class in the third paired class.

TABLE II. THE INFORMATION USED OF ORDERING IN THE EXPERIMENT GROUP BASED ON CCL STRATEGY

Classroom	Selected excessive links	The number of used time information			Percentage of similarity*
		Correctness	Confidence	Location	
1 <sup>st</sup> experiment class	5	5	2	0	100.00
2 <sup>nd</sup> experiment class	5	5	4	0	60.00
3 <sup>rd</sup> experiment class	7	7	6	6	14.29
4 <sup>th</sup> experiment class	5	5	3	2	100.00
5 <sup>th</sup> experiment class	6	6	6	5	16.67

\*The similarity values of selected excessive links ordering between the basic strategy and CCL strategy

Table III presents the gain of averages and its region of each class. Four experiment classes out of five got better the normalized learning gains than their paired control classes. Especially in the fourth- and fifth- paired classes, there were significant differences in normalized learning gains between experiment class and control class.

Moreover, regarding effect size (Cohen’s *d*) as difference of normalized learning gains between control class and experiment class, they are “large” in the 3<sup>rd</sup> and 5<sup>th</sup> paired classes and they are “medium” in the 4<sup>th</sup> one. There results suggest that the experiment classes were better for learning than control classes.

#### D. Discrimination of the Understanding

The discrimination value ( $d_r$ ) represents the recognition of the difference between what they know and what they do not know [1]. The value is measured based on a proportion of the confident correct proportion and the unconfident incorrect proposition against all of the complete propositions in the learner map. The perfect score indicates the learners are able to discriminate according to an appropriate confidence, which implies the learner has confidence on all of the correct understanding and has no confidence on the misunderstanding.

$$d_r = \frac{\text{Correct with confidence} + \text{Incorrect with unconfidence}}{\text{The number of complete propositions in the learner map}} \quad (3)$$

TABLE III. NORMALIZED LEARNING GAIN OF CLASS AND EFFECT SIZE OF EACH PAIRED CLASS

Paired class	Type of class	Number of learners	$\langle g \rangle$	S.D.	Region of $g$	$d$	$p$ -value <sup>a</sup>
1 <sup>st</sup> paired class	Control class	34	0.57	0.48	Medium	0.23	0.5570
	Experiment class	36	0.67	0.38	Medium		
2 <sup>nd</sup> paired class	Control class	24	0.85	0.46	High	0.13 <sup>b</sup>	0.2660
	Experiment class	26	0.79	0.43	High		
3 <sup>rd</sup> paired class	Control class	25	0.50	0.53	Medium	0.83	0.3019
	Experiment class	25	0.93	0.51	High		
4 <sup>th</sup> paired class	Control class	16	0.29	0.23	Low	0.56	0.0389 <sup>c</sup>
	Experiment class	20	0.47	0.41	Medium		
5 <sup>th</sup> paired class	Control class	17	0.18	0.33	Low	1.49	0.0003 <sup>c</sup>
	Experiment class	20	0.71	0.38	High		

<sup>a</sup>. The  $p$ -value of  $g$  between control- and experiment- class of each paired class.

<sup>b</sup>. The value presents  $|d|$  when the control class has the  $\langle g \rangle$  more than the experiment class, which produces a negative value of  $d$ .

<sup>c</sup>. Statically significant difference

TABLE IV. AN AVERAGE OF THE DISCRIMINATION VALUE

Group (N=10)	Formative map	Final map	p-value
Control group (5 classes)	0.6007	0.7624	$p < 0.01$
Experiment group (5 classes)	0.6820	0.8842	$p < 0.01$
p-value	0.0794	$p < 0.01$	

Table IV shows the discrimination value of learners and the significant difference between the control group and the experiment group. There was no significant difference between the formative map of the control- and experiment- group ( $p=0.794$ ), which means that the learners have an ability to discriminate about their knowledge not much different after lecturing. The feedback of instructors improved discrimination of learners in both groups significantly ( $p<0.01$ ). Then, there was a significant difference of final map between the control- and experiment- group ( $p<0.01$ ). These results suggest that the correctness and confidence based feedback can improve the discrimination of their confidence on their understanding better than the correctness based feedback.

E. Certainty of the Understanding

The confidence on the incorrect proposition is the worst situation that the instructors attempt to correct those misunderstanding by providing the supplementary lecture based the diagnosis results. On the other hand, the confidence on the correct proposition is the best situation for representing the certainty of the understanding. The hit rate (HR) represents consistency with the interpretation that if a correct response is covertly selected, then its execution helps the learner to confirm its correctness [1]. The value is measured based on a proportion of the number of confident correct propositions against the number of correct propositions in the learner map.

$$HR = \frac{\text{Correct proposition with confidence}}{\text{The number of correct propositions in the learner map}} \quad (4)$$

Table V shows the hit rate and the significant difference between two learner maps of two groups. There was no significant different between control- and experiment- group ( $p=0.1976$ ) that means learners have not much different confidence on the correct answers after lecturing. Then the feedback of instructors can improve confidence on the correct answers in both groups significantly ( $p<0.01$ ). There was also a significant difference of final map between the control- and experiment- group ( $p<0.05$ ), which suggests that the correctness and confidence based feedback can improve the certainty of the understanding better than the correctness based feedback.

TABLE V. AN AVERAGE OF THE HIT RATE

Group (N=10)	Formative map	Final map	p-value
Control group (5 classes)	0.7430	0.8888	$p < 0.01$
Experiment group (5 classes)	0.6714	0.9587	$p < 0.01$
p-value	0.1976	$p < 0.05$	

F. Changing of Proposition based on the Confidence

For more emphasis on the confidence of learners, Table VI shows a possibility of proposition changing based on the confidence information from the formative map to the final map. The analysis of change of proposition type presents that the propositions with unconfidence are easier to change than the propositions with confidence. Particularly, the changing of unconfidence propositions to confident correct propositions of experiment group is 80.30%, while 69.60% unconfidence propositions of the control group are changed to confident correct propositions. The proposition changing suggests that the correctness and confidence based feedback can help the learners to improve their understanding and get more confidence better than the correctness based feedback.

TABLE VI. A PROPOSITION CHANGING BASED ON THE CONFIDENCE OF LEARNERS FROM THE FORMATIVE MAP TO THE FINAL MAP

Group (N=10)	Percentage of proposition changing	
	Confidence	Unconfidence
Control group (5 classes)	33.07%	66.97%
Experiment group (5 classes)	33.08%	85.40%

G. Results of the Questionnaire

The questionnaire was conducted to the learners who participated in the practical uses, which content of the questionnaire contains three sessions following the overview of the KB map with confidence tagging, emphasizing on the effect of confidence tagging, and the effect of instructor’s feedback. Fig. 15 displays a part of the questionnaire of learners. The positive evaluations received from the learners by the questionnaire. Such as the first questions, 60.70% of learners “strongly agree” enjoy constructing the learner map and tagging of the confidence. 51.26% “strongly agree” and 29.14% “agree” are the results of the second question about constructing the map and tagging confidence are useful for expressing the understanding of lecture content. The confidence tagging as an additional task did not disturb the learners in the structuring task, which 34.67% and 31.66% “strongly agree” and “agree” on they feel free to tagging their confidence respectively as the results of the fourth question. Finally, the results of seventh- and eighth- questions have more than fifty percent on “strongly agree” that the instructor’s feedback in the form of the supplementary lecture can help learners to get more understanding and get more confidence. The results of learner’s questionnaire illustrate the satisfaction of learners that suggests that the learners accepted the mechanism of the KB map with confidence tagging.

The questionnaire of the instructor was also conducted for investigating the aspect of the instructors when the KB map was utilized in their lecture classes. Fig. 16 displays a part of the instructor’s questionnaire. The results of the questionnaire demonstrate the positive satisfaction of the instructors.

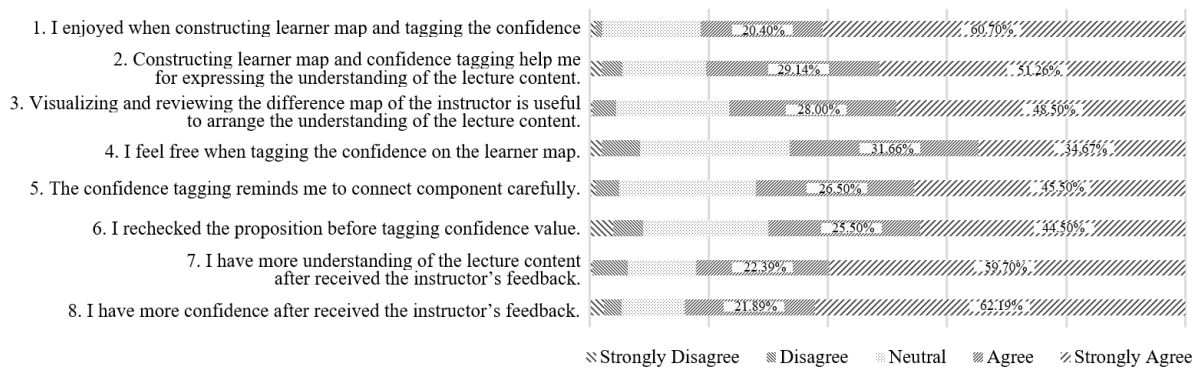


Fig. 15. A part of learner's questionnaire and its results.

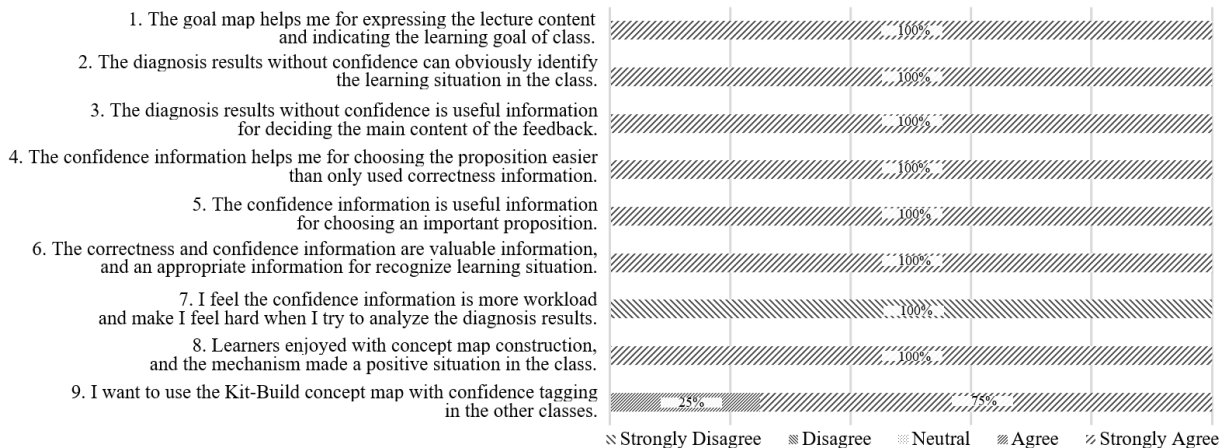


Fig. 16. A part of instructor's questionnaire and its results.

The goal map creating can help the instructors to express the lecture content, and indicate the learning goal as the result of the first question. The results from the second- to sixth-questions present that all instructors gave “strongly agree” to the diagnosis results, which are useful information for visualizing the current learning situation, identifying the critical misunderstanding of learners, until selecting and ordering the supplementary content. Moreover, the instructors also strongly agreed on the eighth question that their learners enjoyed with the mechanism which formed the positive environment for the learning situation. On the other hand, the instructor gave “strongly disagree” on the seventh question that the confidence information was more workload when analyzing the diagnosis results. Thus, the instructor accepted the diagnosis results that include the correctness and confidence information. Notably, the diagnosis results with the confidence information are useful information for selecting and ordering the supplementary feedback, which is more satisfactory than no confidence information.

## V. DISCUSSION

In this study, we present the encouragement of correctness and confidence information with the KB map with confidence tagging for selecting and ordering the supplementary content as the feedback of the instructors in the lecture classes. The KB map creates an opportunity for an instructor to assess a current learning situation, which the instructor can give the feedback to learners for improving learning achievements in the class

period. The different behavior of the instructors was observed when the system provided only the correctness information in the control group, while the correctness and confidence information were provided in the experiment group. The ordering of the supplementary content demonstrates how the instructor used the correctness, confidence information, and location visualization.

The observed evidence of the practical uses can represent the relation between the instructor's behavior and the confidence information of learners. The instructors did not only use the confidence information in selecting and ordering the supplementary content, but we also found the mentioning to the confidence of learners on some selected excessive links in the supplementary lecture of the experiment group when the instructor received the confidence information. Correspondingly, the relation of instructor's behavior and learning evidence suggests that the different behavior of the instructors is positive changing to improve the learning achievements and also improve the confidence of learners. The normalized learning gain of class ( $\langle g \rangle$ ) and effect size (Cohen's  $d$ ) illustrate that the correctness and confidence based feedback of the experiment group is more effective than the only correctness based feedback of the control group. The discrimination value ( $d_r$ ) demonstrates that the learners of experiment group can discriminate the different understanding based on correctness and confidence better than the learners of control group significantly. Similarly, the hit rate ( $HR$ ) shows that the learners of experiment group have an ability to

REFERENCES

represents consistency with the interpretation better than the learners of control group significantly. These results of the practical uses suggest that the confidence information of learners affects the instructor's behavior and then the different behavior of the instructor effects to the learning achievements continuously. In addition, the results of questionnaire present the positive satisfaction of both instructors and learners when the KB map with confidence tagging was utilized in the lecture classes. The learners accepted the mechanism for representing their understanding and their confidence. The instructors accepted that the confidence information of learners is valuable information for recognizing the learning situation. Nevertheless, the content details of the supplementary lecture were not investigated in this experiment such as what kind of feedback was designed from only correctness, or correctness and confidence information.

VI. CONCLUSION

Even the correctness assessment can determine the knowledge of learners, the quality of that knowledge cannot be identified by using only the correctness information. We propose the KB map with confidence tagging that can provide the mechanism to learners for representing their understanding and identifying their confidence on their understanding. The learner map and confidence of each proposition are the learning evidence, which the learner map can represent the understanding of learners in the lecture content and the confidence tagging promotes them to reconsider their propositions again. The system facilitates learners to create learning evidence in a class period and identify the current learning situation through diagnosis results immediately. Subsequently, the learning evidence of learners affects the instructor behavior directly when they accepted the information as a valuable information. The supplementary lecture based on the correctness and confidence information is utilized as evidence-based feedback of the instructor, which is a key of formative assessment to improve learning achievements in the classroom situations.

Moreover, the different behavior of the same instructor illustrates the utilizing of the confidence information on the supplementary lecture that can demonstrate that the instructor accepted the confidence information as the valuable information. The confidence information can encourage the strategy for selecting and ordering the supplementary content. The results of the practical uses suggest that the different feedback of the instructor is important through normalized learning gains and effect size, which the correctness and confidence based feedback can improve the learning achievements and confidence of learners concurrently.

For the future work, the individual feedback will be focused based on the current ability of the KB map with confidence tagging. Even the instructor can improve the learners understanding, some propositions are disregarded such as the correct proposition with unconfidence. Consequently, we aim to direct to all learners and support all their propositions via the KB map with confidence tagging for improving the learning achievements in the form of system feedback.

- [1] D. P. Hunt, "The concept of knowledge and how to measure it," *Journal of intellectual capital*, vol. 4, no. 1, pp. 100-113, 2003.
- [2] W. Bruine de Bruin, A. M. Parker, and B. Fischhoff, "Individual differences in adult decision-making competence," *Journal of personality and social psychology*, vol. 92, no. 5, pp. 938-956, 2007.
- [3] A. Efklides, and A. Tsiora, "Metacognitive experiences, self-concept, and self-regulation," *Psychologia*, vol 45, no. 5, pp. 222-236, 2002.
- [4] A. Efklides, "Metacognition and affect: What can metacognitive experiences tell us about the learning process?," *Educational research review*, vol. 1, no. 1, pp. 3-14, 2006.
- [5] S. Kleitman, and T. Moscrop, "Self-confidence and academic achievements in primary-school children: Their relationships and links to parental bonds, intelligence, age, and gender," In *Trends and prospects in metacognition research*, Springer US, pp. 293-326, 2010.
- [6] S. Kleitman, L. Stankov, C. M. Allwood, S. Young, and K. K. L. Mak, "Metacognitive self-confidence in schoolaged children," In *Self-directed learning oriented assessments in the Asia-Pacific*, Springer Netherlands, pp. 139-153, 2012.
- [7] L. Stankov, and J. Lee, "Confidence and cognitive test performance," *Journal of Educational Psychology*, vol. 100, pp. 961-976, 2008.
- [8] L. Stankov, J. Lee, and I. Paek, "Realism of confidence judgments," *European Journal of Psychological Assessment*, vol. 25, no. 2, pp. 123-130, 2009.
- [9] G. Heron, and J. Lerpiniere, "Re-engineering the multiple choice question exam for social work," *European Journal of Social Work*, vol. 16, no. 4, pp. 521-535, 2013.
- [10] S. M. Cisar, P. Cisar, and R. Pinter, "True/false questions analysis using computerized certainty based marking tests," In *Intelligent Systems and Informatics, 7th International Symposium on*, pp. 171-174, 2009.
- [11] T. Hirashima, K. Yamasaki, H. Fukuda, and H. Funaoi, "Framework of kit-build concept map for automatic diagnosis and its preliminary use," *Research and Practice in Technology Enhanced Learning*, vol. 10, no. 17, pp. 1-21, 2015.
- [12] K. Yoshida, K. Sugihara, Y. Nino, M. Shida, and T. Hirashima, "Practical Use of Kit-Build Concept Map System for Formative Assessment of Learners' Comprehension in a Lecture," *Proc. of ICCE2013*, pp. 892-901, 2013.
- [13] K. Yoshida, T. Osada, K. Sugihara, Y. Nino, M. Shida, and T. Hirashima, "Instantaneous Assessment of Learners' Comprehension for Lecture by Using Kit-Build Concept Map System," In *International Conference on Human Interface and the Management of Information*, Springer, Berlin, Heidelberg, pp. 175-181, 2013.
- [14] J. Pailai, W. Wunnasri, K. Yoshida, Y. Hayashi, and T. Hirashima, "The practical use of Kit-Build concept map on formative assessment," *Research and Practice in Technology Enhanced Learning*, vol. 12, no. 20, pp. 1-23, 2017.
- [15] J. E. Bruno, "Using testing to provide feedback to support instruction: A reexamination of the role of assessment in educational organizations," In *Item banking: Interactive testing and self-assessment*, Springer, Berlin, Heidelberg, pp. 190-209, 1993.
- [16] A. R. Gardner-Medwin, and M. Gahan, "Formative and summative confidence-based assessment," *Proc. of the 7th CAA Conference*, Loughborough: Loughborough University, pp. 147-155, 2003.
- [17] T. Gardner-Medwin, and N. Curtin, "Certainty-based marking (CBM) for reflective learning and proper knowledge assessment," In *REAP Int. Online Conf. on Assessment Design for Learner Responsibility*, 2007.
- [18] G. Yuen-Reed, and K. B. Reed, "Engineering Student Self-Assessment through Confidence-Based Scoring," *Advances in Engineering Education*, vol. 4, no. 4, pp. 1-23, 2015.
- [19] A. R. Gardner-Medwin, "Optimisation of certainty-based assessment scores," 2013, *Proc. of The Physiological Society, The Physiological Society*.
- [20] C. M. Moss, and S. M. Brookhart, *Advancing formative assessment in every classroom: A guide for instructional leaders*, ASCD, 2010.
- [21] D. Wiliam, C. Lee, C. Harrison, and P. Black, "Teachers developing assessment for learning: Impact on student achievement," *Assessment in Education: Principles, Policy & Practice*, vol. 11, no. 1, pp. 49-65, 2004.

- [22] R. Ballantyne, K. Hughes, and A. Mylonas, "Developing procedures for implementing peer assessment in large classes using an action research process," *Assessment & Evaluation in Higher Education*, vol. 27, no. 5, pp. 427-441, 2002.
- [23] J. D. Novak, and A. J. Cañas, "The theory underlying concept maps and how to construct and use them", Technical Report IHMC CmapTools, 2008.
- [24] M. Buldu, and N. Buldu, "Concept mapping as a formative assessment in college classrooms: Measuring usefulness and student satisfaction," *Procedia-Social and Behavioral Sciences*, vol. 2, no. 2, pp. 2099-2104, 2010.
- [25] D. L. Trumpower, and G. S. Sarwar, "Formative structural assessment: Using concept maps as assessment for learning," *Proc. of the Fourth International Conference on Concept Mapping*, vol. 22, pp 132-136, 2010.
- [26] J. Schacter, H. E. Herl, G. K. W. K. Chung, R. A. Dennis, and H. F. O'Neil, "Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving," *Computers in Human Behavior*, vol. 15, no. 3, pp. 403-418, 1999.
- [27] I. L. G. Hsieh, and H. F. O'Neil, "Types of feedback in a computer-based collaborative problem-solving group task," *Computers in Human Behavior*, vol. 18, no. 6, pp. 699-715, 2002.
- [28] C. C. Chiou, "The effect of concept mapping on students' learning achievements and interests," *Innovations in Education and Teaching International*, vol. 45, no. 4, pp. 375-387, 2008.
- [29] P. Chularut, and T. K. DeBacker, "The influence of concept mapping on achievement, self-regulation, and self-efficacy in students of English as a second language," *Contemporary Educational Psychology*, vol. 29, no. 3, pp. 248-263, 2004.
- [30] N. Aghakhani, H. S. Nia, S. Eghtedar, and C. Torabizadeh, "The Effect of Concept Mapping on the Learning Levels of Students in Taking the Course of," *Jundishapur Journal of Chronic Disease Care*, vol. 4, no. 2, pp. 1-5, 2015.
- [31] J. Pailai, W. Wunnasri, Y. Hayashi, and T. Hirashima, "Ongoing Formative Assessment with Concept Map in Proposition Level Exact Matching," *Proc. of ICCE2016*, pp. 79-81, 2016.
- [32] J. Pailai, W. Wunnasri, Y. Hayashi, and T. Hirashima, "Automatic Concept Map Assessment in Formative Assessment Approach," *Engineering Research Council and the 76th Advanced Learning Science Society for Artificial Intelligence*, vol. 5, no. 3, pp. 48-53, 2016.
- [33] E. M. Taricani, and R. B. Clariana, "A technique for automatically scoring open-ended concept maps," *Educational Technology Research and Development*, vol. 54, no. 1, pp. 65-82, 2006.
- [34] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, "Taxonomy of educational objectives, handbook I: The cognitive domain," 1956.
- [35] M. Alkhateeb, Y. Hayashi, T. Rajab, and T. Hirashima, "Comparison between kit-build and scratch-build concept mapping methods in supporting EFL reading comprehension," *The Journal of Information and Systems in Education*, vol. 14, no. 1, pp. 13-27, 2015.
- [36] M. Alkhateeb, Y. Hayashi, T. Rajab, and T. Hirashima, "Experimental Evaluation of the KB-mapping Method to Avoid Sentence-by-Sentence Map-building Style in EFL Reading with Concept Mapping," *The Journal of Information and Systems in Education*, vol. 15, no. 1, pp. 1-14, 2016.
- [37] M. Alkhateeb, Y. Hayashi, T. Rajab, and T. Hirashima, "Experimental Use of Kit-Build Concept Map System to Support Reading Comprehension of EFL in Comparing with Selective Underlining Strategy," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 4, pp. 80-87, 2016.
- [38] H. Funaoi, K. Ishida, and T. Hirashima, "Comparison of kit-build and scratch-build concept mapping methods on memory retention," *Proc. of ICCE2011*, pp. 539-546, 2011.
- [39] R. R. Hake, "Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses," *American journal of Physics*, vol. 66, no. 1, pp. 64-74, 1998.

# Method for Detection of Foreign Matters Contained in Dried Nori (Seaweed) based on Optical Property

## Transparent Foreign Matter Detection by using Bidirectional Reflectance and Polarization Characteristics

Kohei Arai

Science and Engineering Faculty, Department of Information Science  
Saga University  
Saga City, Japan

**Abstract**—Optical property, such as spectral reflectance, bidirectional reflectance distribution functions and polarization characteristics of dried seaweeds is clarified together with an attempt for transparent foreign matter detection considering optical property of dried seaweeds, such as spectral transparency, bidirectional reflectance and transparency as well as polarimetric properties. Through experiments, it is found that transparent foreign matter can be detected by using bidirectional reflectance distribution function as well as polarization characteristics.

**Keywords**—Seaweeds; optical characteristics; BRDF; polarization characteristics; foreign matter detection

### I. INTRODUCTION

Nori (*Porphyra Yezoensis*<sup>1</sup>) of seaweed is the famous product for Japanese traditional food in particular. Seaweed rolls are getting popular in the world. One of the problems in Seaweed production is foreign matter detection in particular, transparent foreign matters. It is relatively easy to detect non-transparent foreign matters such as fishing hook, gravel and so on. By using the difference between transparent of the seaweed products, foreign matters can be detected. Most of the foreign matter detection machines use such difference of the transparency. If the transparency of seaweed is smaller than the previously determined threshold, such seaweed has to be removed from the seaweed product for sale. It, however, is comparatively difficult to detect transparent foreign matters such as a piece of tegus (fishing line fragment).

The optical properties of dried seaweed are not necessarily clarified, and it is not clear whether it is useful for detecting foreign matter mixed with any optical characteristics. In addition, it is difficult for transparent foreign matters such as tegus to detect contaminants contaminated into dried seaweed by optical characteristics.

Conventional methods for foreign matter detection use just reflectance of the dried nori in concerns at a certain wavelength. It, however, is difficult to detect transparent foreign matter, in particular, such as tegus. In this paper, the

spectral reflectance and transparency characteristics, bidirectional reflectance and transparency characteristics and polarization characteristics of dried seaweed is clarified. Then, a detection method of the inclusion of transparent foreign matters is proposed based on these characteristics.

The following section describes fundamental optical characteristics of dried seaweeds. Then the proposed method for foreign matter detection by using such optical properties is described followed by the experimental configuration and experimental results. Ultimately, conclusion is described with some discussions.

### II. OPTICAL CHARACTERISTICS OF A TYPICAL DRIED SEAWEED

#### A. Test Piece of the Typical Dried Seaweed and the Measuring Instrument used as well as Fundamental Optical Characteristics of the Test Piece Measured in the Laboratory

Spectral reflection characteristics and its angle dependency (bidirectional reflectance characteristics: BRDF (Bi-Directional Reflectance Distribution Function)) [1]-[4] as well as, s and p polarized radiance of front and back sides of representative dried seaweed is measured using spectral radiometer MS-720 (spectral sensitivity ranges from 350 nm to 1050 nm). The major specification of MS-720 is shown in Table I while the outlook of the MS-720 is shown in Fig. 1.



Fig. 1. Outlook of the MS-720.

<sup>1</sup> <http://en.wikipedia.org/wiki/Porphyra>

TABLE I. MAJOR SPECIFICATION OF MS-720 OF SPECTRORADIOMETER

Wavelength range	350~1,050nm
Wavelength Interval	3.3nm
Wavelength Resolution	10nm
Wavelength Accuracy	<0.3nm
Aperture Angle(Full)	180°
Straylight	<0.15%
Temperature Dependency	±5%
Output Unit	W/m <sup>2</sup> /μm
Measuring Interval	0.005~5sec(Auto)
Size	W100×D165×H60(mm)
Weight	720g
Interface	RS-232C/USB



Fig. 2. The typical dried seaweed used.

The reflectance and linear polarization degree of the component are measured. A polarizer is attached to the front aperture of the optical system of the spectroradiometer so that the deflection angle can be changed.

Also, incidence angle can be changed. The typical dried seaweed used for experiments is shown in Fig. 2.

It shows the front side of the dried seaweed. Therefore, it is a little bit glossy partially. The reflectance and the degree of polarization is measured at the non-glossy portion of the dried seaweed. The degree of linear polarization DP [5]-[7] can be defined by (1).

$$DP = (R_{max} - R_{min}) / (R_{max} + R_{min}) \quad (1)$$

Where,  $R_{max}$  and  $R_{min}$  denote maximum and minimum radiance, respectively.

### B. Measured Spectral Reflectance, BRDF, Degree of Polarization

The measured spectral reflectance is shown in Fig. 3. In the figure, the observation angle is represented by the zenith angle. Also, Transparency,  $R_t$ ,  $R_h$ ,  $R_s$ ,  $R_p$ ,  $R_{15}$ ,  $R_{30}$ ,  $R_{45}$ ,  $R_{60}$ , DP denotes Transparency, Reflectance of the tail (back) seaweed surface, Reflectance of the head (front) seaweed surface, Reflectance for s polarization, Reflectance for p polarization, Reflectance at zenith angle of 15, 30, 45, and 60 degrees, and Degree of Polarization: DP defined as (1), respectively.

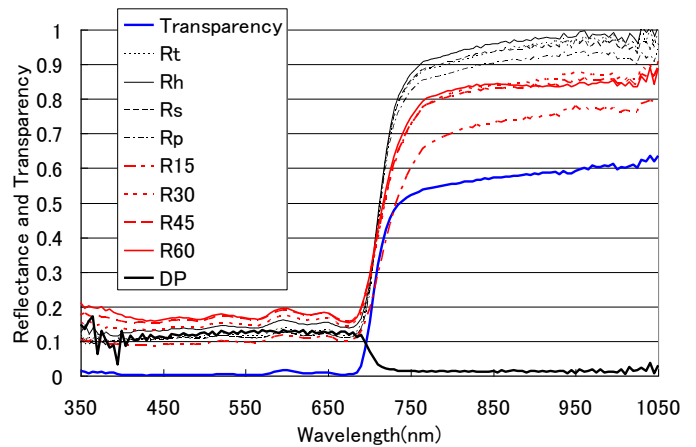


Fig. 3. Spectral reflectance as a function of observation angle and transparency as well as s and p polarized reflectance together with degree of polarization of the typical dried seaweed.

Depending on the moisture content of dried seaweed, in particular, the reflectance in the absorption region of water at 940 nm is largely different. It is also possible to estimate the moisture content using this characteristic.

The reflectance in the wavelength range from the ultraviolet to the red edge is as low as 0.2 or less, but extremely high in the near infrared wavelength range. The transparency is also extremely low in the wavelength range from the ultraviolet to the red edge, but it is understood that it is translucent in the near infrared wavelength range. Therefore, it is possible to utilize the difference in transparency as well as the reflectance of dried seaweed with no foreign matter to detect transparent contaminants. In addition, the difference in reflectance on the back and front side of dried seaweed shows that the reflectance of the glossy surface nori surface is somewhat higher than that of the large surface roughness back-seaweed surface.

BRDF of dried seaweed can be calculated with the measured reflectance at the different observation angles. Fig. 4 shows the calculated BRDF. BRDF may change when the target dried seaweed contains foreign matter. Therefore, there is a possibility of foreign matter detection based on BRDF measurements.

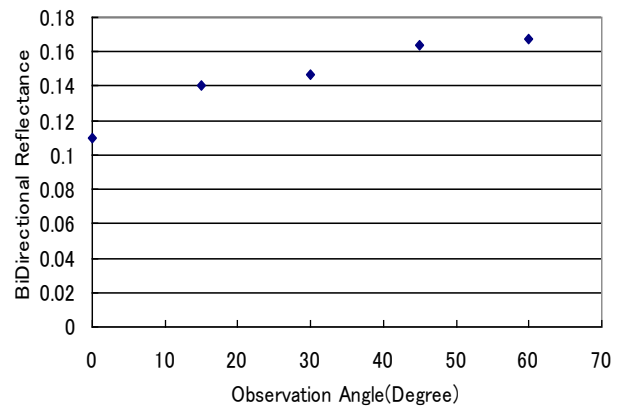


Fig. 4. BRDF of the typical dried seaweed.



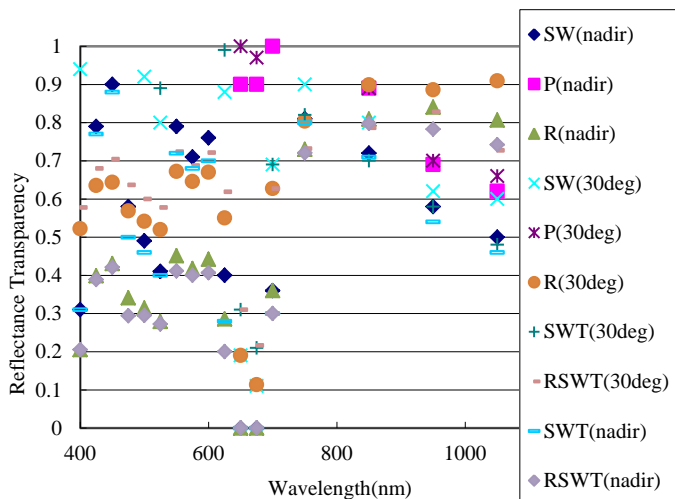


Fig. 5. Reflectance and transparency of dried seaweed at the different viewing angles, nadir and 30 degrees.

Spectral reflectance of the dried seaweed at the different viewing angles, nadir and 30 degrees is shown in Fig. 5. In this case, CCD cameras are situated in the nadir and 30 degree of nadir angle. Therefore, reflectance in the directions of nadir and 30 degree can be measured.

In the figure, SW, P, R denotes the incident and reflected radiance, respectively. Also, The number in the bracket denotes viewing angle, nadir and 30 degrees. For instance, RSWT(nadir) denotes the reflectance of the dried seaweed containing tegus in the nadir direction while RSW denote the reflectance of the dried seaweed without tegus in the nadir direction.

Also, the difference between “delta(nadir): R and RSWT” as well as “delta(30deg): SWT and R” can be evaluated together with “the effect” of the difference between the differences. Also, the difference between “delta'(30deg): RSWT(30deg) and R(30deg)” is shown in Fig. 6. Therefore, the effect of BRDF for foreign matter detection can be assessed.

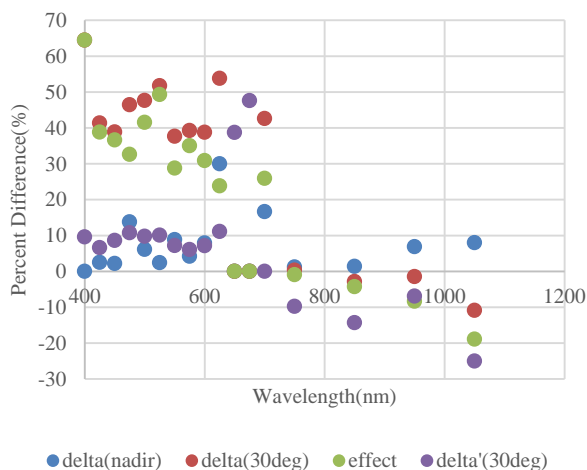


Fig. 6. Effect of BRDF for foreign matter detection.

Also, spectral reflectance of the typical dried seaweed is measured as a parameter of polarization angles ranged from -220 to 100 degrees with 20 degrees of step. Fig.7 shows the result. Then, DP is calculated with the measured reflectance. The DP is also shown in the same figure. From Fig.7, it can be seen that dry seaweed has DP less than 0.1 from ultraviolet to red edge, but has almost no polarization characteristics in the near infrared wavelength range. Furthermore, when the observation angle is varied, it can be confirmed that the reflectance at the observation zenithal angle of 0 degree, that is, the reflectance in the direct downward direction is the lowest, and the reflectance becomes higher as the languidness increases.

When the reflectance at 500 nm is expressed as a function of the observation zenith angle, it is as shown in Fig.8, and it is understood that the BRDF characteristic of the dry seaweed surface is not Lambertia<sup>2</sup>, but follows the Minnerart Reflection Model<sup>3</sup> [1]. The reflectance varied periodically with the 180 degree of polarization angle.

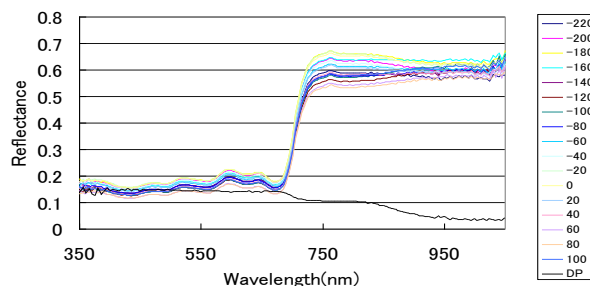


Fig. 7. Spectral reflectance as a function of polarization angle and degree of polarization of the dried seaweed.

These are fundamental optical properties of the dried seaweed. Foreign matter contained in the dried seaweed is intended to detect based on the optical properties. Namely, the BRDF of the seaweed containing foreign matter is different from the BRDF without foreign matter. Also, the DP with foreign matter differ from the DP without foreign matter. Transparency difference between seaweed with and without foreign matter can be used for foreign matter detection. This is the fundamental idea of the proposed method for foreign matter detection.

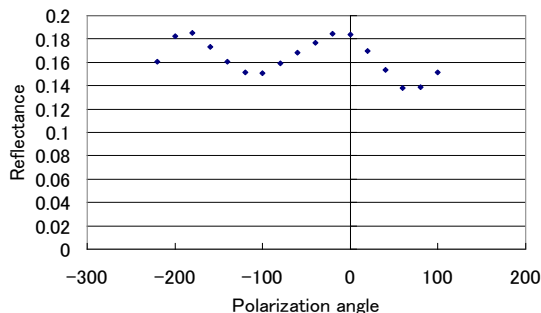


Fig. 8. Reflectance at 500nm as a function of polarization angle.

<sup>2</sup> [https://en.wikipedia.org/wiki/Lambert%27s\\_cosine\\_law](https://en.wikipedia.org/wiki/Lambert%27s_cosine_law)

<sup>3</sup> [http://thesai.org/Downloads/IJARAI/Volume2No9/Paper\\_4-Bi-Directional\\_Reflectance\\_Distribution\\_Function.pdf](http://thesai.org/Downloads/IJARAI/Volume2No9/Paper_4-Bi-Directional_Reflectance_Distribution_Function.pdf)

From these measurement results and estimation results:

- 1) Reflectance and transmittance in the near-infrared wavelength range are particularly effective for detecting foreign matter mixed in dry seaweed.
- 2) The bidirectional reflectance characteristic is useful for detecting transparent foreign matters such as tegus.
- 3) Linear polarization degree is also effective for detection of transparent foreign matters.

### III. PRACTICAL EXPERIMENTS

#### A. Measurement Configuration

Foreign matter containing dried seaweed is placed at 40 degrees for the light source and 70 degrees for the surface reflection observation camera with respect to the dried seaweed surface. The waveforms of the two front reflection cameras are compared. Also, the bidirectional reflectance is measured by similarly arranging the table reflection observation camera at 100 degrees. The measurement configuration is shown in Fig. 9.

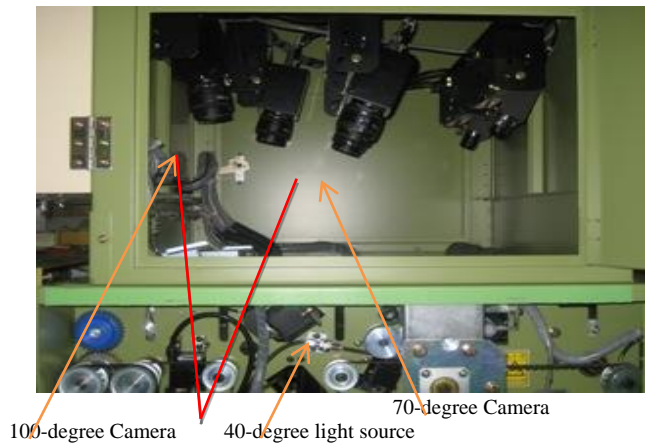


Fig. 9. Measurement configuration (Side view).

The table reflector lights the inverter with a fluorescent lamp of 20W (3 wavelength daylight color). A polarizing film is attached to the glass surface. As shown in the photograph of Fig. 10, when the polarization film is shifted by 90 degrees on the light source, light is not transmitted.

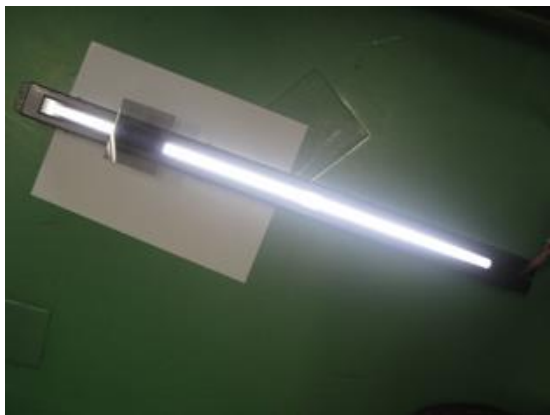


Fig. 10. Light source used (20W three wavelength combined quasi natural light source with polarization film).



Fig. 11. CCD camera used.

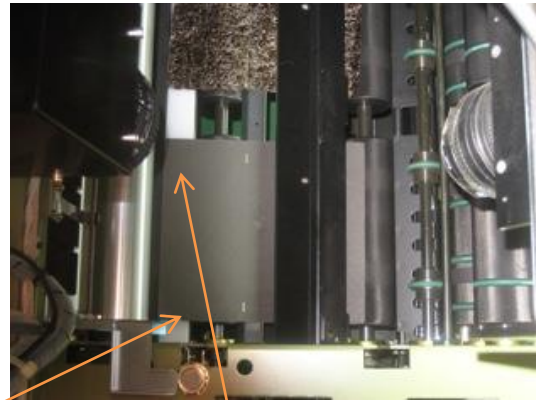


Fig. 12. Measurement configuration (Top view).

For the table reflection camera, the CCD line camera (black and white) in Fig. 11 is used. A polarizing film is attached to the CCD portion to be orthogonal to the fluorescent lamp light source.

A top view of the measurement configuration is shown in Fig. 12. In the figure, standard white paper placed at the reference position for optical axis adjustment is situated. Also, on the top of it is placed the test piece of dried seaweed containing transparent foreign matter. When detecting the contamination of actual dried seaweed contamination, the dried seaweed is moved at a high speed in this.

To match the optical axis of the 70 and 100-degree camera, as shown in Fig. 13, the optical axis adjustment paper is affixed to the reference position on the dried seaweed surface. While monitoring the camera output, the optical axis is adjusted so that the output positions when both cameras saw the same optical axis adjustment paper are matched. Fig. 14 shows the waveform of the camera output in optical axis adjustment.

We set the reference sensitivity reference color as the reference to Japan Painting Association No. C22-40D (Munsell number 2.5Y4 / 2) shown in Fig. 15, put the reference color on the optical axis of the table reflection camera, are adjusted to have the same sensitivity (waveform voltage). At that time, the sensitivity is adjusted to be 33.

Milky white tegus with thickness of about 0.2 mm and length of 35 mm are used for testing. Fig. 16 shows this Tegus. The "U"-shaped tegus is placed on the dried seaweed in the state of 0.2 mm in thickness and 35 mm in length on the optical axis. The camera output waveform is shown in Fig. 17.

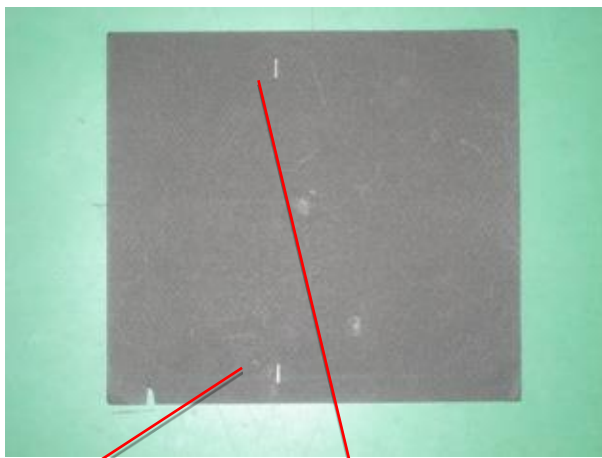
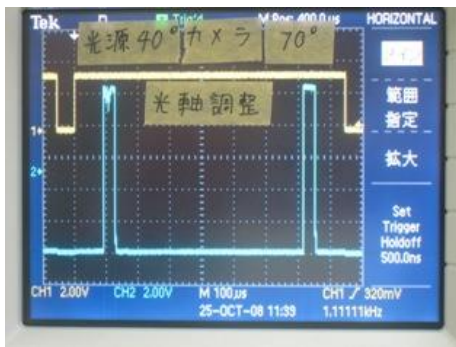
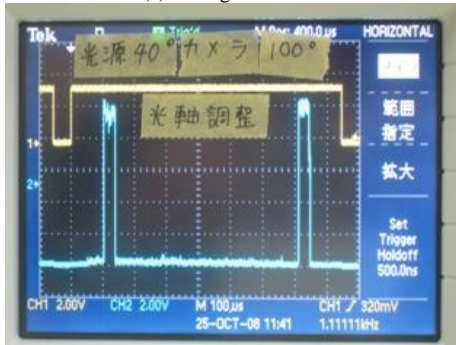


Fig. 13. Alignment adjust for both of cameras with two white alignment adjust papers.



(a) 70-degree camera



(b) 100-degree camera

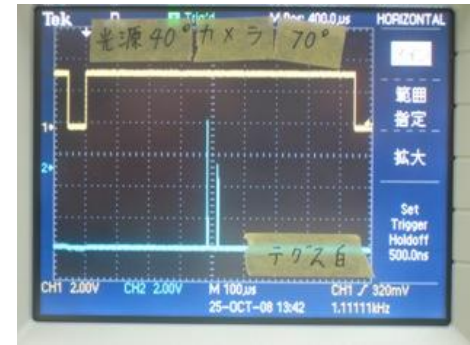
Fig. 14. Camera output when alignment adjustment for both 70 and 100-degree cameras.



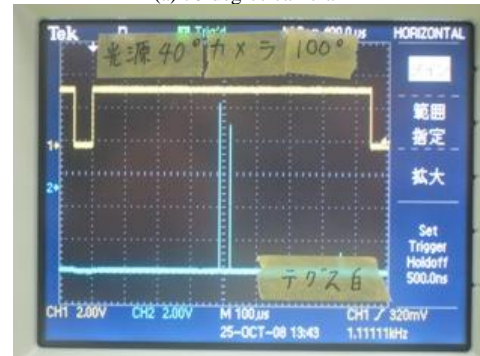
Fig. 15. Standard color of C22-40D.



Fig. 16. Transparent fish line used (in the red circle).



(a) 70-degree camera



(b) 100-degree camera

Fig. 17. Camera output from the transparent fish line on the head surface of the dried seaweed.

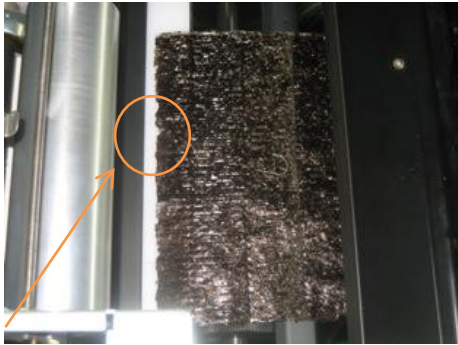
Reflected light from the tegus differs by 1.25 times between the 70-degree camera and the 100-degree camera because the BRDF changes (sensitivity is higher because the output of the 100-degree camera is closer to specular reflection), it is confirmed that foreign matter detection is possible.

Table II also shows the outputs of the 70-degree camera and the 100-degree camera and their ratios when the position of the tegus is changed. From this table, it is found that foreign matter detection is possible regardless of the position of the tegus.

The surface roughness of the backside of dried seaweed is relatively high. Therefore, the foreign matter detection capability is confirmed when foreign matter such as Tegus is mixed in or sticking to it. Fig. 18 shows the dried seaweed that Tegus sticks to the reverse side.

TABLE II. CAMERA OUTPUT DIFFERENCE BETWEEN 70 AND 100 DEGREE CAMERAS WHEN BOTH CAMERAS OBSERVE THE SAME FISH LINE (LOCATIONS ARE DIFFERENT)

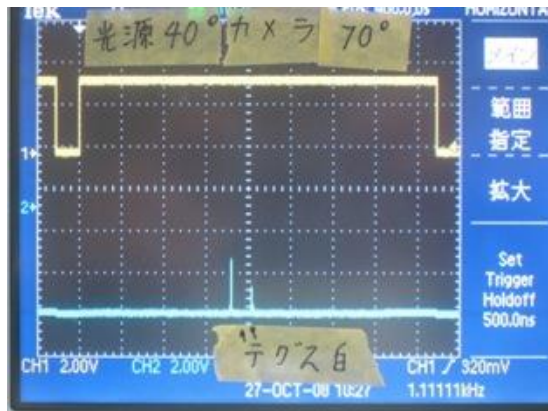
	Sensitivity at the various positions					Mean
	1	2	3	4	5	
70 Deg.	34	20	36	25	34	29.8
100 Deg.	144	25	134	53	114	94
100 Deg./70 Deg.	4.2	1.3	3.7	2.1	3.4	3.2



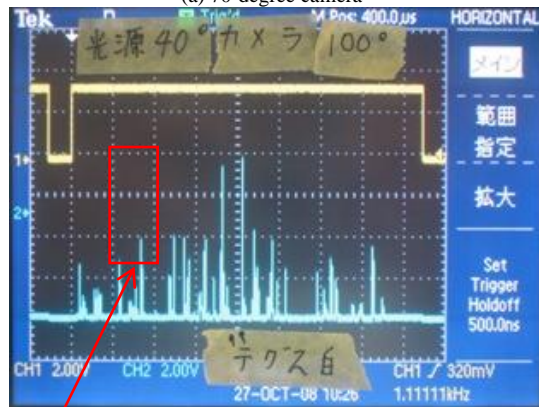
Transparent fish line is hard to see

Fig. 18. Dried seaweed with transparent fish line attached to the tale surface.

Fig. 19 shows the camera output from the transparent fish line on the tale surface (backside) of the dried seaweed. It is clearly seen that the Tegus containing dried seaweed can be detected.



(a) 70-degree camera



Transparent fish line

(b) 100-degree camera

Fig. 19. Camera output from the transparent fish line on the tale surface of the dried seaweed.

In the case of the camera angle of 70 degrees, the waveforms of the tegus are clear compared to other parts. There is no influence on the output waveform of backside roughness of dried seaweed. In the case of the camera angle of 100 degrees, the mountain portion of the roughness of the backside of dry seaweed appears in the corrugation. Therefore, it is difficult to discriminate between the tegus and the dried seaweed. Despite sticking the polarizing film at 90 degrees to the light source part and the camera part to suppress the reflection of the dried seaweed. Therefore, it is not satisfactory to suppress the influence of the roughness of the backside of the dried seaweed.

#### IV. CONCLUSION

As a method for detecting a contaminant accompanied with undulations contained in a sheet-like foreign matter of dried seaweed, a method of judging by using reflectance in two directions (bidirectional reflectance) is proposed. Light is irradiated from the light source to the sheet-like foreign matter of dried seaweed, and the reflected light is received from at least two directions. The moisture content is investigated by using a visible near infrared and short wavelength infrared CCD array for this light receiving element. As a result, foreign matters with undulations contained in the liquid foreign matter of seaweed can also be detected, and the moisture content is also measured.

Provided is a foreign matter detection device capable of accurately detecting a foreign matter which is difficult to detect even when it is mixed in a measurement object of dried seaweed. The reflectance of the object to be measured is obtained by using the oblique direction light receiving element of CCD camera, whereas in the case of the direct light receiving element of CCD camera, the reflectance with substantially no difference is obtained, whereas in the case where a foreign matter which is difficult to detect is mixed. In addition, the reflectance is obtained using the upward direction light receiving element of CCD camera with respect to the portion where the reflectance is obtained by the oblique direction light receiving element of CCD camera, and the reflectance is obtained (oblique direction reflectance/right upward direction (Reflectivity in oblique direction/reflectance in the upward direction) compared with the case in which a foreign matter is not mixed (determination of reflectance) to judge the presence or absence of contamination of foreign matters, and it is simply compared with the case of comparing the reflectance of the measurement object. Therefore, it is confirmed that foreign matters can be accurately detected.

According to the experimental results using the difference in reflectivity between directly nadir and 30 degrees, it is found that the difference in the presence/absence of tegus due to the conventional method (direct reflectance) is insignificant. Enlarged, detection capability of Tegus is improved with the proposed method by about one digit.

Experimental results, however, show that the surface reflection camera installed at 100 degrees can detect tegus when the condition of dried seaweed surface is good, but it is difficult to detect the tegus when affected by dry roughness backside roughness. Further investigation is required for overcoming this matter.

#### ACKNOWLEDGMENT

The author would like to thank Dr. Osamu Yamaguchi and Dr. Masanori Tsurumaru of Nishihatsu Co. Ltd. for their effort to conduct the experiments.

#### REFERENCES

- [1] Kohei Arai, Method for estimation of grow index of tealeaves based on Bi-Directional reflectance function: BRDF measurements with ground based network cameras, International Journal of Applied Science, 2, 2, 52-62, 2011.
- [2] Kohei Arai, Comparison between linear and nonlinear models of mixed pixels in remote sensing satellite images based on Cierniewski surface BRDF model by means of Monte Carlo ray tracing simulation, International Journal of Advanced Research in Artificial Intelligence, 2, 4, 1-7, 2013.
- [3] Kohei Arai, Bi-directional reflectance distribution function: BRDF effect on un-mixing, category decomposition of the mixed pixel (MIXEL) of remote sensing satellite imagery data, International Journal of Advanced Research in Artificial Intelligence, 2, 9, 19-23, 2013.
- [4] Kohei Arai and Long Lili, BRDF model for new tealeaves on old tealeaves and new tealeaves monitoring through B RDF measurement with web cameras, Abstract of the 50th COSPAR(Committee on Space Research/ICSU) Congress, A3.1-0008-08 ,992, Montreal, Canada, July, 2008
- [5] Kohei Arai, Osamu Yamaguchi, Masanori Tsurumaru, Optical property of dried sea weeds, Nori and detection of transparent suspected objects containing in Nori, Research Report of Science and Engineering Faculty, Saga Unievrsity Japan, 37, 2, 1-6, 2008
- [6] Kohei Arai, Hajime Koshiishi, Fukazawa, Saino, Hajime Fukushima, Ishimaru, Okuda, Hiroshi Kawamura, Total System for Ocean Observation Using Remote Sensing Satellites, Journal of Japanese Society of Ocean, 1, 51-62, 1989.
- [7] K.Arai, Monte Carlo ray tracing based sensitivity analysis of the atmospheric and oceanic parameters on the top of the atmosphere radiance, International Journal of Advanced Computer Science and Applications, 3, 12, 7-13, 2012.

#### AUTHORS PROFILE

**Kohei Arai.** He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post-Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a counselor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission-A of ICSU/COSPAR since 2008. He received Science and Engineering Award of the year 2014 from the minister of the ministry of Science Education of Japan and also received the Best Paper Award of the year 2012 of IJACSA from Science and Information Organization: SAI. In 2016, he also received Vikram Sarabhai Medal of ICSU/COSPAR and also received 37 awards. He wrote 37 books and published 570 journal papers as well as 370 conference papers. He is Editor-in-Chief of International Journal of Advanced Computer Science and Applications as well as International Journal of Intelligent Systems and Applications. <http://teagis.ip.is.saga-u.ac.jp/>

# On P300 Detection using Scalar Products

Monica Fira

Institute of Computer Science  
Romanian Academy  
Iasi, Romania

Liviu Goras

Institute of Computer Science,  
Romanian Academy  
“Gheorghe Asachi” Technical  
University of Iasi, Romania

Anca Lazar

University of Medicine and  
Pharmacy “Grigore T. Popa” of Iasi,  
Romania

**Abstract**—Results concerning detection of the P300 wave in EEG segments using scalar products with signals of various shapes are presented and their advantages and limitations are discussed. From the point of view of the computational complexity, the proposed algorithm is a simple algorithm, based on a scalar product and searching for the max value of 6 calculated values. Because we considered that the human subject is not a robot that precisely generates P300 and that there is also a human component of error in the involuntary generation of such waves, we have also calculated the rate of classification of character in the human visual field. To validate the proposed method, electroencephalography recordings from the competition for Spelling BCI Competition III Challenge 2005 - Dataset II have been used.

**Keywords**—Electroencephalographic (EEG); brain computer interface; P300; spelling paradigm; classification; signal processing

## I. INTRODUCTION

The use of electroencephalographic (EEG) signals as a vector of communication between man and machine continues to be a challenge in the “brain-computer interface” (BCI) paradigm aimed at achieving proper interpretation of brain specific electrical activity signals and parameters [1], [2].

A definition of a BCI, subsequently accepted by most researchers in the field, is a communication system in which the messages or the commands sent to the exterior environment by an individual do not pass through the normal output paths of the brain, paths constituted by the peripheral nerves and muscles [3].

The purpose of BCI is to establish a communication system that translates human intentions, represented by appropriate signals - into control signals for an output device, such as a computer or neuroprosthesis.

Since BCI can potentially provide a link between the brain and the physical world without any physical contact [1] its main objectives are to process the EEG waveforms and to generate the necessary signals to control some external systems. The most important application is to stimulate paralyzed organs or bypass disabled parts of the human body. BCI systems may appear as the unique communication mode for people with severe neuromuscular disorders such as spinal cord injury, amyotrophic lateral sclerosis, stroke and cerebral palsy [4], [5].

The correspondence between EEG patterns and computer actions constitutes a machine learning problem since the

computer should learn how to recognize a given EEG pattern. As for other learning problems, in order to solve this problem, a training phase is necessary, in which the subject is asked to perform prescribed mental activities and a computer algorithm is in charge of extracting the associated EEG patterns. After the training phase is finished the subject should be able to control the computer actions with his or her thoughts. This is the major goal for a BCI system.

In this work we investigate the possibility of detecting P300 waves using scalar products between EEG segments and various waveforms with shapes similar to P300 ones on data from the BCI III competition.

## II. SPELLING PARADIGM AND THE DATA SET: DESCRIPTION

The P300 speller paradigm makes use of waves that are expressions of event related potential produced during decision making process.

One of the first examples for BCI is the algorithm proposed by Farwell and Donchin [6] that relies on the unconscious decision making processes expressed via P300 in order to drive a computer.

According to the P300 speller paradigm described in [1] the subject should watch a 6x6 matrix containing all letters and digits (as shown in Fig. 1) and should focus the attention on a character in the matrix. The protocol contains two steps:

Step 1: The matrix is presented to the subject for 2.5 seconds.

Step 2: All lines and all columns are randomly highlighted each for 100 ms with a pause of 75 ms.



Fig. 1. Classical P300 spelling paradigm described by Farwell-Donchin 1988 [1].

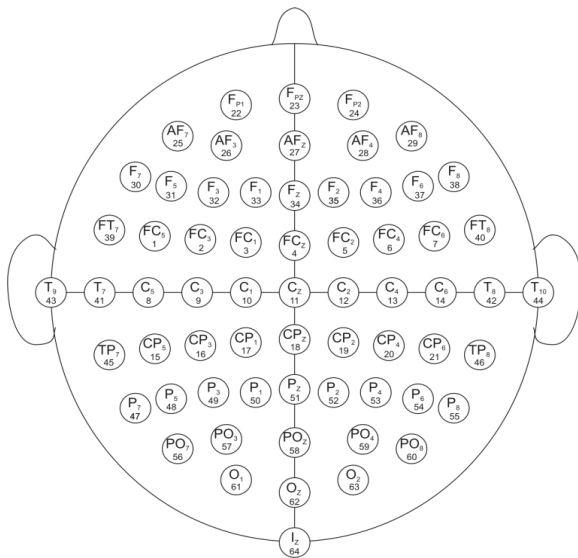


Fig. 2. Placement of the 64 electrodes.

The procedure consists in repeating 15 times Step 2 (15 epochs) for each character, followed by a pause of 2.5 seconds (Step 1). For a given character, there will be  $6 \times 2 \times 15 = 180$  intensifications:  $2 \times 15$  will contain the target character (once when the column is highlighted, second for the line it belongs to, repeated for 15 epochs) and the all the others *are supposed* not to contain it.

The data for the BCI III competition has been recorded from two different subjects in five sessions. The raw signals have been processed with a band-pass filter in the range 0.1 – 60Hz and sampled at 240Hz. During each session, the subjects were asked to spell words and for each word a run was defined. All EEG signals of a 64- scalp electrodes placement (Fig. 2) have been continuously collected during an acquisition session. The data set is split in the train set (containing 85 characters) and the test set (containing 100 characters) for each subject. An extended description of the dataset is available in the BCI competition paper [1].

The winners of the competition proposed a method based on an ensemble of classifiers [7] of linear Support Vector Machine type. They were trained on a reduced part of the available data by performing a channel selection operation. The reported classification rate was 95.5% for 15 sequences and 73.5% for 5 sequences [7].

### III. METHOD

The aim of the investigations presented in what follows is to check the advantages and limitation of finding a *prototype* P300 waveform and to detect the presence of a P300 wave in an EEG segment by using the well-known scalar product. The main idea is that if an EEG segment presents a P300 waveform, it can be considered as a component orthogonal to the subspace of all the others in the segment. Thus, taking the scalar product of an EEG segment with a P300 prototype, we should obtain a high/small value if the segment exhibits/does not exhibit P300. The difficulty of the method consists in the fact that the shape of the P300 components is not the same for all segments.

Thus, even though we did not expect better classification results than those reported in [7] this simple approach offers interesting insights.

Basically the aspects we considered were:

- a) Train and test signals pre-processing.
- b) Generating various “prototype” shapes for the P300 waveform.
- c) Using the above waves to compute scalar products for one or several channels for the test case.
- d) Analyse and compare results.

The pre-processing stage consists of EEG signals segmentation into slices of 1s length beginning at the start of the flashing intervals (100ms at 75ms distance) and averaging of the EEG segments corresponding to a flash i.e., to each line or column that has been intensified 15 times. More precisely, at this stage the pre-processing is based on the average of the selected channels for the 15 repetitions in order to calculate the P300 pattern from all 85 epochs and for calculation of 12 average signals for each epoch (corresponding to the 6 lines and 6 columns). Moreover, averaging has been used not only for one channel but also for several channels as it will be shown later.

After averaging, for each line/column, a pattern of the EEG results which either is known to exhibit or not P300 for the train signals or should be classified as exhibiting or not P300 for the test signals.

Thus, for *classification* on the testing set (epoch) we constructed 12 test signals for each character (corresponding to the 6 lines and 6 columns) by averaging all 15 repetitions. In other words, by averaging all 15 intensifications corresponding to each row or column we obtained a pattern characteristic to the respective line or column. The obtained patterns have been smoothed with a 10 samples window median filter.

In the next steps various shapes for approximating the P300 pattern from the information in the test waveforms have been generated. The aim was to find which shapes give the best classification results using the scalar product with testing signals and to make a comparative analysis not only regarding the best shape (“prototype”) but also choosing the most significant channels for classification.

After the scalar product between the averaged flashing pattern and the P300 “prototype” pattern is computed, the lines and columns containing P300 are determined according to the highest values for the scalar product.

Another aspect that has been investigated is related to the hypothesis that it is not unlikely for subjects to produce P300 signals when adjacent lines to the character the patient was thinking at. Thus we have computed the so-called *Partial Classification* rates representing the classification rate of the lines and of the columns containing P300. Indeed, it is possible to correctly determine the line containing P300 corresponding to a given character while the column is wrong such that the final classification of the character is wrong. In other words, we did not make the intersection of the line and column containing the P300, but separately the classification rate for lines and for columns.

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	-

Fig. 3. Panel for the classical P300 spelling paradigm, a desired character (colored in dark grey) and the characters from visual field (colored in light grey).

We have also evaluated the co-called *Visual\_Field Classification* which has been introduced based on the hypothesis that a subject may also generate P300 at highlights (illuminations) of adjacent lines or columns to the chosen (target) character. Indeed it is not unlikely that there is an error due to the human visual field and attention of the subject when adjacent lines/columns are illuminated. Indeed, since the human subject is not a robot or a machine that responds precisely, each subject has a degree of error in generating P300 depending on the state of fatigue, stress or attention, visual field, etc.

In order to better understand the *Visual\_Field Classification* concept, we have presented in Fig. 3 the character matrix, a desired character and the characters considered in the visual field which are taken into account.

This is why in addition to the final classification rate, there were calculated the *Partial and Visual\_Field Classification* rates.

#### IV. EXPERIMENTAL RESULTS

For the evaluation of the analyzed methods we used the dataset II of the BCI Competition III 2005 -P300 Spelling. The data for the BCI III competition has been recorded from two different subjects but in this paper we used only teh recordings from one subject because we followed the presentations of the method and the presentation of the results from both subjects would make it difficult to follow the ideas.

According to the above discussion we remind the reader that, for the scalar product method, the shape of the P300 “prototype” pattern is the one that mainly influences the final classification results. Thus, several such “prototype” patterns have been generated and tested. To make a global image the types of patterns that have been used are listed below.

- Average P300 Pattern - calculated by averaging all EEG segments containing P300
- Five P300 patterns calculated based on the K-Means algorithm applied to all segments containing P300
- Stylized pattern by summing up three Gaussians
- Stylized pattern by 1 Gaussian
- Stylized P300 Pattern by decimating and then interpolating the average P300 pattern defined first in this list.

In what follows we will discuss the results obtained with the above “prototype” patterns.

#### A. Average P300 Pattern

The first pattern has been obtained by averaging all segments known/supposed to exhibit P300 in the training set.

For the selected channels, there were extracted and normalized all data samples between 0 and 1 sec next to the beginning of a flash. From a training set of 85 spelling characters, for each analysed channel, there were taken  $12 \cdot 15 \cdot 85 = 15300$  signals of 240 samples. These 15 300 signals are split into “P300 signals” and “non-P300 signals”, resulting two subsets, the subset of 2550 signal with P300 and the subset of 12750 signals without P300. From **all** subsets “P300 signals”, by averaging, the P300 pattern shown in Fig. 4 was achieved. Such a pattern can be considered a first prototype for the shape of the P300.

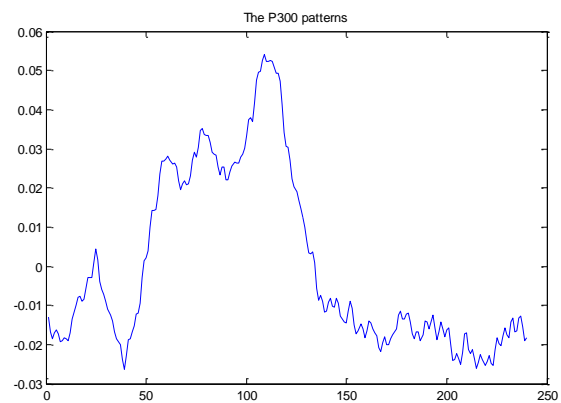


Fig. 4. A P300 pattern obtained by averaging.

At a first thought we might hope that the scalar product of the above wave would make a significant difference between P300 and non P300 segments. The reasons this did not happen are at least two: a. the pattern in Fig. 4 is an average and b. each waveform exhibiting P300 has its own shape which will not necessary give a high scalar product with the above pattern. To keep things as simple as possible, we selected a list of most important channels by making the scalar product with segments extracted from each one [8], [9]. More precisely we wanted to find which channels are the best in terms of the classification rate. At this stage we used one single channel and the proposed algorithm. In Table I the classification rate for several channels are presented.

TABLE I. CLASSIFICATION RATES FOR 18 CHANNELS

Channel	Classification %	Channel	Classification %
F3 - 32	30	Cz - 11	49
F1 - 33	45	C2 - 12	37
F2 - 35	42	CP3 - 16	15
FC3 - 2	40	CP1 - 17	36
FC1 - 3	51	CPz - 18	43
FCz - 4	33	Pz - 51	45
FC2 - 5	50	P1 - 50	28
C3 - 9	22	POz - 58	18
C1 - 10	38	PO7 - 56	15



TABLE II. CLASSIFICATION RATES FOR AVERAGE P300 PATTERN USING VARIOUS COMBINATIONS OF CHANNELS AND VARIOUS DECISION METHODS

Decision method	Channels Selection		
	3, 5, 11	3, 4, 5, 10, 11, 12, 17, 18, 19	3, 5, 11, 17, 19
Final Classification	60%	56.47%	58.82%
Partial classification	89.41%	88.23%	84.70%
Visual_Field Classification	86.47%	87.05%	85.88%

Analysing the results the following channel combinations and averaging have been chosen for classification comparison (3, 5, 11), (3, 4, 5, 10, 11, 12, 17, 18, 19) and (3, 5, 11, 17, 19). Again, we stress the fact that using averages we cannot expect high classification rates. We have tried simulations with a different number of channels, that is, more or less channels, hoping that classification rates will change significantly.

Using scalar products of this pattern with average segments from various channels, the classification rates reported in Table I have been obtained.

We have thus built on the idea that, when multiple-channel EEG signals are averaged, a kind of P300 filter is obtained so that using EEG signals from multiple channels increases the classification performance. Thus, taking into account the results presented in Table I for the proposed algorithm, we have tested various combinations of handled channels the classification performance being presented in Table II.

*B. Five P300 Patterns based on the K-Means Algorithm*

Observing that a single pattern cannot define very well all cases of EEG segments that contain P300, we tested an alternative in which more patterns (in our case five) were used. Thus, by means of the K-mean algorithm and the EEG segments with P300 from the training database, we defined the class centroids as the (five) “prototype” patterns for P300 wave. In the classification stage, we calculated the scalar products between the EEG segments and each of such a pattern and the values of scalar products are summed which were summed in order to find the line and the column containing the P300 wave.

In Fig. 5 the patterns obtained with K-means algorithm and the silhouette plot for the clustered data are presented. It can be seen that the patterns exhibit rather different forms. The silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster. The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

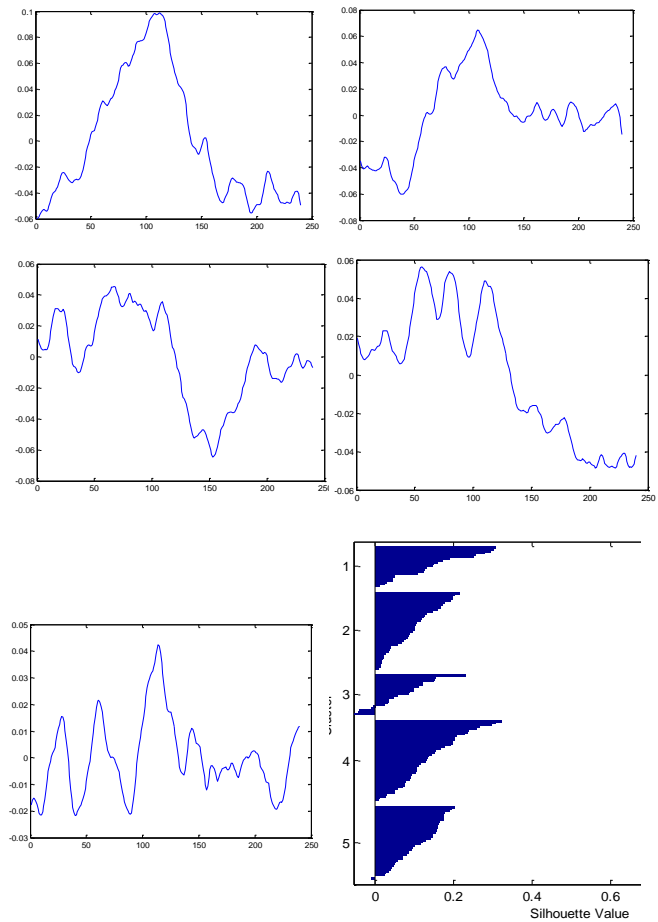


Fig. 5. P300 patterns obtained with the K-means algorithm applied on EEG segment with P300 and silhouette plot from the clustered data.

In Table III the classification results obtained with K-means patterns are presented. It can be noticed that the final classification rate is similar with the case with average P300 pattern but the partial classification rate and Visual\_Field Classification are slightly better comparatively with Table II.

TABLE III. CLASSIFICATION RATES FOR K-MEANS PATTERNS USING VARIOUS COMBINATIONS OF CHANNELS AND VARIOUS DECISION METHODS

Decision method	Channels Selection		
	3, 5, 11	3, 4, 5, 10, 11, 12, 17, 18, 19	3, 5, 11, 17, 19
Final Classification	57.6%	57.64%	57.64%
Partial classification	89.41%	88.23%	87.05%
Visual_Field Classification	86.47%	84.11%	87.64%

*C. Stylized Pattern by Summing up Three Gaussian Functions*

In several papers [10], [11] authors proposed as a P300 pattern a waveform model based on mathematical expressions of Gaussian functions. We have investigated the above approach by defining the model by summing up 3 Gaussian functions (see Fig. 6 and Table IV for results).

TABLE IV. CLASSIFICATION RATES FOR STYLIZED PATTERN COMPOSED OF GAUSSIAN USING ONE COMBINATIONS OF CHANNELS AND VARIOUS DECISION METHODS

Decision method	Channels Selection
	3, 5, 11
Final Classification	7.05%
Partial classification	42.35%
Visual_Field Classification	64.11%

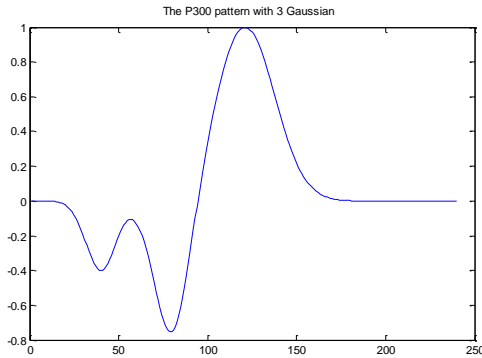


Fig. 6. P300 pattern composed by three Gaussian functions.

D. Pattern Defined by 1 Gaussian Function

We have also investigated the case of one Gaussian function and shown in Fig. 7.

The final classification rate is low compared to the case with average P300 Pattern, but the partial classification rate and Visual\_Field Classification are much better (see Table V).

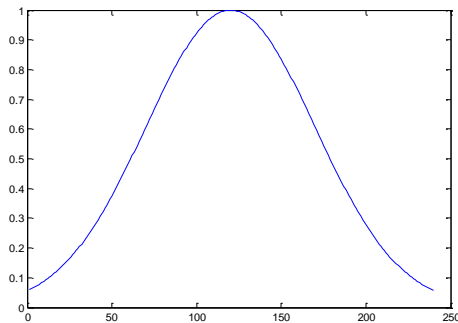


Fig. 7. Gaussian P300 pattern.

TABLE V. CLASSIFICATION RATES FOR GAUSSIAN P300 PATTERN USING ONE COMBINATIONS OF CHANNELS AND VARIOUS DECISION METHODS

Decision method	Channels Selection
	3, 5, 11
Final Classification	30.58%
Partial classification	75.29%
Visual_Field Classification	77.05%

E. Stylized P300 Pattern by Decimating and then Interpolating the Average P300 Pattern

The last attempts to find “prototype” P300 patterns starting from the average P300 pattern from Fig. 4 were to construct a stylized pattern by decimating the average P300 Pattern and rebuilding the pattern waveform by cubic spline data interpolation (see Fig. 8).

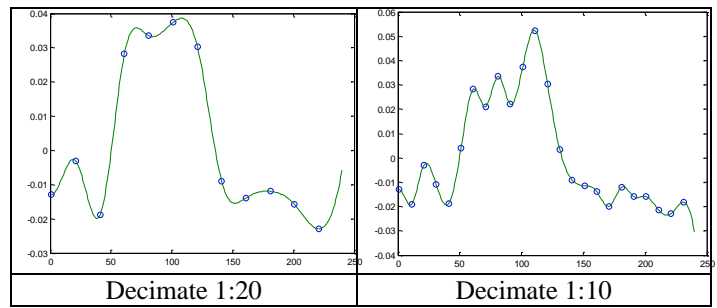


Fig. 8. Stylized P300 Pattern by decimating with factor of 1:20 and 1:10.

TABLE VI. CLASSIFICATION RATES FOR STYLIZED P300 PATTERN BY CUBIC INTERPOLATION USING ONE COMBINATIONS OF CHANNELS AND TWO P300 WVEFORMS

3, 5, 11	Decimate 1:20	Decimate 1:10
Final Classification	56.47%	56.47
Partial classification	87.05%	90.58
Visual_Field Classification	85.29%	85.88

The final classification rate is similar with average P300 Pattern, but the partial classification rate and Visual\_Field Classifications are much better. In fact, the partial classification rate is for 1:10 decimation has the highest values of all.

V. CONCLUSIONS

In this work we have analyzed and tested a very simple algorithm for the P300 detection based on the values of the scalar product between EEG segments and several “prototype” waves. The proposed algorithms are of low complexity and basically reduce to finding a maximum among 6 calculated values. Moreover, we have computed so called partial and visual field classifications. As expected, the results are significantly influenced by the shape of the prototype waveform.

Thus, the best final classification results have been obtained with the averaged P300 wave and average of channels 3, 5 and 11 which gave the best results for the Visual\_Field Classification as well (marked grey in Table II) . The results confirm that a value of 86.47% for the Visual\_Field which is higher than the Final Classification value with about 26%. This value shows that the probability of generating a P300 for characters around the target one is much higher compared to the case when this error would be uniformly distributed. Thus the probability of generating a P300 for characters in in the vicinity of the target one is higher.

Another interesting result is that through decimation and interpolation the results are rather similar to the case of the initial “prototype” shape; moreover, for partial classification with 1:10 decimation rate they are the best of all (marked grey in Table VI).

We envisage further investigations might envisage results of the proposed method applied to other EEG as well as ECG databases.

REFERENCES

[1] Blankertz, B.: BCI competition III webpage, [http://www.bbci.de/competition/iii/desc\\_II.pdf](http://www.bbci.de/competition/iii/desc_II.pdf)

- [2] Blankertz, B., Mueller, K.-R., Curio, G., Vaughan, T., Schalk, G., Wolpaw, J., Schloegl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schroeder, M., Birbaumer, N.: The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials. *IEEE Trans. Biomed. Eng.* 51(6), 1044–1051 (2004)
- [3] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan, *Braincomputer interfaces for communication and control*. *Clin Neurophysiol*, 113, Jun., 767–791, (2002).
- [4] Schalk, G., McFarland, D., Hinterberger, T., Birbaumer, N., Wolpaw, J.: BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering* 51(6), 1034–1043 (2004)
- [5] Serby, H., Yom-Tov, E., Inbar, G.F.: An improved p300-based brain-computer interface. *IEEE Trans. Neural Syst Rehabil Eng.* 13(1), 89–98 (2005)
- [6] L. A. Farwell & E. Donchin, "Talking off the top of your head: A mental prosthesis utilizing event-related brain potentials", *Electroencephalogr. Clin. Neurophysiol.* 70 (6): 510–523, 1988
- [7] A. Rakotomamonjy, V. Guigue, BCI Competition III: Dataset II-Ensemble of SVMs for BCI P300 Speller, *IEEE Transactions on Biomedical Engineering*, Volume:55, Issue: 3, 2008, pp. 1147 – 1154
- [8] K. A. Colwell, D. B. Ryan, C. S. Throckmorton, E. W. Sellers, L. M. Collins, Channel Selection Methods for the P300 Speller, *J Neurosci Methods*, July 30; 232: 6–15, 2014
- [9] Minpeng Xu, Hongzhi Qi, Lan Ma, Changcheng Sun, Lixin Zhang, Baikun Wan, Tao Yin, Dong Ming, Channel Selection Based on Phase Measurement in P300-Based Brain-Computer Interface, *PLOS ONE*, April, Volume 8, Issue 4, 2013
- [10] Lucie Daubigney and Olivier Pietquin, Single-trial P300 detection with Kalman filtering and SVMs, *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges (Belgium), 27-29 April 2011
- [11] Setare Amiri, Ahmed Rabbi, Leila Azinfar and Reza Fazel-Rezai (2013). A Review of P300, SSVEP, and Hybrid P300/SSVEP Brain- Computer Interface Systems, *Brain-Computer Interface Systems - Recent Progress and Future Prospects*, Dr. Reza Fazel-Rezai (Ed.), InTech, DOI: 10.5772/56135. Available from: <https://cdn.intechopen.com/pdfs-wm/44907.pdf>

# Implicit and Explicit Knowledge Mining of Crowdsourced Communities: Architectural and Technology Verdicts

Husnain Mushtaq, Babur Hayat Malik, Syed Azkar Shah, Umair Bin Siddique, Muhammad Shahzad, Imran Siddique  
Department of CS & IT  
The University of Lahore  
Gujrat, Pakistan

**Abstract**—The use of social media especially community Q&A Sites by software development community has been increased significantly in past few years. The ever mounting data on these Q&A Sites has open up new horizons for research in multiple dimensions. Stackoverflow is repository of large amount of data related to software engineering. Software architecture and technology selection verdicts in SE have enormous and ultimate influence on overall properties and performance of software system, and pose risks to change if once implemented. Most of the risks in Software Engineering projects are directly or indirectly coupled with Architectural and Technology decisions (ATD). Advance Architectural knowledge availability and its utilization are crucial for decision making. Existing architecture and technology knowledge management approaches using software repositories give a rich insight to support architects by offering a wide spectrum of architecture and technology verdicts. However, they are mostly insourced and still depend on manual generation and maintenance of the architectural knowledge. This paper compares various software development approaches and suggests crowdsourcing as knowledge ripped approach and brings into use the most popular online software development community/Crowdsourced (StackOverflow) as a rich source of knowledge for technology decisions to support architecture knowledge management with a more reliable method of data mining for knowledge capturing. This is an exploratory study that follows a qualitative and qualitative e-content analysis approach. Our proposed framework finds relationships among technology and architecture related posts in this community to identify architecture-relevant and technology-related knowledge through explicit and implicit knowledge mining, and performs classification and clustering for the purpose of knowledge structuring for future work.

**Keywords**—StackOverflow; architecture and technology verdicts; crowdsourcing; data mining; explicit and implicit knowledge; software repositories; knowledge structuring

## I. INTRODUCTION

Large sharing of data and information on social media and Community Q&A websites have become a potentially valuable and rich source of knowledge for a large spectrum of users across the world. It has become a norm of learning and professional communities to knock at web based social media communities when they feel a need to get insight about a new topic, subject or to solve a particular problem. The fact behind

is that with knowledge relevant web communities, such online forums, social media groups, Community Question Answering (CQA)/crowdsourcing, traditionally known as Question Answering (Q&A) websites gather contributions from a large pool of users with different levels of expertise and experiences. Recent time is witnessed the acute emergence and growing popularity of these Community Q&A sites among software professionals [1]. The use of social media as source of knowledge for software engineering process and practical learning has been on a steady rise in recent years.

Software architecture is a complex process of making technology and design decisions which impose potential risk on software development project success, directly or indirectly. Technology and Architectural Decisions not only affect a system's structural properties but also its quality attributes [15], [19]. It is dying hard to change an architectural solution and Technology decision after it has been implemented as such change become perilous risk for system and project [2]. Software architecting and implementation are different aspects of software projects because changing in technology is more complex and crucial as compared to bug fixing in a system. Architecture and Technology Knowledge plays a vital role regarding Architectural and Technology Decisions [2]. Existing architectural and Technology Knowledge (ATK) management approaches emphasis on fabrication of repositories of ATK to guide software architects and system analysts to make the desired decision [3]. These repositories of architecture and technology are built manually by architects and it is prolonged and mined-numbing process to populate and utilize such repositories. Manually populated architectural and technology knowledge repository can accumulate limited knowledge while, in the meantime, manual evolution and maintenance of these repositories is another laborious task.

Architectural and technology decisions (ATD) imprint their effects on all over software project. Any risk attached with such decisions may spoil the whole project due to dependency of SDLC on ATD. Conceptual ATDs effect the architecture configuration of the system which is not aligned implementation [12]. No doubt, risk is inevitable in large, even in medium and smaller; software projects and ATDs are about selection of concrete technology and architecture solutions, such as design patterns, frameworks and tools through which it

could be easy to implement risk avoidance and risk mitigation strategies [17]. The selection between technologies solutions architectural designs mostly depend upon architecture-technology relationship and its knowledge. About 30% of executive decisions constitute of a system ATDs, and most of them are technology decisions. Moreover, executives, technology experts, system analysts, architects and domain experienced personals are involved in ATDs and organizations spent enormous time in getting the people engaged and spent their dear cost to get valuable opinions. Major purpose of experts' opinion is to identify and avoid risks in start to make rightful ATDs. In addition, technology ATDs are the mostly documented for future use. Physical identification and involvement of experts across the globe and utilization of manual ATDs knowledge repositories is possible but not seems feasible where project time is less and cost is high.

Social software (e.g., forums) offers novel, advanced and mechanized methods to share, capture and utilize knowledge across the world. Now, Software engineers ask questions, and assimilate from the crowdsourced (people on social media, Q&A communities and forums) about architecture and technological related risks. Many architects mentioned social software as a rich source of knowledge for technology solutions where large number of experts and unlimited knowledge repositories and huge technology solutions are available [13]. Recent social media studies about software development gives an insight that developers utilizes the rich libraries of software developers' communities spread across the social media to discuss variety of concepts, such as architecture, design, technology and domain concepts and related software engineering risks [5], [6]. Hence it is revealed that architectural, technology, design and risk knowledge exists in software development communities. Ultimate approach of this paper is to complement architectural and technology management systems with efficient data mining methods for extraction, classification and clustering architecture, technology and risk related knowledge [9]. Clustering of architectural and technological risk with respect to software project domain and build relationship between risks and ATK clusters in developers and experts communities is purpose of this paper.

To achieve this, there is need to consider a well-structured online software development community as a rich knowledge repository of architectural, technology and related risks along with their solutions. StackOverflow (SO), as software development community websites, is large website with Question and Answer (Q&A) structure [5]. SO supports utilitarian knowledge management profile like incorporation of Q&A posts with context details and it ensures quality of post through categorization and tagging the posts. It is an emerging platform for data mining researchers to extract, classify and sort largely distributed useful knowledge through various data mining and machine learning techniques like clustering, k-means algorithm, association, etc. [9]. Architectural and Technology management approaches can be supported with progressive technology-related knowledge and software related risks, evaluated by a pool of technology experts across the globe via virtual means [10]. Fig. 1 shows the overall problem domain, crowdsourced knowledge and insourcing knowledge.

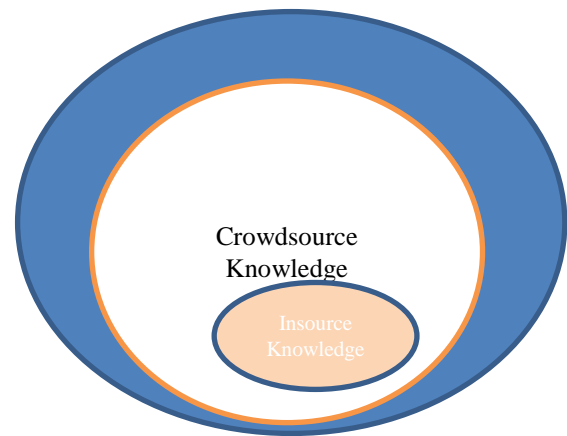


Fig. 1. Knowledge comparison of Crowd and insource w.r.t application knowledge.

The paper provides a novel framework with an objective to identify SO posts related to software architecture, technology and risks management Knowledge. Moreover, the study explores the key differences and relationships among Architecture, technology and programming posts. Based on this study, the paper focuses following Objectives:

- Extraction of raw data in form of SO Posts
- Classify SO posts as architectural, technology and risk through data mining techniques
- Find a relationship and association among SO posts using machine learning approaches
- Analyze triangular based relationships make best architectural and technological choice with minimum risk in software project.

## II. LITERATURE REVIEW

The study organizes the literature review in related areas. This study is a multi-faced study that involves mining of Software Engineering Knowledge from unstructured and unorganized data from community Q&A platform using Data Mining and Machine Learning techniques [8]. This domain guided knowledge is then brought into use to make architectural and technology decisions. Right architectural decisions largely depend upon software architectural knowledge and its utilization in different kinds of applications. Software architects also rely and seek guidance from software architecture based knowledge repositories. These knowledge repositories are part of software repositories which are specifically being built to make their best use in future projects as guideline [2]. Both, architect and technology decisions are considered of highest importance to decide fate of software project [1]. Software project implementation depends upon right and suitable software architect and system components configuration decisions. Technological knowledge follows architectural knowledge in importance regarding software project and technology selection decision also rely on architectural design. Because there exist plus and minus of each technology for different design patterns [4]. Architecture and technology decisions are made by top management and properly documented due their importance in whole project

life. Q&A communities present themselves as rich candidate for conducting knowledge mining research.

The ever increasing volume of data and information on social media has become valuable and rich source of knowledge. Definitely, people look for fastest, largest and reliable platform when they need to learn about some new topics or to solve some specific problem [7]. Therefore, they give priority to consult with some web community like social media, online Question Answering (Q&A) Sites, which welcome data form large pool of users (students, Experienced, Experts, etc.). Recent years are witness of emergence and growing popularity of these sites among academicians and industrialists. Domain of computer sciences, Software engineering and Information Technology are major contributors and beneficiaries of these Q&A Sites Stackoverflow is good example of such community where millions of users from similar domain are registered for Quo pro quid [10]. Pearson's latest annual report on use of social media has revealed that there is an acute rise in use of social media by software developer community and graph of their dependency on these social Q&A sites has gone up tremendously[21]. Some researchers even say that online assistance has become indispensable for programmer and software engineers [25].

To date knowledge mining research on community Q&A has directed in line to predict answer quality, crowd participation rewards and users ranking and profiling expert finding, and success factors of community, subject related analysis [8]. Advances in Knowledge Discovery and Data Mining brings together the latest research in statistics, databases, knowledge discovery, machine learning, and artificial intelligence that are part of the exciting and rapidly growing field of Knowledge Discovery and Data Mining [13]. We believe, there is an apparent lack of knowledge mining of Community Q&A sites especially in context of decision making process of software engineering process: opinion mining. Mining community Q&A sites to extract knowledge (implicit and explicit) to refine decision making process [26][27]. This study describes a first level attempt to gather implicit and explicit knowledge form stackoverflow and make it useful in architectural and technology related verdicts.

### III. SOFTWARE DEVELOPMENT APPROACHES V/S CROWDSOURCING APPROACH

#### A. Insourcing

In insourcing, organizations accomplish project goals by using internal expertise. Organization opt to hire new qualified human resource, shift staff from one project to another or train staff to complete a project, Instead of subcontracting to third parties [11]. Insourcing improves communication among staff members and internal IT resources are also innovated. Cost is the main challenges of, insourcing due to hiring new staffs and software licenses.

Comparing crowdsourcing with insourcing in software development, user participation, flexibility, openness, scalability and flexibility in insourcing are said to be lower than crowdsourcing. On the other hand, development time, development cost, trustworthiness, license requirement, business risk and operational control of insourcing are higher

than crowdsourcing [11]. Control over software development process is stronger in insourcing software development compared to outsourcing and crowdsourcing [11].

#### B. Outsourcing

In case, sufficient in-house expertise is not available then organizations choose outsourcing. It means, organizations contract with external companies for accomplishment of project. It not only reduces burden to find out human resource but reduces cost also. Finding the right service provider according to its expertise is the main task of software development via outsourcing [11].

Crowdsourcing and outsourcing are intermingled with each other [18]. Crowdsourcing utilizes open calls to large number of geographically distributed people to achieve tasks from volunteer workers (with all levels of expertise) while outsourcing makes contract with other companies or professional organizations. Besides, outsourcing performs is business relationships [20], while crowdsourcing is all about participation and motivation. Factors like development time and cost, confidentiality, software license issues, business related risks and level management control of outsourcing are said to be higher than crowdsourcing [11]. However, Transparency, ability to have tailored product, user participation, scalability of crowdsourcing factors are greater than outsourcing.

#### C. Opensourcing

Open source and crowd source are two different software development approaches. Projects developed by crowdsourcing are not distributed in public for free, while open source software development contains freely distributed software [16]. People may work independently or collaboratively in crowdsourcing tasks while people work in collaboration during open source software development [20]. Factors like user participation, openness, scalability and flexibility are mostly lower in open source software development as compared to crowdsourcing. On the hand, development time, development cost, confidentiality, license issues, business risks and management control are higher in opensourcing than crowdsourcing [11].

#### D. Nearshoring

It geographic proximity between client and sourcing locations [22], i.e. nearshoring is associated with outside of client country but proximate to sourcing countries. Client countries achieve their tasks at lower wages in sourcing countries by utilizing the proximity of geography, culture, language and economic characteristics between countries [22]. There are three major nearshoring clusters in the world: the USA and Canada, wealthy nations of Western Europe and Korea and Japan [23].

#### E. Offshoring

Project is performed between clients and supplier organizations located at different countries. The driving force behind offshoring is cost reduction. Communication limitation, language barriers, cultural differences and political issues are few drawbacks of offshore software development approach [24].

TABLE I. COMPARISON OF DIFFERENT SOFTWARE DEVELOPMENT PARAMETERS TO FIND THE BEST AND SUITABLE SOFTWARE DEVELOPMENT APPROACH

Software Development Parameters	In sourcing	Out sourcing	Open sourcing	v/s Crowdsourcing
User Participation	Lower -1	Lower -1	Lower -1	Higher +1
Flexibility	Lower -1	Lower -1	Lower -1	Higher +1
Openness	Lower -1	Lower -1	Lower +1	Higher +1
Scalability	Lower -1	Lower -1	Lower -1	Higher +1
Time	Higher -1	Higher -1	Higher +1	Lower +1
Cost	Higher -1	Higher -1	Higher +1	Lower +1
Trustworthiness	Higher +1	Higher +1	Higher +1	Lower -1
Licenses	Higher -1	Higher +1	Higher -1	Lower +1
Business related Risks	Higher -1	Higher +1	Higher -1	Lower +1
Management Control	Higher +1	Higher -1	Higher +1	Lower -1
Transparency	Higher +1	Higher -1	Higher +1	Lower -1
Risk Prediction	Lower -1	Lower -1	Lower -1	Higher +1
Participants	Lower -1	Lower -1	Lower -1	Higher +1
<b>Points Gain</b>	<b>-7</b>	<b>-7</b>	<b>-1</b>	<b>+7</b>

#### F. Crowdsourcing

Crowdsourcing has become an emerging platform for both academics and industrial world. It is applicable to software development approach in which open calls are utilized in order to have the tasks done by a large group of volunteer people connect through a platform like Stack overflow (SO). There are three main models of Crowdsourcing: peer production, competitions and micro-tasking [10]. Peer production crowdsourcing is the oldest model where people work collaboratively without any reward expectation. Competitions crowdsourcing approach constitutes that workers compete with each other for achieve projects' goals in order to gain monetary rewards. The requirements are submitted to the crowdsourced platforms. The copilot/platform manager splits the project into sub tasks which are competition tasks with different rewards. The large group of workers i.e. community propose diverse solutions for these tasks. The best solution is chosen among all solutions and winning solution is rewarded [10]. The micro-tasking crowdsourcing is last approach which divides works into several self-contained and small tasks to complete in a short time period by a large group of people via scalability feature of software works.

Table I shows that among all software development approaches, crowdsourcing seems better to adopt as most suitable and broad approach. Emergence of social media and bog data has changed the trend of software development. Crowdsourcing is a hidden potential that can be used in software development process especially in architectural and design decision that are considered more important and pivotal in SDLC. This study unveils the idea that how to mine big data and community Q&A platforms to utilize implicit and explicit knowledge scattered across the internet. The paper takes stackoverflow as targeted Q&A community and gives a framework to mine opinions and utilize that implicit and explicit knowledge in architectural and technology verdicts for software development.

#### IV. STACKEXCHANGE AND SO AS CROWDSOURCING COMMUNITY

StackExchange is a large network of Q&A community websites and each covers a particular topic at vast level like technology, science, business, etc. Currently, StackExchange is ranked at 170 in global traffic graph. It covers 119 topic websites and as well as 119 Meta websites. StackExchange is a platform where users create, vote and edit questions and answers and filter the answering posts using popularity voting mechanism. It also utilizes gamification to boost up user participation and game design elements.

SO was created in 2008 as part of StackExchange network and at now it is the most famous and collaborative website in the SE network. It is a free Q&A platform which facilitates exchange of knowledge among fresh, experts and experienced programmers. SO have over 3.5 million users registered with it. Since 2008, almost over 8 million questions and over 14 million answers have been posted on the SO website. All these posts have, collectively, turned the SO into a large repository of computer programming, software development and other related knowledge. This is an evidence about popular usage of SO for discussions and exchange of information about a particular technology and revealed that SO encompasses a wide spectrum of technologies. Almost 7,000 questions are posted on the SO website daily. Furthermore, SO maintains a complete record of each user including ones badges, points, and scores which may be utilized for various research purposes. Attributes of SO:

- Extensive coverage and comprehensiveness
- Up-to-dateness
- Rich Description.

## V. CONCEPTUAL FRAMEWORK

### A. Implicit Knowledge Mining/ Knowledge Discovery from Community Data

1) Tacit or implicit knowledge is the kind of knowledge that is not easy to transfer from person to person by means of writing it down or verbalizing it. Here in this paper, implicit knowledge refers to the useful information that is to be extracted from crowdsourcing Q&A platform and then to be used to make architectural and technology decisions for software projects. The knowledge explored from questions and answers posts posted on SO is implicit because data is posted for particular user in a specified sense but here we are using SO data to make different decisions of architecture and technology in software development process. Hence the knowledge extracted from such data can be called as implicit knowledge.

2) *Determine Knowledge Domain / Q&A Community*: The first and the foremost step involve in knowledge discovery is determination of knowledge domain. This study focuses the two main domains of software engineering, architecture and technology decisions and risk involved with these aspects of software engineering projects. As mentioned above, architecture and technology verdicts bear highest importance because any loop-hole in any of these decisions often poses risks in almost every phase of software development. Therefore, for right decision right knowledge is required to achieve right goal.

3) *Select Data Mart*: SO is among biggest Q&A networks online. One can use the data dump for research and mining MSR which contains all data since 2008 from inception of SO. Here the term data mart is used for specific kind of posts posted over SO. Data marts can be distinguish among IT, CS, software development, tools, technology, architect, programming, framework, etc. this paper selects architectural and technology data marts for data mining as shown in Fig. 2. Data marts can be import as XML data into database with ease for with further processing. The data dump includes all the questions and their answers including partially anonymized user data, user tags and actions logs and their rewarded reputation points.

4) *Data Preprocessing*: Creation of Target Data: It is selection of data of interest from data set.

*Data cleaning*: Data cleaning is process of making the selected or targeted data by removing the noisy, redundant and irrelevant data [14]. The ability to understand and to correct quality of data is imperative in getting to accurate final analysis. Data cleaning in data mining gives the user an insight to discover inaccurate or incomplete data before analysis phase.

*Data reduction and transformation*: This step involves the cleaned data set to be reduced by choosing dimension of interest as in this study, architecture and technology, and transforming to the format that can be easily and properly interpreted by data mining methods.

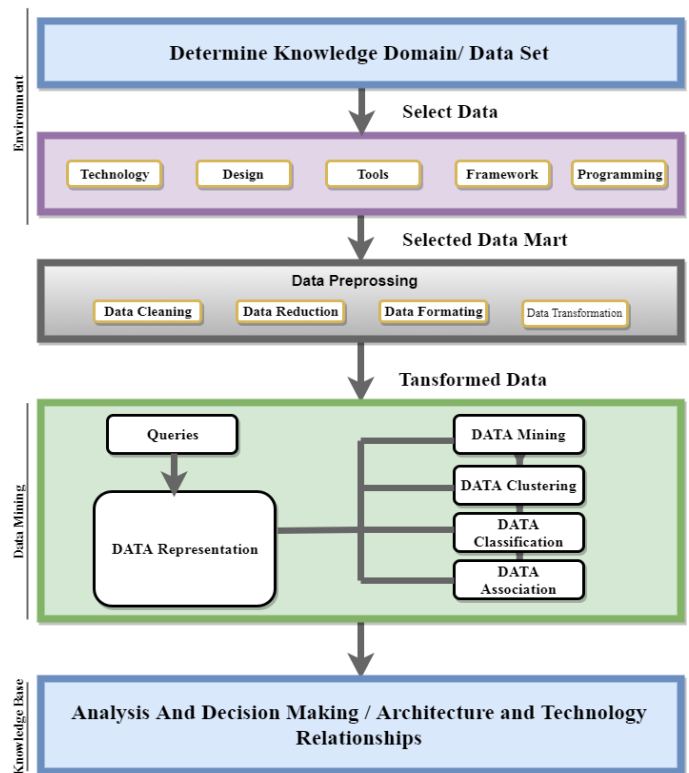


Fig. 2. Implicit knowledge mining framework.

5) *Queries*: The framework provides two different ways to extract usable data for knowledge presentation. First and short cut method is querying the CSV formatted data. CSV formatted data file is opened in excel and query is made to filter desired data. This is simple way just to get quick and to the point data. Limitation of query is that it does not give relationship and association among data that is the ultimate purpose of the study. Hence data mining is applied on formatted data to find knowledge patterns that are not possible in CSV query based method.

*Application of data mining*: According to knowledge determinants and goals of knowledge discovery, any of the mining intelligent methods can be applied according to data mining expertise and needs of knowledge discovery process. Data patterns are result of data mining process.

*Association*: Association technique is also named as relational technique. Data mining discovers patterns of relationship among data set in the transaction through if/then statements that help to unveil relationship between seemingly unrelated data in relational database or some other information repository this paper ponders to find association rules to find architectural and technology relationships.

*Classification*: This method classifies the data set into predefined classes or concepts. There are several mathematical techniques like decision-tree, neural network and other are utilized to classify the items in data set.

*Clustering*: Clustering is grouping technique of data mining that organizes the items of similar characteristics in clusters whose classes are known or unknown.



*Regression:* This method finds relationship between the data variables like dependent and independent variables in dataset.

*Sequential patterns:* This analysis discovers the regular and frequent patterns in data set.

6) *Evaluation of patterns:* Patterns discovered by data mining methods are then analyzed and evaluated according to the knowledge.

7) *Knowledge presentation:* Presentation of knowledge patterns to represent the decisions in form of metaphors like graphs. This is visualization stage of knowledge so that the discovered knowledge might be used in decision making of architectural and technology verdicts.

8) The knowledge discovered by using machine learning and data mining techniques using dump data of SO can be termed as Implicit knowledge because it is an inference from a large pool of data. The knowledge is not delivered in response of an explicit query rather it is extracted from dump data, formatted, cleaned and then summarized in form of patterns to make it useful in particular sense for which it might not be produced originally.

### B. Explicit Knowledge Discovery / Knowledge Discovery from Experts

Contrary to implicit knowledge, explicit knowledge is directly discovered from experts in response of direct and to the point problem statement. This study proposes a three-tier or tri level mechanism to find experts from data set. Here, a question arises that to whom should put a problem statement to discover explicit knowledge. Each post in data set point a user. All these users are producer of knowledge repository by asking and answering question knowledge sharing communities like SO. Experts finding among Crowd or Expert Finding Mechanism is given as follows:

1) *Post acceptance level:* Each post on SO gets acceptance or rejections through ups and downs respectively from registered users across the community. Post quality, reliability and authenticity is measured from its ups and downs. A post with positive response reflects that user, who posted that post or comment, bears good knowledge about that domain.

2) *User badge:* SO also involve in gamification of community to attract more people to register themselves and to gear up activities of registered users across the community to share knowledge. Each badge reflects knowledge and experience level of SO users according to criteria of each badge set by community admin. Each user across the platform strive his best to achieve highest badge through raising quality questions and producing quality answers of difficult problems. This study utilizes user badge criteria to select user for gathering explicit knowledge about technology and architecture as shown in Fig. 3.

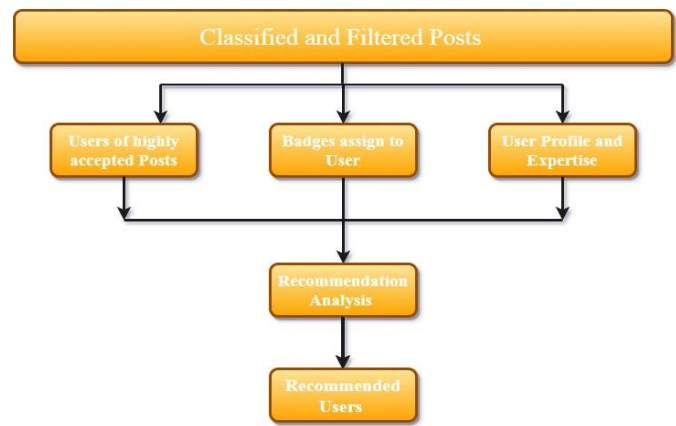


Fig. 3. Explicit knowledge mining framework.

3) *User profile and expertise:* Each registered user across the SO distinguishes himself through unique capabilities and expertise. User profile on the SO is a direct way to determine user expertise in required domain. This is the third criteria to select a user from data posts for explicit knowledge mining.

This study presents a triangular approach in selection of SO user for explicit knowledge gathering is done through three evaluation criterion. Only User posts in selected domain, user badge or user profile is not enough to select a user for inquiring about architectural or technology knowledge rather this technique gives weightage to the all three aspects of user selection framework to select right person.

## VI. CONCLUSION

Goal of this study is to unveil the potential of crowdsourced experts and communities across the internet to support architectural and design management approaches and decision makings with a more efficient method for capturing architecture and technology related knowledge. Moreover, this paper distinguishes the crowdsourcing approach of software development as more suitable and rip with knowledge. The proposed framework of implicit and explicit knowledge checks if SO could be viable source for reusable architecture and design knowledge. The framework utilizes the data mining techniques to support software development team to get wide spectrum of opinions in form knowledge patterns discovered by the crowdsourcing knowledge mining framework. This study focuses both qualitative and quantitative aspects of SO knowledge repository. In our future work, we will implement this framework to find architectural and technology relationships and pattern through classification, clustering and association techniques. Moreover, we will extend our study to remaining core knowledge base area over SO and on other Q&A communities.

## REFERENCES

- [1] (n.d.). Retrieved from [https://en.wikipedia.org/wiki/Educational\\_data\\_mining](https://en.wikipedia.org/wiki/Educational_data_mining)
- [2] Achmad Arwan, S. R. (2015). Source Code Retrieval on StackOverflow Using LDA. 3rd International Conference on Information and Communication Technology (ICoICT), (pp. 295-299).
- [3] Arash Joorabchi, M. E. (2015, August ). Text mining stackoverflow, An insight into challenges and subject-related difficulties faced by computer

- science learners. *Journal of Enterprise Information*, Vol. 29 No. 2, 2016, 255-275.
- [4] ASLI SARI, G. I. (2017). An Overview of Crowdsourcing Concepts in Software Engineering. *International Journal of Computers*, 106-114.
- [5] Bosch, A. J. (2005). Software architecture as a set of architectural design decisions. *WICSA*, (pp. 109-120). A. Jansen and J. Bosch, “;” in *WICSA*, 2005, pp. 109–120.
- [6] Fabio Calefato, F. L. (2015). Mining Successful Answers in Stack Overflow.
- [7] J. L. C. Ramos, R. E. (2016). A Comparative Study between Clustering Methods in Educational Data Mining. *IEEE LATIN AMERICA TRANSACTIONS*, (pp. 355-3361).
- [8] Juan Yang, S. P. (2016). Finding Experts in Community Question Answering Based on Topic-Sensitive Link Analysis. *IEEE First International Conference on Data Science in Cyberspace*, (pp. 55-60).
- [9] Katiyar, N. U. (2014). A Survey on the Classification. *International Journal of Computer Applications Technology and Research*, 725-728.
- [10] L. Bass, P. C. (2012). *Software Architecture in Practice* (3rd ed.). (A.-W. Professional, Ed.)
- [11] L. Bass, P. C. (2012). *Software Architecture in Practice*. Addison-Wesley Professional.
- [12] Latifa Guerrouj, O. B. (2016). Software Analytics: Challenges and Opportunities. *IEEE/ACM 38th IEEE International Conference on Software Engineering Companion*, (pp. 902-903). Austin, USA.
- [13] M Amala Jayanthi, R. L. (2016). Research Contemplate on Educational Data Mining. *IEEE International Conference on Advances in Computer Applications (ICACA)*, (pp. 110-114).
- [14] M Amala Jayanthi, R. L. (2016). Research Contemplate on Educational Data Mining. *IEEE International Conference on Advances in Computer Applications (ICACA)*, (pp. 110-114).
- [15] M. Soliman, M. R. (2015). Enriching architecture knowledge with technology design decisions. *WICSA*.
- [16] Maalej, D. P. (2013). How do open source communities blog. *Empirical Software Engineering*, vol. 18, no. 6, (pp. 1090–1124).
- [17] McConnell, S. (2004). *Code Complete* (2nd ed.). (Microsoft Press.
- [18] Meiyappan Nagappan, E. S. (n.d.). *Future Trends in Software Engineering Research* for.
- [19] Mohamed Soliman, M. G. (2016). Architectural Knowledge for Technology Decisions in Developer Communities. *IEEE/IFIP Conference on Software Architecture*, (pp. 128-133).
- [20] Muhammad Ahsanuzamman, M. A. (2016). Mining Duplicate Questions in Stack Overflow. *IEEE/ACM 13th Working Conference on Mining Software Repositories*, (pp. 402-412).
- [21] Neelamadhav Gantayat, P. D. (2015). The Synergy Between Voting and Acceptance of Answers on StackOverflow, or the Lack thereof. *12th Working Conference on Mining Software Repositories*, (pp. 406-409).
- [22] O. Zimmermann, J. K. (2009). Managing architectural decision models with dependency relations, integrity constraints, and production rules,”. *Journal of Systems and Software*, vol. 82, 1249–1267.
- [23] Philipp Berger, P. H.-P. (2016). A Journey of Bounty Hunters: Analyzing the Influence of Reward Systems on StackOverflow Question Response Times. *IEEE/WIC/ACM International Conference on Web Intelligence*, (pp. 644-649).
- [24] Piatetsky-Shapiro, G. (2017). *Advances in Knowledge Discovery and Data Mining*. (G. P.-S. Usama M. Fayyad, Ed.) America: American Association for Artificial Intelligence Press.
- [25] Questions, M.-c. M.-t. (2015). Jos´e R. Cede˜no Gonz´alez, Juan J. Flores Romeroy, Mario Graff Guerreroz and Felix Calder´onx. *IEEE / ROPEC 2015 - Computing*.
- [26] Thiago B. Procaci, B. P.-F. (2016). Finding Topical Experts in Question & Answer Communities. *IEEE 16th International Conference on Advanced Learning Technologies*, (pp. 407-411).
- [27] Yunxiang Xiongy, Z. M. (2017). Mining Developer Behavior Across GitHub and StackOverflow.

# Web-Based COOP Training System to Enhance the Quality, Accuracy and Usability Access

Amr Jadi

Department of Computer Science and Engineering  
College of Computer Science and Engineering, University  
of Hail, Hail, Saudi Arabia

Eesa A. Alsolami

Department of Information Technology  
College of Computing and Information Technology  
University of Jeddah, Saudi Arabia

**Abstract**—In this paper, a web based COOP training system is demonstrated to ensure usable process of task interactions between various participants. In the existing method various issues related with the paper work, communication gap, etc. raised serious issues between the colleges and industries while implementing the COOP training programs. The primary data was collected by conducting interviews with the supervisors and also by taking the opinion of students to improve the proposed COOP system. The proposed system is capable of reducing the complexity of operations to a greater extent by avoiding overlapping of the information, reducing the communication gap and by increasing the accuracy of the information. The outcomes of the proposed system proved to be very fruitful in terms of results obtained from the point of view of all the participants in the COOP system. The performance, accuracy, quality and assessment of the student reports found to be improved to deliver excellent results.

**Keywords**—COOP training; web applications; integration; quality; accuracy

## I. INTRODUCTION

The training system in Kingdom of Saudi Arabia (KSA) is considering various modifications to meet the demands of current trends and to enhance the quality of education system under various trending circumstances. The introduction of Cooperating (COOP) training for the graduates made mandatory to improve the skill set as required by any industrial unit. In this COOP training, the student need to attend relevant industry to learn and work as a temporary employee and then they need to perform the task as per the guidelines of the company.

There are various objectives of the COOP training system includes:

- to develop different types of practical skills those are needed by real-time applications,
- to implement various applications from the knowledge acquired from different domains,
- to evaluate the attitude and working environment,
- to interpret the learning outcomes of the university studies with the applications of real-time activities,
- to prepare a comprehensive report on the overall training learning outcomes, and

- to present the working skills obtained along with the experience obtained out of the learned skills.

In the Section II, a detailed activity of COOP training system is explained for an easy understanding of each participant and their roles.

## II. CONCEPT OF COOP TRAINING SYSTEM

To obtain the above objectives of the COOP training system various participants with defined responsibilities are being allotted to perform by various educational managements and industry authorities as shown in Fig. 1. The list of participants includes a) Employer, b) Site Supervisor, c) COOP Training Coordinator, d) Faculty Advisor, and e) COOP Examining Committee.

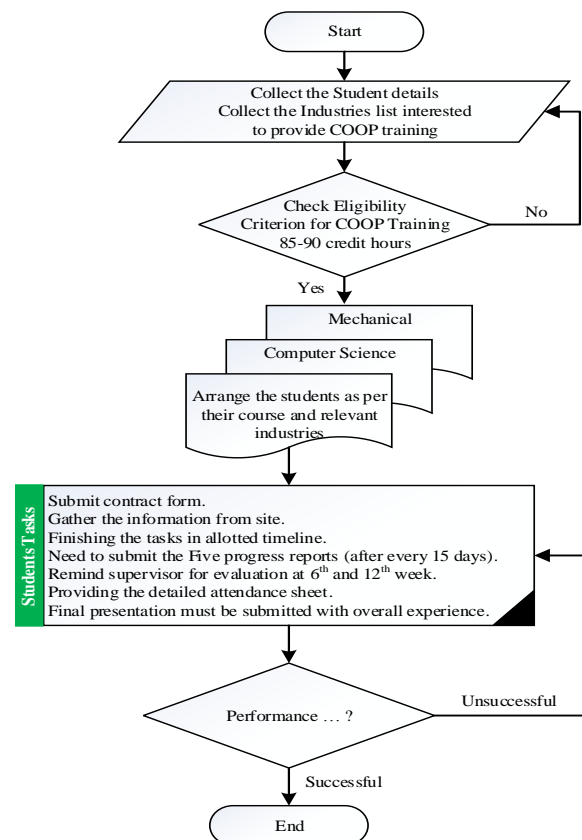


Fig. 1. Functional flow diagram of the COOP training system.

#### A. Employer

An employer is considered to be the key player in the whole COOP training system for promoting a successful training program for the students and thereby achieve common goals of an educational institution with following aspects [1].

1) Need to provide a training / task plan for a period of 12 weeks related with the student's area of study and it has to be approved by faculty advisor.

2) To accomplish the given task a site supervisor must be assigned for making the training program more meaningful and effective.

3) Needs to monitor and report any kind of irregularities by the students in the organization with the faculty advisor so as to implement the corrective actions are taken by the students advisor and the same will be reflected in the evaluation reports.

4) A proper communication and visiting hours must be allotted for the faculty advisors.

#### B. Site Supervisor

A site supervisor is a professional within the area of student area of study. Plays a critical role in developing the student in the assigned project area with the given tasks and in given time durations [1].

1) Ensures a proper direction for the students in the training area for the guidelines given by company and faculty advisor.

2) Communicates the performance of student with employer and faculty advisor for corrective actions.

3) Encourages the students for preparing technical reports and to conduct oral presentations.

4) Ensures quality and accuracy for the bi-weekly reports submitted by students.

5) Helps the students to prepare mid-point (6 weeks) and final (12 weeks) evaluation reports for the COOP training period.

#### C. COOP Training Coordinator

The COOP training coordinator plays an important role in providing various bridging activities between the educational institutions and industries with relevant subject areas for different student courses [1].

1) A coordinator will contact the companies to ensure the slots are available and booked for the COOP training sessions.

2) Assigns a meaningful coordination between the students and qualified experts from industry to meet the requirements of student courses and given tasks.

3) Necessary information will be gathered and shared with students for the employment in COOP training program.

4) Communication between the students and employers will be established through proper channel to ensure timely training sessions.

5) Coordinates the evaluation forms from employers and faculty advisors.

6) Collects, compiles and submits the final grades to concerned departments.

#### D. Faculty Advisor

The faculty advisor is selected by the COOP training committee based on the subject, assignment, and the experience in the subject area [1]. The role of a faculty advisor includes.

1) To ensure proper work assignments, assessment of task plans, progress and student activities, and finally taking appropriate decisions during the course time by providing feedback to students, management and industry with the progress report results.

2) Providing the guidelines for the students with proper report submitting guidance and professional formats.

3) Involving positively and evaluating student presentations in necessary.

#### E. COOP Examining Committee

Two or more faculty members are assigned to examining the student performance during COOP training. This committee evaluates the performance reports, presentations and feedback from the COOP coordinators [1].

In Section III, a detailed study on the existing method of implementing the COOP training system is discussed with considerable factors. Based on the discussions a suitable and comparably faster approach to realize an accurate COOP training system is proposed in this work.

### III. SYSTEM CONCEPT

Functioning of COOP training system representation is shown in Fig. 2 with existing method and in Fig. 3 with proposed web based method. Both the methods are explained in detail in the following discussions.

#### A. Existing COOP Training System

In the existing COOP training system (see Fig. 2) the interaction between each participant can be independent and might lead to a worst scenario at times due to lack of proper communication or missing information. The data sharing between each participant right from faculty advisor to site supervisor and to students may get into a murkier situation whereby both institution and organization providing the facilities for COOP training may get into serious conflicts.

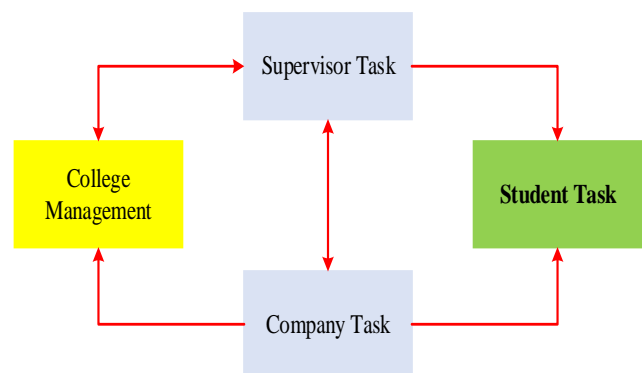


Fig. 2. Representation of existing COOP training system.

The managements of universities and companies get to know unrealistic information from the ground truths. The probability of biased evaluations is always a threat and impact (i.e. the quality of training or improving the student skills for real time problems will affect seriously) is reasonably known to all the participants involved in education system and industry. Now such issues must be dealt with care and need to ensure an efficient COOP training system.

**B. Proposed Web-based COOP Training System**

The proposed COOP system as shown in Fig. 3 consists of a web-based internal communication system which collects all the information from different participants and allows them to submit their activity information using the system. A detailed tasks and communication established between all the participants are shown in the sequence diagram (see Fig. 4) using the proposed web based COOP training system.

The integration of each activity in the proposed method will be stored in a database with appropriate measure to secure the information of all the participants in the COOP system. The list of companies selected by the supervisors will be uploaded into the website by the supervisor using his login ID and the same information can be communicated to the companies, college management and students if it is related with them. Similarly, the training coordinator from the company can provide the details of the training procedures, objectives, requirements, etc. and upload them on the web portal to communicate with the relevant participants. In the process a student also can upload the project reports, etc.

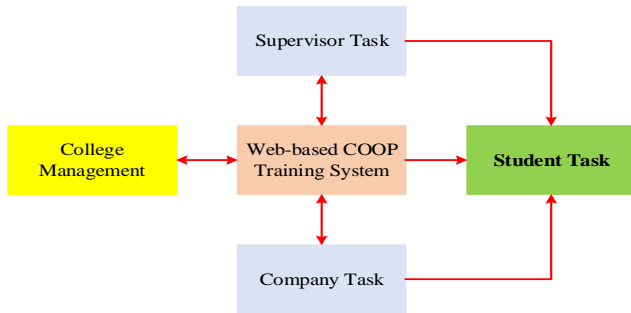


Fig. 3. Proposed web based COOP training system.

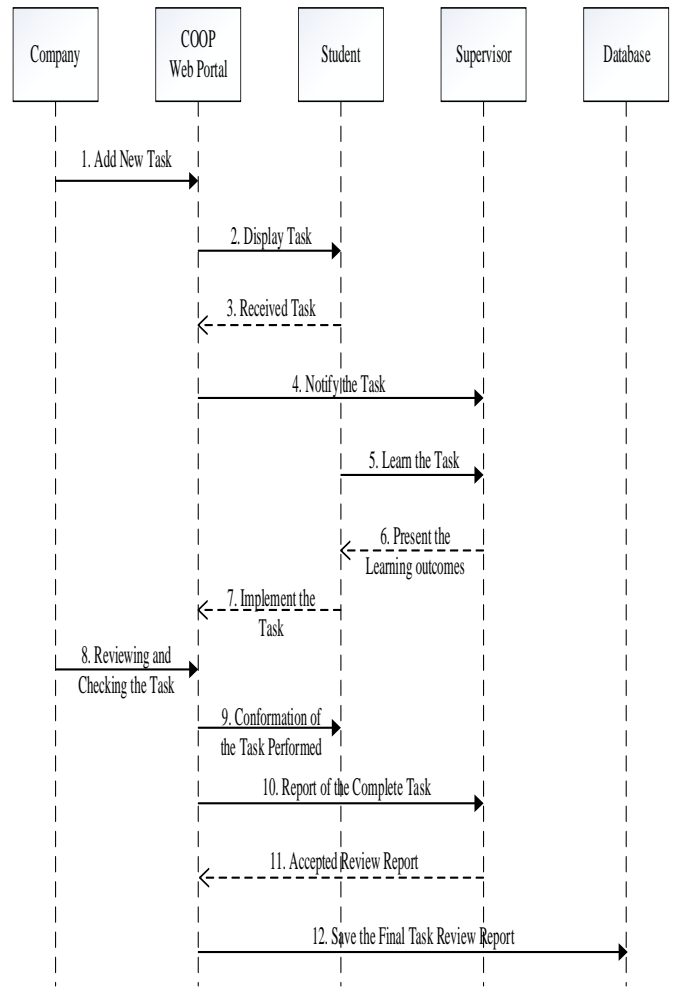


Fig. 4. Sequence diagram representing the tasks and communication between different participants using web based COOP Training System.

Similarly, the role of each participant in the web based COOP training system is summarized in Table I.

TABLE I. THE ROLE OF EACH INDIVIDUAL IN COOP TRAINING SYSTEM

COOP Training System		
Supervisor Task	Company Task	Student Task
<ul style="list-style-type: none"> <li>• Login to system</li> <li>• Can add new student</li> <li>• Can edit task or revive information from the company</li> <li>• Can search the status of student performance, attendance and Discipline.</li> <li>• Can update the information (deleting old student information, entering new data, etc.)</li> </ul>	<ul style="list-style-type: none"> <li>• Login to system</li> <li>• Can add new student or task</li> <li>• Can edit task or revive task information</li> <li>• Can search the status of student performance, attendance and Discipline.</li> <li>• Can update the information (deleting old student information, entering new data, etc.)</li> </ul>	<ul style="list-style-type: none"> <li>• Login to system</li> <li>• Can add his project reports</li> <li>• Can edit and view his own reports</li> <li>• Can search the schedules, tasks, grades, etc.</li> <li>• Can communicate with the site supervisors, employers, training coordinators, examining committee, etc.</li> </ul>

The proposed COOP training system based on Web 2.0 [2] is proved to be a well suited tool for teaching and training using social networking as a tool. Various perceptions, attitude and acceptance for the Web 2.0 in the COOP training system were discussed in their work to improve the learning quality of the students. At the same time the information sharing, learning experiences, communicating information, assessment requirements and moral support will increase using web based system. It helps both industries and college staff members to identify, understand and communicate different tasks and roles of jobs to do under the COOP system. The proposed web based system helps them to develop and communicate their decisions towards other participants in the COOP system very effectively so that the quality of training and education can be evaluated easily. Most importantly, the management of both industry and university can access the proceedings of training program without any kind of delays as seen in the case of traditional methods, where the information will be in distributed form with different people and hence it used to take a lot of time as discussed earlier. Aleisa and Alabdulahfez reported spectacular growth of the student's performance and employee satisfaction in their research findings [3]. However, the same report also highlighted the challenges in terms of securing the placements in various sectors along with financial resources and personal information. The challenges were highlighted in the work of Weber using web technologies include the computer literacy, lack of Arabic language learning objects, interoperability issues, cross-platform issues, etc. in the Gulf countries [4]. The important factor for a rejection for the COOP training system by most of the people found to be due to lack of awareness of internet usage and various applications related to internet and computer based training activities in various Gulf countries [4]. Such a situation of the students completing the graduation with no computer and internet skills lead to a worst situation for most of the private employees. In a recent report, world economic forum criticized the inadequate educated workforce as a major problem in Qatar to do any kind of business. Hence, having COOP training programs will encourage both the industries and colleges to obtain encouraging results out of present demand for the qualified workforce [4]. The usage of web based learning has seen a grater improvement according to Gulf Cooperation Council (GCC) as shown in Fig. 5.

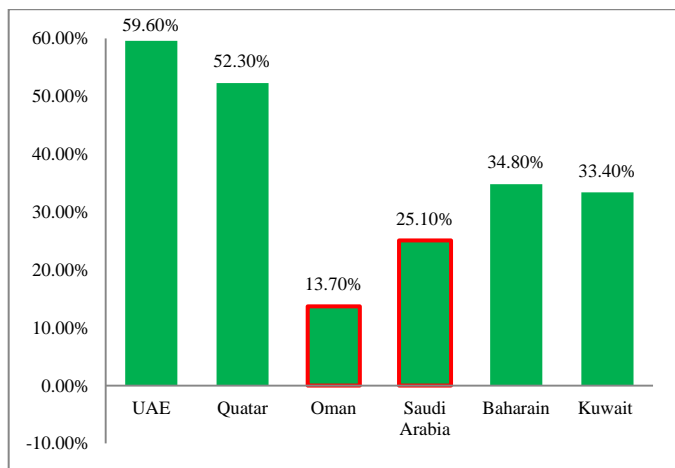


Fig. 5. Response of the students using web-based applications for learning.

It can be seen from Fig. 5 that most of the Gulf nations are penetrating towards internet as a medium of training. However, Oman is performing in a poor way as compared to the world average of 23% and Saudi Arabia (SA) also needs to enhance its percentage like in the case of UAE and Qatar. The usage of COOP training system in SA increases by implementing suitable environment for the students and parents to consider various internet based communication techniques such as emails, messaging using various social networking apps, etc.

Various reasons with respect to the companies for encouraging COOP training systems are listed in their work include:

- improvement of the image and brand of the company,
- reducing the burdens related with recruitment of talented and trained employees for featuring projects,
- reducing the time and expenditure,
- increasing the production at the reduced cost,
- to reduce the training cost and money by training the suitable employees,
- improving the retention rate,
- streamlining the company employees and employer policies at the root level by training the students,
- to ensure the professional improvement,
- to include the social and ethnic considerations at the work place, and
- to enhance the ability to implement the new ideas at the work place.

Various international institutions such as UNESCO (United Nations Educational Scientific and Cultural Organization) are involved in various technical and vocational training (TVT) strategies to envisage Saudi Arabia as a developing leadership in technical training. The complete training and objectives structure is designed by the technical and vocational training corporation (TVTC) and is approved by the council of Ministers Resolution No. 779 of 17 December 1969 [5]. The role of TVTC is to develop the procedures, programs and training plans for various industrial and vocational institutes (IVT). The SA government is aimed to have almost 180 industrial secondary institutes and 50 technical college colleges for both boy and girls separately so as to achieve the target of creating training and placements for ~ 500,000 students in the country [5]. Here the government also trying to empower girls by giving a 50 – 50 chance for both boys and girls by this training programs with a plan and aim to reduce the foreign workers on its soil in various technical and vocational professions [6].

A detailed assessment report on COOP training using ABET (accreditation board for engineering and technology) was carried out by Faiz and Al-Multain. A detailed experience demonstration was given in their work using faculty course assessment report (FCAR) method for both direct and indirect assessment of course learning outcomes (CLO) [7]. A detailed comparison of student performance between Harf Al-Batin Community College (HBCC) was carried out with respect to the students of Saudi Aramco, Saudi Electricity Company

(SEC) and Saudi Telecom Company (STC). Various influential factors were highlighted in this work as a part of strengths and weaknesses of COOP training system.

In the next section, a detailed project analysis has been carried out among the students and staff to understand the response for COOP training system in University of Hail.

#### IV. REQUIRED ELEMENTS OF COOP SYSTEM

In the process of designing the web-based COOP system there are three key aspects to be realized are listed:

- 1) Functional and non-functional requirements
- 2) Usability requirements
- 3) Preparing the questionnaires for collecting primary data

##### A. Functional and Non-Functional Requirements

In a web-based COOP system, the *functional requirements* includes login, logout, edit, delete and search functions.

- **Login Function:** Used to ensure a secured gateway for the user to access the relevant information as shown in Fig. 6. Based on the participant the restrictions and access for the information can be managed by the admin. The management of University of Hail can monitor the total activity information of each participant in a secured way. Any kind of suspicious login activity will be informed to the participant and the system admin to follow further security measures to protect the personal data.



Fig. 6. Security gateway using login function

- **Logout Function:** This is very much essential to avoid the misuse of the information and helps to secure the privacy of the participant. The care has been taken to avoid multiple system usage and the system will logout even when the user is not active after getting into their personal account. Such provision will be useful to secure the personal data as shown in Fig. 7.



Fig. 7. Security provided using the logout function

- **Add Function:** Helps the participant to add the information to be shared with all other participants. For example, the supervisor will add the time-tables,

project details, etc. to communicate with students and COOP training coordinator. In the similar fashion, the course coordinator can submit the assessment reports and a student can add the project reports.

- **Edit Function:** Helps each participant to edit the submitted information within the course time-line. This helps to update any kind of changes or event activities in an effective manner as shown in Fig. 8.

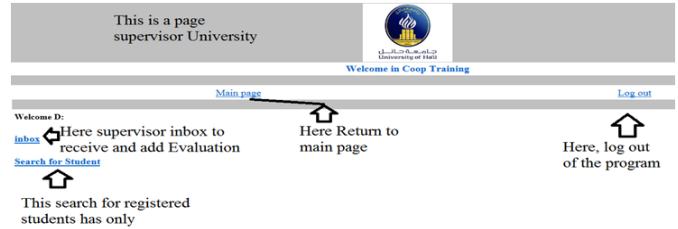


Fig. 8. Main page of the proposed COOP System

- **Delete Function:** Helps the participants to remove unwanted information from their account and also can avoid errors by using this function effectively.
- **Search Function:** It helps to find the required information from the bulk of materials in the web portal and helps in reducing the searching time for the user as shown in Fig. 9.



Fig. 9. Example for the Supervisor's Page

Whereas in the case of *non-functional requirements* it include flexibility, manageability and quality of the attributes involved in the system. These attributes are involved in controlling the authentication, security levels, usability, reliability and performance of the system.

- **Flexibility:** Here the flexibility does not mean that every participant can access entire information but must be able to get the defined information easily according to his status. Flexibility for accessing the information, sending and receiving the emails, uploading and downloading the information, etc.
- **Manageability:** Each participant must be able to manage the information according to his choice but without effecting or influencing the stipulated regulations. Students must be able to manage their search results, access the grades and reports from supervisors and trainers. Similarly, the supervisors and training coordinators must be able to manage the assessment reports and students performance summary to the seniors for necessary actions.

- *Quality:* With the flexibility provided for supervisors and training coordinators will help them to understand the performance of the students with time to time updates using this web based COOP training system.

### B. Usability Requirements

The usability requirements consider the easiness of system usage, matching the system with the requirements of users, effectiveness, efficiency and satisfaction towards security, performance and reliability.

### C. Preparing the Questionnaires for Collecting Primary Data

In this work the primary data was collected by conducting interviews with the COOP trainers and supervisors to understand their experience and suggestions. In the interview carried out with the supervisors revealed that 6 sections (of which 3 male and 3 female) from software engineering, computer science and computer engineering were participated in the COOP training. Also the opinion of all students was collected at the end to assess the performance of the proposed method and its performance.

## V. FINDINGS AND ANALYSIS

Most of the supervisors highlighted the issues related with the old system were included with paper work, communication gap between students and supervisors, requirement of additional COOP coordinators and finally the communication gap between academic supervisor and COOP trainer. One of the supervisors criticized for not having an appropriate structure to deal with the transactions and paper work always changes with supervisor to supervisor which in turn created too many issues at the time of assessment of student performance. Too much of paper work between

colleges and local companies involved serious challenges to deal with and always used to miss some of the important information or lead to the confusing scenarios. The suggested specifications for the web based COOP training system needs to include a user friendly atmosphere, security for the data, safety for the personal information, trustworthy and with good quality, good performance and must be maintainable by all the participants.

On the other hand, the students participated in COOP training from University of Hail are asked to participate in a survey for which the students are provided with brief survey questionnaire and the response of the students shown in Fig. 10 towards COOP training system with the questions as well. Most of the participants in this survey expressed their satisfaction towards the COOP training system and almost all the students enjoyed the training sessions at various industries according the survey results and are willing to attend it on the regular basis. However, some students criticized this method to have an impact on their academic grades. However, most of the students were very much positive towards the training system. Apart from that the assessment method using the web based COOP system helped the students to understand the accuracy of the system assessment. However, some people are having few doubts about the success rate of the COOP training system due to lack of computer literacy in the region of Hail and expressed most of the institution they come from are not well equipped with computer and internet facilities. However, all students agreed with the accuracy, quality and effectiveness of the proposed web based COOP system to provide the best results as compared to the existing method.

Most of the students participated in the COOP training got an advantage of expressing their opinions and feedbacks.

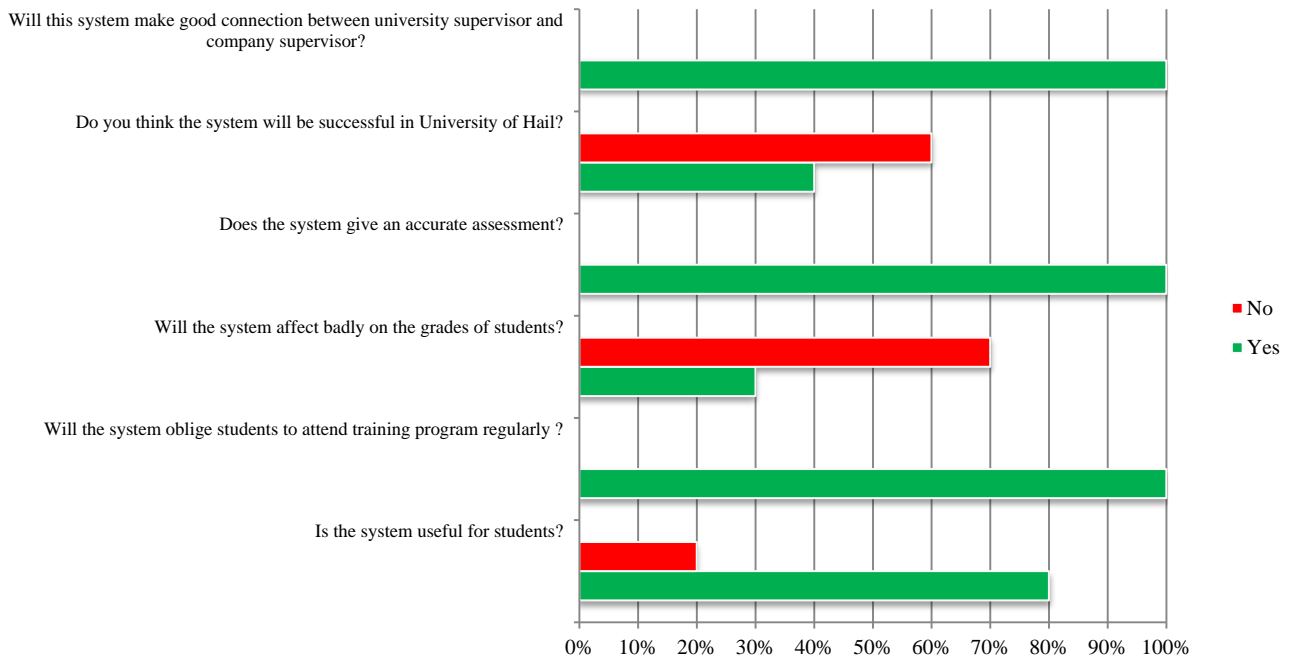


Fig. 10. Response of the students towards implementing the COOP training system.



## VI. CONCLUSIONS

The proposed method proves to be very accurate and delivered expected results by reducing the burden of communication gap between various participants. This method allowed all the participants to interact in an effective manner by using the proposed web based COOP system for respective activities and usage. Considering the inputs of various supervisors and training coordinators the present system has been developed to help all the participants.

### Merits of the proposed system

- The interaction or communication between various participants increased in an effective way.
- The companies can post their requirements and training instructions on the web portal by using the provided login account details.
- Eligible candidates list can be displayed by the college managements on the web portal for the view of both students and industries to select appropriate companies and candidates respectively.
- The paper work has been reduced and instructions are subjected to avoid overlapping of the information.
- Can consider any kind of review at any time by the management or by the participants involved within the system.
- The standard processing steps or procedures for all supervisors, training coordinators and student activities can be implemented.
- A crystal clear format for each step can be submitted and can reduce the man made errors to a greater extent.
- The supervisors and training coordinators can communicate the information of student performance or feedbacks with ease.

- In the case of indiscipline students the alerts can be raised and corrective actions can be implemented with ease.
- Project submission deadlines can be monitored by students and can submit their project reports online.
- A fair assessment of the project reports can be considered and the list of the successful candidates can be displayed on the web portal.
- Based on the previous reports, performances and feedbacks the corrective measures for the next COOP batches can be improved.

### ACKNOWLEDGMENT

The authors gratefully acknowledge the support and facilities provided by the Management and Department of Computer Science and Software Engineering, University of Hail.

### REFERENCES

- [1] CCSIT. CO-OP Training Guidelines. KSA – Ministry of Higher Education. 2014.
- [2] R. Echeng, A. Usoro and G. Majewski. "Acceptance of Web 2.0 in Learning in higher education: an empirical study of a Scottish university." In *WBC July Conference Proceedings on E-learning*. 2013.
- [3] A. M. Aleisa and M. A. Alabdulahfez. "Cooperative education at the Riyadh College of Technology: Successes and challenges." *Asia-Pacific Journal of Cooperative Education* 3, no. 2 (2002): 1-7.
- [4] A. S. Weber. "Web-based learning in Qatar and the GCC states." (2016).
- [5] UNESCO-UNEVOC. World TVT Database – Saudi Arabia. June 2012.
- [6] G. A. Khan. Technical training for half a million youth. In *Arab News*, April 2011.
- [7] M. M. Faiz and M. S. Al-Mutairi. "Assessment of a cooperative training course using faculty course assessment report in an ABET accredited engineering technology program." In *Frontiers in Education Conference (FIE), 2015 IEEE*, pp. 1-7. IEEE, 2015.

# Standard Intensity Deviation Approach based Clipped Sub Image Histogram Equalization Algorithm for Image Enhancement

Sandeepa K S , Basavaraj N Jagadale  
Department of Electronics  
Kuvempu University  
Karnataka, India

J S Bhat  
Department of Physics  
Karnataka University  
Karnataka, India

**Abstract**—The limitations of the hardware and dynamic range of digital camera have created the demand for post processing software tool to improve image quality. Image enhancement is a technique that helps to improve finer details of the image. This paper presents a new algorithm for contrast enhancement, where the enhancement rate is controlled by clipped histogram approach, which uses standard intensity deviation. Here standard intensity deviation is used to divide and equalize the image histogram. The equalization processes is applied to sub images independently and combine them into one complete enhanced image. The conventional histogram equalization stretches the dynamic range which leads to a large gap between adjacent pixels that produces over enhancement problem. This drawback is overcome by defining standard intensity deviation value to split and equalize the histogram. The selection of suitable threshold value for clipping and splitting image, provides better enhancement over other methods. The simulation results show that proposed method out performs other conventional histogram equalization (HE) methods and effectively preserves entropy.

**Keywords**—Standard intensity deviation; histogram clipping; histogram equalization; contrast enhancement; entropy

## I. INTRODUCTION

Digital imagery plays an important role in the fields of medical, industry, civil, security, astronomy, animation, forensic, web design. Image processing, helps human visual perception, visual quality and is used in many areas of image enhancement, image compression, image de-noising, image sharpening etc. Image enhancement is the process of making digital image more suitable for visualization or for further analysis and identifying key features of image. Contrast enhancement of an image is one of the well-known techniques in image enhancement. The contrast is created by the difference in luminance reflectance from two adjacent surfaces [1], [2] and enhancement is a technique of changing the pixel intensity of the input image. The quality of image contrast reduces, due to various factors, like of poor and ambient light conditions, aperture size and shutter speed of camera [3]. Histogram equalization is a technique that improves image contrast by adjusting image intensity and is used in wide range of applications as it is a simple method can be implemented easily. Its performance is limited because it tends to change the mean brightness of the image to the

middle of the gray level and creates undesirable effects that lead to over enhancement [4]. This method doesn't preserve image brightness because it is global operation and introduces the noise artifacts. To overcome this problem, different methods of histogram equalization were proposed to enhance the image brightness [5].

To preserve mean brightness of the image, Brightness preserving bi-histogram equalization (BBHE) method was proposed [6]. Here, mean value is used to bisect the histogram and then equalize both sub images independently. Another method, Dualistic sub image histogram equalization (DSIHE), follows BBHE and it differs in that it uses median value instead of mean to create sub images [7]. But, these methods fail to preserve the brightness of the image effectively. To improve the preservation of brightness, minimum mean brightness error bi-histogram equalization [MMBEBHE] was tried [8]. This method uses histogram separation based on threshold value and is an extension of BBHE, however, it fails to control the over enhancement.

To improve the visual quality of image, multi-histogram equalization approaches have come into existence. They are, recursive mean separate histogram equalization [RMSHE] [9], which performs BBHE recursively and recursive sub image histogram equalization [RSIHE], that performs division of histogram based median value [10]. From both methods, selection of number of iterations is an annoying issue and may lead to over enhancement. To overcome, over enhancement problem, histogram clipping approach was used and it helped in avoiding saturation effect and preserved the details of the image by controlling high frequency bins [11]-[13].

The exposure based sub image histogram equalization [ESIHE] performs enhancement of low exposure image by using exposure threshold value for image sub division [14]. This method equalizes, sub images individually and uses clipped histogram for controlling, over enhancement. It doesn't consider the variation of the exposure value from each gray value and stretches the contrast at high intensity region. This method is well suited for low exposure images and produce better visual quality images.

This paper refers to the ESIHE and proposes a contrast enhancement algorithm by defining new standard intensity deviation value instead of exposure threshold value. This

improves the effect of enhancement in terms of average information contents. The performance of proposed method was analyzed by enhancing low exposure images and low exposure underwater images. We find that the proposed method helps in enhancing contrast, visual quality and average information contents of the images, as compared to others methods.

The paper is structured as follows: Section 2 describes the proposed method. Experimental results and discussion are given in Section 3 and the conclusion is given in Section 4.

## II. PROPOSED ALGORITHM

The conventional histogram equalization methods improve the image contrast by stretching the dynamic range of the image using cumulative distribution function. Fig. 1(a), (c), (e) and (g) represent original image, HE image, BBHE image and ESIHE image, respectively. Fig 1(b), (d), (f) and (h) are the histograms of respective images. Fig. 1(c), (e) and (g) illustrates over enhancement problem, where, road, top of the truck and soil portions of image texture are very bright, which is highlighted in red square boxes. The reason for this problem is large gap between two adjacent gray values of the histogram (Fig. 1(d), (f) and (h)). The gap denotes the number of gray levels between two neighboring gray values, and large gap between the adjacent pixels leads to over enhancement problem [15].

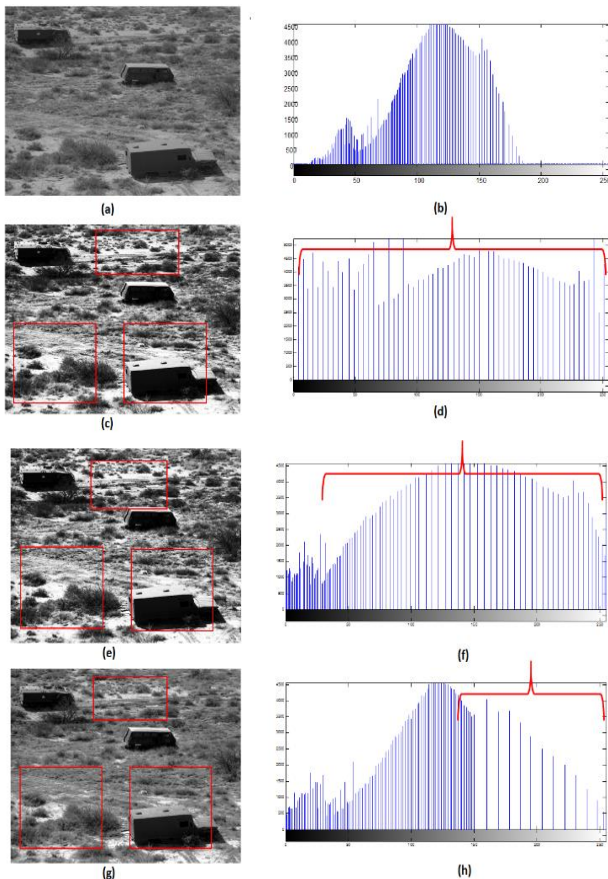


Fig. 1. Result of field image and its histogram (a) original image; (b) original image histogram; (c) HE image; (d) HE histogram; (e) BBHE image; (f) BBHE histogram; (g) ESIHE image; (h) ESIHE histogram.

From Fig. 1, it is found that, gap between adjacent pixels, affect the enhancement quality and selection of threshold value to divide the image histogram also plays an important role. To overcome these problems explained earlier, the standard intensity deviation based clipped sub image histogram equalization (SIDCSIHE) algorithm is presented by defining new threshold value. This value is being used to divide the image histogram, as it helps to enhance image contrast by preserving maximum information and minimizing the gap between adjacent pixels. The algorithm consists of three steps, namely standard intensity deviation value calculation, histogram clipping and histogram equalization.

### A. Standard Intensity Deviation Value Calculation

To measure volatility of the image intensity, the standard deviation function  $\sigma$ , by finding the variance between corresponding intensity and mean image histogram is given by (1) [16].

$$\sigma = \left( \frac{\sum_{i=1}^L (i - H_{\mu})^2 xH(i)}{\sum_{i=1}^L H(i)} \right)^{1/2} \quad (1)$$

The mean image histogram of the low contrast image is given by (2).

$$H_{\mu} = \frac{\sum_{i=1}^L H(i) i}{\sum_{i=1}^L H(i)} \quad (2)$$

where  $H(i)$ , is image histogram with its corresponding intensity  $i$  and  $L$  is its total number of gray levels. The normalized standard deviation value is expressed as in (3) and its range is  $[0 \ 1]$ . Another parameter  $X_{SID}$  is defined in (4), by using normalized standard deviation value and it also used to modify each input gray level by dividing the image into two sub images.

$$\sigma_{norm} = \left( 1 - \left( \frac{\sigma}{L} \right) \right) \quad (3)$$

$$X_{SID} = L * \sigma_{norm} \quad (4)$$

### B. Histogram Clipping

The process of clipping histogram controls the enhancement rate. The threshold value  $T_C$  given in (5) is calculated as an average number of grey level occurrences. The histogram bins are clipped, and is greater than the clipping threshold as given by (6). The histogram is then clipped by clipping threshold as shown in Fig. 2(b).

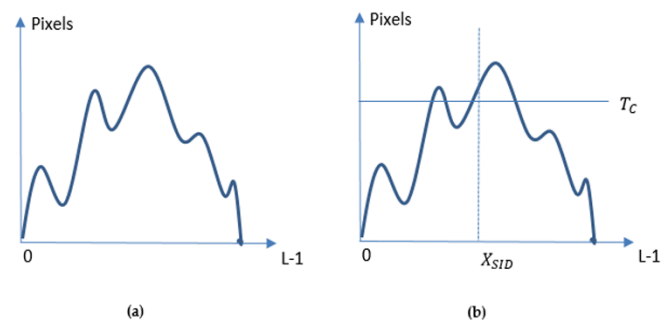


Fig. 2. Sub division process (a) Image Histogram sub division and clipping; (b) Image Histogram sub division and clipping.

$$T_c = \text{mean}(H(i)) \quad (5)$$

$$H_c(i) = T_c \text{ for } H(i) \geq T_c \quad (6)$$

where  $H(i)$  and  $H_c(i)$  are the original and clipped histogram, respectively. The histogram clipping consumes lesser time as it needs less number of computations [14].

### C. Clipped Histogram Sub Division and Equalization

Based on standard intensity deviation value  $X_{SID}$ , the clipped histogram is divided into two sub images  $I_{low}$  and  $I_{up}$  with ranges varying from 0 to  $X_{SID}$  and  $X_{SID}+1$  to  $L-1$  respectively. The probability of these two sub images are  $P_{low}(i)$  and  $P_{up}(i)$ , respectively

$$P_{low}(i) = \frac{H_c(i)}{N_{low}} \text{ for } 0 \leq i \leq X_{SID} \quad (7)$$

$$P_{up}(i) = \frac{H_c(i)}{N_{up}} \text{ for } X_{SID} \leq i \leq L-1 \quad (8)$$

where  $N_{low}$  and  $N_{up}$  are total number of pixels in each sub images and its cumulative distribution function  $C_{low}(i)$ ,  $C_{up}(i)$  can be defined as:

$$C_{low}(i) = \sum_{i=0}^{X_{SID}} P_{low}(i) \quad (9)$$

$$C_{up}(i) = \sum_{i=X_{SID}+1}^{L-1} P_{up}(i) \quad (10)$$

The histogram equalization is done individually for two sub images using the transfer function  $F(i)$  as expressed in (11):

$$F(i) = \begin{cases} \frac{C_{low} * X_{SID}}{(X_{SID} + 1)} & \text{for } 0 \leq i \leq X_{SID} \\ \frac{C_{up} * (L - X_{SID} + 1)}{(L - X_{SID} + 1)} & \text{for } X_{SID} + 1 \leq i \leq L - 1 \end{cases} \quad (11)$$

The sub images are combined into one final image by the transfer function  $F(i)$  for further analysis. Fig. 3(a) is the processed image appears cleaner and overcome over enhancement in the highlighted boxes. Fig. 3(b) represents its histogram, has overcome the gap between the adjacent pixels which lead to better image quality.

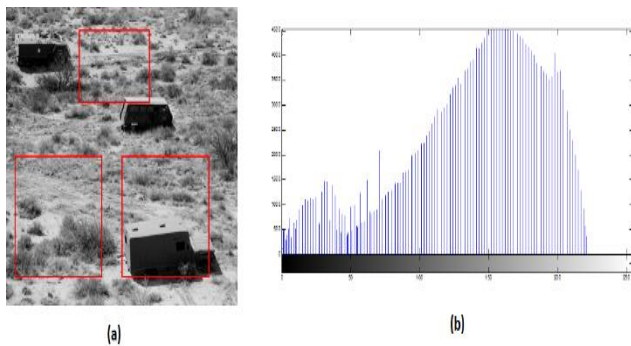


Fig. 3. Result of field image and its histogram (a) SIDCSIHE image; (b) SIDCSIHE image histogram.

### D. Algorithm of SIDCSIHE

- Compute input image histogram  $H_i$ .
- Compute standard intensity deviation value and threshold parameter  $X_{SID}$ .
- Compute Histogram clipping  $H_c(i)$  based clipping threshold  $T_c$ .
- The input clipped histogram splits into two sub images based  $X_{SID}$ .
- Histogram Equalization is applied on individual sub image histogram.
- Integrate both images into final image.

## III. RESULTS AND DISCUSSION

The pre-eminence of the proposed method is illustrated by comparing both objective and subjective assessments with well-known methods.

### A. Objective Assessment

To verify the effect of enhancement, the objective assessment has been carried out and compared on the basis of entropy. Entropy means average information content that is the measure of richness of image details. A higher value indicates the availability of more information content and is perceived to have better quality of the image. Equation (12) defines entropy [17].

$$\text{Entropy}(p) = - \sum_{k=0}^{L-1} p(k) \log p(k) \quad (12)$$

where  $p(k)$ , is probability density function at the intensity level  $k$  and  $L$  is total number of gray levels of the image.

The different types of 15 test images with low exposure underwater images and low exposure images are used. The performances of the proposed method is compared with other existing methods HE, BBHE, DSIHE, MMSICHE, NMHE and ESIHE for better entropy results. From Table I, it is evident that the proposed (SIDCSIHE) method has the highest entropy values as compared to other methods and its value is very near to the input average entropy value (6.269), which indicates that more information is extracted from an input image.

TABLE I. AVERAGE ENTROPY VALUE AND EXECUTION SPEED OF COMPARISION METHODS

Methods	Entropy	Execution speed in sec
HE	5.534	0.0688
BBHE	6.139	0.0888
DSIHE	6.120	0.0818
MMSICHE	6.193	1.2493
NMHE	6.095	0.1435
ESIHE	6.221	0.1465
SIDCSIHE	<b>6.227</b>	<b>0.0966</b>

To check its robustness, the proposed method execution time is compared with other methods because most of the studies focus on image quality as well as execution time. The execution of the method is carried out on the computer with 64bit, windows 8 and Intel i3 processor with 4GB RAM. The average execution time as compared to other methods has been tabulated in Table I. The execution time of the proposed method is compared with advanced methods MMSICHE, NMHE and ESIHE as the conventional methods HE, BBHE and DSIHE does not possess better visual quality.

**B. Subjective Assessment**

The performance of algorithm in contrast enhancement, provides the information about over enhancement and unnatural look of the image, by inspecting the visual appearance and corresponding histogram. Although objective assessment provides quantitative information but its quality evaluation can be accomplished by subjective assessment and this is the most direct approach to judge the quality of image from an observer. To prove the robustness and versatility of the proposed methods, the standard images are chosen from different fields, like, underwater images and low exposure images (tank, fish, elk, plane, sanctuary) as shown in Fig. 4 to 8.

Fig. 4(a) is a low contrast image. Fig. 4(b), (c) and (d) are processed by HE, BBHE and DSIHE leads over enhancement, can be seen in the visual appearance and corresponding histogram. Fig. 4(e), (f) are the results of MMSICHE and NMHE individually. These images seem to be gloomy and Fig. 4(f) is too dark, as compared to other images. Fig. 4(g), processed by ESIHE, has better enhancement over other methods. However, its histogram detail seems to equalize more in lower intensity region due to which the tank appears dark as compared to Fig. 4(h) proposed by SIDCSIHE. The image in Fig. 4(h) looks more natural with better visual quality than other methods.

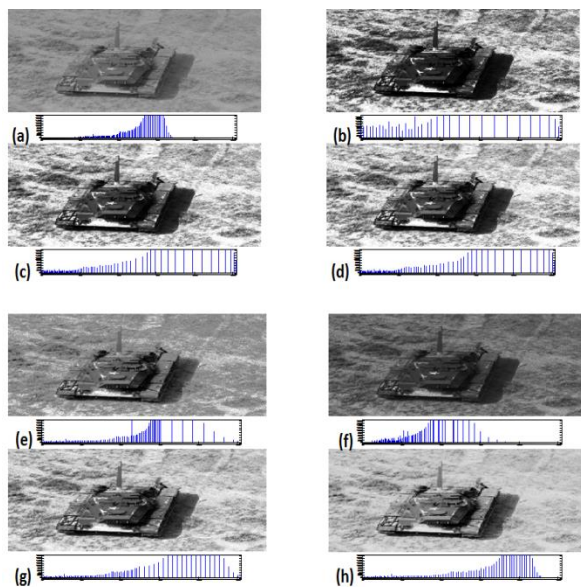


Fig. 4. Result of Tank image (a) original image; (b) HE image; (c) BBHE image; (d) DSIHE image; (e) MMSICHE image; (f) NMHE image; (g) ESIHE image; (h) SIDCSIHE image.

Fig. 5(a) is an example of underwater low exposure fish image. Fig. 5(b) is processed by HE, which enhances the image in great way but the white pebbles of image have become brighter and cannot be visible clearly.

Fig. 5(c), (d), (e) and (f) are processed by BBHE, DSIHE, MMSICHE and NMHE methods respectively and the methods fail to distinguish the fish from the background. Fig. 5(g), (h) are the results of ESIHE and SIDCSIHE and the images have better enhancement over other methods.

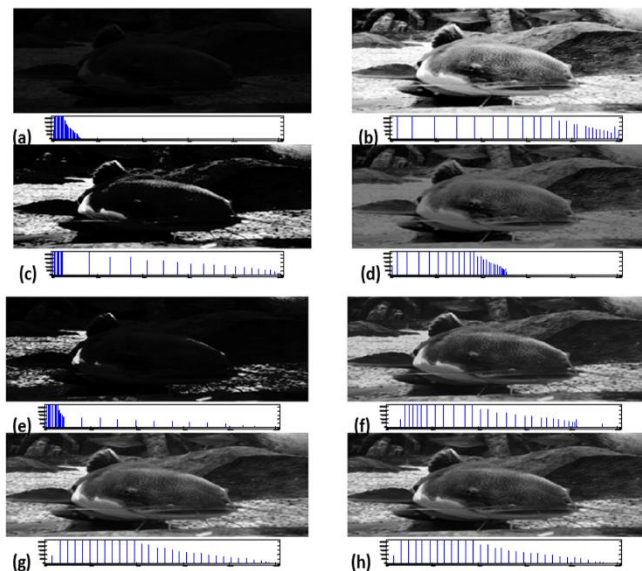


Fig. 5. Result of under water fish image (a) original image; (b) HE image; (c) BBHE image; (d) DSIHE image; (e) MMSICHE image; (f) NMHE image; (g) ESIHE image; (h) SIDCSIHE image.

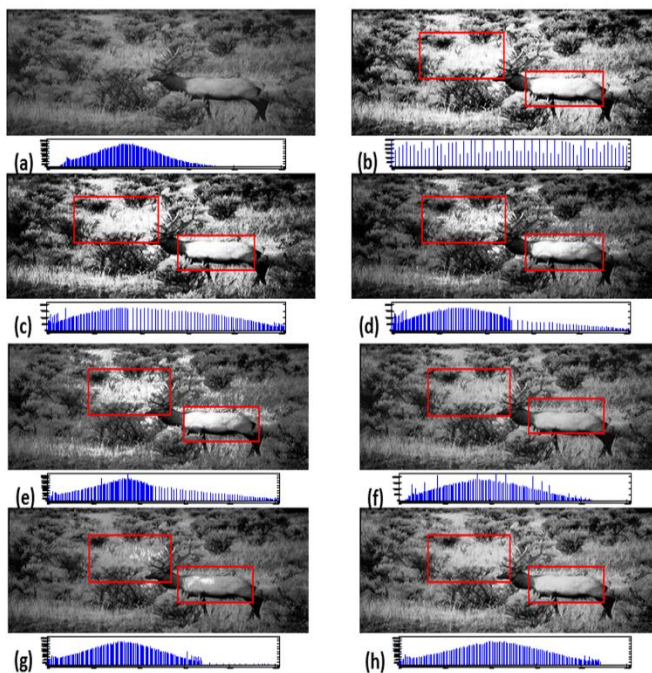


Fig. 6. Result elk image (a) original image; (b) HE image; (c) BBHE image; (d) DSIHE image; (e) MMSICHE image; (f) NMHE image; (g) ESIHE image; (h) SIDCSIHE image.

The low contrast image is shown in Fig. 6(a). The supremacy of the proposed method (Fig. 6(h)) can be analysed by comparing the Fig. 6(b), (c), (d), (e) and (g) of HE, BBHE, DSIHE, MMSICHE and ESIHE respectively. Due to over enhancement problem information is unclear especially in the skin of elk and grass, which is highlighted square boxes. Fig. 6(f) is the result of applying NMHE, appears dark as compared to Fig. 6(h) obtained by the proposed SIDCSIHE method. In addition to that, Fig. 6(h), have better contrast and visual quality resembling its natural look.

Fig. 7(a) is low contrast plane image. Fig. 6(b), (c), (d), (e) and (g) are results of HE, BBHE, DSIHE, MMSICHE and ESIHE. The texture of the plane and surrounding cloud area in the images are not clear and have unpleasant visual artefacts. Fig. 7(f) processed by NMHE has clear look but lacks the visual effect as compared to Fig. 7(h) the result of proposed method. The proposed method gives image with clear outer surface, which is highlighted in square boxes. The resulting image of the proposed method have natural look by preserving maximum information.

Fig. 8(a) is low contrast sanctuary image. Fig. 8(b)-(g) are results of HE, BBHE, DSIHE, MMSICHE, NMHE and ESIHE, respectively. The methods fail to enhance the texture of mountain, grass and road in all the images except Fig. 8(f) which is highlighted in the square boxes. The same texture information in SIDCSIHE fig 8h is clearly visible and the image looks more natural without any unpleasant artefact.

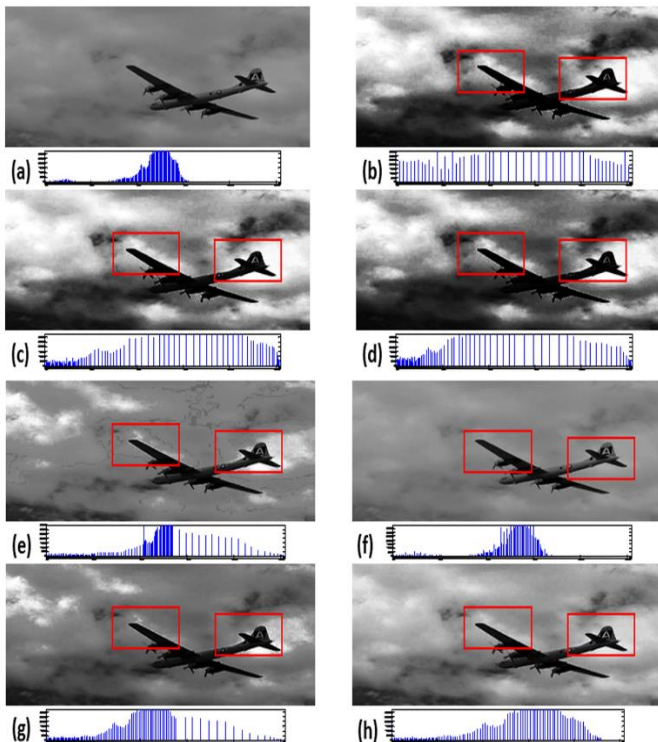


Fig. 7. Result of plane image (a) original image; (b) HE image; (c) BBHE image; (d) DSIHE image; (e) MMSICHE image; (f) NMHE image; (g) ESIHE image; (h) SIDCSIHE image.

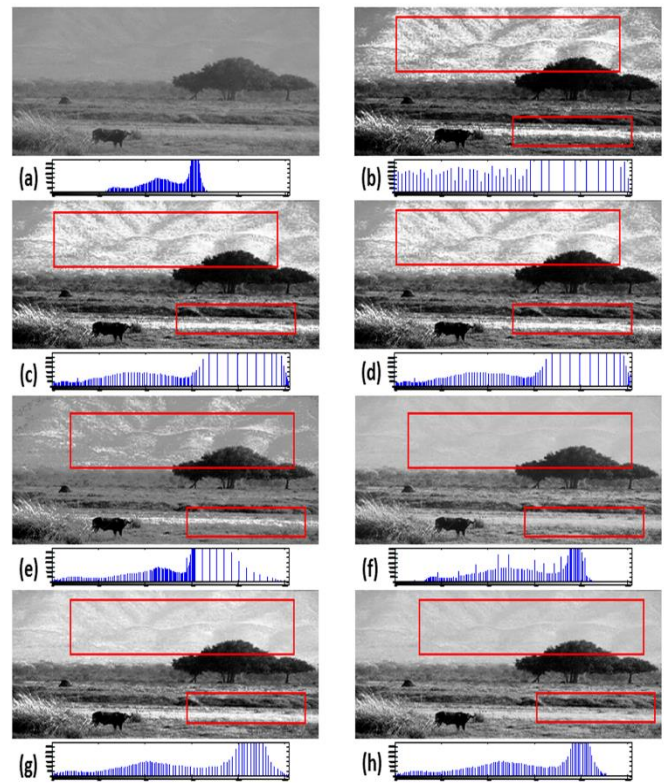


Fig. 8. Result of sanctuary image (a) original image; (b) HE image; (c) BBHE image; (d) DSIHE image; (e) MMSICHE image; (f) NMHE image; (g) ESIHE image; (h) SIDCSIHE image.

#### IV. CONCLUSION

The proposed method has promising performance in terms of both entropy and overall visual quality. The selection of standard deviation intensity value provides new optimal threshold value to split the clipped histogram and equalize sub image effectively and gives control on over enhancement rate. The visual quality, entropy value and execution speed shows the robustness of the proposed method as compared to existing algorithm for low exposure of images.

Looking at the efficiency of the proposed method by protecting detailed information, especially for low exposure underwater image, the proposed method can be extended further to improve the average information by decomposing histogram into multiple segment based on suitable threshold and equalizing each segment independently.

#### ACKNOWLEDGMENT

This research work is supported by UGC-MRP, New Delhi, India.

#### REFERENCES

- [1] Y. S. Chiu, F. C. Cheng and S. C. Huang, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution", IEEE Transactions on Image Processing, **2013 Mar**, 22(3), 1032-41.
- [2] A. Raju, G. S. Dearakish and A. Venkat Reddy, "comparative analysis of histogram equalization based technique for contrast enhancement and brightness preserving", International journal of signal processing, image processing and pattern recognition. **2013**, 6, 353-366, 10.14257.

- [3] M. Hanmandlu, O. P. Verma, N. K. Kumar and M. A. Kulkarni, “novel optimal fuzzy system for color image enhancement using bacterial foraging”, IEEE Transactions on Instrumentation and Measurement, **2009 Aug**, 58(8):2867–79.
- [4] Gonzalez, R.C, R. Woods, Digital Image Processing, 3rd ed, Pearson, INDIA, 2014, pp. 144–166.
- [5] R. Garg, B. Mittal and S. Garg. “Histogram equalization technique for image enhancement”, IJECT. **2011 Mar**, 2(1), 107-11.
- [6] Y. T. Kim, “Contrast Enhancement using Brightness preserving bi-histogram equalization”, IEEE Trans. Consumer Electron, (1997), 43: 1-8.
- [7] Y. Wan, Q. Chen and B. M. Zhang. “Image enhancement based on equal area dualistic sub image histogram equalization method”, IEEE Trans. Consumer Electron, **1999**, 68-75.
- [8] S. D. Chen and A. R. Ramli. “Minimum mean brightness error bi-histogram equalization in contrast enhancement”, IEEE Trans. Consumer Electron, 2003, **49(4)**, 1310-1319.
- [9] S.D. Chen and A.R Ramli, “Contrast enhancement using recursive mean separate histogram equalization for suitable brightness preservation”, IEEE Trans. Consumer Electron, 2003, 49, 1301-1309.
- [10] K. S. Sim, C.P. Tso and Y. Y. Tan, “Recursive sub image histogram equalization applied to gray scale images”, Pattern Recognition Letters, **2007**, 28, 1209-1221.
- [11] T. Kim and J. Paik, “ Adaptive contrast enhancement using gain controllable clipped histogram equalization”, IEEE Trans. Consumer electron, ( **2008**), 54 (4), 1803-1810.
- [12] C.H. Ooi and N.S.P. Kong, “Ibrahmin, H. Bi-histogram equalization with plateau limit for digital image enhancement”, IEEE Trans. Consumer Electron ,( **2009**), 55 (4), 2072-2080.
- [13] Q. Wang and R. K. ward, “ Fast image/video contrast enhancement based on weighted threshold histogram equalization”, IEEE Trans. Consumer Electron ,( 2007), 53(2), 2072-2080.
- [14] K. Singh Kapoor R. “Image enhancement using exposure based sub image histogram equalization”, Pattern Recognition Letters. 2014, 36(1), 10–4.
- [15] Chung-Cheng Chiu and Chih-Chung Ting, Contrast enhancement algorithm based on Gap adjustment for histogram equalization, Sensors 2016, 16, 936; doi:10.3390/s16060936.
- [16] Shin-Chia Huang and Chien-Hui Yeh. “Image contrast enhancement for preserving mean brightness without losing image features”. Engineering Application of Artificial Intelligence, ( 2013), 26:1487-1492.
- [17] S. D. Chen. “ A new image quality measure for assessment of histogram equalization based contrast enhancement”, Digital signal processing; **April**, 22:640-647.

# Quality Ranking Algorithms for Knowledge Objects in Knowledge Management Systems

Amal Al-Rasheed

Information Systems Department  
Princess Nourah Bint Abdulrahman Univesity (PNU)  
Riyadh, Saudi Arabia

Jawad Berri

Information Systems Department  
King Saud University (KSU)  
Riyadh, Saudi Arabia

**Abstract**—The emergence of web-based Knowledge Management Systems (KMS) has raised several concerns about the quality of Knowledge Objects (KO), which are the building blocks of knowledge expertise. Web-based KMSs offer large knowledge repositories with millions of resources added by experts or uploaded by users, and their content must be assessed for accuracy and relevance. To improve the efficiency of ranking KOs, two models are proposed for KO evaluation. Both models are based on user interactions and exploit user reputation as an important factor in quality estimation. For the purpose of evaluating the performance of the two proposed models, the algorithms were implemented and incorporated in a KMS. The results of the experiment indicate that the two models are comparable in accuracy, and that the algorithms can be integrated in the search engine of a KMS to estimate the quality of KOs and accordingly rank the results of user searches.

**Keywords**—Knowledge Management System (KMS); Knowledge Object (KO); knowledge evaluation; quality indicator; recommender system

## I. INTRODUCTION

The ever-increasing volume and diversity of knowledge in Knowledge Management Systems (KMSs) has required users to spend more time searching for the information they need. Searches of such knowledge repositories often yield a large number of results, making it difficult for users to choose items that will actually meet their requirements [1]-[3]. Ranking of knowledge objects (KOs) in search results is based on measurement of the degree of similarity between the query submitted by the user and topics in the knowledge repository, regardless of any consideration of quality [4]. Without an evaluation process that can determine the relevance significance and quality of KOs, most searches will be weak and of limited benefit [5]. For that reason, some knowledge bases have resorted to the use of expert evaluations. Although these are efficient, they necessarily encompass only a limited number of KOs because of the limited number of experts and the tediousness of manual evaluation [6]. Moreover, these evaluations are implemented individually, which limits their validity, owing to bias and differences of opinion. As a consequence, when searching KOs, resources that have not been evaluated will appear at the bottom of the list of search results, regardless of their actual quality.

To overcome the problem of the large number of KOs that remain unevaluated, there is a need to measure their quality automatically. To alleviate the problem of unbiased

evaluations, the evaluation process needs to be based on collaboration, through which participants converge on more accurate evaluations [7]. To date, few studies have focused on the issue of quality of information in virtual communities. As members of these communities have more freedom to add new KOs, the quality of available knowledge tends to be lower than in knowledge repositories in organizations [8]. This paper proposes a general framework for quality evaluation that includes two models of quality measurement. The first of these recommends KOs for online communities on the basis of quality indicators that are grouped into four dimensions. The second model is based on estimates of user reputation and exploits a content-based recommender technique for estimating the quality of KOs.

The remaining of this paper is structured as follows. Next section provides a review of research work in the field. Section 3 proposes two models for KO evaluation. Section 4 explains the data set, evaluation metrics, evaluation procedure and the result of the experiment. Then, Section 5 illustrates the application of the models through a case study. Finally, last section concludes this paper and presents future work that can be done.

## II. BACKGROUND AND RELATED WORK

### A. User Reputation

According to the American Heritage dictionary, reputation is “the general opinion or judgment of the public about a person or thing” [9]. Reputation scores are utilized to motivate users to actively participate [10]. To enable users to comprehend them easily, these scores are generally simple and count-based. In practice, reputable users are considered to be among the most important assets of websites [11]. The present study focuses on user reputation in an online collaborative KMS, where users contribute, share, and rate knowledge. Online collaboration has become an important means of creating and organizing knowledge, but the approach presents challenges for both content creation and content use [12]. The process of content creation is open to abuse, and content consumers may have difficulty in distinguishing between high- and low-quality content. Reputation systems can help to prevent abuse and bring order to indications of content quality. One of the main objectives of a reputation system for collaborative content is to provide indications of content quality to users [12]. Reputation scores are computed



according to the quality and quantity of contributions made by individual users [10].

### B. Knowledge Evaluation

The process of evaluating knowledge quality is difficult and complicated because multiple aspects must be considered. To the best of our knowledge, few existing studies have focused on the automatic evaluation of the quality of KOs; in this section, we review those of immediate relevance. One important evaluation and ranking algorithm is EigenRumor, which uses link analysis to calculate scores for community contributions, based on links from contributors to information objects. These scores can be used to classify information and contributors and are used as incentives to stimulate ongoing contributions to the community [13]. In [6], another evaluation algorithm assesses the value of knowledge and contributors according to common contribution actions, in which the dissimilar evaluation capabilities of contributors are reflected in the weightings of evaluated items. In [14], a general model is advanced for the automatic calculation of reputation scores based on the ratings given to knowledge resources, integrating these reputation scores to create value-added information about the rated resources. Other studies in related fields regarding evaluation of online resources are also of relevance here; for instance, in the context of e-learning, Ochoa and Duval proposed a number of quality metrics for the automatic evaluation of metadata characteristics [15]. In a subsequent study [16], the same authors proposed another set of metrics for deriving measures of personal, topical, and situational relevance [17].

## III. PROPOSED EVALUATION MODELS

### A. Quality Indicators-Based Evaluation Model (QIEM)

The first proposed model captures four dimensions of quality to recommend knowledge objects to users. To our knowledge, this model is the first to consider contributor reputation as a measure of quality to enhance evaluation of KOs. We propose that reputation score can enhance quality and alleviate the effects of shortages of ratings or usage that undermine other models. The aggregation of all the scores from previous indicators support computing a general rating for a specific KO.

1) *Quality indicators*: Quality indicators are “statistical measures that give an indication of output quality;” the quality of outputs is best defined in terms of how well outputs address user needs, or whether they fit the user’s purpose [18]. Given the increasing importance of properly defined quality indicators in the knowledge management field, academics from Loughborough University developed and facilitated a workshop entitled “The Use of Indicators for Monitoring and Evaluation of Knowledge Management and Knowledge Brokering in International Development,” in association with the Institute for Development Studies. The workshop brought together thirty knowledge researchers and academics from twenty organisations to discuss quality indicators, and a resource pool of 100 such indicators are presented in the workshop report [19]. Of these, the indicators for quality of knowledge in virtual community include the following:

- Number of created knowledge objects
- Percentage of users who rate knowledge objects
- Number of citations of knowledge objects
- Number of downloads of knowledge objects
- Number of views of knowledge objects
- Percentage of readers of knowledge objects
- Number of items of relevance to one’s work
- Number of channels that provide a knowledge object
- Availability of discussions of knowledge objects
- Number of recommendations of knowledge objects
- Usefulness of knowledge objects as perceived by target audience (5-point Likert scale)
- Number of examples where work has been cited.

It is noteworthy that while the majority of indicators are quantitative measures of how much or how many, relatively few are qualitative indices of how or why. In practice, indicators gain in strength when used as part of a basket of indicators [19]. For that reason, we will use some indicators from the above list that meet our requirements here and benefit from other indicators in adjacent fields, integrating them in a quantitative measure that serves to clarify the level of KO quality in online communities.

2) *Quality indicators for KO evaluation*: In evaluating the quality of KOs, four major dimensions will be taken into account: social, usage, characteristic, and contributor (Fig. 1).

#### a) Indicators of social quality

Social indicators are metrics that track users’ explicit feedback and interaction [20]. Common forms of explicit feedback include comments, vote up, and vote down, as well as star-based ratings [21]. Previous examples of this kind give a good indication of the quality of content [22]. Explicit rating is the most effective way of capturing the user’s judgments of KOs because it reflects the user’s own evaluation of importance and quality. However, explicit feedback requires users to perform extra rating actions, which users may find inconvenient. Arising from this, the main limitations of explicit rating are that (i) new KOs that have not been evaluated appear at the bottom of the list of search results, as if their quality is low; and (ii) users have little interest in rating content because of a lack of incentives. To mitigate these limitations, we assign a neutral rating (3 out of 5) to new KOs until they attract user evaluations. Moreover, according to their interaction with the application, we award users extra points to encourage them to evaluate KOs.

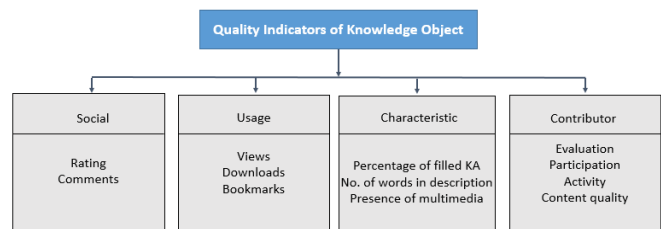


Fig. 1. Proposed indicators for knowledge object evaluation.

b) Indicators of usage quality

Usage indicators are metrics that track users' implicit feedback and behaviors. Claypool et al. reported that implied data acquired from user behaviour is very effective for sorting lists of search results [23]. As implicit feedback is based on search behaviour, there are many possible sources of such information. A number of studies have looked at the classification of possible sources of implicit feedback [23]-[26]. Implicit feedback systems commonly use such measures as document reading time, interaction, and scrolling, as these measures reflect the concerns of users and their satisfaction level and cost less than explicit evaluation [23], [27]. However, these systems are built on the assumption that relevant documents will be viewed and interacted with more than those that are less relevant.

In the particular case of knowledge bases. Data such as popularity, number of views, or number of bookmarks can be utilized to complement information on the quality of knowledge objects.

c) Indicators of characteristic quality

The characteristic dimension includes quality indicators based on the capability of information to describe a knowledge attribute. Some quality characteristics proposed in [28] include accuracy, provenance, completeness, consistency and coherence, timeliness and accessibility, and conformance to expectations. For instance, to evaluate the completeness of a knowledge attributes record (quality characteristic), we can check how many attributes have been filled with information (metric).

It is important to note that these metrics relate to the quality of knowledge attributes but not to the quality of the KO itself. The completeness metric has been selected for present purposes, as it is convenient to implement for the available information in real knowledge repositories. Completeness is the extent to which knowledge attributes contain all the information needed to provide a complete representation of the labelled resource.

When all the attributes have non-null values, the value of this metric will be 1 (maximum value); in cases of an empty resource, the value will be 0 (minimum value). While this seems straightforward, not all knowledge attributes are equally relevant for all resources. For this reason, it is better to use a weighting factor to illustrate the significance of the attribute.

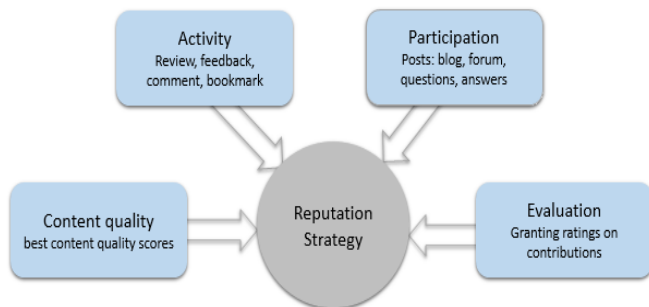


Fig. 2. Reputation strategy.

To measure the characteristic dimension, we define indicators based on the more general indicators discussed in the preceding section. In addition to the number of filled attributes, we added the number of words in the description attribute and the presence of multimedia.

d) Indicators of contributor quality

Reputation is defined by Alfarez et al. as “an expectation about an agent’s behaviour based on information about or observation of its past behaviour” [29]. A number of previous studies have demonstrated that user reputation is a good indicator of the reliability and quality of content [30]. Within an online community, users can build their reputation through condensing their activities. User reputation score can be calculated by reference to features that denote the user’s authority and influence within the community [31]. In the present case, we have identified four features to determine user reputation score (Fig. 2): contributor’s evaluation, participation, activity, and content quality. During the evaluation process, there is a clear association between a contributor’s reputation score and knowledge value score, as contributors with high reputation scores are likely to supply the knowledge repository with high quality KOs [32]. The four features used to estimate user’s reputation can be described as follows:

- Evaluation: The average of all ratings of the contributor for their contributions
- Participation: The quantity of contributions made by the user, weighted according to type (article, blog, question, answer, forum post)
- User activity: The weighted average of three metrics: user ratings, user bookmarks, and user comments
- Content quality: The average of the contributor’s best content quality scores

3) Evaluation model: Fig. 3 illustrates inputs and outputs of the evaluation system. Using collected data for predefined quality indicators and statistical measures, the quality manager assigns a score to each search result. First, values are individually calculated for each of the four indicators. These values are then added together, using their respective (and configurable) weighting percentages. After that, they are normalized using a scale between 0 and 1 to adjust for the disparity in the values measured on different scales. Finally, quality manager automatically re-rank the search result according to the calculated scores.

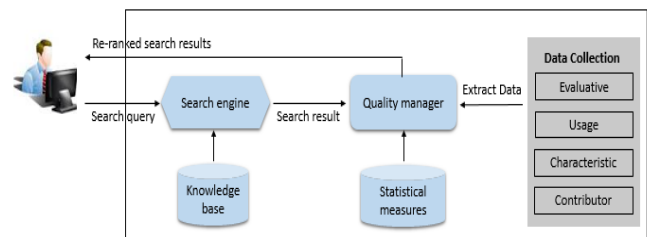


Fig. 3. Evaluation system.

The resulting score is an outline of numerical values assigned to all predefined quality indicators [14]. Measurement of the estimated quality score entails the following steps:

**Step 1: Normalize indicators.** The original data for all indicators should be normalized to eliminate the influence of any dimension over the others. The normalized value of  $e_i$  for indicator E is calculated as follows:

$$\text{Normalized } (e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}} \quad (1)$$

where:  $E_{min}$  = minimum value for indicator E and  $E_{max}$  = maximum value for indicator E If  $E_{max}$  is equal to  $E_{min}$  then Normalized ( $e_i$ ) is set to 0.5.

**Step 2: Weight indicators.** To measure the effect of indicators in an evaluation system, assigning weights to all indicators is an essential procedure. An indicator whose weight is high will exert a greater effect on overall quality; otherwise, its effect is lower. In statistical terms, because standard deviation measures the distribution of numbers, it is one of the best weighting methods. The basic principle of standard deviation is that when the data of one indicator present large differences among multiple evaluated objects, the standard deviation of this indicator must be high [33]. The value of an indicator's standard deviation is directly proportional to its contribution to the integrated formula of all indicators.

Standard deviation determines the weights of indicators by means of the following equations:

$$w_j = \frac{\sigma_j}{\sum_{j=1}^n \sigma_j} \quad j = 1, \dots, n \quad (2)$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^m (x_{ij} - x_j)^2}{m}} \quad j1, \dots, n \quad (3)$$

where  $w_j$  is the weight of a criteria and  $\sigma_j$  is the standard deviation. Table I sets out the weight assigned to each quality indicator using this method.

**Step 3: Integrate indicators.** Now, the information about quality indicators can be integrated in one score, where each indicator makes a measured contribution to overall quality. The estimated quality score combines all quantitative information about quality indicators of a KO, which means that if a quality indicator does not exist, the estimated score can be calculated automatically from the existing indicators. Moreover, the new KOs will be assigned a neutral rating (3 out of 5) to resolve the critical problem of new KOs without ratings appearing at the end of the list of search results, increasing the reliability of recommendations. Estimation of the quality score of a knowledge resource is described in (4):

TABLE I. STANDARD DEVIATION VALUES AND WEIGHTS OF QUALITY DIMENSIONS

	Social	Usage	Characteristic	Contributor
SD	0.144822	0.299779	0.096784	0.07456
Weighted SD	0.235122	0.486698	0.157132	0.121049

$$\text{Score} = \sum_{i=1}^m a_i * \text{Social} + \sum_{j=1}^n b_j * \text{Usage} + \sum_{k=1}^o c_k * \text{Characteristic} + \sum_{l=1}^p d_l * \text{Participant} \quad (4)$$

where a, b, c, and d represent the respective weights of social, usage, characteristic, and participant indicators, and m, n, o, and p represent the indicator number in each quality dimension. Where any of these data are missing, the weights are adapted to compensate for this absence in calculating estimated quality. In addition, all indicators are normalized by scaling between 0 and 1 as described in (2), and their mean values are included in the final score.

### B. Reputation-Based Evaluation Model (REM)

This second model assesses the quality of knowledge objects automatically by exploiting the capabilities of recommender systems and user reputation scores. The proposed model is based on the concept of recommending KOs that are similar in content and specifying KO quality on the basis of ratings posted by reputable users to help other users to select the best KOs. This model is based on the premise that users with high reputation points are more reliable in evaluating KOs. Intuitively, as reputable users can be expected to submit high quality contributions and to attract high ratings from the user community, users can benefit from reputation scores identifying good contributions [10].

1) *Evaluation model and algorithm:* Recommending KOs involves two phases. In the first phase, relevant KOs are retrieved according to the keywords entered by the user. The second phase re-ranks search results according to the estimated quality score for each KO.

Algorithm 1 describes the reputation-based evaluation approach.

#### Algorithm 1. Reputation-based evaluation

**Input:**  $S$  = vector of KOs, as returned by search engine  $S$

$UP$  = users' profiles

$R$  = ratings of KOs

**Output:** Vector of KOs, re-ranked according to  $SCORE$

**For each**  $KO \in S$

Set  $n = 0$

**For each** rater  $I$

Set  $E$  = the average of all ratings the rater  $I$  gains

Set  $P$  = the number of contributions made by the rater  $I$

Set  $A$  = the weighted average of: user ratings, user bookmarks, and

user comments

Set  $Q$  = the average of best content scores the rater  $I$  gains

$REP = (E + P + A + Q)$

**If**  $REP > threshold$  **Then**

$n = n++$

**End For**

**If**  $n \geq 3$

$SCORE = R$  of reputable users /  $n$

**Else**

Calculate cosine similarity between the  $KO$  and  $KOs$  in database

Set  $X$  = most similar  $KO$

$PRED = R$  of reputable users on  $X$

$SIM$  = cosine similarity between the user and the  $KO$  creator

$SCORE = PRED + SIM/2$

**End If**

**End For**

**Sort**  $S$  based on  $SCORE$  in descending order

The evaluation process begins by retrieving ratings of the KO by reputable users. In the absence of sufficient ratings, the recommendation process will start by measuring the similarity of the KO in the search result and other KOs in the database, based on their attributes. The most similar KOs exceeding the threshold of similarity will be utilized to predict the ratings of reputable users. Where more than one KO has the same similarity value, the result will be improved by calculating the similarity between the user and the creator of the KO. Below, we describe the REM in detail, using the content-based collaborative recommender and user reputation.

**Step 1: Find ratings.** To begin, historical ratings of reputable users are used to estimate the quality score that will assist the user in selecting high quality KOs. The quality estimation strategy begins by scanning the search results. If the KO returned by the search has attracted some ratings, the system will interrogate the repository of ratings for the reputation points of users who have rated that KO. In calculating the average rating for a particular KO, the system will consider only reputable users' ratings. Equation (5) specifies how the average rating is calculated.

$$Q_j = \frac{\sum_{i=1}^n r_{i,j}}{N_j} \quad (5)$$

where  $r_{i,j}$  is the rating of  $user_i$  of  $KO_j$ ,  $N_j$  is the total number of reputable users who rated  $KO_j$ , and  $Q_j$  is the estimated quality score.

**Step 2: Retrieve the most similar KOs.** If no ratings of users with high reputation points can be found for a particular KO, the system will calculate predicted ratings, using the content-based filtering algorithm to calculate similarity and to make quality predictions. The predicted rating is computed on the basis of (i) the similarity between the characteristics of the KO in the search result and other KOs in the database and (ii) neighbors of the user whose profiles reveal similar characteristics. In determining the user's neighbors and finding similar KOs, we use the cosine similarity measure to calculate both user-user similarity and item-item similarity [34]. Finally, the overall quality estimate is the linear combination of predicted rating and user similarity. Prediction of a KO's quality is then computed by performing a linear weighted average.

A vector space model [35] is used to represent a KO as a vector of attributes. Weight is then calculated and included in the vector. In the vector, attribute value is (1) for presence or (0) for absence (0) of a term; binary weights are utilized to compute similarity between two KOs. Following the weighting of knowledge attributes, the similarity between two KOs can be computed using the following cosine similarity formula:

$$similarity = \cos(\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} = \frac{\sum_{i=1}^n (A_i * B_i)}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (5)$$

where  $\vec{A}$  and  $\vec{B}$  are treated as vectors of attributes of KOs and  $\|\vec{A}\|$  and  $\|\vec{B}\|$  are the magnitudes of vectors  $\vec{A}$  and  $\vec{B}$ . The angle between the two vectors indicates their degree of similarity; a smaller angle signifies greater similarity.

For more accurate results in the absence of similar KOs with sufficient ratings, users' neighbors with similar characteristics can be identified on the basis of their profiles. In the same way, the system calculates the similarity of user and creator of a KO in the search result on the basis of their user profiles.

**Step 2: Predict the quality score.** The predicted quality score is computed using the similarity value for the user and creator of the KO and average ratings of users with high reputation points on the most similar KO. The equation is defined as follows:

$$Q_s = \left( \frac{\sum_{i=1}^n r_{i,s}}{N_s} + Sim_{u,c}/2 \right) \quad (7)$$

where  $r_{i,s}$  is the rating of  $ruser_i$  with high reputation points on the most similar KO, and  $Sim_{u,c}$  is the similarity between the user who performs the search and the creator of KO.

**Step 4: Re-rank the search results.** Search results can now be re-ranked. Each KO in the search results is assigned a score, representing the estimated quality score in helping users to find valuable KOs, calculated either by (5) or (7), according to the availability of ratings. Search results are ranked in descending order from the highest score to the lowest for presentation to the user as the list of recommended KOs.

#### IV. EXPERIMENTS

In experiments to examine the performance of our proposed quality recommendation models, the following were the main objectives: 1) to assess the effectiveness of the proposed quality recommendation models and 2) to evaluate the accuracy of the proposed models in order to select an appropriate model for adoption in our KMS. The results are presented and discussed below, following a description of the data sets.

##### A. Data Sets

One of the challenges facing the implementation of KMS is to find appropriate data sets for experimentation. Although there are many available sources of data, most of these have not been defined or documented. Most KM websites allow users to see part of the data but do not generally offer open data sets or provide evaluation of KOs. Additionally, no data set contains all the quality indicators under consideration here. The two available options for constituting the data set, then, were to use a real data set that might (imperfectly) match the characteristics of the target domain and task, or to synthesize a data set specifically to match the required properties. The proposed models entail data sets that contain information about users and their action types, reputation scores, and knowledge resources, as well as explicit and implicit ratings. Fortunately, our search identified a website ([www.teachability.com](http://www.teachability.com)) containing some indicators from the four quality dimensions mentioned above, enabling us to run the experiments using a real data set. The website in question is a collaborative online sharing space for teachers, enabling them to connect, learn, and improve their capabilities.

The data set contained 217 resources and 58 users. The resources gathered were available on the website in the period between May 2011 and November 2015. The data set contained some quality indicators from each of the dimensions (social, usage, characteristic, and contributor). Each resource had a title, description, keywords, and information about the creator. User actions recorded on the Teachability website included accessing learning resources, bookmarking resources, adding a comment, adding a rating, and accessing user pages. These actions provided useful implicit and explicit knowledge about the quality of the resource. Teachability awards points to users for their actions, which is a pivotal factor for present purposes. To examine the proposed models, we could not use all the quality indicators as they were not found in a unique data set. Instead, we used the available indicators, which included KO description, author, reputation points, views, bookmarks, ratings, and comments).

**B. Evaluation Metrics**

In respect of the proposed models, we were interested in ordering the list of search results according to estimated quality. This process is usually referred to as the ranking of items, and the appropriate order of a set of search results can be determined using a reference ranking [36]. A reference of this kind is essential in order to evaluate a ranking algorithm. In the case of the proposed models, where such a reference does not exist, it may be appropriate to compose a reference ranking by asking an expert to order the search results according to estimated quality.

*1) Spearman’s rank correlation coefficient*

Rank correlation measures such as Spearman’s  $\rho$  can be used as a reliable and fairly simple method of testing both the strength and direction (positive or negative) of any correlation between two variables [37]. Spearman’s  $\rho$  also takes account of problems with ties. The relevant equation is:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \tag{8}$$

where  $d_i$  is the difference between ranking of the reference and ranking of object  $i$ ,  $i \in \{1, 2, \dots, n\}$ . Spearman’s  $\rho$  is normalized in the interval  $[-1, 1]$  (see Table II). When both rankings are identical,  $\rho = 1$ ; while in case one ranking is opposite in order to the other,  $\rho = -1$  [38].

TABLE II. INTERPRETATION OF VALUES OF CORRELATION COEFFICIENT

Correlation coefficient	Dependence between variables
1	absolute
0.9 - 1	very high
0.7 - 0.9	high
0.4 - 0.7	medium
0.2 - 0.4	low
0 - 0.2	very low
0	none

To test whether a perceived value of  $\rho$  is significantly different from zero, the t-test is among the most commonly used approaches [39], where

$$t = I - \frac{\rho}{\sqrt{\frac{1-\rho^2}{(n-2)}}} \tag{9}$$

*2) Kendall’s coefficient of concordance (W)*

Where there are more than two rankings of the same domain, Kendall’s coefficient of concordance (Kendall’s W) can be used to assess agreement between them. This coefficient ranges in value from 0 to 1, where 0 denotes no agreement and 1 denotes complete agreement. Kendall’s W is given by [38]:

$$W = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}{\frac{1}{12} \cdot k^2 \cdot (n^3 - n)} \tag{10}$$

where  $x_i$  is the sum of ranks for object  $i$ ,  $k$  is the number of rankings, and  $n$  is the number of objects. The statistical significance of Kendall’s W can be assessed using a  $\chi^2$  test with  $n - 1$  degrees of freedom [38]:

$$\chi^2 = W \times k \times (n - 1) \tag{11}$$

**C. Experimental Procedure**

Users searched for KOs using the Solr search engine. Solr is “an open source enterprise search platform, runs as a standalone full-text search server” [40]. It uses the Lucene Java search library for text indexing and searching. Solr supports advanced customization, using plugin architecture [40]. To query a specific domain, users must first enter keywords. In general, the search results from a query are ranked according to the degree of similarity between keywords. For the purposes of the experiment, we first searched the database using the phrase “technology in teaching.” In response to that query, the system retrieved 11 resources. Users can choose how they want search results to be ordered; one of the options is “recommended,” which ranks search results in descending order according to the estimated quality score. To begin, we ranked results according to the first model and recoded the order and scores. The same procedure was followed for the second model. A teacher was asked to rank the search results in descending order on the basis of his teaching experience. The ranking strategy should prioritize those KOs that are most valuable to the user. After comparing each approach with the expert list, we compared the two approaches and measured the significance level of scores for each.

**D. Experimental Results**

*1) Spearman’s rank correlation coefficient for QUIM and REM*

To evaluate the effectiveness of QIEM and REM, we recorded the system’s ranking order and the score assigned to each search result for comparison with the reference rank list (Table III).

TABLE IV. SEARCH RESULT RANK AND SCORE USING QIEM AND REM

KO Title	Expert rank	QIEM rank	QIEM Score	REM rank	REM Score
YouTube Launches Site Specifically for Teachers	1	5	2.222	2	3.868
Differentiated Instruction with Technology	2	1	2.714	1	4.000
Pocket Genius Teachers Guide	3	4	2.268	7	3.600
The Power of Documentation	4	3	2.506	5.5	3.668
How can I teach this student?	5	2	2.511	3.5	3.800
Setting expectations in the 21st century	6	8	1.775	10	1.333
Impacts of the Digital Ocean on Education	7	6	1.824	3.5	3.800
ELL Technology Integration and Tips	8	7	1.801	8	3.120
Education’s Guide to Mobile Learning Devices	9	9	1.736	5.5	3.668
Airboat Lesson Activity	10	10	1.545	9	2.624
WW technology	11	11	1.536	11	1.226

TABLE V. SPEARMAN’S COEFFICIENT FOR THE PROPOSED MODELS

Model	Spearman’s coefficient	n	Significance level	Critical value	t
QIEM	0.85	11	0.05	0.5273	4.75
REM	0.71	11	0.05	0.5273	3.01

Spearman’s  $\rho$  ranking coefficient was used to measure similarity between the expert ranking and the system ranking for both models (Table IV).

As noted from Table III, there is a high level of similarity between the rankings of both models and the expert rankings, indicating that the models agree with the expert in most cases.

The next step was to test whether this agreement was accidental. The null hypothesis ( $H_0$ : “Agreement between both rankings is accidental”) can be tested using t-values for both models [41]. For  $n = 11$  and significance level = 0.05, the

critical value is 0.5273 [42]. From (11),  $t = 4.75$  for QIEM, and for REM,  $t = 3.01$ . As this exceeds the critical value for both models, the null hypothesis was rejected, indicating that agreement between the rankings was statistically significant (i.e., not accidental).

2) Kendall’s coefficient ( $W$ ) for QIEM, REM, and expert rankings

Table V summarizes agreement among QIEM, REM, and expert rankings.

Sum of  $x_i = 198$ ; sum of  $x_i^2 = 4402$ ;  $k = 3$ ;  $n = 11$ . From (10):

$$W = \frac{4402 - \frac{39204}{11}}{\frac{1}{12} \cdot 3^2 \cdot (11^3 - 11)} = \frac{838}{990} = 0.846 \quad (12)$$

This result indicates very high agreement between rankings.

TABLE VI. EXPERT AND SYSTEM RANKINGS

KO Title	Expert rank	QIEM model	REM model	Sum of $x_i$	Sum of $x_i^2$
YouTube Launches Site Specifically for Teachers	1	5	2	8	64
Differentiated Instruction with Technology	2	1	1	4	16
Pocket Genius Teacher’s Guide	3	4	7	14	196
The Power of Documentation	4	3	5.5	12.5	156.25
How can I teach this student?	5	2	3.5	10.5	110.25
Setting expectations in the 21st century	6	8	10	24	576
Impacts of the Digital Ocean on Education	7	6	3.5	16.5	272.25
ELL Technology Integration and Tips	8	7	8	23	529
Education’s Guide to Mobile Learning Devices	9	9	5.5	23.5	552.25
Airboat Lesson Activity	10	10	9	29	841
WW technology	11	11	11	33	1089

E. Discussion

The main purpose of the experiment was to measure and contrast the efficiency of the suggested quantitative evaluation models in augmenting automatic knowledge sharing and dissemination services in a KMS. The two proposed evaluation models were QIEM (quality indicators-based evaluation model) and REM (reputation-based evaluation model). Fig. 4 compares the ranking performance of both models against the expert ranking, showing that the results from both models align significantly with the expert ranking.

To further assess performance, the rank correlation measure for each model was calculated, in which a higher positive value indicates a more effective model. Table VI summarizes rank correlation values for the proposed models.

Table VI indicates that both models provide highly accurate quality estimation of KOs. However, the results suggest that the QIEM provides higher accuracy and outperforms the REM. Although the performance of the REM model is 0.71, error arose from insufficient ratings of KOs. Because of the novelty of the system, the resources have not yet gained enough ratings. As the model searches for similar KOs with sufficient ratings, those with more ratings may have a lower similarity score. In addition, users can choose whether to provide their information, resulting in lower similarity scores when comparing the searcher to a set of incomplete users' profiles in the database. Further investigation revealed that the QIEM's superior performance is accounted for by the use of diversity indicators that make up for the absence of other data. Quality estimation can be roughly predicted using only a set of interactions with KOs and characteristics of KOs. However, there is no guarantee that users who interact with the same set of KOs will always return a similar ranking.

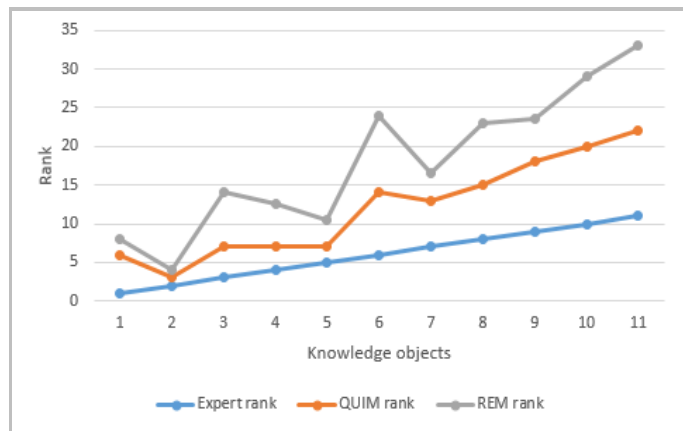


Fig. 4. Comparison of the two methods of re-ranking with expert ranking.

TABLE VII. EVALUATION OF THE PROPOSED MODELS

Model	Correlation coefficient	Strength of the correlation
QIEM	0.85	Very strong
REM	0.71	strong

V. APPLICATION OF THE EVALUATION MODELS

In developing a KMS to support participation in knowledge sharing among university instructors, two problems were encountered. The basic problem in designing the system was how to assess the usefulness of a given KO. In addition, some mechanism was needed to encourage the reuse of knowledge. To address these issues, previous algorithms for recommending KOs were further developed, exploiting reputation scores previously assigned to each user to assess KO quality and encouraging user involvement by awarding extra points for interactions. The KMS was implemented as a knowledge portal using the Drupal content management system. The portal runs on a platform that supports Apache, PHP, and MySQL to store content and settings. The knowledge portal maintains a dynamic graphical user interface running on the client side that handles all user requests and collaborative activities. It facilitates knowledge acquisition, storing, and sharing, enabling users to submit documents, share ideas, work collaboratively, and store knowledge in searchable repositories. Fig. 5 shows the knowledge portal homepage.

To resolve the issue of identifying valuable knowledge, the proposed system integrates QIEM to provide for the automatic evaluation of knowledge, assessing its quality, recommending the qualified experience in terms of various measures, providing a quantitative score for overall rating of knowledge objects, and re-ranking search results based on the quality score.



Fig. 5. Knowledge portal homepage.

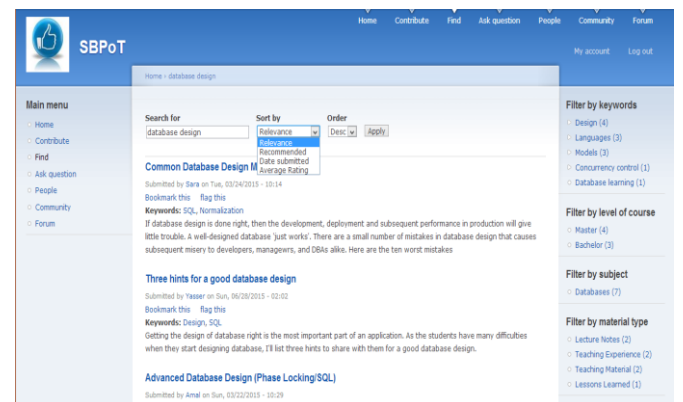


Fig. 6. Ranked list of search results.

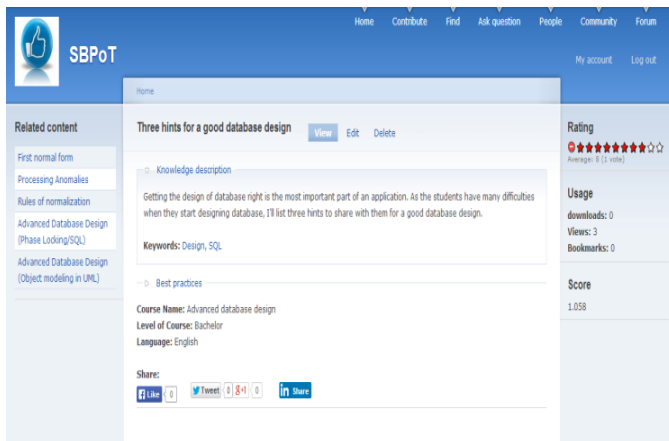


Fig. 7. Detailed resource information.

Members can log into the system and search any chosen topic. Once submitted, the query is forwarded to the search engine and database, and search results are compiled and presented. The user receives a wide-ranging set of search results of different types, presented as a ranked list. Additionally, the user can select the ranking method (by relevance, recommended, date submitted, and average rating). Fig. 6 shows an example of a list of matching objects, ranked to help users to find the most valuable KOs.

When the user selects one or several KOs, they can then criticize or rate them to offer the community explicit feedback. Statistical information about users' views, downloads, shares, and bookmarks is recorded for future evaluation of the resource as illustrated in Fig. 7.

## VI. CONCLUSION

The present research highlights the importance of assessing the quality of knowledge objects in knowledge management systems and proposes a quantitative model for automatically evaluating that quality, based on a number of metrics. For this purpose, two quality models were introduced. The first of these exploits knowledge quality indicators to recommend quality knowledge objects for online communities, integrating the indicators into a measure and ranking the results according to estimated scores. The second model exploits a content-based recommender technique and user reputation scores for quality estimation. The results show that the proposed models perform well when integrated into the implemented KMS and tested using real data. Additionally, the use of indicators for quality estimation showed better accuracy than the ratings of reputable users.

The findings suggest several directions for future research. As the initial testing was conducted offline using a predefined data set, it is planned to run the experiment online to compare the performance of the two models with real data sets. The reputation-based evaluation model depends on explicit ratings by reputable users, and it is planned to incorporate implicit feedback within the evaluation framework for better performance. It is also planned to adapt the proposed evaluation systems as program modules that can be consolidated into any web-based knowledge management system.

## REFERENCES

- [1] Al-Rasheed and J. Berri, "Effective Reuse and Sharing of Best Teaching Practices," *Computer Applications in Engineering Education*, vol. 25, no. 2, pp. 163-178, 2017.
- [2] B. Fan, L. Liu, M. Li and Y. Wu, "Knowledge recommendation based on social network theory," in *Advanced Management of Information for Globalized Enterprises*, 2008. AMIGE 2008, Tianjin, 2009.
- [3] W. Zhao, J. Wang and G. Liu, "A Knowledge Recommendation Algorithm Based on Content Syndication," in *Fourth International Conference on Computer Sciences and Convergence Information Technology, ICCIT '09*, Seoul, 2009.
- [4] H. Niu and H. Chen, "An Improved Recommendation Algorithm in Knowledge Network," *JOURNAL OF NETWORKS*, pp. VOL. 8, NO. 6, 1336-1342, 2013.
- [5] A. Vizcaíno, J. Portillo-Rodríguez, J. P. Soto, M. Piattini and O. Kusche, "A Recommendation Algorithm for Knowledge Objects based on a Trust Model," in *Proceedings of the 3rd International Conference on Research Challenges in Information Science*, pp. pp. 93-102 , 2009 .
- [6] F. Dai, X. Gu, L. Zeng and Y. Ni, "An Enterprise Knowledge Management System (EKMS) Based on Knowledge Evaluation by the Public," *Knowledge Engineering and Management, Advances in Intelligent and Soft Computing*, pp. vol. 123, pp. 267-272, 2011.
- [7] A. Al-Rasheed and J. Berri, "Knowledge Management of Best Practices in a Collaborative Environment," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, no. 3, 2016.
- [8] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan and Z. Zhang, "Expert rank: A topic-aware expert finding algorithm for online knowledge communities," *Decision Support Systems*, pp. vol. 54, no. 3, pp.1442-1451, 2013.
- [9] "American Heritage dictionary of the English Language," 2015. [Online]. Available: <https://ahdictionary.com/word/search.html?q=reputation>. [Accessed 6 12 2015].
- [10] B.-C. Chen, B. Tseng, J. Yang and J. Guo, "User Reputation in a Comment Rating Environment," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2011.
- [11] B.-C. Chen, J. Yang, J. Guo and B. Tseng, "A survey of trust and reputation systems for online service provision.," *Decision Support Systems*, vol. v. 43, pp. pp. 618-644, 2007.
- [12] B. Adler, A. Kulshreshtha, L. d. Alfaro and I. Pye, "Reputation Systems for Open Collaboration," *Communications of the ACM* , vol. Volume 54, no. Issue 8, pp. pp. 81-87, 2011.
- [13] K. Fujimura and N. Tanimoto, "The EigenRumor Algorithm for Calculating Contributions in Cyberspace Communities," in *Trusting Agents for Trusting Electronic Societies*, Berlin Heidelberg, Springer, 2005, pp. pp. 59-74.
- [14] M. Chen and J. P. Singh, "Computing and using reputations for internet ratings," in *Proceedings of the 3rd ACM conference on Electronic Commerce, EC '01*, New York, NY, USA, 2001.
- [15] X. Ochoa and E. Duval, "Quality Metrics for Learning Object Metadata," in *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 2006.
- [16] X. Ochoa and E. Duval, "Relevance Ranking Metrics for Learning Objects," in *IEEE Transactions on Learning Technologies*, pp. vol. 1, no. 1, pp. 34-48, 2008.
- [17] A. Zapata, V. Menendez, Y. Eguigure and M. Prieto, "Quality Evaluation Model for Learning Objects from Pedagogical Perspective. A Case of Study," in *ICERI2009 Proceedings*, 2009.
- [18] Eurostat, "Handbook on improving quality by analysis of process variables," produced by ONS-UK, INE Portugal, NSS of Greece and Statistics Sweden, 2004.
- [19] W. Mansfield and P. Grunewald, "The use of Indicators for the Monitoring and Evaluation of Knowledge Management and Knowledge Brokering in International Development," Report of a workshop held at the Institute for Development Studies 8th March, 2013.
- [20] E. R. N. Valdez, J. M. C. Lovelle, O. S. Martínez, C. E. M. Marín, G. I. Hernández and S. Verma, "Social Voting Techniques: A Comparison of



- the Methods Used for Explicit Feedback in Recommendation Systems," International Journal of Interactive Multimedia and Artificial Intelligence, pp. vol. 1, no. 4. p. 61, 2011.
- [21] J. Bian, Y. Liu, E. Agichtein and H. Zha, "Finding the right facts in the crowd: Factoid Question Answering over Social Media," in in Proceeding of the 17th international conference on World Wide Web - WWW '08, 2008.
- [22] E. Agichtein, C. Castillo, D. Donato, A. Gionis and G. Mishne, "Finding high-quality content in social media," in in Proceedings of the International Conference on Web Search and Web Data Mining, New York, NY, 2008.
- [23] M. Claypool, P. W. M. Le and D. Brown, "Implicit interest indicators," in Proceedings of the 6th International Conference on Intelligent User Interfaces, 2001.
- [24] D. M. Nichols, "Implicit ratings and filtering," in Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering, 1997.
- [25] D. Oard and J. Kim, "Modeling information content using observable behaviors," in Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology, 2001.
- [26] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference. , 37 (2),," SIGIR Forum, pp. vol. 37, no.2, pp. 8-28, 2003.
- [27] S. Fox, K. Karnawat, M. Mydland, S. Dumais and T. White, "Evaluating Implicit Measures to Improve Web Search," ACM Transactions on Information Systems, pp. Vol. 23, No 2, pp. 147-168, 2005.
- [28] X. Ochoa and E. Duval, "Automatic evaluation of metadata quality in digital repositories," International Journal of Digital Libraries, pp. vol. 10, no. 2, pp. 67-91, 2009.
- [29] A.-R. Alfarez. and S. Hailes, "Supporting Trust in Virtual Communities," in in 33rd Hawai'i International Conference on System Sciences (HICSS 33), 2000.
- [30] A. El-korany, "Integrated Expert Recommendation Model for Online Communities," International Journal of Web & Semantic Technology (IJWesT), pp. Vol.4, No.4, 2013.
- [31] M. J. Blooma, D. H. Goh and A. Y. Chua, "Predictors of high-quality answers," Online Information Review, pp. vol. 36, no. 3. pp. 383-400, 2012.
- [32] h.-h. j. Gang xu, "Overview on the structure and information diffusion of the online social networks," Journal of Theoretical and Applied Information Technology , p. Vol. 51. No. 3 , 2013.
- [33] N. Zardari, K. Ahmed, S. Shirazi and Z. Yusop, Weighting Methods and their Effects on Multi-Criteria Decision Making Model Outcomes in Water Resources Management, Springer, 2015.
- [34] Y. Shih and D. Liu, "Product Recommendation Approaches: Collaborative Filtering via Customer Lifetime Value and Customer Demands," Expert Systems with Applications, ,, pp. vol. 35, nos. 1/2, pp. 350-360, 2008.
- [35] M. Eleni and K. John, "Utilizing vector space models for user modeling within a learning environments," Computers & Education , pp. 51(2), 493-505, 2008.
- [36] G. Shani and A. Gunawardana, "Evaluating Recommendation Systems," in Recommender Systems Handbook, Springer US, 2011, pp. pp 257-297.
- [37] G. A. Fredricks and R. B. Nelsen, "On the relationship between spear spearman's rho and kendall's tau for pairs of continuous random variables," Journal of Statistical Planning and Inference, p. 137(7):2143-2150, 2007.
- [38] J. Mazurek and K. slova, "EVALUATION OF RANKING SIMILARITY IN ORDINAL RANKING," Acta academica karviniensia, pp. pp. 119-128, 2011.
- [39] M. G. Kendall and A. Stuart, The Advanced Theory of Statistics, Volume 2: Inference and Relationship, Griffin. ISBN 0-85264-215-6. (Sections 31.19, 31.21), 1973.
- [40] D. Smiley, E. Pugh, K. Parisa and M. Mitchell, Apache Solr 4 Enterprise Search Server (1st ed.), Birmingham: Packt Publishing, p. 451, 2014.
- [41] H. Abdi, "The Kendall Rank Correlation Coefficient," in In: Neil Salkind (Ed.). Encyclopedia of Measurment and Statistics, 2007.
- [42] J. White, A. Yeats and G. Skipworth, Tables for statisticians, United Kingdom: Nelson Thornes Ltd, 1979.

# The Effect of Music on Shoppers' Shopping Behaviour in Virtual Reality Retail Stores: Mediation Analysis

Asim Munir Dad, Assistant Professor  
Department of Management Sciences  
University of Islamabad-Institute of Engineering & Sciences  
(UoI)  
Islamabad, Pakistan

Asma Abdul Rehman  
Cardiff School of Management  
Cardiff,  
United Kingdom

Andrew Kear  
Senior Lecturer in Digital Media and Communication  
Bournemouth University,  
United Kingdom, Talbot Campus

Barry J. Davies  
Emeritus Professor and Dissertation Supervisor  
The Business School, University of Gloucestershire  
Cheltenham,  
United Kingdom

**Abstract**—The aim of this study is to investigate the effect of music, as an atmospheric cue of 3D virtual reality retail (VRR) stores, on shoppers' emotions and behaviour. To complete this research, a major empirical study was conducted in Second Life (SL) which is one of the most mature virtual worlds (VWs). The effect of the music on shoppers' emotions was experimentally tested in computer labs. Pre-test and post-test were conducted to evaluate the emotion levels before and after experiencing 3D VRR stores. Detailed mediation analysis was done with the PROCESS tool at the later stage of the analysis. This research confirmed 'music' as an atmospheric cue of 3D Servicescape. Results of this research determined the effect of music on shoppers' arousal, pleasure and consequent shopping behaviour. Further, this research could not identify the direct effect of arousal on shoppers' behaviour, however, it was a major source of inducing pleasure and increasing shoppers' positive approach behaviour. This paper contribute to better understanding the 3D VRR store atmospheric, role of music in it, shoppers' emotions and behaviour.

**Keywords**—Music; retail atmospherics; 3D virtual reality retailing; second life (SL); mediation analysis

## I. INTRODUCTION

As a result of the huge technological developments in information technology (IT), shoppers are now able to shop from their homes using the Internet. This provided shoppers with another retail channel known either as web retail stores, online retail stores or web 2.0 retail stores. In this research, such first-generation offerings are called 'traditional' online/web retail stores. Beyond studying the conventional retail environment, academics have also explored the online environment and its effect on online shoppers. The first research in this area started in 1999 and was published in 2001 by Eroglu, Machleit, and Davis [14]. In this research, they posited that the online retail store environment affects shoppers in the same way as the brick and mortar retail environment [14]. This notion was later supported by a large

number of researchers [6], [8], [15], [26], [30], [31], [45], [53]. Though it has been just one and half decade passed since researchers started exploring the online retail atmospherics and its effect on different number of behavioural variables, there are now a large number of researches available.

The online retail atmospheric is not same as the brick and mortar retail [8]. There are many environmental cues that are missing in an online retail environment (physical layouts, temperature, olfaction etc.) and there are also many that are offered in the online retail environment but are not present in the brick and mortar retail environment such as content and navigation, etc. [38]. Different environmental cues in the online retail environment have been explored such as music and colours [53], layout and design [38], [64], web stores' quality and brand [6], and web graphics, links, and colours [31].

Today's shoppers have the new retail channel of 3D VRR in which to shop. A plethora of research has been done in the setting of brick and mortar & online retail environments and their effect on shoppers' emotions and behaviour [6], [10], [28], [38], [40], [64]. However, there are only four known studies, to date, conducted in the setting of 3D virtual reality retail atmospherics and their effect on virtual shoppers' behaviour [8], [24], [32], [65].

Online and web retail stores do not provide the same environment as brick and mortar or physical retail stores. Various atmospheric cues are missing in online retail stores e.g. face-to-face interaction. This can cause a lack of trust between customers and retailers [1]. However, such discrepancies are, or can potentially be, overcome in 3D VRR stores where virtual shoppers can see other avatars (either customers or employees) around them. Moreover, VRR stores provide a sense of walking, crowding, flying etc. that are not present in traditional online retailing [8], [65].

Previous research [8], [21], [24], [32], [65] indicate that the presence of virtual shoppers and their spending in 3D virtual stores is increasing day-by-day; however, research in the context of 3D VRR atmospheric is quite limited.

Hence, this study provides research parallel to that already undertaken in brick and mortar and traditional web/online/2D retail environments, but in the context of 3D VRR environments. This study consider Dad's et al. *3D Servicescape model* and aims to investigate the effect of 3D VRR stores' *background music* on virtual shoppers' behaviour, through the mediating variables of emotions (arousal and behaviour).

## II. LITERATURE REVIEW

### A. Virtual Reality and Virtual Worlds

Virtual worlds (VWs) are gifted by the merger of two technology based concepts: virtual reality and gaming world [58]. Researcher further argued that if the father of the VW is virtual reality then the mother is the gaming world [58]. However, these worlds are not only based on these two concepts but also on economy, sociology, business, law, biology, computer science and mathematics [58]. Messinger et al. [44] agree with Sivan [58] and believe that online gaming and social networks led to today's VWs.

Virtual reality is computer based electronic environment that provides immersion, interaction and imagination simultaneously. Virtual world is not a new concept, and it has been used by armed forces since 1962 [58]. Messinger et al. [44] also posited that antecedents of VWs are gaming worlds. Gaming worlds have been known to the world since 1978, and the first multi user game was MUD (Multi User Dungeon). MUD is a well-known first multi-user game, but it had no graphics and was totally text based. Other well-known gaming worlds are Ultima Online, EverQuest, The Sims Online, World of Warcraft, There.com, and Second Life.

At the beginning there were few users of VWs; however, this number grew gradually along with the gradual development in technology and greater broadband access with higher speeds at lower prices [44]. Today, VWs have been matured comparatively and a long list of VWs exists. Real life businesses are also considering these virtual worlds for their commercial activities [52]. According to Kzero [33], the number of VW users had reached around 671 million around the globe and their disbursement were \$1.8 billion of virtual assets. Users (residents) of Second Life (one of the most mature virtual world) alone did trade of \$150 million, only in the third quarter of 2009 [36]. This number later hit 21.3 million users [41], [56].

According to Kzero more and more individuals and businesses are taking interest in VWs and by 2012 the number of VW users has increased up to 2.1 billion [34]. This increasing interest by individuals and businesses has prompted VWs themselves to consider various segments of everyday life such as education, entertainment, health, business and special interests of individuals. It is proposed that sooner or later VWs will one day become a necessity for individuals and organizations [9].

### B. Virtual Reality Retailing (VRR)

Three dimensional (3D) retail stores, also called 3D virtual reality retail (VRR) stores, provide a new and innovative mediums of shopping; such stores are full of opportunities for both retailers and shoppers [65]. These 3D stores, or VRR stores, are available in above discussed 3D virtual worlds. Arrival of such new shopping mediums facilitate shoppers with an alternative, enhanced and amended shopping experience, where shoppers (Avatars) purchase items for their virtual and real lives as well [8]. Avatars spend virtual money within these 3D virtual reality retail stores i.e. Linden Dollars are spent in one of the virtual world called Second Life (SL).

These 3D VRR stores are at developmental stage yet still closely providing a real world (brick and mortar) simulated retail environment [35]. 3D VRR stores, with the support of computer graphics, are built with majority of the brick and mortar retail environmental cues such as walls, colours, lighting, floors, background music, ceiling, layout and design [8]. One of the key environmental cues (social cue) of the brick and mortar retail environment was missing in traditional web retail environment. However, this limitation was covered in 3D VRR stores where virtual shoppers can experience other shoppers' avatars shopping around them. Moreover, sales staff can also be experienced in many 3D VRR stores [8].

Traditional web or online stores were using web 2.0 technology and hence had many discrepancies [21]. Such as product image in web 2.0 technology base online retail stores was not a true picture of the product, and social cue was missing in them and customers were not fulfilling their hedonic needs of shopping [8], [65], [67]. However, 3D VRR stores are providing an enhanced experience to the shoppers where they can pick a 3D electronic object, in the hands of their avatars, which is a close resemblance of the real world product. Furthermore, avatars can be customised (depends on the user's expertise) up to the exact appearance of the users in their real lives (such as their height, body shape, facial appearance, etc.); and hence, shoppers can try any cloth on their avatar before making the actual purchase.

Researchers in this research claims that though there is a large number of research in brick and mortar and traditional web retail atmospherics [3], [7], [11], [14], [15], [27], [31], [37], [53], [54], [59], [61], [64], [66], [69], [71], [74]. However, virtual worlds and virtual reality retail stores have existed since 2003 but, research in the context of 3D VRR atmospherics is still at its initial stage [8], [24], [32], [65]. There are studies investigating other realms within the virtual worlds but research to assess 3D VRR store atmospherics and their effect on shoppers' behaviour have been ignored [8]. To date there are only four known studies discussing about the 3D VRR store atmospherics [8], [24], [32], [65]. However, among these studies only Vrechopoulos et al. [65] actually explored *layout* as a 3D VRR store atmospheric cue and its effect on shoppers' behaviour. Other three named studies in this realm focused only to define and explain the VRR environment. Dad et al. coined the term '3D Servicescape' where they defined the whole 3D virtual reality retail environment [8]. 3D Servicescape model consists of 21 environmental cues, from virtual air to compatibility [8]. Researchers further called for a future empirical research to investigate the effect of

atmospheric cues of 3D Servicescape on virtual shoppers' behaviour [8].

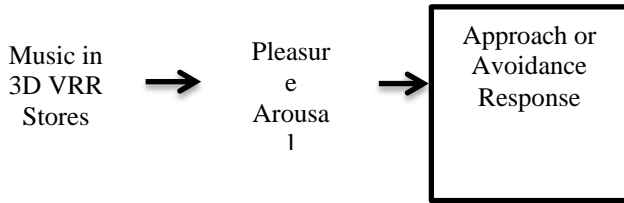


Fig. 1. Conceptual Model to Investigate the Effect of Music on Shoppers' behaviour in 3D VRR Stores.

This paper aims to investigate the effect of one of the environmental cue of 3D VRR store 'Music' on virtual shoppers' behaviour. Music has been focused a lot more than any other environmental cue, in physical [2], [16], [28], [29], [45], [46], [62], [72]-[75], and web retail atmospheric [47], [53] studies. Furthermore, in this research Mehrabian and Russell's affect model [42] will be adapted to investigate the effect of background music in 3D VRR stores' on shoppers' behaviour through the mediating variables of pleasure and arousal (see Fig. 1).

There are three principal rationales in adapting Mehrabian and Russell affect model: firstly, it provides a perfect theoretical framework to investigate the effect of any environment on human behaviour through the mediating variables of emotions; secondly, it benefits in measuring the possible emotional responses; and thirdly, this affect model empirically claims that it can measure the effect of any built environment on human behaviour [18], [42], [43], [55], [57]. On the bases of these three stated rationales it is believed that the Mehrabian and Russell's affect model would be appropriate to measure the effect of music, in 3D VRR stores, on shoppers' behaviour.

### III. RESEARCH METHODOLOGY

The purpose of this research is to test the proposed conceptual model and the assumed relationships between variables. Researchers with theoretical interests (testing the conceptual model) should give their first priority to internal validity [5], but laboratory experiments are assumed to be appropriate for testing variable relationships, according to Wang [66]. However, the selected research design followed the 'natural field experiment' approach [22]. Experimental and natural conditions were utilized in computer labs; in this research, participants experienced 3D VRR stores through the electronic virtual world of Second Life, a 'natural setting', but under controlled conditions (an experimental setting). Controlled conditions were necessary as the responsiveness and robustness of Second Life depends on available Internet speed, the efficiency of the graphic cards and (generally) the age of the computer. The researcher needs control over extraneous variables and laboratory experiments were necessary to achieve the research objectives.

In this project, computer laboratories were used as a base for the research. This was because they allowed a highly controlled environment. Participants were controlled and guided through the research, but the 3D retail stores' environments were not manipulated at all. The studied real 3D

VRR stores were being visited repeatedly in a pattern, over a period of one and a half months.

A final questionnaire (being modified after taking experts' opinion and pilot study), used for the final study of this research, consisted of the five parts which contains items to measure arousal, pleasure, effect of 3D VRR background music and approach/ avoidance behaviour.

The first page of the questionnaire had a short description about the researcher, research purpose and how to fill the questionnaire. The first 12 items measured participants' emotions (pleasure and arousal). All 12 items were orthogonal (e.g. unhappy – happy), and participants had to rate their emotion on a six-point symmetric scale. These 12 items were adapted from Mehrabian and Russell [42], Donovan and Rossiter [11] and Newman [50]. Part 2 of the questionnaire repeated these two scales to measure pleasure and arousal (six items for each), to measure pre-test and post-test differences in the participants' emotions.

The third part of the questionnaire had items to investigate the effect of 3D VRR store environment cue (background music). These items to measure the effect of music on shoppers' behaviour were adapted from Vida [63]. The fourth part of the questionnaire had four items to measure participants' behaviour towards the 3D VRR store environment (either positive or negative). These four items were adapted from Newman [50]. A six-point symmetric scale (1 - 6: 1 for strongly disagree and 6 for strongly agree) was provided to answer the 3<sup>rd</sup> and 4<sup>th</sup> part of the questionnaire. The fifth and last part of the questionnaire had four demographic-related questions, such as age, gender, year of study and field of study.

In this research to examine 3D VRR stores, present in Second Life, the 'freebie'<sup>1</sup> stores in that VW were selected. None of the environmental cues would be manipulated and stores would be visited as they are, without any modifications, allowing all those environmental cues (stimuli) in each store could be explored.

#### A. Industry Selection

A convenience sample of university students would be used in the laboratory experiments. Most of the university students were between 17 and 34 years old. Apparel stores from the Fashion and Style category in Second Life were selected for this study, as such stores were thought likely to be of interest to the students (rather than, say, stores from the home and garden or land and estates categories). Prior visits yielded a number of locations in Second Life where students could find fashion apparel VRR stores. These selected stores are located in 'London Regent Soho Park', 'London City Shopping Centre', and the 'New York Shopping Mall' in Second Life.

#### B. Participants Selection

There were two significant groups in this study: respondents and retail environments. The determination of the

<sup>1</sup> Freebie stores, in Second Life, are offering (completely free) comprehensive essentials for avatars. Normally designers add new items in such stores on weekly basis.

sample size in classical inferential studies hinges on variability in the underlying characteristics of the population, and also on the desired degree of confidence in the outcome. There is therefore no available sample frame for the consumer population and a judgemental approach is necessary. Donovan and Rossiter [11] used a sample of 30 students in their study, but there is no standard sample size evident in retail environment studies. It varies from 30 [11] to 2098 [49]. Wakefield and Blodgett has 1836 participants [68]; Wang, Minor and Wei had 400 [70]; Noone and Mattila had 198 [51]; Ward et al. (2007) had 429 [71]; Nath had 2098 [49] and Krasonikolakis et al. had 104 [32], while exploring VRR atmospherics.

Convenience sampling is one of the simplest techniques for selecting accessible subjects (Marshall, 1996). Convenience sampling has been used in this research with 200 students from the five universities. The five Universities were: COMSATS, Institute of Information Technology, Wah Cantt campus; Riphah International University Islamabad, NUFAST, and Mirpur University of Science and Technology (MUST). These five institutes were chosen because of their IT reputation and because of their available infrastructure. The researcher also considered the feasibility of travelling between these five universities.

### C. Data Collection in Computer Labs

In October 2014, the convenience sample members were invited to participate in the research, to begin in November 2014. The intention was to recruit 200 participants for the study. Participants were asked to register their accounts in Second Life and to familiarize themselves with the virtual world features, before actually taking part in the experiment. Laboratory visits were started at the beginning of November 2014 in a computer laboratory of the Computer System Engineering and Information Technology Department in Mirpur University of Science and Technology. The laboratory consisted of the latest computer systems and with 80 mbps Internet speed, sufficient enough to run Second Life. Students were invited in groups; each session consisted of 8 to 10 students. The researcher and one faculty member (Assistant Professor from the Power Engineering department, who had already been a user of Second Life for two years) acted as sales representative avatars, greeting and guiding students into the VRR stores. 63 students from MUST participated in total. 27 students from COMSATS Institute of Information Technology also participated. Even though the Internet speed in COMSATS was only 40 mbps, it did not make any difference in the running of Second Life.

In the cases of the 14 students from Riphah International University, Islamabad and 7 students from NUFAST, though these universities had the latest computer systems in their computer labs, the Internet speed was only 24 mbps. Therefore, the program could only run with a maximum of 4 students simultaneously in these locations.

A total of 118 students participated in this research. This was the number from the initial 200 participants who had been requested to open Second Life accounts and undertaken familiarization activities. However, 13 questionnaires had missing values on completion of the experiments, so a total of

105 questionnaires were considered and the 13 questionnaires with missing values were ignored.

## IV. RESULTS AND ANALYSIS

### A. Paired Sample T-Test

Measuring participants' emotions at two points: before experiencing the 3D VRR store *background music*, and after experiencing it should ensure that the particular environmental influences on the participants' emotions could be isolated. Otherwise, it was possible that participants' positive or negative emotions were not influenced by the 3D VRR store environment (*music*), but established before they came into this environment. The t-test is applied whenever comparing two means is required [17] and its requirements are met. There are two kinds of t-test; independent-samples t-test and paired-samples t-test. The paired-sample t-test pertains in a situation where there are two experimental conditions, and the same group of participants are assigned to those two experimental conditions [17]. The paired-sample t-test was applied here.

Researcher further argues that the difference in mean value of the two different situations should be different to zero, because it shows that there was a difference between the effects of two different situations [17]. If the mean value of the same participants' emotions at two levels (pre/post exposure) of this research were different, it meant that the 3D VRR store *background music* had a significant effect on participants.

Table I (see Appendix) shows that the mean value of pleasure levels of the participants' pre exposure was 4.6667 ('Pre-pleasure'). The mean value of pleasure levels of the same 105 participants after experiencing VRR store environment was 5.0635 ('Post pleasure'). Likewise, the mean value of the pre-arousal levels was 4.3643, whereas the mean value of the post-arousal levels was 4.7917.

Paired sample t-test correlation of Pair 1 (pre/post pleasure) is 0.539, which is highly significant at 0.000. For Pair 2 (arousal), the paired sample correlations value is 0.395, which is again significant at 0.000 (see Table II in Appendix).

The paired sample t-test (Table III in Appendix) shows mean difference between the pleasure levels pre and post at 0.39683 (standard deviation 0.80476), and 0.42738 (s.d. 0.91885) for arousal. The standard error mean for pleasure is 0.7854 and 0.8967 for arousal. As, by default, the confidence interval of SPSS is set at 95%, so the 95% confidence interval of the difference for pre and post pleasure is from 0.24108 to 0.55257. For pre and post arousal the confidence interval is from 0.24956 to 0.60520. It can be seen that the t-values of pleasure and arousal respectively are 5.053 and 4.766, with 104 degrees of freedom. The statistical significance (2-tailed p-value) of the paired t-tests for Pair 1 and Pair 2, (**Pr** ( $|T| > |t|$ ) **under Ha: mean (diff) = 0**), which is 0.000. It can be seen clearly in Table III (see Appendix) that the p-value is 0.000, i.e.,  $p < 0.05$ , for both pleasure and arousal (Pair 1 and Pair 2).

If the p-value is less than 0.05, there is a significant difference between two variable scores [17]. This paired sample t-test demonstrates that the effect of a VRR stores

*background music* on participants' emotion was highly significant.

### B. Mediation Analysis

Mediation analysis is a contemporary approach to analyse the effect of independent variables on dependent variables, where there is a potential effect from intervening variables [20]. Hayes argued that any researcher who wants to investigate the effect of an X variable on a Y, may well postulate one or more intervening variables M between X and Y [23]. These intervening variables M are known as 'mediating variables'.

In this research, the *music*, as a cue of 3D VRR store environments, is independent variables, X, and the goal is an investigation into its effect on behaviour, which is the dependent variable Y. Here, the conceptualization is that 3D VRR store *music* (X) is affecting shoppers' behaviour (Y) through two mediating variables of emotions ( $M_1 = \text{Pleasure}$ , and  $M_2 = \text{Arousal}$ ) (see Fig. 2 in Appendix).

Hence, instead of using Structural Equation Modelling (SEM), it was decided to run the PROCESS tool to check the relationships of independent, mediating and dependent variables, as it has advantages in this type of situation. The PROCESS tool is highly recommended by Field [17] for use when multiple regression analysis is required and the model also has mediating variables.

Mediation analysis shows researchers the direct effect of independent variables on dependent variables. Unlike AMOS (the SPSS SEM tool), it does not indicate the relationship between independent, mediating and dependent variables in 'one go', nor does it show if a conceptual model has more than one independent and dependent variable. Hayes suggested using PROCESS when accuracy is desired, even though more time will be taken on the experiment than by using SEM [23].

Mediation analysis is provided with a number of conceptual models, supported by appropriate statistical models in the software. There are two chief kinds of mediation analysis outlined by Hayes [23]. In simple mediation modelling, there is only one mediating variable between independent and dependent variables. However, in multiple mediation modelling there are two or more intervening variables.

Where there is more than one intervening variable, two forms of 'multiple mediation' are possible: parallel multiple mediation and serial multiple mediation. In parallel multiple mediation, the independent variable X affects dependent variable Y through two or more mediating variables. However, the model does not assume that the mediating variables affect each other. In serial multiple mediation, the model allows that one or more mediating variables are correlated with other mediating variables.

In this research, parallel multiple mediation was initially used. Here, the PROCESS tool was run, *music's* effect on the two mediating variables of pleasure and arousal, and the dependent variable of behaviour - approach or avoidance.

When PROCESS was run to find out the effect of *music* on behaviour, through the mediating variables of  $M_1 = \text{pleasure}$  and  $M_2 = \text{arousal}$ , it was found that *Music* (X) affected pleasure and arousal at the same time, but arousal never affected behaviour.

As parallel mediation did not uncover the expected effects from arousal, serial multiple mediation modelling was then used. Serial modelling was found to give a better result, i.e. some of the expected effects were then discerned. It should be noted that the PROCESS command for parallel multiple mediating models and serial multiple mediating models looks the same. Hayes provides a series of model templates, conceptual and statistical, on which the mediation analysis may be based [23].

For serial multiple mediating, Hayes' Model 6 is used instead of Hayes' Model 4. Whilst using Hayes' Model 6 of the PROCESS command, the order of the mediating variables matters, but it does not whilst using his Model 4. Using Hayes' Model 6 of the PROCESS command, a sequential rationale is followed: arousal is taken as the primary mediating variable, and then pleasure levels are taken as the secondary mediating variable. Arousal is taken as the primary mediating variable, when the objective is to determine where (if anywhere) its effect is 'going', i.e. it is getting effect from the independent variable (*music*), but not affecting dependent variables behaviour.

The summary of the results is given below. The serial multiple mediator model consists of four indirect effects; whose values are the products of regression coefficients relating X to Y. The first indirect effect estimated is the specific effect of  $X \rightarrow M_1 \rightarrow Y$  (*Music*  $\rightarrow$  *Arousal*  $\rightarrow$  *Shoppers' Behaviour*); the second indirect effect estimated is the effect of  $X \rightarrow M_1 \rightarrow M_2 \rightarrow Y$  (*Music*  $\rightarrow$  *Arousal*  $\rightarrow$  *Pleasure*  $\rightarrow$  *Shoppers' Behaviour*); the third indirect effect estimated is the effect of  $X \rightarrow M_2 \rightarrow Y$  (*Music*  $\rightarrow$  *Pleasure*  $\rightarrow$  *Shoppers' Behaviour*), and finally the fourth indirect effect estimated in the serial multiple mediator model is the total indirect effect that is estimated as the sum of all the specific indirect effects. These four indirect effects are output in the PROCESS results, alongside 95% bias-corrected bootstrap confidence intervals, based on 10,000 bootstrap samples [23]. If the bootstrap value does not include zero (between lower and upper limit confidence intervals – LLCI & ULCI) then the p-value is assumed to be less than or near to 0.05, which means an effect is significant.

### C. Results of Mediation Analysis

*Music* significantly predicts arousal ( $M_1$ ); as  $b = 0.2743$ ,  $p = 0.0006$  and  $t \text{ value} = 3.5629$  (see Outcome 1 in Appendix). *Music* does not affect pleasure ( $M_2$ ) significantly, as  $b = 0.0903$  with  $p \text{ value} > 0.05$  ( $p = 0.1017$ ) and  $t = 1.6514$  (see Outcome 2 in Appendix). Arousal, in Outcome 2, has a significant effect on pleasure with  $b = 0.6450$ ,  $p = .0000$  ( $< .05$ ) and  $t = 9.7640$  (see Outcome 3 in Appendix). In Outcome 3, arousal's effect on shoppers' behaviour is insignificant with the statistical values of  $b = 0.0383$ ,  $p = 0.7754$  and  $t = 0.2861$ . However, pleasure affected shoppers' behaviour significantly in PROCESS Outcome 3, with the statistical values of  $b = 0.4484$ ,  $p = 0.0024$  and  $t = 3.1101$ .

'Indirect effect path 1' (see Outcome 4 in Appendix), shows the indirect effect of music on arousal, then on shoppers' behaviour, which equates as music  $\rightarrow$  arousal  $\rightarrow$  behaviour. The first indirect effect is estimated as  $0.2743(0.0383) = 0.0105$ . This path of influence is not significant because the bootstrap confidence interval straddles zero (-0.0531 to 0.1119).

The second indirect effect is labelled as 'Indirect effect path 2', which shows the effect of music on behaviour in serial (music  $\rightarrow$  arousal  $\rightarrow$  pleasure  $\rightarrow$  shoppers' behaviour). The 2<sup>nd</sup> indirect effect is estimated as  $0.2743(0.6450) 0.4484 = 0.0793$ . This path of influence can be interpreted as significantly positive because the bootstrap confidence interval is above zero (0.0249 to 0.1826).

The third indirect effect, labelled as 'Indirect effect path 3', estimates the effect of music on pleasure, which in turn affects behaviour. The 3<sup>rd</sup> indirect effect is estimated as  $0.0903(0.4484) = 0.0405$  (virtual air  $\rightarrow$  pleasure  $\rightarrow$  shoppers' behaviour). This path of relationships can also be interpreted as significant, because the bootstrap confidence interval does not straddle zero (0.0083 to 0.1089).

The serial multiple mediator model gives the sum of all specific indirect effects, which is known as 'total indirect effect'. The total indirect effect is 0.1304; this can be interpreted as significantly positive, because the bootstrap confidence interval is above zero (0.0402 to 0.3052).

Although the first indirect effect is not significant, the second, third, and total indirect effects are significant for music as an environmental cue of 3D VRR stores. Therefore, assuming total indirect value, which is significant, it is said that the total indirect effect is significant and even the p-value is deemed to be very close to 0.05. See Fig. 3 in Appendix for the effect of *music* on behaviour through the mediating variables of Arousal and Pleasure.

## V. DISCUSSION OF THE RESULTS

Music is a cue that has been studied many times in previous research [4], [12], [13], [15], [19], [25], [28], [29], [39], [47], [53], [60], [62], [63], [66], [74], [75]. These researchers investigated different styles of music, different tempos, volume, whether to use the top 40 or classical music, and effect of music on shoppers' behaviour. However, all these studies were conducted either in brick and mortar retail environments or in traditional online retail environments. None of the research was conducted in an immersive 3D VRR store environment that explored the effect of music on virtual shoppers. However, Krasonikolakis et al. [32] suggested that music might have less or no effect on shoppers' behaviour in 3D VRR stores as shoppers can turn off the music while making their visit.

Despite their suggestion, this research found the comparable result that music has a significant effect on shopping behaviour, which is linked to shoppers' emotions. Among the three indirect effects of music on shoppers' shopping behaviour, two are found to be significant and positive. The sum of all indirect effects (total indirect effect) is also found to be positive and significant (see Outcome 4).

Previous research also found that different styles and tempos of music have an effect on shoppers' behaviour through shoppers' perceptions, attitudes, feelings and pleasure arousal levels. This research broadly confirms the previous studies and determines that, in 3D VRR stores, music is an important environmental cue that affects shoppers' emotions and behaviour in a positive way. It was found that the presence of music in 3D VRR stores affected shoppers' arousal and pleasure levels in a serial, and then affected approach behaviour. Music was found to significantly increase shoppers' arousal levels, which increases shoppers' pleasure levels, and then their approach behaviour. Pleasure also had a positive effect, separately, from music and then went on to increase shoppers' approach behaviour. However, arousal on its own was found not to affect shoppers' behaviour significantly in the presence of music in 3D VRR store environments.

Although most previous studies reported a significant and positive effect of music over shoppers' behaviour, Wang found that music was a source of irritation for shoppers if present in (traditional) web stores [66]. VRR stores are also a kind of online store, but they have features that make them different to traditional online stores. These 3D VRR stores are closer to the brick and mortar stores; shoppers' can experience a very similar electronic environment. That is why this study contradicts Wang [66], but confirms other studies [28], [62], [39]. Based on this study, music should be considered by all those retailers who wish to move towards 3D retailing, or for all those retailers who are already running their retail stores in these 3D VWs.

## VI. CONCLUSION

This study is concluded as 3D VRR stores' background *music*, which significantly and positively affect shoppers' emotions, and subsequent behaviour. Though a previous study [65] did not find the effect of 3D VRR store environment (*layout*) on shoppers. This study also confirms Morrison, Gan, Dubelaar, & Oppewal [48] results in which it was found that the effect of environmental cues (music and aroma) on arousal, and high arousal levels themselves induced pleasure and behaviour positively. In this study, the results show that the second indirect effect path is significant which means that *music* induces shoppers' arousal, which subsequently increases shoppers' pleasure levels and positive behaviour.

Importantly, this research determines that in 3D VRR store environments, whilst adapting the M-R (1974) model, 'arousal' did not induce shoppers' behaviour directly.

### a) Managerial Implications

The retail environment and its effect on shoppers' shopping behaviour is an important area of research among retail researchers. Brick and mortar and conventional high street retail environments have been investigated in a great deal since 1974. Moreover, online retail environments have also been explored thoroughly since their development. Newer VWs provide an innovative way of shopping (3D VRR stores), full of opportunities for both retailers and shoppers. One implication of this study is that retail management needs to maintain strategic perspective on the potential of 3D VRRs

to influence current business, both positively and negatively. 3D VRR should not be ignored.

By determining that 3D VRR stores' music have a great effect on shoppers' arousal and pleasure levels, it shows consequently the potential impact on their shopping behaviour. Therefore, 3D retailers are advised that they should be very careful when designing their retail stores in VWs and should follow the pattern of application that results here suggest. This study found that music, to be the environmental cues that induced the greatest approach behaviour. Although music here had a positive effect on shoppers' arousal and pleasure levels and subsequent shopping behaviour, this research also determined that it is possible that, if not manipulated properly, music could have a negative effect on shoppers' emotions and behaviour in 3D VRR stores. Therefore, future researchers and retailers should explore further why music has a positive effect on shoppers' emotions and behaviour.

#### b) Limitations and Future Studies

As with all research, this study is not free from limitations. The first limitation of this research is that only one type of speciality retail (apparel stores) was used to test the conceptual model. It is quite possible that other types of retail stores may have given different results. Therefore, in future studies other types of retail stores should be investigated.

Secondly, the sample used in this research was a convenience group of university students. Although they had good knowledge of VWs and most of them were technology-oriented, they were not actual users of the VW. Therefore, in future studies, actual users should be used in a sample to study their behaviour in 3D VRR store environments. Moreover, this small sample size (105 participants) does not allow the research to be generalized. Because of the time limitations and health and safety issues, the sample size and characteristics were acceptable, but future research should be done with actual visitors of 3D VRR stores, questionnaires should be filled in within 3D VRR environments, and a larger sample size should be used for generalization of the results. It would also be useful to pursue a different research approach, starting with the current population of web customers and studying their reactions to 3D VRRs. Thirdly, in this study, 3D VRR environments were investigated without any manipulation of the environment. Manipulation of the environment in an experimental approach could give better results for understanding and improving the 3D VRR store environments.

Another limitation of this research is the usage of limited dependent variables (arousal, pleasure and behaviour). Previous studies have explored satisfaction, feelings and other dependent variables induced in retail environments. In the future, researchers should employ other dependent variables other than arousal, pleasure and behaviour.

#### REFERENCES

- [1] Aldiri, K., Hobbs, D. & Qahwaji, R. (2010). Putting the Human Back into e-Business: Building Consumer Initial Trust through the Use of Media-Rich Social Cues on e-commerce Websites. In *Transforming E-Business Practices and Applications: Emerging Technologies and Concepts*. (pp. 13-43). Western Illinois University, EUA: In Lee.
- [2] Andrus, D. (1986). Office Atmospherics and Dental Service Satisfaction. *Journal of Professional Services Marketing*, 1, 77-85.
- [3] Areni, C. S. & Kim, D. (1993). The influence of background music on shopping behavior: Classical versus Top-Forty Music in a wine store. In L. McAlister & R. L. Michael (Eds.), *Advances in Consumer Research* (pp. 336-340). UT Provo: Association for consumer research.
- [4] Broekemier, G., Marquardt, R. & Gentry, J. W. (2008). An exploration of happy/sad and liked/disliked music effects on shopping intentions in a women's clothing store service setting. *Journal of Services Marketing*, 22(1), 59-67.
- [5] Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 223-326). Chicago: Rand McNally.
- [6] Chang, H. H. & Chen, S. W. (2008). The impact of online store environment cues on purchase intention: Trust and Perceived risk as a mediator. *Online Information Review*, 32(6), 818-841.
- [7] Chebat, J. C., Gelinias-Chebat, C. & Filiatrault, P. (1993). Interactive effects of musical and visual cues on time perception: an application to waiting lines in banks. *Perceptual and Motor Skills*, 77(3), 995-1020.
- [8] Dad, A. M., Davies, B., & Rehman, A. A. (2016). 3D Servicescape Model: Atmospheric Qualities of Virtual Reality Retailing. *International Journal of Advanced Computer Science and Applications*. 7 (2). 25-38
- [9] Daley, J. (2010). Should we meet in another world?. *Entrepreneur*, 38 (6), 46-46.
- [10] Dijkstra, K., Pieterse, M. E. & Pruyn, A. T. H. (2008). Individual differences in reactions towards color in simulated healthcare environments: the role of stimulus screening ability. *Journal of Environmental Psychology*, 28 (3), 268-77.
- [11] Donovan, R. J., & Rossiter, J. R. (1982). Store atmosphere: an environmental psychology approach. *Psychology of Store Atmosphere*, 58(1), 34-57.
- [12] Dube, L., Chebat, J. C. & Morin, S. (1995). The Effects of Background Music on Consumers' Desire to Affiliate in Buyer-Seller Interactions. *Psychology and Marketing*, 12 (4), 4, 305-319.
- [13] Dube, L. & Morin, S. (2001). Background music pleasure and store evaluation Intensity effects and psychology mechanisms. *Journal of business Research*, 54,107-113.
- [14] Eroglu, S. A., Machleit, K. A., & Davis, L. M. (2001). Atmospheric qualities of online retailing: a conceptual model and implications. *Journal of Business Research*, 54, 177-184.
- [15] Eroglu, S. A., Machleit, K. A., & Davis, L. M. (2003). Empirical testing of a model of online store atmospherics and shopper response. *Psychology and Marketing*, 20(2), 139-50.
- [16] Eroglu, S. A., Machleit, K. A. & Chebat, J. C. (2005). The Interaction of Retail Density and Music Tempo: Effects on Shopper Responses. *Psychology and Marketing*, 22(7), 577-589.
- [17] Field, A. (2013). *Discovering Statistics using IBM SPSS Statistics* (4<sup>th</sup> Ed.). London: Sage Publication Ltd.
- [18] Graa, A., & Dani-elKebir, M. (2011). Situational factors influencing impulse buying behavior of algerian consumer. *RRM*, 2, 52-59.
- [19] Gulas, C. S. and Schewe, C. D. (1994). Atmospheric segmentation: Managing Store Image with Background Music, In R. Acrol & A. Mithcell (Eds.) *Enhancing Knowledge Development in Marketing* (pp. 325-330). Chicago, IL: American Marketing Association.
- [20] Gunzler, D., Chen, T., Wu, P. & Zhang, H. (2013). Introduction to mediation analysis with structural equation modeling. *Shanghai archives of psychiatry*, 25 (6), 390-394
- [21] Haenlein, M., & Kaplan, A.M. (2009). Flagship brand stores within virtual worlds: the impact of virtual store exposure on real-life attitude toward the brand and purchase intent. *Research Applications en Marketing*, 24(3), 57-79.
- [22] Harrison, G. W. & List, J. A. (2004). Field Experiments. *Journal of Economic Literature*, 42 (4), 1009-1055
- [23] Hayes, A. F. (2013). *Introduction to Mediation, Moderation and Conditional Process Analysis: A Regression-Based Approach*. New York: The Guildford Press



- [24] Hassouneh, D., & Brengman, M. (2015). Retailing in social virtual worlds: developing a typology of virtual store atmospherics. *Journal of Electronic Commerce Research*, 16(3), 218-241.
- [25] Herrington, J. D. & Capella, L. M. (1996). Effects of Music in Service Environments: A Field Study. *Journal of Services Marketing*, 10 (2), 26-41
- [26] Huang, M-H. (2003). Modeling Virtual exploratory and shopping dynamics: an environmental psychology approach. *Information & Management*, 41, 39-47.
- [27] Hussain, R., & Ali, M. (2015). Effect of Store Atmosphere on Consumer Purchase Intention. *International Journal of Marketing Studies*, 7 (2), 35-43. DOI: <http://dx.doi.org/10.5539/ijms.v7n2p35>
- [28] Iyiola, O. & Iyiola, O. (2011). "Interpretation and Effect of Music on Consumers' Emotion, *Journal of Business Diversity*, Vol. 11, No. 1, pp. 56-65.
- [29] Kellaris, J. J. & Kent, R. J. (1992). The influence of Music on Consumers' Temporal Perceptions: Does Time Fly When You're Having Fun?. *Journal of Consumer Psychology*, 1 (4), 365-376.
- [30] Kim, H., & Lennon, S. J. (2010). E-atmosphere, emotional, cognitive, and behavioral responses. *Journal of Fashion Marketing and Management*, 14(3), 412-428.
- [31] Koo, D. & Ju, S. (2010). The interactional effects of atmospherics and perceptual curiosity on emotions and online shopping intention. *Computers in Human Behavior*, 26, 377-388
- [32] Krasnikoulakis, I. G., Vrechopoulos, A. P. & Pouloudi, A. (2011). Defining, Applying and Customizing Store Atmosphere in Virtual Reality Commerce: Back to Basics?. *International Journal of E-Services and Mobile Applications*, 3 (2), 59-72
- [33] Kzero. (2009). *Virtual World Accounts Q2 2009: 5 to 10*. Retrieved August 19, 2013 from Kzero Worldwide. <http://www.kzero.co.uk/blog/virtual-world-accounts-q2-2009-5-to-10/>
- [34] KZero. (2012). *Virtual worlds/MMOs: industry and user data, universe chart for Q4, 2012*, Retrieved from: <http://www.slideshare.net/nicmitham/kzero-universe-q4-2012>
- [35] Lau, H., Kan, C., & Lau, K. (2013). How Consumers Shop in Virtual Reality? How It Works?. *Advances in Economics and Business*. 1 (1). 28-38. DOI: 10.13189/aeb.2013.010104
- [36] Linden. (2009). *The Second Life Economy-Third Quarter 2009 in Detail*. Retrieved September 2, 2013 from Second Life Blogs. <https://blogs.secondlife.com/community/features/blog/2009/11/02/the-second-life-economy-third-quarter-2009-in-detail>
- [37] Machleit, K. A., Kellaris, J. J. & Eroglu, S. A. (1994). Human versus Spatial Dimensions of Crowding Perceptions in Retail Environments: A Note on Their Measurement and Effect on Shopper Satisfaction. *Marketing Letters*, Vol. 5, No. 2, pp. 183-194.
- [38] Manganari, E. E., Siomkos, G. J., Rigopoulou, I. D. & Vrechopoulos, A. P. (2011). Virtual Store Layout effects on Consumer Behaviour. *Internet Research*, 21(3), 326-346. Marshall, M. N. (1996). Sampling for Qualitative Research, *Oxford University Press*, 13 (6), 522-525
- [39] Mattila, A. S. & Wirtz, J. (2001). Congruency of scent and music as a driver of in-store evaluations and behavior. *Journal of Retailing*, 77, 273-89.
- [40] Massara, F. (2003). Store Atmosphere: Still a fledgling art. *ECR Journal*, 3(2), 47-52.
- [41] Melancon, J. P. (2011). Consumer profiles in reality vs fantasy-based virtual worlds: implications for brand entry. *Journal of Research in Interactive Marketing*, 5 (4), 298-312.
- [42] Mehrabian, A., & Russell, J. A. (1974). *An Approach to Environmental Psychology*. Cambridge, MA: MIT Press.
- [43] Mehrabian, A. (1976). *Public Spaces and Private Spaces: The Psychology of Work, Play and Living Environments*. New York: Basic Books, Inc.
- [44] Messinger, P. R., Eleni, S., Lyons, K., Bone, M., Niu, R., Smirnov, K. & Perelgut, S. (2009). Virtual worlds - past, present, and future: New directions in social computing. *Decision Support Systems*, 47 (3), 204-228
- [45] Menon, S. & Kahn, B. (2002). Cross-category effects of induced arousal and pleasure on the Internet Shopping Experience. *Journal of Retailing*, 78(1), 31-40.
- [46] Milliman, R. E. (1982). Using background music to affect the behavior of supermarket shoppers. *Journal of Marketing*, 46, 86-91.
- [47] Morin, S., Dube, L. & Chebat, J. (2007). The role of pleasant music in servicescapes: A test of the dual model of environmental perception. *Journal of Retailing*, 83, 115-130.
- [48] Morrison, M., Gan, S., Dubelaar, C. & Oppewal, H. (2011). In-store Music and Aroma Influences on Shopper Behavior and Satisfaction. *Journal of Business Research*, 64 (6), 558-564
- [49] Nath, C. K. (2009). Behaviour of Customers in Retail Store Environment- An Empirical Study. *Vilakshan, XIMB Journal of Management*, 63-74.
- [50] Newman, A. J. (1997). *Consumption and the inanimate environment: The airport setting* (Doctoral dissertation). Manchester Metropolitan University, Manchester
- [51] Noone, B. M. & Mattila, A. S. (2009). Consumer reaction to crowding for extended service encounters. *Managing Service Quality*, 19(1), 31-41.
- [52] Parmentier, G., & Rolland, S. (2009). Consumers in Virtual Worlds: Identity Building and Consuming experience in second life. *Recherche et Applications en Marketing*, 24(3), 43-55.
- [53] Price-Rankin, K. (2004). *Online Atmospherics: An investigation of feeling and Internet purchase intention*. Unpublished Doctoral Dissertation, The University of Tennessee, Knoxville.
- [54] Quartier, K., Christiaans, H. & Cleempoel, K.V. (2009). Retail Design: lighting as an atmospheric tool, creating experiences which influence consumers' mood and behaviour in commercial spaces. In: *Undisciplined! Design Research Society Conference 2008*, Sheffield Hallam University, Sheffield, UK, 16-19 July 2008.
- [55] Quartier, K. (2011). *Retail design: lighting as a design tool for the retail environment* (Doctoral dissertation). Retrieved from ProQuest database.
- [56] Renaud, C., & Kane S. F. (2008). Virtual Worlds Industry Outlook 2008-2009. *Technology Intelligence Group*, Retrieved 15 August, 2013, from <http://blog.techintelgroup.com/2008/08/announcing-thetig-virtual-worlds-industry-outlook-2008-2009.html>
- [57] Russell, J. & Pratt, G. (1980). A description of the affective quality attributed to environments. *Journal of personality and social psychology*, 38, 311-346
- [58] Sivan, Y. (2008). 3D3D real virtual worlds defined: The immense potential of merging 3D, community, creation and commerce. *Journal of Virtual Worlds Research*, 1 (1)
- [59] Spangenberg, E. R., Crowley, A. E. & Henderson, P. W. (1996). Improving the store environment: Do olfactory cues affect evaluations and behaviors?. *Journal of Marketing*, 60, 67-80.
- [60] Spangenberg, E., Grohmann, B. & Sprott, D. (2005). It's beginning to smell and (sound) a lot like Christmas: the interactive effects of ambient scent and music in a retail setting. *Journal of Business Research*, 58, 1583-1589.
- [61] Spangenberg, E., Sprott, D., Grohmann, B. & Tracy, D. (2006). Gender-congruent ambient scent influences on approach and avoidance behaviors in a retail store. *Journal of Business Research*, 59, 1281-1287.
- [62] Sweeney, J. C. & Wyber, F. (2002). The role of cognitions and emotions in the music approach avoidance behaviour relationship. *Journal of Service Marketing*, 16 (1), 51-69.
- [63] Vida, I. (2008). The Impact of Atmospherics On Consumer Behaviour: The Case of The Music Fit In Retail Stores. *Economic and Business Review for Central and South-Eastern Europe*, 10(1), 21-35.
- [64] Vrechopoulos, A. P., O'Keefe, R. M., Doukidis, G. I. & Siomkos, G. J. (2004). Virtual store layout: an experimental comparison in the context of grocery retail. *Journal of Retailing*, 80, 13-22.
- [65] Vrechopoulos, A., Apostolou, K., & Koutsouris, V. (2009). Virtual reality retailing on the web: emerging consumer behavioural patterns. *The International Review of Retail, Distribution and Consumer Research*, 19(5), 469-482.
- [66] Wang, C. (2003). *The Role of Atmospherics in E-Tailing* (Doctoral dissertation). Retrieved from ProQuest database.

- [67] Wang, L. C., Baker, J., Wagner, J. A. & Wakefield, K. (2007). Can a Retail Web Site be Social?. *Journal of Marketing*, 71 (3), 143 – 157
- [68] Wakefield, K. L. & Blodgett, J. G. (1996). The effect of servicescape on customers' behavioral intentions in leisure service settings. *Journal of Services Marketing*, 10 (6), 45-61
- [69] Wakefield, K. L. & Baker, J. (1998). Excitement at the mall: Determinants and effects on shopping response. *Journal of Retailing*, 74 (4), 515-539.
- [70] Wang, Y. J., Minor, M. S. & Wei, J. (2011). Aesthetics and the online shopping environment: Understanding consumer responses. *Journal of Retailing*, 87 (1), 46-58
- [71] Ward, P., Davies, B.J. & Kooijman, D. (2007). Olfaction and the retail environment: examining the influence of ambient scent. *Service Business*, 1, 295-316.
- [72] Yalch, R. & Spangenberg, E. (1988). *An Environmental Psychological Study of Foreground and Background Music as Retail Atmospheric Factors*. Chicago, IL, American Marketing Association, pp. 106-110.
- [73] Yalch, R. & Spangenberg, E. (1990). "Effects of Store Music on Shopping Behavior", *Journal of Consumer Marketing*, Vol. 7, pp. 55-63.
- [74] Yalch, R. & Spangenberg, E. (1993). Using store music for retail zoning. In L. McAlister & M. Rothschild (Eds.), *Advances in consumer research*, (pp. 632-636). Provo, UT: Association for consumer research
- [75] Yalch, R. F., & Spangenberg. (2000). The Effects of Music in a Retail Settings on Real and Perceived Shopping Times. *Journal of Business Research*. 49 (2) 139-147

APPENDIX

TABLE I. PAIRED SAMPLES T-TEST STATISTICS

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Post Pleasure	5.0635	105	0.80840	0.07889
	Pre-Pleasure	4.6667	105	0.86510	0.08443
Pair 2	Post Arousal	4.7917	105	0.86839	0.08475
	Pre-Arousal	4.3643	105	0.79853	0.07793

TABLE II. PAIRED SAMPLES CORRELATIONS

		N	Correlation	Sig.
Pair 1	Post Pleasure & Pre Pleasure	105	0.539	0.000
Pair 2	Post Arousal & Pre Arousal	105	0.395	0.000

TABLE III. PAIRED SAMPLES TEST

		Paired Differences					T	df	Sig. (2-tailed p-value)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Post Pleasure – Pre-Pleasure	0.39683	0.80476	0.07854	0.24108	0.55257	5.053	104	0.000
Pair 2	Post Arousal – Pre-Arousal	0.42738	0.91885	0.08967	0.24956	0.60520	4.766	104	0.000

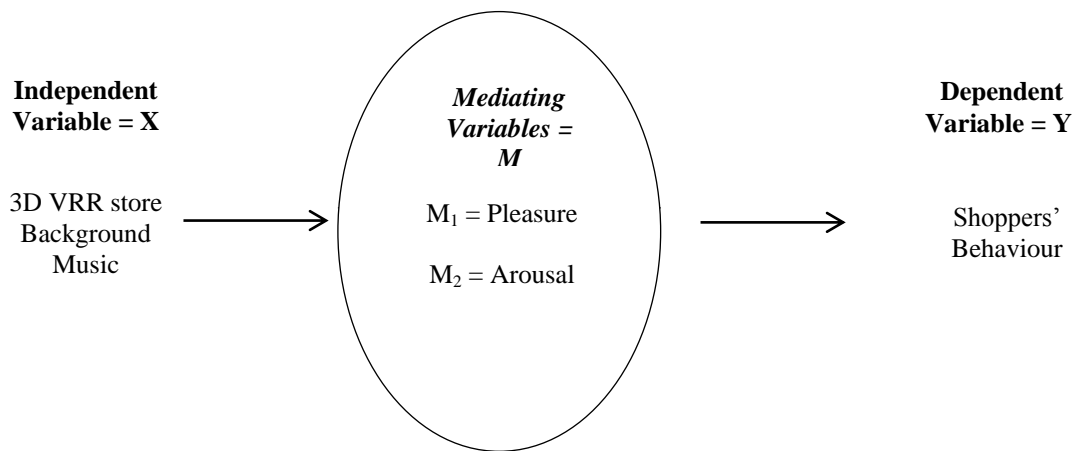


Fig. 2. Conceptualized 3D VRR store environment affecting shoppers' behaviour through mediating variables of Pleasure (M<sub>1</sub>) and Arousal (M<sub>2</sub>).

Outcome 1: Arousal						
Model Summary						
R	R-sq	MSE	F	df1	df2	p
0.3312	0.1097	0.6746	12.6942	1.0000	103.0000	0.0006
Model						
	Coeff	SE	T	P	LLCI	ULCI
Constant	3.4742	0.3706	9.3739	0.0000	2.7392	4.2093
Music	0.2743	0.0770	3.5629	0.0006	0.1216	0.4270

Outcome 2: Pleasure						
Model Summary						
R	R-sq	MSE	F	df1	df2	p
0.7382	0.5449	0.3032	61.0741	2.0000	102.0000	0.0000
Model						
	Coeff	SE	T	P	LLCI	ULCI
Constant	1.5663	0.3383	4.6305	0.0000	0.8954	2.2373
Arousal	0.6450	0.0661	9.7640	0.0000	0.5140	0.7761
Music	0.0903	0.0547	1.6514	0.1017	-0.0182	0.1989

Outcome 3: Behaviour						
Model Summary						
R	R-sq	MSE	F	df1	df2	p
0.5765	0.3324	0.6429	16.7606	3.0000	101.0000	0.0000
Model						
	Coeff	SE	T	P	LLCI	ULCI

Constant	1.1402	0.5418	2.1043	0.0378	0.0653	2.2151
Arousal	0.0383	0.1338	0.2861	0.7754	-0.2271	0.3037
Pleasure	0.4484	0.1442	3.1101	0.0024	0.1624	0.7344
Music	0.2745	0.0807	3.4011	0.0010	0.1144	0.4346

Outcome 4: Total, Direct and Indirect Effects						
Total Effect of X on Y						
Effect	SE	T	P	LLCI	ULCI	
0.4049	0.0819	4.9441	0.0000	0.2425	0.5673	
Direct Effect of X on Y						
Effect	SE	T	P	LLCI	ULCI	
0.2745	0.0807	3.4011	0.0010	0.1144	0.4346	
Indirect Effect(s) of X on Y						
	Effect	Boot SE	BootLLCI	BootULCI		
<b>Total</b>	0.1304	0.0649	0.0402	0.3052		
<b>Indirect effect path 1</b>	0.0105	0.0387	-0.0531	0.1119		
<b>Indirect effect path 2</b>	0.0793	0.0384	0.0249	0.1826		
<b>Indirect effect path 3</b>	0.0405	0.0243	0.0083	0.1089		

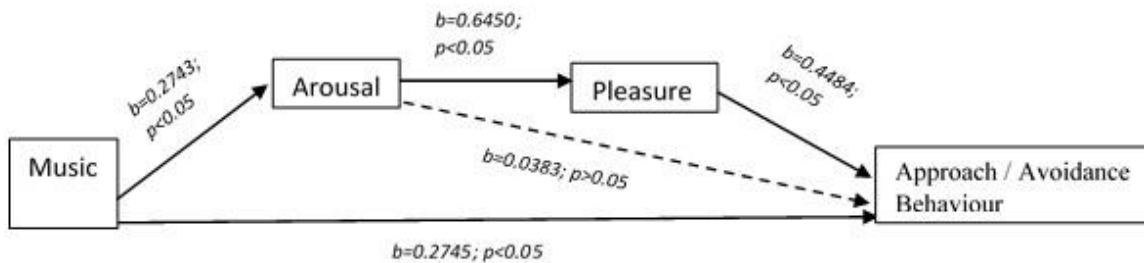


Fig. 3. Statistical outcome of 'music' affecting shoppers' behaviour through two mediating variables of arousal and pleasure.

# A Web Service Composition Framework based on Functional Weight to Reach Maximum QoS

M.Y. Mohamed Yacoab, Abdalla AlAmeen, M. Mohemmed Sha  
Department of Computer Science, College of Arts and Science,  
Prince Sattam Bin Abdulaziz University

**Abstract**—The recent trend in the web world is to accomplish almost all the user services in every field through the web portals of the respective organizations. But a specific task with series of actions cannot be completed by a single web service with limited functionality. Therefore, multiple web services with different functionalities are composed together to attain the result. Web service composition is an approach that combine various services to fulfill the Web related tasks with preferred quality. Composition of such services will become more challenging when these web services are with similar functionalities, varying Quality and from several providers. Hence, the overall QoS (Quality of Service) could be considered as the major factor for composition. Moreover, in most of the compositions the expected QoS cannot be attained when the task is finished. Sometimes the complete task may have affected by a poor performed single web service. So, while composition, at most care should be taken to select a particular web service. Composing web services dynamically is the main method used to overcome these difficulties. However, to reach the actual functionality of the specific task the quality of each individual service is very much necessary. The QoS of a web service normally evaluated using the non-functional attributes, such as response time, availability, reliability, throughput, etc. Also, while composition, the same level of quality is not expected for individual web services that are included in the chain. So, a framework proposed in this research paper, for web service composition by setting appropriate weightage for the non-functional parameters. Experimental results show that implementation of this method will definitely pave the way to reach the maximum performance of the composition with improved QoS.

**Keywords**—Web service; composition of services; non-functional parameters; QoS

## I. INTRODUCTION

Web-oriented services are considered as an application component that makes itself available over the internet. A single web service cannot useful to attain the desired specific task in all the cases. So different services are grouped together to get the work done. The existing web services are grouped together as a chain in some order to attain the target. The main parameters considered while composing such web services may vary, but the target is to reach best result with the expected level of quality. While selecting a particular web service, the quality of the service is measured and its performance in the real time also taken care to maintain the quality [1], [6]. Sometime the external factor such as network traffic can also influence the performance of the service. Any how the ultimate aim to reach the genuine functionality of the desired web-oriented service. When it is a single web service

the QoS is evaluated mainly from non-functional parameter and which is considered as the base factor [11]. It is more complex when combining the services together in some order to reach the entire functionality. A single web service in the group may affect the overall performance of the composed services. So extreme care should be taken to compose the web services together to achieve the expected outcome. Also, it is interesting to see that, the same level of quality for all the web services are not required to gain the actual functionality. It means that the same level of QoS is not essential for all the web services needed to compose [10]. Different frameworks and methodologies are suggested for composing the web services with varying QoS. Most of the methods consider the overall QoS is the major factor after composition of web services rather than considering the individual service quality. So, this work considers the problem in two different aspects. First a pool of web services is grouped based on the individual QoS if these services are up to the mark. Here the QoS is evaluated after setting the appropriate weightage for the non-functional parameters of the service. Secondly, the best arrangement of composition is selected based on the real-time availability of the services from each pool that make up the composition.

The proposed framework and its related methods to compose the web services are defined in Section III. The results are vastly discussed in Section IV and finally concluded in Section V.

## II. RELATED WORK

Web services are always termed with it functional and non-functional properties by its providers. Web service composition is the organization of those services in some order to perform a particular task. Service providers sell their dynamic web service components in the international market so that these services can be used by the customers [7]. The important property to explain a web task is that takes account of the signature, session states and the requirements of specific criteria. Non-functional values are made to assess the cost involved, the quality of service provided and to track issue related to security factors of the web composition [14]. While composing a web service, these measures are considered to check whether the service is up to the expected level of functionality. Several approaches are proposed to compose the services based on its QoS, but most of them focusing on the overall quality after composition.

Yu, T. et al. [1] modelled the problem in two different models such as the analytical model and the chart prototypical

model. The analytical design discusses the key issues on n-dimensional zero-one Knapsack Problem. The chart prototypical model expresses the issues as the Multiple Restriction Best Route value process. These algorithms also proposed to insight their achievements by test cases performance.

L.Zeng [2] proposed a method to build a high quality web task while designing a web composition. First the user requirement is considered, then the quality of web tasks is calculated, finally web tasks are selected to reach the quality as estimated for the composition of web service. The parameters of QoS viz., response instance, cost expenditure, execution duration, reput, successful execution rate and readiness are used here.

Berbner [4], [5] present a structural design called Web task merit of Service Design Extension that helps dynamic linking and binding of Web tasks at dynamic compilation along with monitoring mechanisms. Heuristics problem solving methods are proposed to make use a collection procedure calculates the total task combination. The sequential form procedure that uses several QoS attributes is proposed in this approach.

Jiuyun et al. [12] suggested an immune algorithm to handle the composition problem. Here they consider multiple quality parameters with different business process flow variables to make the attribute such as cost involved in service; it's time for respond, task availability and consistency.

Alrifai et al. [9] propose a heuristics algorithm divides the most important problem to number of small tasks then by finding solution to these as a best possible result can be generated. Finally, the algorithm use data gathering procedure for calculating the total Quality of Service Parameters.

The papers [8], [12] used Genetic algorithms to solve the composition problem. The fitness function used here compare solutions by considering all forms of workflow in business process. The approaches [3], [9], [13], [14] also present the composition of web services based on QoS evaluated from the non-functional parameters. Most of them focus on composing of web services in different aspect mainly the overall QoS, but not much concentration given to the individual web service performance based on its functionality.

### III. COMPOSITION METHDOLOGY

Web services are published by its providers along with the attributes such as signature, states and non-functional parameters. Always non-functional parameters such as response time, reliability, availability, etc. are mainly used for evaluating the quality of a web service [6]. Fig. 1 illustrate the proposed framework to compose the services that help to attain the best possible arrangement of web task to execute an appropriate work. Initially a pool of web services is grouped based on its functionality and individual QoS, to understand whether it will be used in composition chain to complete a specific task. Next appropriate weight is assigned for the non-functional parameters for evaluating the QoS of the service. These services are ranked and the services at the satisfaction level are added in the group.

All these services with specific category are stored in the task service repository. The procedure generator took the functionality attributes of a particular task as initial data input, and generate the system representation that explains the combined task services. The system design model holds the combination of chosen dynamic tasks and their different flow models associated with those services. In contrary we might have found that, many of web task services use to have the identical or equal structure to define a particular web service. In that situation the web architect will design a number of web service tasks that meet out necessary plan of requirement [1]. At this moment, the complex tasks are estimated by the total overall quality from the value stated from the requirement of specific tasks parameters. Normally the person one who request the service, supposed to identify the weight values to each individual requirement of specific tasks parameters.

Finally, the composite service with best quality will ahead in the rank list. Following the evaluation, the individual qualities of the service are maintained as a group. To compose the services, the individual services from each group is selected based on rank and availability. The process generator uses appropriate algorithms that consider the services from each individual group to set the best composition. Once a good composite process is identified and selected, the web service is in the ready state and the attained service to be executed. Carrying out the implementation of a composite web task the sequence of message passing across the service are well explained in the process model [1], [6]. The dataflow in the composition is defined as the actions that the output data from the executed service transfers to the input to the following service.

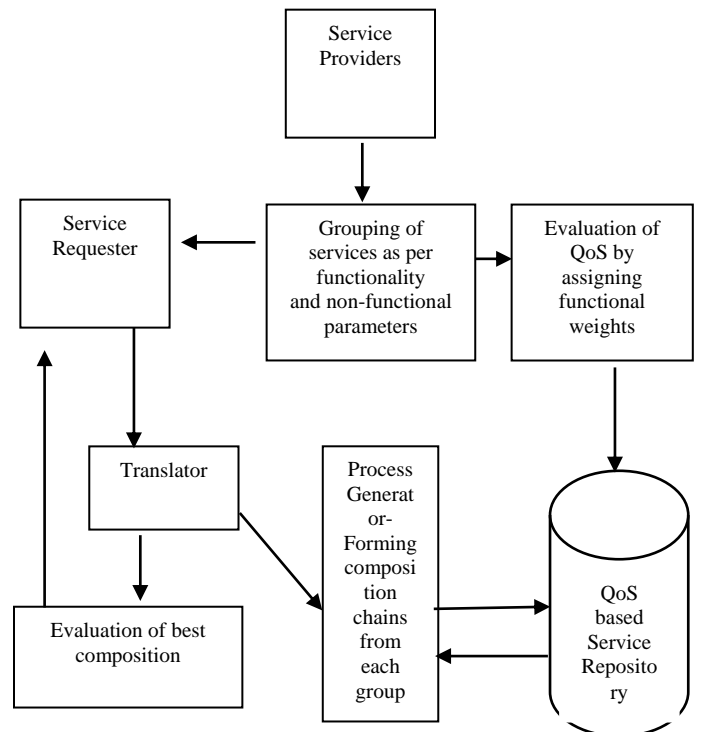


Fig. 1. Functionality based web service composition framework.

The methodology of functionality based composition of web services include the following steps.

**A. Identify the Functionality Equivalent Web Services in each Group that Are Needed to Compose**

The first step is to identify the group of web services that are to be included in the chain of composition. In each group the services that can satisfy functionality only be added. In this step the atomic web services collected into a specific category based on the minimum expected QoS and functionality. For that the non-functional parameter values are verified with the expected level to reach the actual functionality [11].

For example, the services that need to complete the task for online book purchasing system may be as follows:

*Book Ordering service* → *Warehouse service* → *Payment service* → *Delivery service*

The services with the above four categories are grouped based on its QoS and evaluated by fixing appropriate functional weights.

Consider  $G_1WS_1, G_1WS_2, \dots, G_1WS_n$  are the services in the first category. Similarly, in each category the services in other groups are also collected.

The non-functional parameters  $P_1, P_2, P_3, P_4$  and  $P_5$  are considered to evaluate the QoS for individual services are as follows:

- $P_1$  - Response time     $P_2$ - Availability
- $P_3$  - Throughput       $P_4$ - Successibility
- $P_5$  - Reliability

Table I lists all the average non-functional attribute values for 10 web services in the group  $G_1$  are as follows:

TABLE I. WEB SERVICE NON-FUNCTIONAL VALUES

S.no	Web services	P1 (ms)	P2 (%)	P3 (Inv/s)	P4 (%)	P5 (%)
1	$G_1WS_1$	290	90	5	95	80
2	$G_2WS_2$	210	95	4	99	60
3	$G_3WS_3$	135	65	8	65	75
4	$G_4WS_4$	140	80	3	80	70
5	$G_5WS_5$	255	80	9	80	75
6	$G_6WS_6$	145	99	25	99	70
7	$G_7WS_7$	165	85	30	85	80
8	$G_8WS_8$	130	90	13	99	75
9	$G_9WS_9$	160	95	2	95	80
10	$G_{10}WS_{10}$	125	90	8	90	70

**B. Ranking of the Web Services based on QoS after Setting the Weight to the Non-Functional Parameters**

The performance of the composition is directly depending on the extent to which the atomic services reach its functionality. So, the appropriate the weightage is fixed by the requester of the service to the individual non-functional parameter [14]. The same weight is assigned in each category but not required to be for the other web services in different categories.

For example,  $G_1$  is a group that contains the services with similar functionalities. Here priority is given to those services based on a particular non-functional parameter that influence the service to reach the actual functionality. Based on that a particular weightage is given to all the non-functional parameter with expected level [5].

Let  $G_1WS_i, G_2WS_i, \dots, G_5WS_i$  are the categories of services included in the composition in order.

The weights assigned to the non-functional parameters  $P_1, P_2, P_3, P_4$  and  $P_5$  are represented as in Table II.

TABLE II. WEB SERVICE NON-FUNCTIONAL WEIGHTS

S. no	Service Name	W1 (0-1)	W2 (0-1)	W3 (0-1)	W4 (0-1)	W5 (0-1)
1	$G_1WS_i$	0.4	0.69	0.76	0.52	1
2	$G_2WS_i$	0.55	1	0.68	0.91	0.84
3	$G_3WS_i$	0.91	0.87	0.93	0.78	1
4	$G_4WS_i$	1	0.73	0.64	0.82	0.85
5	$G_5WS_i$	0.83	0.69	0.84	1	0.85

The QoS of each web service in the categories can be calculated as follows:

$$QoS(G_iWS_j) = \frac{1}{m} \sum_{j=1}^m W_j \cdot P_j$$

Where,  $m (1 \leq j \leq m, 1 \leq i \leq n)$ .

After evaluating the QoS, the services are ranked in each category and stored in the service repository.

**C. Dynamic Composition of Web Services from the Pool based on the Availability**

In this step the process generator generates all possible combinations of web services from each group that are considered to complete the task (Fig. 2). As a result, based on the availability of services, the maximum possible compositions are prepared by the process generator. Also, the overall QoS of the generated combinations are prepared and the best available composition is selected to complete the task.

G1                      G2                      G3                      Gn

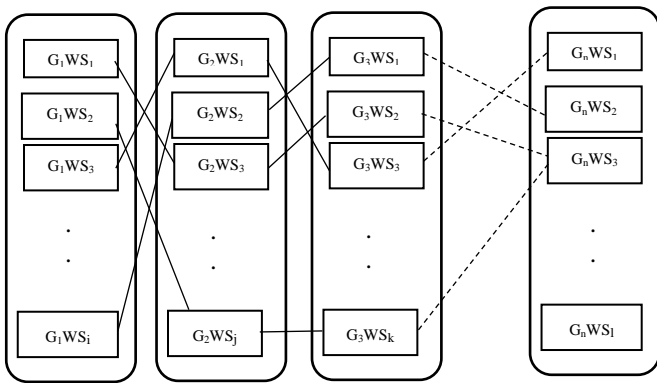


Fig. 2. Generation of composition chains.

Let  $C_1, C_2, C_3 \dots C_n$  are the possible compositions generated by the process generator.

From that, each  $C_i$  may contain the services  $G_1WS_j, G_2WS_j \dots G_nWS_j$ .

Here the  $WS_j$  is selected from each group with best QoS and its availability.

The overall QoS of the composition is:

$$QoS(C_i) = \frac{1}{n} \sum_{m=1}^n QoS(G_i, WS_j)$$

Where  $m (1 \leq i \leq n)$

The process generator algorithm ranks all these compositions as per the evaluated QoS. The best composition is selected from this and executed to complete the task.

#### IV. EXPERIMENTAL RESULTS

The number of services in each category may depend on the non-functional parametric values that satisfy the functionality of the web services. The QoS of the services are evaluated and grouped in specific categories to form the composition.

Table III shows the evaluated QoS of all the services that are in the group  $G_1$ .

TABLE III. EVALUATED QoS OF WEB SERVICES

$G_1WS_1$	$G_1WS_2$	$G_1WS_3$	$G_1WS_4$	$G_1WS_5$
0.338	0.384	0.646	0.502	0.515
$G_1WS_6$	$G_1WS_7$	$G_1WS_8$	$G_1WS_9$	$G_1WS_{10}$
0.707	0.447	0.665	0.706	0.585

TABLE IV. OVERALL QUALITY OF COMPOSITIONS

$C_1$	$C_2$	$C_3$	$C_4$	$C_5$
0.562	0.658	0.455	0.682	0.732
$C_6$	$C_7$	$C_8$	$C_9$	$C_{10}$
0.714	0.566	0.489	0.457	0.791
$C_{11}$	$C_{12}$	$C_{13}$	$C_{14}$	$C_{15}$
0.356	0.671	0.744	0.622	0.693

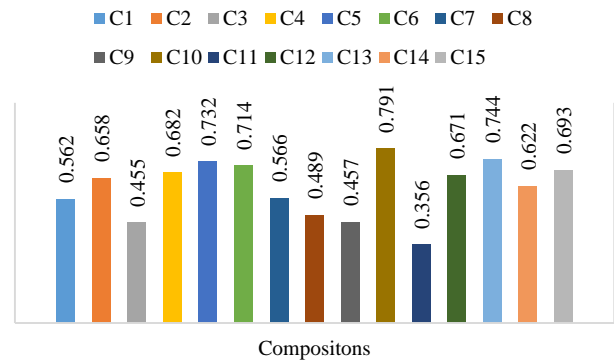


Fig. 3. Performance of composition chains.

These services are ranked as per the QoS and listed in the group. The same process will be followed for all the services that are grouped in the categorized chain. Finally, the process generator develops all the possible compositions  $C_1, C_2 \dots C_n$ . From this, top listed compositions that are up to the expectation form the requester are selected by the evaluator. When the task is invoked the evaluator executes the best composition based on its availability. Table IV lists the overall QoS values evaluated for the compositions.

The performance level of all the combinations generated by the process generator is shown in Fig.3.

#### V. CONCLUSION

The functional requirement of the web services involved in a particular task may vary from service to service based on its performance. Also, it is not sure that all the web services in the composition are at the same level of QoS because it will be the overhead in terms of cost. So appropriate QoS is fixed before composing the web services. Selecting an individual service up to the expectation means attaining the actual functionality of that service. This work presents the method that fix weightage to the non-functional parameters that influence the actual performance of the service. After selecting a pool of such service from each domain they are ranked based on quality. When composing a particular chain, the services are dynamically created based on its availability. Finally, the best composition is executed to complete the task to achieve the maximum performance.

#### ACKNOWLEDGEMENTS

This research project was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University, KSA under specialized research project grant no. 2017/01/7561.

#### REFERENCES

- [1] T. Yu "Service Selection Algorithms for Composing Complex Services with Multiple QoS," In Proc of ICSOC, pp. 130-143, 2015.
- [2] L. Zeng, "Quality Driven Web Service Comp.," Proc. of 12th Intl WWW Conference, 2013.
- [3] Yu, and S. R.-Marganec, "Service composition using selective appr. Based on backwards composition context Service", IEEE Intl, pp. 419 – 426, Oct 2009.
- [4] R. Berbner, "QoS-aware Web Services Composition using Heuristics approaches," Intl, pp. 72 – 82, Dec 2006.



- [5] R. Berbner, "Design Arch. for a Quality Service management of Web task oriented Work flow," Networking and Electronic Communication Research Conference (NAEC 2005).
- [6] Jinghai Rao and Xiaomeng Su, A Survey of Automated Web Service Comp. Methods, Semantic Web Services and Web Process Comp., First International Workshop, SWSWPC 2004, San Diego, CA, USA, July 6, 2004
- [7] Ikbel Guindra, "Heuristic Based Time-Aware Service Selection Approach," International Conference on Web Services, Jun 2015, New York, United States. 2015, Web Services (ICWS), 2015 IEEE Intl Conf.
- [8] Xu, S. Reiff-Marganiec, "Immune Algorithm implemented using Empirical Web Task Arrangement," IEEE -Web Services ICWS '08Intl, pp. 238 – 245, Nov 2008.
- [9] M. Alrifai, T. Risse, "Combining Global Optimization with Local Selection for Efficient QoS-aware Service Comp.," Proc of the 18th international Conference on World Wide Web, ACM, 2009.
- [10] Sha MM, Manesh T, Mohamed Mustaq, A Structure for Assessing the Quality of Service and Cost of Web Task Built on Its Well-designed Presentation Intl Journal of Computer, Electrical, Automation, Control and Engineering Vol:10, No:1, 2016.
- [11] C. Zhao, "A Active Web Task Algorithm. Built on the Grouping of Ant Colony algorithm and Genetic procedure," Journal of Comp Information Sys, pp. 2617-2622, 2010.
- [12] Jiuyun Xu Stephan, Reiff-Marganiec, ICWS '08. IEEE Intl Conference on, Proc, pp. 238-245. September 23 - 26, 2008
- [13] W. Zhen-wu, "An Method for Web task built on QoS and Separate Unit Swarm Optimization," Eighth ACIS Intl Conference on Software Engineering, IEEE, 2007.
- [14] Mohemmed Sha, Vivekanandan K. "Selection of Web Services Based on Providers Reputation," International Journal of Innovation Technical Expl Eng. 2013; 3(2):140-143.

# Encrypted Fingerprint into VoIP Systems using Cryptographic Key Generated by Minutiae Points

<sup>1</sup>Mohammad Fawaz Anagreh

Department of Computer and Self Development, Prince  
Sattam Bin Abdulaziz University,  
Kharj, KSA

<sup>2</sup>Anwer Mustafa Hilal

Department of Computer and Self Development, Prince  
Sattam Bin Abdulaziz University, KSA Kharj, KSA  
Faculty of Computer Science and Information Technology,  
Omdurman Islamic University, Khartoum, Sudan

<sup>3</sup>Tarig Mohamed Ahmed

Department of MIS, Prince Sattam Bin Abdulaziz University,  
KSA, Kharj, KSA, Department of Computer Sciences,  
University of Khartoum, Khartoum, Sudan

**Abstract**—The transmission of the encryption voice over IP is challenging. The voice is recorded, eavesdropping, change and theft voice, etc. The voice over IP is encrypted by using Advance Encryption Standard (AES) Algorithm. AES key is generated from Minutiae Points in fingerprint. By other way, we talk about biometric-cryptosystem, which is hybrid between one of the cryptosystems and biometric systems, such as fingerprint using for authentication as well as to generate cryptographic key to encrypt voice over IP by applying AES. In this paper, we define a new term which is Fingerprint Distribution Problem (FDP) based on Key Distribution Problem. Also, we suggest a solution for this problem by encrypted fingerprint before sending between users by using one of the public key cryptosystems which is RSA Algorithm.

**Keywords**—IP; cryptography; fingerprint; minutiae Advance Encryption Standard (AES); RSA; information security

## I. INTRODUCTION

Now a day, the computer networks allow sending different types of data via communication channels which connected with each other, audio, voice, text and video are different examples of data. Almost types before sending via communication channel will converted to the bits encapsulated in the packets. Voice over internet protocol (VoIP) is a generally term for a many transmission technologies to deliver voice over internet protocol (IP) using applications designed for this purpose [8], [11], VoIP should be known as IP Telephony also Voice over IP protocols carry telephony signals as digital audio.

Deliver voice over IP commonly used between massive numbers of users around of the world, but the problems are begin when attacker eavesdropping calls between users by attack communication channel, to avoid these problems many of researchers suggested solutions for this shortcoming by encrypt the voice before deliver via communication channels, one of these solutions is a system which encrypt the VoIP data packets using Advanced Encryption Standard (AES), AES key

is extracted from minutiae points from fingerprint authentication. There are some of shortcomings demonstrate when apply this technology, the main shortcoming is when send the fingerprint between users via unsecure communication channel because these fingerprint is not just for authentication but it will be used to generate key for AES algorithm [1], [2].

In this paper we suggest solution for this shortcoming. The rest of the paper is organized as follows: Section 2 provides a brief introduction to VoIP and its security vulnerabilities. As well as it provides a brief introduction to AES and RSA Algorithms also a brief Cryptographic Key Generated from Biometrics. Section 3 describes what the Fingerprint Distribution Problem is. Section 4 presents a solution for the fingerprint distribution Problem. Section 5 is the conclusion for our work.

## II. RELATED WORK

### A. VoIP Security Vulnerabilities

Internet Protocol Version 6 (IPv6) has been reached to become internet protocol for next generation. Especially that IPv6 consist of more updated compared with IPv4 in terms, number of new IP addresses ( $2^{128}$ ). As well as add more improvements in area specially, IPv6 support a new mechanism which called flow label that allows to support traffic such as real-time audio and video, more than in IPv4, Support for more security about encryption and authentication options and a new options for additional functionalities [11]. Problems at voice over IP are more and different, we can divide the problems into two categories based on situation of occurred. The first one is threats to the network; the second is threats to end users. Voice over IP is converted to the packets before send to other users in the network by communications channels, the problems here if the data unencrypted, then anyone can access to the data when sending between sender and receiver. Therefore, the attackers can listen to the calls and they can record the conversation [8].

### B. RSA and AES Algorithms

Encryption is one of security technology for computer, Encryption is based on transformation data or messages from original status called plaintext to a new status called cipher text, the features of a new status are unreadable for anyone except those possessing special knowledge, which are encryption algorithm and secret key. Decryption is the conversion cipher text into original status (plaintext), and also nobody can convert that just who has possessing secret key and encryption algorithm. There are two essentially types for encryption based on keys, the first one is public key cryptosystem which have two keys (public and private), public key is used to encryption, the private key is used for decryption. The other type is a private key cryptosystem, which have one key used for both encryption and decryption.

**AES Algorithm:** Is a block cipher. Advance Encryption Standard (AES) was established in 2001 by the National Institute of standards and Technology as a development on DES, AES takes a fixed data block size of 128 bits and unfixed key size of 128 bits, 192 bits or 256 bits. In encryption phase, AES depend on round transformation from plaintext to cipher text, number of rounds depend on key length. If the key is 128 bits then uses 10 rounds, 192 bits uses 12 rounds and 256 uses 14 round. The stages of all rounds are Sub Bytes, Shift Rows, Mix Columns and Add Round Key [7].

**RSA Algorithm:** Is a first published by the three researchers Ron Rivest, Adi Shamir and Len Adleman in 1977 [12]. The name of algorithm came from the initials of surnames for the researchers. However, the RSA Algorithm using both digital signature and public key encryption, as any algorithm in this filed, RSA consist of two keys, the public key is used for encryption and private key for decryption.

RSA algorithm consists into three phases: key generation phase, encryption and decryption phases. The user of RSA applies the key generation phases to generate keys based on two big prime numbers. The prime number must be kept secret, then apply other steps in key generation phase to get the keys of RSA. As a rest of public key cryptosystems, any one has the public key can encrypt the plain text to get the cipher text, then send the cipher text to the user who generated keys to decrypt the cipher text to get the plain text. The first user (generator keys) transmit his public key ( $n, e$ ) to the second user via communication channel, the private key  $d$  is never distributed any way. Suppose Bob (first user) would like to send messages (Plain text)  $M$  to the Alice (second user), Bob converts the Plain text  $M$  to the cipher text using Public key of Alice  $e$  according to the equation:

$$C = M^e \text{ mod } n$$

After get  $C$ , Bob send the encrypted message  $C$  to the Alice. While Alice received the Cipher text  $C$ , She decrypts the cipher text  $C$  to get the Plain text  $M$  by using her private key exponent  $d$  according to the following equation:

$$M = C^d \text{ mod } n$$

To generate the public and private keys of RSA algorithm by apply following steps:

- Select two prime numbers  $p, q$

- Compute  $n = p * q$
- Calculate  $\phi = (p-1)*(q-1)$
- Choose an integer number  $e$ , by  $1 < e < \phi$ ,  $\text{gcd}(e, \phi) = 1$
- Compute  $d$ , by  $1 < d < \phi$ ,  $e * d = (1 \text{ mod } \phi)$
- Obtain the keys, Public key ( $n, e$ ), Private key ( $d, p, q$ )

The key length of RSA is referring to the modules  $n$ , it is now 1024 bits, 2048 bits or more. Key length with 512 bits is now no longer recommended secure. Therefore, the recommendation is to generate two big numbers  $p$  and  $q$  to insure a big modules number  $n$ . Key length with 1024 bits is a round 300 decimal digits as following example:

$n=778777413433370950905552740560125564964460406$   
96615275036985244819549430568511503338363159570377  
1562029730000007708466899615108922122454571180605  
78888989517080042203063427376322274266393116193517  
83957077350545520309668112192733747397322031251259  
90248513222506060062600665575382385175753906212629  
20956913963

Generate  $n$  above is composed of two big random prime numbers  $p$  and  $q$ :

$P=445571661151720883066847154799984650223454138$   
74567112127345628767000822584313029655212749702453  
44793522942129064489358577701861556582847914646983  
63257581748

$q=861492264535438176093706088214174899339429981$   
01549682098342251385596444849727109106169673491102  
31723734078976011179021708289824396553412180514827  
9973690446

### III. CRYPTOGRAPHIC KEY GENERATED FROM BIOMETRICS

The secret key generated from biometric is common used recently, easy to generate and no need to remember the strong secret key. As well as, key is a big size cause difficult for some people to manage the cryptographic key [5], [9]. Recently, many proposals have been suggested many methods to generate cryptographic key based on biometric such as fingerprint [3]-[5], [10], [13]-[15]. According to (Arul and Shanmugam), they selected fingerprint as the biometrics features to generate a cryptographic key, that can be done by extract minutiae points from the finger print. The group of points are managed together by some methods into seven phases to generate the cryptographic key of AES Algorithm.

#### A. Fingerprint Distribution Problem (FDP)

In this section, will assume two person, Alice as Sender and Bob as recipient. Suppose, Alice wants to begin calling Bob, the voice is delivering over internet protocol.

#### Assumptions

- FP<sub>Alice</sub> → Alice Fingerprint
- K<sub>Alice-Bob</sub> → AES key for both Alice and Bob
- P<sub>Alice-voice</sub> → Original Voice Packets-Alic
- C<sub>Ciphervoice</sub> → Cipher Voice Packets-Alice

G → Generate AES Key from Minutiae Points in Fingerprint  
 E → Encryption  
 D → Decryption

**Step 1:** Alice send fingerprint to the Bob via communication channel.

**Step 2:** Bob receive the fingerprint from Alice and will compare automatically with fingerprints stored in database of Bob. When fingerprints are same authentications is done.

**Step 3:** Alice Generates AES Key from fingerprint by apply Arul and Shanmugam method [1] as denoted:

$$FP\_Alice \rightarrow G(K\_Alice-Bob)$$

**Step 4:** Alice encrypts the voice packets using AES algorithm by use AES Key generated from minutiae points in Alice fingerprint as denoted:

$$C\_Ciphervoice = E(P\_Alice-voice) \text{ by using } K\_Alice-Bob$$

Next, Alice sends the cipher voice to the Bob.

**Step 5:** Bob Generates AES Key from fingerprint by apply Arul and Shanmugam method as denoted:

$$FP\_Alice \rightarrow G(K\_Alice-Bob)$$

**Step 6:** Bob decrypts the cipher voice using AES algorithm by use AES Key generated from minutiae points in Alice fingerprint as denoted:

$$P\_Alice-voice = D(C\_Ciphervoice) \text{ by using } K\_Alice-Bob$$

Bob will apply last steps before begin send voice over IP. According to steps above, we can summarize and detect some things. No need to generate key as a statically because the key automatically generated. As well as, there is no need to store the key because fingerprint generates AES key when needed. Consequently, we consider that as a solution for the key distribution problems because there is no need to send key via communication channel.

Both sender and recipient use the fingerprint to generate AES key for both encryption and decryption. AES key generated from minutiae points from fingerprint. By other way, the purpose of fingerprint is to generate cryptographic key and for authentication. The fingerprint should be saved in database for both sender and recipient for authentication purpose also to generate key for encryption\decryption voice before\after deliver over IP. Consequently, if attacker gets the fingerprint, he can generate the AES key by apply the same algorithm (see Fig. 1), here a new problem incoming, deliver fingerprint in unsecure communication channel allows attacker to exploit fingerprint to generate key by apply same algorithm like look sender or receiver. By other word, sending fingerprint in unsecure communication channel is similar to send a private key, then attacker can generate key and decrypt the cipher voice over IP. Here, we name a new term, which is a Fingerprint Distribution Problem (FDP).

To solve the Fingerprint Distribution Problem, we suggested encrypt fingerprint before sending between users also keep the fingerprint encrypted in the database. We solve

FDP by using public key cryptosystem because this technique was proposed to solve key distribution problem, we mentioned that a sending fingerprint in the unsecured communication channel is same a key distribution problem.

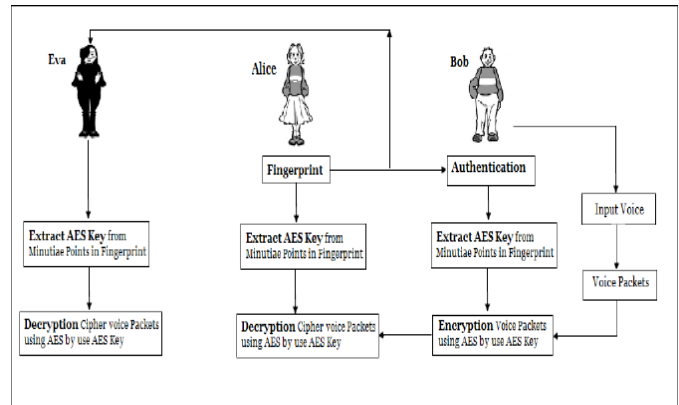


Fig. 1. Fingerprint distribution problem.

### B. Proposed Solution of Fingerprint Distribution Problem

Public key cryptography was invented in 1976 by Diffie and Hillman [6], the main goal of public key cryptography is to solve Key Distribution Problem that obtain a protocol to pass the public key and ciphertext between users in communication channel and use private key for decryption, we solve the fingerprints distribution problem by encrypt the fingerprint before sending via unsecure communication channel. We use RSA Algorithm as a one of public key cryptography (see Fig. 2). Sending fingerprint in unsecure communication channel is as send private key to the other user because attacker can extract the AES key from fingerprint by apply the algorithm to extract Minutiae Points from fingerprint. Assume two users Bob and Alice as following steps:

#### Assumptions

$K_{e\_Bob}$  → Public key for Bob generated by apply RSA algorithm

$K_{d\_Bob}$  → Private Key for Bob generated by apply RSA algorithm

$K_{e\_Alice}$  → Public key for Alice generated by apply RSA algorithm

$K_{d\_Alice}$  → private key for Alice generated by apply RSA algorithm

$C\_Cipherfingerprint-Alice$  → Encrypt Alice fingerprint by apply RSA algorithm using Bob public key

$C\_Cipherfingerprint-Bob$  → Encrypt Bob fingerprint by apply RSA algorithm using Alice public key

**Step 1:** Alice and Bob generate keys by apply RSA Algorithm.

$$Bob \rightarrow G(K_{e\_Bob}, K_{d\_Bob})$$

$$Alice \rightarrow G(K_{e\_Alice}, K_{d\_Alice})$$

**Step 2:** Both Bob and Alice send the public key for each other.

$$K_{e\_Bob} \rightarrow \text{to Alice}$$

$K_{e\_Alice} \rightarrow$  to Bob

**Step 3** Encrypt fingerprint by apply RSA algorithm using public key, send cipher fingerprint to each other and save in database as denoted:

$C\_Cipherfingerprint-Alice = E(FP\_Alice)$  by apply RSA Algorithm using  $K_{e\_Bob}$   
 $C\_Cipherfingerprint-Alice \rightarrow$  Send to Bob

Also Bob apply as denoted:

$C\_Cipherfingerprint-Bob = E(FP\_Bob)$  applying RSA Algorithm using  $K_{e\_Alice}$   
 $C\_Cipherfingerprint-Bob \rightarrow$  Send to Alice

Both Alice and Bob receipt the cipher fingerprint from the other and save in their database .

Bob saves in database  $\rightarrow C\_Cipherfingerprint-Alice$

Alice saves in database  $\rightarrow C\_Cipherfingerprint-Bob$

**Step 4:** Each one has encrypted fingerprint in their database, assume Alice want call Bob:

Alice send cipher finger print to Bob

$C\_Cipherfingerprint-Alice \rightarrow$  Send to Bob

Bob Decrypt the cipher finger print of Alice by use RSA algorithm using Bob private key  $K_{d\_Bob}$  as denoted:

$FP\_Alice = D(C\_Cipherfingerprint-Alice)$  by apply RSA algorithm using  $K_{d\_Bob}$ .

**Step 5:** Bob compare the finger print  $FP\_Alice$  which is came from Alice with fingerprint of Alice saved in Bob database, if two fingerprints are same then the authentication is done.

**Step 6:** Begin calling between Alice and Bob by apply AES algorithm (Section 3) to send encrypted voice packets via communication channel.

The goal from all above to avoid transmit fingerprint using to generate cryptographic key via unsecure communication channel.

#### IV. CONCLUSION

This paper proposed a method to stream cipher voice packets encrypted by using key generated from Minutiae Points from fingerprint also define a Fingerprint Distributed Problem. We suggested a solution for the FDP. This approach has reduced of probability attack this system by increasing layers of security, especially by suggested using RSA algorithm to encrypt fingerprint before send them in unsecure communication channel.

#### REFERENCES

- [1] Arul, P., Shanmugam, A. (2009). Generate A Key for AES Using Biometric for Voip Network Security. Journal of Theoretical and Applied Information Technology, Vol, 15, No.2, pp. 107-112.
- [2] Arul, P., Shanmugam, A. (2008). "New-Fangled Fingerprint Engendered Key For A Secured VoIP". In proc. 7th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications, pp. 45-48, England, UK.
- [3] Bais, R., Mehta, K. (2012). Biometric Parameter Based Cryptographic Key Generation, International Journal of Engineering and Advanced Technology (IJEAT), Vol.1, pp.157-160.
- [4] Ballard, L., Kamara, L., Monrose, F. (2008). "Towards Practical Biometric Key Generation with Randomized Biometric Templates", Proceedings of the 15th ACM conference on computer and communication security, pp. 235-244, Alexandria.
- [5] Balakumar, P., Venkatesan, R. (2011). Secure Biometric Key Generation Scheme for Cryptography using Combined Biometric Features of Fingerprint and Iris, IJCSI International Journal of Computer Science Issues, Vol. 8, No. 2, pp. 349-356.
- [6] Diffie, W., Hellman, M. (1976). A new directions in cryptography, IEEE Transactions on Information Theory Vol. 6 Num, 22 PP. 644-65
- [7] FIPS (2001). 197: Announcing the advanced encryption standard (AES).
- [8] Goode, B., (2002). "Voice over Internet protocol (VoIP)," Proceedings of the IEEE , vol.90, no.9, pp.1495-1517, USA.
- [9] John, J., Rajesh, T. (2013). Multiple Key Generation using Elliptic Curve Cryptography Fusion Algorithm for Biometric Source. International Journal Of Engineering And Computer Science, Vol.2, Issue.3, pp. 732-740.
- [10] Murugesh, R., (2012). "Advanced biometric ATM machine with AES 256 and steganography implementation," Advanced Computing (ICoAC), 2012 Fourth International Conference on, vol., pp.1-4, Munich, Germany.
- [11] Nisar, K., Said, M., Hasbullah, H. (2010) "Enhanced performance of packet transmission using system model over VoIP network", proceeding of Information Technology (ITSim), 2010 International Symposium, vol.2 . pp. 1005-1008. Malaysia.
- [12] Rivest, R., Shamir, A. Adleman, L. (1978). A Method for Obtaining Digital Signatures and Public-Key Cryptosystems, COMMUNICATIONS OF THE ACM, pp. 120-126.
- [13] Sharma, R., (2012). Generation of Biometric Key for use in DES. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue.6, No.1, pp.312-315.
- [14] Uludag, U., Pankanti, S., Prabhakar, S., Jain, A., (2004) "Biometric cryptosystems: issues and challenges," Proceedings of the IEEE , vol.92, no.6, pp.948-960, USA.
- [15] Abuguba, S., Milosavljevic, M. M., & Macek, N. (2015). An efficient approach to generating cryptographic keys from face and iris biometrics fused at the feature level. International Journal of Computer Science and Network Security (IJCSNS), 15(6), 6.

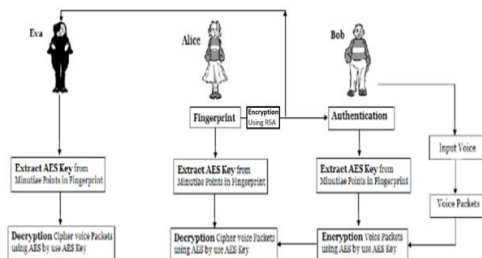


Fig. 2. Fingerprint distribution resolved.

# General Characteristics and Common Practices for ICT Projects: Evaluation Perspective

Abdullah Saad AL-Malaise AL-Ghamdi<sup>1</sup>, Farrukh Saleem<sup>1,2</sup>

<sup>1</sup>Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

<sup>2</sup>Faculty of Computing, Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

**Abstract**—In today's business world, organizations are more dependent on Information and Communication Technologies (ICT) resources. Cloud services, communication services and software services are most common resources, enterprises are spending large amount. To install new services and upgrade existing services, ICT project are essential part of organization's business strategies. Researchers highlighted the real problem for the organization is to initiate new ICT projects and its evaluation after implementation. This research investigated the common approaches organizations using to start with ICT projects and how to evaluate its impact on after implementation. For this, we have extracted the number of steps with the help of literature review. To validate those steps, six case studies are selected for collecting the samples. The findings of this study elaborate that every ICT project has list of objectives i.e. strategic, informational, IT infrastructure and others. Furthermore, the results highlight that organizations believe on both financial and non-financial evaluation methods based on the type of organization i.e. public or private. Moreover, measurement process applied on project wise, monthly and yearly bases. Importantly, we have found that currently outsourcing plays significant role in success of ICT projects. The results of this study can be helpful for the organization to understand the type of ICT investments, approaches and possible impact on the organizations goals.

**Keywords**—ICT project; ICT evaluation; measurement process; case studies; common practices

## I. INTRODUCTION

Information and communication technologies (ICT) are considered one of the main pillars in building business architecture in any organization. The ICT has been categorized in different perspectives such as; hardware and software resources [1], IT infrastructure [2], cloud computing [3], information management tools [4] and different kinds of information systems [5]–[7]. In order to improve the services and business processes, organizations intend to plan, build and implement different kinds of ICT projects every year. Based on the report published by Gartner<sup>1</sup> the ICT spending has been reached to billions of US Dollars. Although, the main purpose of organization to spend large amount of ICT investment is to improve employee productivity [8], customer satisfaction [6], [9], enhance data management [10] and last but not the least, to align ICT resources with business strategies [11].

Due to this large investment on ICT resources, measuring the performance and underutilization of ICT resources are

major concern of the enterprises. This research is actually highlights the common practices for ICT investment and measurement process practicing in an organization. In addition, how to measure the benefits from already implemented ICT projects. The first part of this study is to highlight the ICT project investment process. Normally, ICT investment are implemented based on list of objectives such as; to improve information management process [10] enhance the scalability of IT infrastructure [12], [13] to enhance business structure/process transformational phase [14] or to increase the transactional capabilities [11]. Whereas single ICT projects can help to achieve one or many objectives based on its implementation.

Second major aim of this study is to investigate the common approaches enterprise following to measure those kinds of investments. Evaluation of ICT project's impact on business values/benefits is a complex process and involves multiple stages [15]. The vast literature review suggested number of approaches, techniques, and phases required to accomplish this task. Each approach has its own characteristics and context to be used, which is based on the type of investment and list of expected benefits for which a firm has initiated in the ICT project. Due to complexities involve during evaluation process, organizations are still struggling to know the optimal technique for ICT project post evaluation. Therefore, the idea in this study is to investigate the current status/procedures of organizations, following for ICT project implementation and evaluation. To do this, number of questions extracted from literature review related with ICT projects investment and evaluation. Furthermore, the questions have been asked from major representative of the selected organizations. The extraction procedure of the topics and details of case studies are presented in Section 2. In addition, Section 3 discusses the major findings, limitations and analysis on collected data.

## II. ICT INVESTMENT AND MEASUREMENT PROCEDURE EXTRACTED FROM LITERATURE REVIEW CONFIRMED THROUGH CASE STUDIES

Researchers elaborated ICT investment and evaluation in different ways. There are number of researcher published already supported the idea of using case studies for ICT investment and measurement the post implementation benefits [4], [10], [11], [14], [16]. This study highlights the organization's point of view in finding out the objectives behind every ICT investment and how organization evaluates the performance of those projects. This section categorized in eight sub-sections, whereas each sub-section aimed upon the

<sup>1</sup> www.gartner.com

major topics extracted from literature review and further asked in case studies. Section A discussed about case study characteristics and participants overview. An overview of ICT usage and motivation toward projects in the case studies presented in Section B. Section C elaborate the common practices for ICT investments and evaluation found in case studies. Generally, ICT investments are consisting of different kinds of input and output resources as extracted from literature, discussed in Section D, while Section E, described the major objectives behind ICT project investment. Moreover, Section F talks about types of measurement methods examined from case studies. The case studies are following which kind of evaluation approach inspected in Section G. Rather they are outsourcing their investment and evaluation using third party assistance discussed in Section H.

### A. Case Study Characteristics

The discussion of the results begins with the summary of the six case studies from Saudi region, selected in this research. It especially helps us to understand the participation of each case study, number of employees, and their experiences in ICT field. In order to ascertain the appropriateness and guarantee of the data sources as well as participants from each case studies. The participant were majorly selected for this study are top management, IT project director, supervisors, software developers, IT executives, IT project managers, IT project team members, IT and business users. The number of questionnaire received with the percentage of whole, associating from each case study as depicted in Table I. The sample size is better than the size used in related studies which was 143 [10]. Also the sample size meeting with the condition explained by [17] that the minimum of 5 questionnaires are essential for each variable.

Table I categorizes the ratio of participants under each case who have recorded their responses. It highlights the high responses collected from airline and University is 20.5% and 17.9%, respectively. While other four cases fall around 13.8% to 16.4% which altogether counted as fair average of responses per case. The variation in collected responses collection was based on data-collection timing and other circumstances inside the organization. In order to gauge the size of the organization, a question was asked about the approximate number of employees working at an organization, as demonstrated in Table I. It appears that a large number of responding firms (38.4%) have the largest number of employees as compared with all six case studies. This indicates that the number of participants taken the part in this research belonged to the company, which has more than 1000 employees. On the other side, 13.8% of the participants associated from the bank that have lowest number of employees as compare with all other companies selected in this research. The contribution in this research was essential, especially from participants who have good experience in ICT. Therefore, a question was included to obtain the information about their experiences in this field. It is apparent from the Table II that 43.32% of participants have the experience of between 6 and 10 years, while 30.3% were part of this field since around five and less years. The indicators of experience highlight the significance of the collected data, as 69.7% of participants have been associated with the ICT field.

TABLE I. PARTICIPANTS OVERVIEW IN CASE STUDIES

Companies	Participants in Numbers	Participants in Percentage	Number of Employees
Airlines	40	20.5	More than 1000
University	35	17.9	More than 1000
Food Industry	31	16.0	Around 800
Water & Electric	32	16.4	Around 600
Bank	27	13.8	Around 50
Telecommunication	30	15.4	Around 100
<b>Total</b>	<b>195</b>	<b>100.0</b>	

TABLE II. NUMBER OF YEAR EXPERIENCE IN ICT

No. of Years' Experience in ICT	In Percentage (%)
5 or Less	30.3
6 to 10	43.32
More than 10	26.38
Total	100.0

### B. Motivation towards ICT Project's Investments

Recent developments in ICT have heightened the organizational vision to invest more and get more benefits in terms of profit, intact values to the business, and eventually more customers. As from the literature review, researchers have found the usage of ICT common in almost every organization. In [18], highlights the major spending in ICT categorized by sectors. We tried to confirm those sectors of ICT from the companies, whether they are investing and using those resources or not. Table III proves the findings of literature review, where all selected companies have been investing and using different types of resources mentioned in the table. The advanced technologies always put pressure on the companies to invest more and keep updated the ICT resources, all of the participants agreed upon it. They refused to provide the particular amount for each sector; rather they provided their answers in three categories "High," "Medium," and "Low", which highlights how frequently they invest in different categories of ICT resources.

TABLE III. ICT INVESTMENT IN DIFFERENT SECTORS IN CASE STUDIES

Companies	Data Center Systems	Software	Devices	IT Services	Comm. Services
Airlines	High	Medium	Medium	High	High
University	High	Low	Low	High	High
Food Industry	Medium	Medium	High	Medium	Medium
Water & Electric	Medium	Medium	High	Medium	Medium
Bank	High	Medium	Medium	High	High
Tele-communication	Medium	Medium	High	High	High

TABLE IV. GENERAL PURPOSES OF ICT INVESTMENTS IN CASE STUDIES

Companies	Stimulant		To Get Tangible Benefits				To Get Intangible Benefits			
	Internal Pressure	External Pressure	Technology Upgrades	Security	Training	Data Integration	Improvement Sales	Improvement Communication	Process Efficiency	Improve Customer Service
Airlines	✓	✓	✓	✓		✓	✓	✓	✓	✓
University	✓		✓	✓		✓		✓		✓
Food Industry	✓		✓	✓		✓			✓	
Water & Electric	✓		✓	✓		✓			✓	
Bank	✓	✓	✓	✓		✓		✓	✓	✓
Tele-communication	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

It is evident from the table that different companies have different objectives for ICT projects, as mentioned earlier. Airlines, university, and banks are investing in a high amount for data center systems, as their businesses are more dependent on data integration and validation. On the other side, food industry, water and electric, and telecommunication are investing in a high amount of purchasing devices to support their business processes.

Data integration, improved communication, and many others are the significant purposes for which companies are investing on ICT [19]. Based on the ICT investment and their related tangible and intangible benefits, the organization faces internal and external pressure to invest in a particular sector, where each sector can provide a selected list of benefits, as mentioned in Table IV. Receiving more benefits is another motivational factor investigated from case studies, which keep

creating stress (internal or external) on the organization and highlight reasons for investing on ICT. The result is based more generally on which types of benefits are based on ICT investments in a particular organization. The telecommunication company, which mainly considers ICT services providers, received the most benefits out of it. On the other side, “training” is the least interested item in the case studies that are associated with ICT investment.

Table IV highlights for what purposes these organizations invest more generally. The list of intangible and tangible benefits they can achieve from ICT projects. For example, technology update is very common tangible benefit which organization needs to update their software and license every year. For this they can feel pressure from internal (employees) or external (customers). On the other side intangible benefits such as process efficiency is required when system are running out of order several time during transactions. Most of the organization agreed that for getting those types of updates we always feel pressure from their employees, partners, collaborators and customers as well.

*C. ICT Project’s Investments: Common Practices in Case Studies*

An investigation was conducted to explore the common practice in case studies for each ICT project’s investment they want to implement. The findings in Table V can be used as a base for drawing general conclusions on the ICT project investment practices in the investigated organization. The analysis reveals that most of the organization is invested mostly based on projects, where a project can for strategic, informational, or other purpose [11] and summarized in Table VII. Specifically, the studied cases invested in three kinds of resources technology, human and relationship resources as described in [20], although all of them are critical for investing in technological resources.

The investment needs to be evaluated to find out the return from the investment, financial and nonfinancial types of methods used for evaluation in ICT investment in different cases. University, as being a government organization, is not supportive for measuring financial returns. Altogether, organizations are using different types of approaches, where ROI is being used most except in University. On the other side, airlines, banks, telecommunications, and university are using multidimensional methods as well for measuring business values or nonfinancial factors as a measuring return from the investments. As extracted from literature review post implementation measurement fall in two different categories; first return on investment (ROI) using non-financial benefits [10], [21]–[23], while second one is by evaluating using non-financial measuring factors [4], [11], [14]. It is understood from the case studies that post-multidimensional (for measuring nonfinancial return) and ROI (for measuring financial return) are the most common approach organization for using ICT evaluation. In addition, airline and telecommunication also mentioned the approach of building proposals using the pre- evaluation method such as IT portfolio management.



TABLE V. ICT PROJECT'S COMMON PRACTICES IN CASE STUDIES

Companies	Investment's Type	Input Resources	Evaluation Type	Time of Evaluation	Evaluation Approach	Measuring Factors
Airlines	Project Based	Technology Human Relationship	Financial Non-Financial	Pre & Post	ROI Portfolio Multi-Dimensional	Cost Business Values
University	Upgrading on regular bases	Technology Human Relationship	Non-Financial	Post	Multi-Dimensional	Business Values
Food Industry	Project Based	Technology Human Relationship	Financial	Pre & Post	ROI ROA	Cost
Water & Electric	Project Based	Technology Human Relationship	Financial	Pre	ROI ROA	Cost
Bank	Project Based	Technology Human Relationship	Financial Non-Financial	Pre & Post	ROI Multi-Dimensional	Cost Business Values
Tele-communication	Upgrading on regular bases	Technology Human Relationship	Financial Non-Financial	Pre & Post	ROI Portfolio Multi-Dimensional	Cost Business Values

*D. ICT Project's Investment: Input Resources and Output/objectives*

ICT investment is about implementation of resources for a particular sector to support the enterprises. Referring to [20] the input resources are of three types; human, technology and relationship, while each input resource generates multiple output [20]. Table VI is the combination of input resources and corresponding output generated from those resources investigated from case studies. All of the studied companies are investing in human and technology type of input resources to improve their company's performance and technical skill for their employees. The integration of technology and human resources create IT business value for the organization. Relationship resources are defined as those factors that can increase the business value of the firm. Corresponding on this point, airlines, University, banks, and telecommunication are the companies who are measuring business values from nonfinancial perspectives. The major factors used for evaluating relationship resources are accountability, leadership,

and organizational learning. In human resources, none of the organization is relating to ICT investment for supporting business understanding factor. Most of the participants do not relate ICT investment with business understanding perspectives.

The two approaches defined for measuring ICT investment are based on cost return (financial factors) and business value return (nonfinancial factors). All studied case studies are invested in human and technology resources. Technical resources can be evaluated using return cost factors, but human resources need to be evaluated through nonfinancial factors such as IT staff reduction, employee involvement, etc. Table VI indicates that all organizations are investing for human resources such as for improving technical skills, but it was shown, as in Table V, that only airlines, University, banks, and telecommunication are using non-financial approaches for evaluating their investment. Food Industry and water & electric needs to employ such nonfinancial techniques for evaluating their human resources more properly.

TABLE VI. ICT PROJECT'S INVESTMENT – INPUT RESOURCES AND OUTPUT/OBJECTIVES IN CASE STUDIES

Input Resources →	Human Resources			Technology Resources				Relationship Resources		
Output/objectives →	Technical Skills	Business Understanding	Problem Solving Orientation	Shareable Platforms	Databases	ICT Applications	Data & Security Standards	Business Partners Ownership	Accountability	Leadership
Companies										
Airlines	√		√	√	√	√	√	√	√	√
University	√			√	√	√	√			√
Food Industry	√			√	√	√	√			
Water & Electric	√			√	√	√	√			
Bank	√			√	√	√	√		√	√
Telecommunication	√		√	√	√	√	√	√	√	√

E. ICT Projects Objectives

ICT projects based on different types of objectives to accomplish particular goals are based on organizational requirements. The literature review suggested that ICT practitioners have placed them into eight categories [10], [11]. The current study consolidated those objectives for basic requirements while measuring ICT investments examined using case studies, as illustrated in Table VII. Competitive advantage and aligning of ICT strategy with business strategies are common objectives, which fall under the category of strategic types of investment. Most of the organizations relate ICT investments with their strategic objectives. Another objective investigated is informational type of investment on which all case studies regularly invest. The purpose of this investment is to enable the data to be accessed easily and faster. Information accuracy is the fundamental requirements of each case study examined. IT infrastructure is the most common term, which normally considers a pure ICT investment rather than some scholars, who have excluded it from the list of dimensions, which can create business value of ICT [10].

TABLE VII. ICT PROJECT’S INVESTMENT OBJECTIVES IN CASE STUDIES

Companies	Strategic	Informational	Transactional	Transformational	Organizational	Operational	IT Infrastructure	Managerial	Others
Airlines	√	√	√	√	√	√	√	√	
University									Upgrading on regular basis
Food Industry	√	√	√			√	√		
Water & Electric	√	√	√	√		√	√		
Bank	√	√	√		√		√	√	
Tele-communication									Upgrading on regular basis

For the purpose of reducing operating and communication cost, airlines, food industry, water & electric, and banks are doing ICT investments. They mentioned that installing advanced technologies, web servers, and data storage servers help to reduce operating and communication costs. This type of investment helps them to reduce work load and time consumption while performing different types of activities. While University and telecommunication did not mention this type of investment, they do regular updates to support transactional benefits as well. Most of the case studies did not provide enough evidence for ICT investment’s impact on organizational types of benefits. They are not specifically investing for getting benefits for organizational support such as work pattern, empowerment, and building common vision. Organizational learning, business understanding, and creating common vision are not really supported by investing on ICT, which most of the participants agreed upon. Only airlines and banks mentioned investment for organizational learning to increase employee morale and satisfaction toward business goals of the organization. Moreover, based on collected evidence, most of the time, telecommunication offers an investment plan based on a new service or enhancing previous services for their customers and employees. In summary, the case studies offer different types of aforementioned investments to support their business objectives and goals. They have shown their high interest toward ICT, thus having a positive impact on different types of category for creating business value.

F. ICT Project’s Evaluation Status: Common Practices in Case Studies

ICT project’s evaluation for large enterprises is an iterative process. There are different types of measurement they are following based on the evaluation requirements. Refining the findings from literature review the evaluation categories are known as financial and non-financial which further divide depending upon the time when organization are evaluating it before or after ICT project investment [21], [24]. The combination of type of measurement, time of measurement, and how frequently they are evaluating their ICT investment at cases studied are depicted in Table VIII.

As per the discussion, measuring through financial factors is the most commonly used technique we have found in case studies. University, which is a government organization, is not interested in financial returns of any investment. The budgeting and funding allocated per year by government, which is the main reason they are not acquiring return on investment (ROI) from any ICT projects. University is sometimes for nonfinancial measuring methods using feedback forms or by using other surveying methods, they examine the thinking and views of the stakeholders for any particular ICT services already offered. Airlines, banks, and telecommunication are also performed nonfinancial measurement factors using different factors such as knowing the competitive advantage in the market through information retrieval, workflow, and process efficiency.

TABLE VIII. ICT PROJECT'S EVALUATION STATUS IN CASE STUDIES

Companies	Type of Measurement		Time of Measurement		Frequency of Measurements		
	Financial	Non-Financial	Ex-Ante	Ex-Post	Six Monthly	Yearly	Project Wise
Airlines	√	√	√	√			√
University		√		√		√	
Food Industry	√		√	√			√
Water & Electric	√		√				√
Bank	√	√	√	√			√
Tele-communication	√	√	√	√		√	

Organizations believe in ex ante (pre) and ex post (post) evaluation based on criteria of measurement. From the literature review, we have extracted different types of measurement techniques offering pre- and post-evaluation [22], [25]–[27]. ROI is most commonly used method for measuring return on investment considered post-evaluation methods for assessing hard benefits or financial returns, to know exactly the cost and benefit ratio. Pre- and post-evaluation are common in each case study, while pre-evaluation is time-consuming and assessment is based on previous documents or reports and predicting the future. Risk measurements and possible outcomes discussed in pre-evaluation process during the planning phase of every new ICT project. On the other side, post-evaluation is based on the results and outcomes achieved already. For the given time period, proper measuring factors can provide the analysis and achieving objectives from the invested ICT project. Food Industry and water & electric case studies have mentioned investing on technology resources (Table VI), which can easily be assessed using financial methods to know the return of each ICT project. Some of the resources such as human resource have to employ techniques that can measure hard benefits as well. Airlines, banks, and telecommunication based their objectives believing in pre- and post-evaluation using financial and nonfinancial factors. This means that each of their investment has provided risk assessments and expected cost, which ultimately generates expected benefits in the form of return cost and business values.

The frequency of the ICT project measurement is investigated in the organization, where most organizations evaluate their investments project-wise. As airline, food industry, water & electric, and bank investing is particularly based on the ICT project with lists of objectives and input resources. It has been explored from the literature review and acquired through case studies that it makes it easier to measure ICT project return project-wise, as organizations are well known about objectives of investment, resources, and expected outcomes. Their mapping can provide enough evidence for measuring financial return as well as business values. Apart from an University-use yearly basis approach for ICT evaluation, where the return is measured using nonfinancial factors such as quality of University, faculty member, and course surveys, feedback for e-services facilities is provided to different stakeholders.

### G. ICT Project Evaluation Approaches

Four major approaches used for evaluating ICT investments extracted from literature review are traditional financial [21], [28], IT portfolio management [29], [30], multi-criteria [31], [32], and multidimensional [11], [15], [27], [33]. The most common financial techniques such as ROI and ROA are methods used to measure return on post-investment. It will provide the analysis using variable spending amount versus return amount after a given period of time. Portfolio management is the kind of approach used for building a pre-assessed plan for new IT investments. This technique is used to provide multiple options for new investment while mapping the investment objectives with expected list of outcomes. On the other side, multi-criteria (mostly pre-assessment), and multi-dimensional (mostly post-assessment), which has similarities in the sense of nonfinancial factors.

In addition to the previous discussion section, Table IX indicates investigation on case studies in order to understand their priorities in using ICT evaluation approaches. The evidence from the case studies highlights the most common approach they follow is the financial approach. This kind of method can provide some part of analysis for particular investment to get only a return amount from the investment, but it cannot provide comprehensive analysis regarding business values as criticized by ICT practitioners. Most case studies also rely on a financial approach for specific reasons of knowing profit ratio; otherwise, all of the studies use other approaches as well.

The portfolio management approach used by airlines and telecommunication for some specific ICT projects is joint venture with an outsource agency (Table X) such as ICT services for cargo and catering in airlines and transmission and routing services in telecommunication. Those case studies indicate that the portfolio management approach from an outside consulting agency is feasible where partners are involved. University is a government profile, most of the time depending on multi-criteria and multidimensional approaches to get the value of university assets and resources from a stakeholder's point of view. Food Industry and water & electric are exceptional in this list by using financial approaches mostly for measuring return on their ICT projects.

TABLE IX. ICT PROJECT'S EVALUATION APPROACHES IN CASE STUDIES

Companies	Traditional Financial Approach	Portfolio Management Approach	Multi-Criteria Approach	Multi-Dimensional Approach
Airlines	√	√		√
University			√	√
Food Industry	√			
Water & Electric	√			
Bank	√			√
Telecommunication	√	√		√

H. Third Party Assistance in Investment and Evaluation in Case Studies

Innovation in technologies is happening every day; companies that cannot cope with these changes cannot survive in the market. Organizations have to keep the pace of technology change and regularly upgrade requirements of users and market as well. ICT project outsourcing is based on several reasons; it may be because of low storage so hire cloud computing service and relative cost of outsourcing is cheaper than in-house building, and also time flexibility [34]. The companies can outsource any ICT service due to any reason as discussed above.

The investigated case studies explore that every company, other than in-house building and maintenance, also believe in outsourcing ICT projects as shown in Table X. Notably, some companies mentioned their contract such as in airline system catering and cargo outsources, where all ICT services related to those departments are outsourced, too. In university learning management system, course registration, and other administrative system are outsourced. They discussed saving their time and cost, and it's better than in-house building. For evaluation purposes, airlines typically hire consultancy agencies to evaluate their ICT service performance; they also use help for building proposals for new investments using third-party assistance. On the other side, university typically use an in-house evaluation process; they use some tools and survey instruments for measuring services most of the time at the end of each calendar year. Food Industry, water & electric, and banks and telecommunication take assistance from third parties to outsource systems for their daily routine work. Telecommunications involve collaborative work to boost their services and improve their channels of data communication. Cloud services and data banks are outsourced by food industry and water & electric, while banks have taken assistance for their online services. As far as evaluation is concerned, only telecommunication asked for third-party assistance to help in evaluating some services while others not.

TABLE X. THIRD PARTY ASSISTANCE FOR INVESTMENT AND EVALUATION IN CASE STUDIES

Companies	ICT Projects		ICT Project Evaluation	
	Internal	Outsource	Internal	Outsource
Airlines	√	√	√	√
University	√	√	√	
Food Industry	√	√	√	
Water & Electric	√	√	√	
Bank	√	√	√	
Telecommunication	√	√	√	√

Outsourcing of ICT investments and evaluation can be a better choice for the organizations because it can typically save on time and cost. Most companies seek third-party assistance due to lack of in-house expertise and resource availability. There are some disadvantages for outsourcing, which is the risk of data exposed to outside sources. Switching from one vendor to another can have some complexities and loss of money. Outsourcing will also allow outsiders to learn business models and processes of a particular organization, which can be risky for them. Pre-evaluation planning and assessment approach can be feasible to be outsourced because new development can progress under the supervision of professional ICT experts and consultants offering solutions to different business problems.

III. DISCUSSION

Investigation into the above-mentioned studies had drawn conclusions on measuring business value of ICT projects; thus, investment and evaluation can be realized. Two ICT investment approaches are common practices in six case studies; the companies are investing project-wise associated with the list of objectives or up-gradation on a regular basis using a time frame. In both ways, the finding suggests that, behind every investment, there are specific objectives. Four major objectives strategic, informational, transactional, and IT infrastructure are the main reasons associated with ICT investment. ICT investment can have an impact on those types of a firm's objectives. This study showed that ICT has a significant role and usage in all case studies. The major sectors found in this study are data center systems, IT services, communication services, and IT devices. The majority of respondents said that the investment was motivated by internal pressure while airlines, banks, and telecommunication also highlighted the exceptional cases where external pressure stresses them to invest on ICT resources. Internal matters such as improving access to the information, data server, web servers, and maintaining data warehouses, and external matters such as suppliers, customers, partners, and online transactions are basic motivation to invest more on ICT resources.

The most popular resources companies acquire in this study are technology resources such as hardware, software, ICT applications, and databases where firms invest. Human resources related with ICT are another indication for which the proper budget is allocated to provide technical and support facilities. The relationship and integration between ICT and human resources are the most common method that can generate ICT business values. The most popular approach for measuring ICT investment in our study is a traditional financial approach. Every organization except University uses financial approaches to measure return on investment for specific ICT resources or projects. In the category of a nonfinancial approach, multidimensional practices are used in the case studies. The process of the nonfinancial approach is based on the objective of the investment. Thorough understanding of objectives and expected outcomes can elaborate measuring factors. Time management, process efficiency using process life cycle, technical skill, employee productivity, response time, competitive advantage are major factors used in a multidimensional evaluation approach.

The time and approach of evaluation is critical for examining the case studies. With respect to the time of evaluation, the two common approaches investigated from the case studies are pre- and post. Pre-evaluation and post-evaluation have different purposes. Pre-evaluation, which is mostly used in the planning phase, is where companies try to build a portfolio of new investment with risk analysis, the list of objectives, and expected outcome of the investment. The portfolio provides alternative options to select the optimal and best characterized project. The preliminary work helps them to identify outcomes, and it helps them to link project objectives. All participants agreed that planning and portfolio building can take long time for pre-assessment, where urgent implementation of ICT projects cannot afford this type of pre-evaluation.

On the other side, post-assessment methods help a business to identify the list of benefits achieved after applying an ICT project. The multidimensional post-evaluation method, which is most commonly investigated in the case studies, is measured through several stages. Participants highlighted several questions that need to be investigated in the multidimensional post-evaluation method. It is time-consuming but depends more on initial investigation; however, this approach is better for knowing actual values (nonfinancial) acknowledged by the receivers. The outsourcing of ICT projects is an attractive option in today's turbulent business environment. All case studies include private and government organizations to support their choice for outsourcing the ICT projects as per requirements. It will allow them to activate service on time to accomplish the goal of particular strategy. The approval depends on ICT decision-makers and executive-level management. Accordingly, some organizations hire outside consultants or assessment firms to evaluate their projects. The disadvantages have also been mentioned from all participants are in regards to privacy and data exposed to an outside source, which is an ultimate risk for the organization.

#### IV. CONCLUSION

In this research, researchers have reviewed the most common approaches used for ICT project's investment and evaluation extracted from literature review and investigated in six companies. Based on the similarities, the investment type is categorized as project based (strategic, informational, and so on) and regular upgrading on specific time. The most common input resources are based on technology and human and relationship resources. Finding the most common practices for ICT project evaluation type is essential, which is investigated in case studies as financial and nonfinancial. The investment, which can be measured before and after implementation of the project, depends on the time and requirements. Initial investigation is important to know the objectives, expected outcomes, measuring factors, stakeholders, and time duration for evaluation. Cost and business values are the common factors involved during assessment. In addition to the most common financial methods, ROI and ROA, the multidimensional post-implementation measuring approach is the most common practice in the case studies used for assessing nonfinancial ICT business values. Based on the findings of this research the list and steps of the evaluation processes can be developed in future which is lacking in this research. Using the expected framework in future, the companies can evaluate the performance of their ICT project based on the list of objectives defined in this research.

#### REFERENCES

- [1] F. Saleem, N. Salim, A. G. Fayoumi, A. Alghamdi, and Z. Ullah, "Comprehensive Study of Information and Communication Technology Investments: A Case Study of Saudi Arabia," *Inf. J.*, vol. 16, no. 11, pp. 7875–7893, 2013.
- [2] P. Weill and M. Broadbent, *Managing IT infrastructure: a strategic choice*. Pinnaflex Educational Resources, Inc., 2000.
- [3] P. Maresova and B. Klimova, "Investment evaluation of cloud computing in the European business sector," *Appl. Econ.*, vol. 6846, no. May, pp. 1–14, 2015.
- [4] S. Shang and P. B. Seddon, "Assessing and managing the benefits of enterprise systems: the business manager's perspective," *Inf. Syst. J.*, vol. 2000, pp. 271–299, 2002.
- [5] Z. Ullah, A. S. Al-Mudimigh, A. A. L.-M. Al-Ghamdi, and F. Saleem, "Critical success factors of ERP implementation at higher education institutes: A brief case study," *Inf.*, vol. 16, no. 10, 2013.
- [6] A. S. Al-Mudimigh, F. Saleem, Z. Ullah, and F. N. Al-Aboud, "Implementation of Data Mining Engine on CRM - Improve customer satisfaction," in *2009 International Conference on Information and Communication Technologies, ICICT 2009*, 2009.
- [7] F. Al-Mudimigh, A. S., Ullah, Z., & Saleem, "Data mining strategies and techniques for CRM systems. In System of Systems Engineering, 2009. SoSE 2009. IEEE International Conference on (pp. 1-5). IEEE., *Syst. Syst. Eng. 2009. SoSE 2009. IEEE Int. Conf. (pp. 1-5). IEEE.*, 2009.
- [8] S. Arvanitis and E. N. Loukis, "Information and communication technologies, human capital, workplace organization and labour productivity: A comparative study based on firm-level data for Greece and Switzerland," *Inf. Econ. Policy*, vol. 21, no. 1, pp. 43–61, 2009.
- [9] M. Gammelgård, M. Ekstedt, and P. Gustafsson, "A Categorization of Benefits From IS / IT Investments," *Proc. 13th Eur. Conf. Inf. Technol. Eval.*, no. October, pp. 1–11, 2001.
- [10] A. C. G. Maçada and M. M. Beltrame, "IT business value model for information intensive organizations," *BAR-Brazilian ...*, pp. 44–65, 2012.
- [11] F. Saleem, N. Salim, A. H. Altalhi, Z. Ullah, & AL-Malaise AL-Ghamdi, A., and Z. Mahmood Khan, "Assessing the effects of information and communication technologies on organizational

- development: business values perspectives,” *Inf. Technol. Dev.*, pp. 1–35, 2017.
- [12] A. H. Altalhi, A. AL-Malaise AL-Ghamdi, Z. Ullah, and F. Saleem, “Developing a framework and algorithm for scalability to evaluate the performance and throughput of CRM systems,” *Intell. Autom. Soft Comput.*, 2016.
- [13] S. Aral and P. Weill, “IT Assets, Organizational Capabilities, and Firm Performance: How Resource Allocations and Organizational Differences Explain Performance Variation,” *Organ. Sci.*, vol. 18, no. 5, pp. 763–780, 2007.
- [14] S. Gregor, M. Martin, W. Fernandez, S. Stern, and M. Vitale, “The transformational dimension in the realization of business value from information technology,” *J. Strateg. Inf. Syst.*, vol. 15, no. 3, pp. 249–270, 2006.
- [15] AGIMO, “Demand and Value Assessment Methodology,” Canberra, Australia, 2004.
- [16] P. Gustafsson, J. Hultdt, and H. Lofgren, “Improving the value assessment of IT investments: A case study,” *PICMET '09 - 2009 Portl. Int. Conf. Manag. Eng. Technol.*, pp. 3167–3175, 2009.
- [17] J. Hair, W. Black, B. Babin, R. Anderson, and R. Tatham, *Multivariate data analysis*, 7th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2010.
- [18] J. Lovelock, K. Hale, A. O'Connell, W. Hahn, R. Atwal, C. Graham, M. Dornan, “Forecast Alert: IT Spending, Worldwide, 3Q15 Update,” Gartner Webinars, High-Tech Tuesday Webinar Series, 2015.
- [19] P. Mentzelou and T. Kyriakidou, “A Model for Measuring the Relation ‘Information-Value’ in Companies,” *Exploring Quantifiable IT Yields. EQUITY '07. IEEE International Conference on. EQUITY '07. IEEE International Conference*, 2007. .
- [20] J. Ross, C. Beath, and D. Goodhue, “Developing Long-Term Competitiveness Through Information Technology Assets,” *Sloan Manage. Rev.*, vol. 38, no. 1, pp. 31–42, 1996.
- [21] L. Dadayan, “Measuring return on government IT investments,” in *Proceedings of the 13th European Conference on Information Technology Evaluation*, 2006, no. September, p. 12.
- [22] Saleem, F., Salim, N., Altalhi, A. H., Abdullah, A. L., Ullah, Z., Baothman, F. A., & Junejo, M. H. (2016). Comparative study from several business cases and methodologies for ICT project evaluation. *International Journal of Advanced Computer Science & Applications*, 1(7), 420-427..
- [23] F. Saleem, N. Salim, A. AL-Ghamdi, and Z. Ullah, “Building Framework For ICT Investments Evaluation: Value On Investment Perspective,” *ARNP J. Eng. Appl. Sci.*, vol. 10, no. 3, pp. 1074–1079, 2015.
- [24] F. Saleem, N. Salim, A. G. Fayoumi, and A. Alghamdi, *A General Framework for Measuring Information and Communication Technology Investment: Case Study of Kingdom of Saudi Arabia*, vol. 322. 2012.
- [25] M. McShea, “Return on infrastructure, the new ROI,” *IT Prof.*, vol. 11, no. 4, pp. 12–16, 2009.
- [26] D. Hurley, “Changing the View of ROI to VOI—Value on Investment,” 2001.
- [27] IDA-VOI, “IDA Value Of Investment,” 2003.
- [28] T. A. Pardo, “Public ROI - Advancing Return on Investment Analysis for Government IT Case Study Series Service New Brunswick This Page Intentionally Left Blank,” *Current*, no. 518, 2006.
- [29] S. Bonham, *IT Project Portfolio Management*. ARTECH HOUSE, INC. 685 Canton Street Norwood, MA 02062, 2005.
- [30] R. Cooper, S. Edgett, and E. Kleinschmidt, “Portfolio management in new product development: Lessons from the leaders-1,” *Res. Technol. Manag.*, vol. 40, no. 5, p. 16, 1997.
- [31] V. Graeser, L. Willcocks, N. Pisaniyas, and B. Intelligence, “Developing the IT Scorecard: A Detailed Route Map to IT Evaluation and Performance Measurement Through the Investment Life-Cycle,” 1998.
- [32] M. Parker and R. Benson, “Information Economics,” *Inf. Econ.*, no. C, pp. 1–15, 1989.
- [33] VMM, “The Value Measuring Methodology,” 2002.
- [34] K. Han and S. Mithas, “Information technology outsourcing and non-IT operating costs: An empirical investigation,” *MIS Q.*, vol. 37, no. 1, pp. 315–331, 2013.

# Identification of Toddlers' Nutritional Status using Data Mining Approach

Sri Winiarti, Herman Yuliansyah, Aprial Andi Purnama

Department of Informatics,  
Universitas Ahmad Dahlan,  
Yogyakarta, Indonesia

**Abstract**—One of the problems in community health center or health clinic is documenting the toddlers' data. The numbers of malnutrition cases in developing country are quite high. If the problem of malnutrition is not resolved, it can disrupt the country's economic development. This study identifies malnutrition status of toddlers based on the context data from community health center (PUSKESMAS) in Jogjakarta, Indonesia. Currently, the patients' data cannot directly map into appropriate groups of toddlers' malnutrition status. Therefore, data mining concept with k-means clustering is used to map the data into several malnutrition status categories. The aim of this study is building software that can be used to assist the Indonesian government in making decisions to take preventive action against malnutrition.

**Keywords**—Data mining; k-means clustering; malnutrition status of toddler

## I. INTRODUCTION

Data mining is a process of extracting large amounts of data to know the data pattern. Some topics in data mining are association rule mining, data clustering and data classification. Association rule mining is data mining techniques for finding associative rules between combinations of items. Several studies apply the association rule mining is to identify the risk factors of early childhood caries [1], to determine the pattern feedback of data alumni tracer study at the university [2] and to visualisation of financial Arabic text [3]. Some studies propose clustering method to solve the problems in their research for example a basic health screening system using Bayesian methods [4], detection of heart disease using decision tree methods [5], and to clasify of Alzheimer Disease using K-Nearest Neighbors (KNN) [6]. Several studies also implement the clustering method to perform automatic color segmentation [7], to perform clustering and analysis of earth-quake epicenter [8], and to decrease the load of computation in high dimensional data [9].

Health and nutritional status of children is one of the measure that reflects the public nutrition situation. Malnutrition is not only a burden to the family, but also a burden for the country. Therefore, Indonesian government through the community health center (PUSKESMAS) has conducted data collection of toddlers' nutritional status by using Excel based application. However, the results cannot show the data grouping of nutritional status automatically. The data that

available in PUSKESMAS still not able to determine the nutritional status of toddler, according to the standards set by the Indonesian government. When there is a demand for data related to the community's nutritional status, then the mapping process is done manually. This process becomes not optimal as it will require a long process and can occur duplication of data if thousands of existing data are processed manually.

Previous researches have studied malnutrition in elderly, mothers and toddlers [10]-[12] and Child Care Health Consultation [13]. Malnutrition is the cause and consequence of many geriatric diseases that cause a very significant proportion of state expenditure on health [14]. In [15], author analyzes malnutrition using logistic regression methods and growth charts to reduce the number of children with malnutrition status. This study aims to optimize the data transactions of under five years patients who have malnutrition. The malnutrition patients are grouped according to the nutritional value of children under five years using data mining method with k-means clustering algorithm. Data mining approach is used in this research because data mining are widely used in predicting the various procedures and validity of data. In addition, data mining can improve decision making by finding patterns and trends in complex data [16].

K-means clustering algorithm is also widely implemented in medical science field such as applying k-means clustering to analyze identification of individual characteristics using brainwave signal [17], to identify new candidate drug compounds that have relation with lung cancer drugs [18], to make recommendation of antiarrhythmic drugs [19], and extraction cancer signatures [20]. The other studies are clustering medical data to find direction and effectiveness of the research work [21], enhance cancer subtype prediction [22], color-converted segmentation algorithm for magnetic resonance imaging (MRI) brain images [23] and EEG analysis to detect drowsy driving [24].

Based on the literature review, it is important to continue the research collaboration between data mining and medical science field. The data used in this study refer to nutrition report data from PUSKESMAS Umbulharjo Yogyakarta in 2016. Specification of toddlers' data used in this research is 6 months to 72 months old infants. Parameters that used for the grouping of nutritional status of toddlers namely; height, weight and age.

This research is expects that the PUSKESMAS can access data and data to monitor the nutritional status of children in

This research is supported by Ministry of Research, Technology and Higher Education in the research scheme Higher Education Research Cooperation (Penelitian Kerjasama Antar Perguruan Tinggi/PKAPT) grant number No: 118/SP2H/LT/DRPM/IV/2017 and PEKERTI-058/SP3/LPP-UAD/IV/2017 on 17 April 2017.

every region easily and quickly. This study aims to determine and develop a software that can be used by PUSKESMAS to identify the nutritional status of toddlers using data mining approach to be analyzed in the decision-making process.

## II. METHODOLOGY

This research studies the data mapping of malnutrition patients of children under five using data mining approach. The grouping technique uses k-means clustering. The k-means clustering algorithm is the simplest and most common algorithm used to group objects by attribute/feature into k number of clusters, where k is a positive integer and defined by the user. Grouping is done by minimizing the sum of squares distance between the data and the appropriate centroid cluster. The procedure of k-means clustering is shown in Fig. 1 [25].

As shown in Fig. 1, the procedure of k-means clustering can be explained as follows:

- Step 1. Begin by defining k = number of clusters.
- Step 2. Enter each initial partition that classifies the data into the cluster k. It can be done by randomly sampling the data, or systematically as follows: Take the first training data sample k as a single element cluster. Each of the remaining training samples (N-k) collect on the cluster with the nearest centroid. When finished, recompute centroid from the newly acquired cluster.
- Step 3. Perform each sample in a sequence and calculate the distance from the centroid center of each group. If a sample is currently incompatible with the cluster closest to the centroid, replace the sample in this cluster and update the centroid point with the new sample and the sample loss cluster.

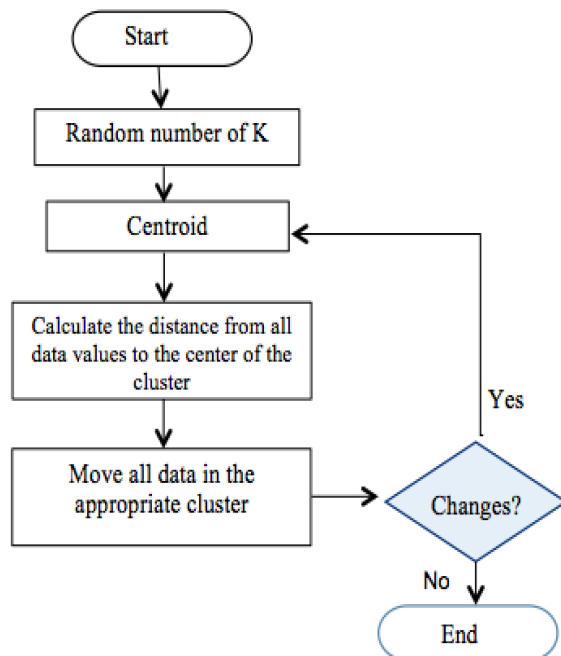


Fig. 1. The procedure of k-means clustering algorithm.

- Step 4. Repeat Step 3 until the target value is reached, i.e. until the training sample matches and there is no new task. If the amount of data is less than the number of clusters, then assign each data as the centroid of the cluster. Each centroid will have a number of clusters. If the amount of data is greater than the number of clusters, for each data, calculate the Distance to all centroids and get the minimum distance. This data is said to belong to a cluster that has a minimum distance value of this data. If you are not sure about the centroid location, you need to be centroid based on your current location. Then set all data to this new centroid. This process is repeated until no data is moved to another cluster again. The k-means algorithm works by using (1).

$$arg_s \min \sum_{i=1}^k \sum_{X_j \in S_i} ||X_j - \mu_i ||^2 \quad (1)$$

Information:

- (X1, X2 ... Xn): the observation results represent a cluster element with a real d dimensional vector.
- n: Number of observations where the observed value to k set (k <= n) S = {S1, S2, ... Sk}.
- $\mu_i$ : the mean value of the point at Si.

## III. RESULT AND DISCUSSION

Why use data mining concept? Because the concept of data mining can analyze and classify the database so that every organization can make decisions based on this classification and can improve their plan in the future. There are many data mining techniques available where we can detect hidden patterns in the database [26].

Referring to the data mining stage in Fig. 1, for the case of nutrition status identification with k-means clustering algorithm, the procedure begins by obtaining patient data from patient's medical record database. Table I shows the patient data with parameters of body height, weight and age of children under five.

TABLE I. TODDLER DATA BASED ON PUSKESMAS LOCATION

No	Age	Weight	Height
1	42	12.7	91
2	41	12.8	94
3	39	16.8	98
4	33	13.4	94
5	24	10.8	85
6	24	10.3	103
7	24	10.3	103
8	48	16.3	104
9	45	15.7	100
10	44	26.5	104



After the data are loaded, the initial centroids are determined according to 5 groups of toddler's nutritional status, that is Bad with value 0.96, Medium with value 0.73, Good with value 0.73, Over with value 0,355 and Obesity with value 0,04. The data is normalized using the normalization equation.

$$\text{Normalized value} = (\text{initial value} - \text{minimum value}) / (\text{max value} - \text{minimum value}) \quad (2)$$

As shown in Table I, the High Body data have minimum value 85 and maximum value 104, the Weight data have minimum value 10.3 and maximum value 26.5, while the Age data have minimum value 24 and maximum value 48. The data normalization is shown in Table II.

If the obtained data are not consistent, they will change the data centroid through the iteration process. The iteration process will stop if the new ratio value is less than the ratio value in the previous iteration. If the condition has not been achieved, the iteration process will be repeated. The iteration result from the normalization of toddlers' data are shown in Table III.

TABLE. II. NORMALIZATION OF TODDLER DATA

No	Age	Weight	Height	Means
1	0.75	0.148148148	0.315789474	0.404645874
2	0.708333333	0.154320988	0.473684211	0.445446177
3	0.625	0.401234568	0.684210526	0.570148365
4	0.375	0.191358025	0.473684211	0.346680745
5	0	0.030864198	0	0.010288066
6	0	0	0.947368421	0.315789474
7	0	0	0.947368421	0.315789474
8	1	0.37037037	1	0.790123457
9	0.875	0.333333333	0.789473684	0.665935673
10	0.833333333	1	1	0.944444444

TABLE. III. ITERATION RESULT I

No	Status				
	Bad	Medium	Good	More	Obesity
1	0.555354126	0.325354126	0.085354126	0.049645874	0.364645874
2	0.514553823	0.284553823	0.044553823	0.090446177	0.405446177
3	0.389851635	0.159851635	0.080148365	0.215148365	0.530148365
4	0.613319255	0.383319255	0.143319255	0.008319255	0.306680745
5	0.949711934	0.719711934	0.479711934	0.344711934	0.029711934
6	0.644210526	0.414210526	0.174210526	0.039210526	0.275789474
7	0.644210526	0.414210526	0.174210526	0.039210526	0.275789474
8	0.169876543	0.060123457	0.300123457	0.435123457	0.750123457
9	0.294064327	0.064064327	0.175935673	0.310935673	0.625935673
10	0.015555556	0.214444444	0.454444444	0.589444444	0.904444444

TABLE. IV. DISTANCE DATA ON THE FIRST ITERATION

Distance Data		
Membership	Min Distance	Min Squared Distance
More	0.049645874	0.002464713
Good	0.044553823	0.001985043
Good	0.080148365	0.00642376
More	0.008319255	6.921E-05
Obesity	0.029711934	0.000882799
More	0.039210526	0.001537465
More	0.039210526	0.001537465
Medium	0.060123457	0.00361483
Medium	0.064064327	0.004104238
Bad	0.015555556	0.000241975
	Wcv	0.0228615

TABLE. V. CLUSTER CENTER DISTANCE DATA D

Cluster Center Distance d		
C1	C2	0.23
C1	C3	0.47
C1	C4	0.605
C1	C5	0.92
C2	C3	0.24
C2	C4	0.375
C3	C4	0.135
C3	C5	0.45
C4	C5	0.315
BCV		3.74

TABLE. VI. NEW CLUSTER CENTER DATA ON THE FIRST ITERATION

New Cluster Center				
Bad	Medium	Good	More	Obesity
			0.404645874	
		0.445446177		
		0.570148365		
			0.346680745	
				0.010288066
			0.315789474	
			0.315789474	
	0.790123457			
	0.665935673			
0.944444444				
0.944444444	0.728029565	0.507797271	0.345726392	0.010288066

The distance data on the first iteration are presented in Table IV. The center distance data d are presented in Table V and the new cluster center data are presented in Table VI.

By using equation Ratio = BCV / WCV, the ratio result is 163.594. When the ratio is compared with the previous ratio, the value of the new ratio is greater than the value of the previous ratio. Therefore, the iteration process is still continued. Fig. 2 and 3 show the interface of the developed software.

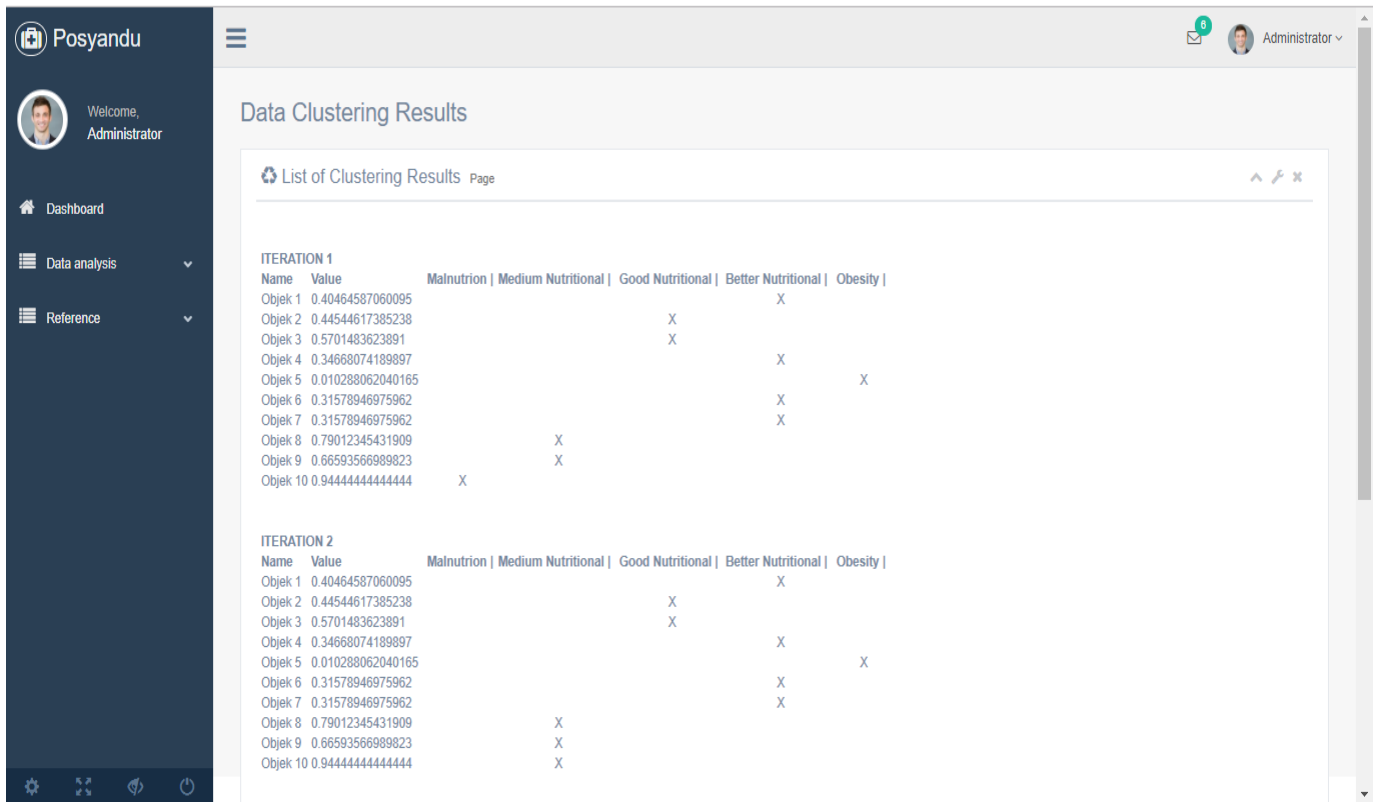
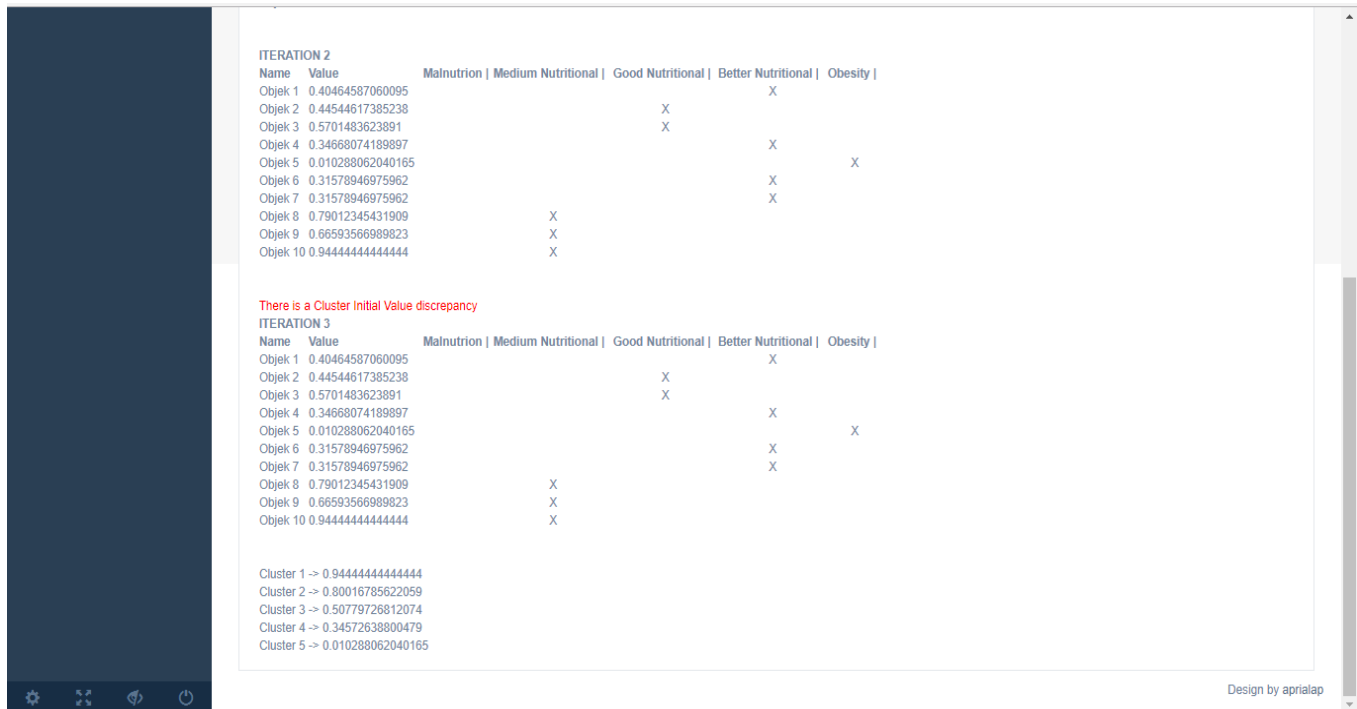


Fig. 2. Iteration process interface 1.



Iteration process interface 2.

#### IV. SYSTEM TEST

The system is tested using cross validation method by comparing the calculation result of k-means manually and with

the result of developed system. Based on the patient data that shown in Table I, the system calculation result is presented in Table VII and the manually calculated result is presented in Table VIII.

TABLE. VII. RESULTS OF CALCULATIONS WITH THE SYSTEM DEVELOPED

Name	Status				
	Bad	Medium	Good	More	Obesity
Dest	0.55535412 6	0.32535412 6	0.08535412 6	0.04964587 4	0.36464587 4
Aura	0.51455382 3	0.28455382 3	0.04455382 3	0.09044617 7	0.40544617 7
Nazwa	0.38985163 5	0.15985163 5	0.08014836 5	0.21514836 5	0.53014836 5
Arisa	0.61331925 5	0.38331925 5	0.14331925 5	0.00831925 5	0.30668074 5
Evelyn	0.94971193 4	0.71971193 4	0.47971193 4	0.34471193 4	0.02971193 4
Amira	0.64421052 6	0.41421052 6	0.17421052 6	0.03921052 6	0.27578947 4
Farah	0.64421052 6	0.41421052 6	0.17421052 6	0.03921052 6	0.27578947 4
Hasna	0.16987654 3	0.06012345 7	0.30012345 7	0.43512345 7	0.75012345 7
Kania	0.29406432 7	0.06406432 7	0.17593567 3	0.31093567 3	0.62593567 3
Aira	0.01555555 6	0.21444444 4	0.45444444 4	0.58944444 4	0.90444444 4

TABLE. VIII. RESULTS OF CALCULATIONS MANUALLY

Name	Status				
	Bad	Medium	Good	More	Obesity
Dest	0.55535412 6	0.32535412 6	0.08535412 6	0.04964587 4	0.36464587 4
Aura	0.51455382 3	0.28455382 3	0.04455382 3	0.09044617 7	0.40544617 7
Nazwa	0.38985163 5	0.15985163 5	0.08014836 5	0.21514836 5	0.53014836 5
Arisa	0.61331925 5	0.38331925 5	0.14331925 5	0.00831925 5	0.30668074 5
Evelyn	0.94971193 4	0.71971193 4	0.47971193 4	0.34471193 4	0.02971193 4
Amira	0.64421052 6	0.41421052 6	0.17421052 6	0.03921052 6	0.27578947 4
Farah	0.64421052 6	0.41421052 6	0.17421052 6	0.03921052 6	0.27578947 4
Hasna	0.16987654 3	0.06012345 7	0.30012345 7	0.43512345 7	0.75012345 7
<b>Kania</b>	<b>0.27405432</b> <b>1</b>	<b>0.05414425</b> <b>3</b>	<b>0.15593555</b> <b>3</b>	<b>0.21094467</b> <b>3</b>	<b>0.52598767</b> <b>3</b>
Aira	0.01555555 6	0.21444444 4	0.45444444 4	0.58944444 4	0.90444444 4

V. CONCLUSION

This research builds a software that can be used to identify the nutritional status of toddlers using data mining technique, with k-means clustering algorithm. The test is conducted by performing cross validation and gives 90% validation that the system can determine nutritional status of toddler by producing 5 clusters, namely, good nutrition, moderate nutrition, malnutrition, more nutrition and obesity.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their comments and suggestions that helped to improve the quality and presentation of this paper. This research is supported by Ministry of Research, Technology and Higher Education in the research scheme Higher Education Research Cooperation (Penelitian Kerjasama Antar Perguruan Tinggi/PKAPT) grant No: 118/ SP2H/ LT/ DRPM/ IV/ 2017 and PEKERTI-058/ SP3/ LPP-UAD/ IV/ 2017 on 17 April 2017.

REFERENCES

- [1] V. Ivančević, I. Tušek, J. Tušek, M. Knežević, S. Elheshk, and I. Luković, "Using association rule mining to identify risk factors for early childhood caries," *Comput. Methods Programs Biomed.*, vol. 122, no. 2, pp. 175–181, 2015.
- [2] H. Yuliansyah and L. Zahrotun, "Designing web-based data mining applications to analyze the association rules tracer study at university using a FOLD-growth method," *Int. J. Adv. Comput. Res.*, vol. 6, no. 27, pp. 215–221, 2016.
- [3] H. AL-Rubaiee, R. Qiu, and D. Li, "Visualising Arabic Sentiments and Association Rules in Financial Text," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 2, pp. 1–7, 2017.
- [4] D. Phongphanich, N. Prommuang, and B. Chooprom, "Basic Health Screening by Exploiting Data Mining Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 9, pp. 79–85, 2017.
- [5] A. Aziz and A. U. Rehman, "Detection of Cardiac Disease using Data Mining Classification Techniques," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 7, pp. 256–259, 2017.
- [6] A. M. Taqi, F. Al-Azzo, and M. Milanova, "Classification of Alzheimer Disease Based on Normalized Hu Moment Invariants and Multiclassifiers," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 11, pp. 10–18, 2017.
- [7] A. Prahara, I. T. R. Yanto, and T. Herawan, "Histogram Thresholding for Automatic Color Segmentation Based on k-means Clustering," in *Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016)*, Bandung, Indonesia, August 18-20, 2016. Proceedings, T. Herawan, R. Ghazali, N. M. Nawati, and M. M. Deris, Eds. Cham: Springer International Publishing, 2017, pp. 344–354.
- [8] P. Novianti, D. Setyorini, and U. Rafflesia, "K-Means cluster analysis in earthquake epicenter clustering," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 2, pp. 81–89, 2017.
- [9] D. Ismi, S. Panchoo, and M. Murinto, "K-means clustering based filter feature selection on high dimensional data," *Int. J. Adv. Intell. Informatics*, vol. 2, no. 1, pp. 38–45, 2016.
- [10] J. Studnicki, A. R. Hevner, D. J. Berndt, and S. L. Luther, "Comparing alternative methods for composing community peer groups: a data warehouse application," *J. public Heal. Manag. Pract.*, vol. 7, no. 6, pp. 87–95, 2001.
- [11] D. Berndt, A. Hevner, and J. Studnicki, "Data warehouse dissemination strategies for community health assessments, informatik/informatique," *J. Swiss Informatics Soc.*, vol. 1, pp. 27–33, 2001.
- [12] S. Winiarti, S. Kusumadewi, I. Muhimmah, and H. Yuliansyah, "Determining the nutrition of patient based on food packaging product using fuzzy C means algorithm," in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2017, no. September, pp. 1–6.
- [13] R. Johnston, B. A. DelConte, L. Ungvary, R. Fiene, and S. S. Aronson, "Child Care Health Consultation Improves Infant and Toddler Care," *J. Pediatr. Heal. Care*, vol. 31, no. 6, pp. 684–694, Nov. 2017.
- [14] S. N. Jang, S. I. Cho, J. Chang, K. Boo, H. G. Shin, H. Lee, and L. F. Berkman, "Employment status and depressive symptoms in Koreans: results from a baseline survey of the Korean Longitudinal Study of Aging," *J. Gerontol. B. Psychol. Sci. Soc. Sci.*, vol. 64, no. 5, p. 677, 2009.
- [15] M. Ohlyver, J. V. Moniaga, K. R. Yunidwi, and M. I. Setiawan, "Logistic Regression and Growth Charts to Determine Children Nutritional and Stunting Status: A Review," *Procedia Comput. Sci.*, vol. 116, pp. 232–241, Jan. 2017.
- [16] D. Thangamani and P. Sudha, "Identification of Various Deficiencies Using Data Mining Techniques – A Survey," *Int. J. Sci. Res.*, vol. 3, no. 7, pp. 1270–1274, 2014.
- [17] A. Azhari and L. Hernandez, "Brainwaves feature classification by applying K-Means clustering using single-sensor EEG," *Int. J. Adv. Intell. Informatics*, vol. 2, no. 3, pp. 167–173, 2016.
- [18] J. Lu, L. Chen, J. Yin, T. Huang, Y. Bi, X. Kong, M. Zheng, and Y. D. Cai, "Identification of new candidate drugs for lung cancer using chemical-chemical interactions, chemical-protein interactions and a K-

- means clustering algorithm,” *J. Biomol. Struct. Dyn.*, vol. 34, no. 4, pp. 906–917, 2016.
- [19] J. Park, M. Kang, J. Hur, and K. Kang, “Recommendations for antiarrhythmic drugs based on latent semantic analysis with K-means clustering,” 2016 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., pp. 4423–4426, 2016.
- [20] Z. Kakushadze and W. Yu, “\*K-means and cluster models for cancer signatures,” *Biomol. Detect. Quantif.*, vol. 13, no. July, pp. 7–31, 2017.
- [21] S. V and G. H. A, “Appraising Research Direction & Effectiveness of Existing Clustering Algorithm for Medical Data,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 343–351, 2017.
- [22] N. Nidheesh, K. A. Abdul Nazeer, and P. M. Ameer, “An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data,” *Comput. Biol. Med.*, vol. 91, pp. 213–221, Dec. 2017.
- [23] L.-H. Juang and M.-N. Wu, “MRI brain lesion image detection based on color-converted K-means clustering segmentation,” *Measurement*, vol. 43, no. 7, pp. 941–949, Aug. 2010.
- [24] N. Gurudath and H. B. Riley, “Drowsy Driving Detection by EEG Analysis Using Wavelet Transform and K-means Clustering,” *Procedia Comput. Sci.*, vol. 34, pp. 400–409, Jan. 2014.
- [25] S. Shinde and B. Tidke, “Improved K-means Algorithm for searching Research papers,” *Int. J. Comput. Sci. Commun. Networks*, vol. 4, no. 6, pp. 197–202, 2014.
- [26] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

# A Comparative Study on Steganography Digital Images: A Case Study of Scalable Vector Graphics (SVG) and Portable Network Graphics (PNG) Images Formats

Abdulgader Almutairi

College of Sciences and Arts in ArRass,  
Qassim University, Kingdom of Saudi Arabia

**Abstract**—Today image steganography plays a key role for exchanging a secret data through the internet. However, the optimal choice of images formats for processing steganography is still an open issue; therefore, this research comes into a table. This research conducts a comparative study between Scalable Vector Graphics (SVG) image format and Portable Network Graphics (PNG) image format. As results show, SVG image format is more efficient than PNG image format in terms of capacity and scalability before and after processing steganography. As well, SVG image format helps to increase simplicity and performance for processing steganography, since it is an XML text file. Our comparative study provides significant results between SVG and PNG images, which have not been seen in the previous related studies.

**Keywords**—Image steganography; data hiding; raster and vector images; Scalable Vector Graphics (SVG) and Portable Network Graphics (PNG) images format

## I. INTRODUCTION

Nowadays, steganography is used widely for delivering a secure message among different parties distributed across the world. Steganography is a Greek term, which means a concealed writing. Steganography comprises four main categories: video steganography, audio steganography, image steganography, and text steganography. It has been observed that image steganography is the most popular steganography type due to images frequencies in the internet, and therefore image steganography type is identified for discussion and review in this research [1]-[4].

A digital image is a numeric form of a two dimensional image that is made of image components known as pixels. Usually, pixels are structured in well-arranged array. The columns number of the array defines width of the image, while the rows number of the array defines image height. Broadly, digital images are two types, namely raster digital images and vector digital images [5].

The raster digital images are bitmaps that are defined through a grid of specific pixels, which jointly structure the digital image. It concentrates digital images as a pool of uncountable small squares called pixels. Every single pixel is coded in a particular sort. It can be produced by an illustration program or generated from a scanner. Thus, raster images can

encompass millions of different colors, with each one being represented by a single pixel.

The raster digital images are commonly used for non-line digital images, since they normally involve complicated composition, thin chromatic gradations, and indefinite lines and shapes. The most common spread raster digital images include Joint Photographic Experts Group (JPEG), Windows Bitmap (BMP), Graphics Interchange Format (GIF), and Portable Network Graphics (PNG) images formats [5]-[7].

On the contrary, vector digital images are constructed based on mathematical forms that state geometric features like rectangles, curves, lines, circles, and polygons. In addition, these components are loaded with gradients, color, blends, and tints. Besides that, the lines project a stroke merit like various chunkiness and colors for a solid or dashed line. The vector digital images are used to construct more organized digital images, such as fonts, logos, letterhead. The most commonly used vector digital images are Vector Markup Language (VML) and Scalable Vector Graphics (SVG) images formats [8]-[11]. Therefore, the difference between vector images and raster images is that the vector images are object based and the others are pixel based. The rest of the paper is organized as follows: Section II starts with literature review. Section III gives comparative study on steganography digital images between portable network graphics (PNG) and scalable vector graphics (SVG) images formats. Results discussion and analysis is presented in Section IV. And the paper has been summarized in Section V.

## II. LITERATURE REVIEW

Broadly, digital images are divided into two main types, namely raster digital images and vector digital images. The former are bitmaps that are well-defined through a net of specific pixels in order to form digital images. The most common spread raster digital images are Joint Photographic Experts Group (JPEG), Graphics Interchange Format (GIF), and Portable Network Graphics (PNG) images formats. The latter are constructed based on mathematical forms that state geometric features like rectangles, curves, lines, circles, and polygons. In addition, these components are loaded with gradients, color, blends, and tints. Besides that, the lines project a stroke merit like various chunkiness and colors for a

solid or dashed line. The most commonly used vector digital images are Vector Markup Language (VML) and Scalable Vector Graphics (SVG) images formats [2], [5]-[8], [10], [11], [15], [16].

Amongst the various raster’s digital images type, a Portable Network Graphics (PNG) image format is the most widely spread raster digital image, which can be attributed to its exceptional features, which are absent in the others raster image formats. Hence, PNG image format is nominated in the current research to conduct a comparative study on steganography digital images. PNG image format is designed to be exchanged smoothly through network, and to work well in online applications like web browsers. The PNG image format provides integrity checking and transmission errors detection. PNG image format is a free and open source format, which was an alternative to GIF. PNG image format supports 8-bit, 16-bit, 24-bit, 32-bit, and 48-bit images, while GIF image format supports only 256 colors and a single transparent color. In addition, PNG image format excels JPEG image format, particularly in case of a big sized image, which losses compression, which means when compressed, it does not lose any data. The animated images format of PNG is MNG and APNG images formats. PNG image format is commonly used for graphs, diagrams, and anywhere to display flat colors and lines, not needing scaling up [2], [5]-[8], [10], [11], [15], [16].

The RGB (Red, Green and Blue) color model is an essential tool, to edit an image’s pixels in order to hide a secret data into PNG image formats. The RGB color model is a collective color model, which collects red, green and blue colors together in several means to reproduce a various colors. Fig. 1 presents PNG24 image in RGB color model [6], [17].

Amongst the various vector’s digital images type, a Scalable Vector Graphics (SVG) image format is the most widely spread vector digital image. Hence, SVG image format is selected in the current research in order to conduct a comparative study on steganography digital images. SVG image format is an open standard and based on Extensible Markup Language (XML), which is defined in XML text file using any text editor. It presents the qualities of being searchable, scriptable, compressible, and also can be zoomed and indexed. It supports animation and scalability, since its data is saved as a geometric description, instead of the description of each single image pixel in a raster digital image. Table I presents basic comparisons between SVG and PNG images format [2], [5]-[8], [10], [11], [15], [16].



Fig. 1. PNG24 image in RGB color model.

TABLE I. BASIC COMPARISONS BETWEEN SVG AND PNG IMAGE FORMATS.

No	Property	SVG Images Format	PNG Images Format
1	Type	Vector Image	Raster Image
2	Resolution	Unlimited	Limited to image’s pixels
3	Speed	Fast, since its file’s size is small.	Slow, since its file’s size often is larger
4	Animation	Support animation (ECMAScript, CSS, and SML)	Support animation (MNG and APNG images formats)
5	Compressed	It can be compressed	It can be compressed
6	Zoomable	Zoomable without degradation	Zooming effects shadow in the image.
7	Open Standard	Standardized by W3C	Standardized by ISO/IEC.

### III. A COMPARATIVE STUDY ON STEGANOGRAPHY DIGITAL IMAGES BETWEEN PORTABLE NETWORK GRAPHICS (PNG) AND SCALABLE VECTOR GRAPHICS (SVG) IMAGES FORMATS

This section, presents a comparative study on steganography digital images between PNG and SVG images formats. Firstly, it presents SVG and PNG images format sizes prior to steganography processing. This is followed by a discussion on SVG and PNG images format sizes after steganography processing [10]-[15], [18], [19].

#### A. Comparing SVG and PNG Images Format Sizes before Steganography Processing

In order to conduct this comparison, we selected the most commonly used three types of image editing software, namely Inkscape image editing software [20], Visio image editing software [21], and Dia image editing software [22]. Afterwards, we used these three image editing software to create and edit the three famous images namely, Lena image, Baboon image, and Pepper image, and to save each image twice; one in SVG image format, and second in PNG image format. After that, we calculated image size for all saved images, and the difference in image size between SVG and PNG image formats. Table II below presents the corresponding findings [5], [6], [7], [17].

TABLE II. A COMPARISON BETWEEN SVG AND PNG IMAGES FORMAT SIZES BEFORE STEGANOGRAPHY PROCESSING

Image Name	Inkscape Image Editing Software			Visio Image Editing Software			Dia Image Editing Software		
	SVG Size (Bytes)	PNG Size (Bytes)	Size Diff. (Bytes)	SVG Size (Bytes)	PNG Size (Bytes)	Size Diff. (Bytes)	SVG Size (Bytes)	PNG Size (Bytes)	Size Diff. (Bytes)
Lena	562,514	1,275,975	713,461	288,569	841,419	552,850	434	336,087	335,653
Baboon	881,448	1,161,746	280,298	1,083,259	1,322,223	238,964	436	629,707	629,271
Pepper	18,652	645,490	626,838	187,391	1,046,081	858,690	437	102,015	101,578

**B. Comparing SVG and PNG Images Format Sizes after Steganography Processing**

In order to perform this comparison, we used OpenStego software, with the purpose to hide a secret data into PNG images format, and notepad text editor software to hide the same secret data into SVG images format [23]. This choice was particularly underpinned by the idea that OpenStego software manipulates images through GRB color model, which is required for processing PNG images format since they are bitmaps images file, and they are only processed through GRB color model. On the other side, notepad text editor software manipulates SVG images format since they are written in XML file format that is a text and not bitmaps, which can be processed through notepad software [1], [24]-[29]. The secret data is “My Visa Credit Card number is 4539962848122779, CCV number is 483, and Expiration Date is 12/17”. First, in the study, a secret data was encrypted using AES cryptographic algorithm, which was then hidden (embedded) into SVG and PNG images. As stated earlier, OpenStego software hid the secret data written inside Microsoft word file, into PNG image format, while the notepad software added XML tag into SVG image format. The XML tag merely included an external another XML file, which contained the real encrypted secret data as shown in Fig. 2 below. Hence, SVG image format itself was not be affected when hiding bigger secret data. Besides that, it reflected the simplicity of steganography processing, which is easy for SVG images format compared with PNG images format requiring RGB color model. Moreover, the simplicity of steganography processing usually provides a better performance [1], [16], [24], [30].

Table III below shows the corresponding sizes of SVG and PNG images format after steganography processing.

```

<!-- dh xmlns:inc="http://www.w3.org/2001/XInclude" -->
    <!-- inc:include href="dh.xml"/ -->
<!-- /dh -->
    
```

Fig. 2. XML Tag includes an Encrypted Data in SVG Image Format.

**IV. RESULTS DISCUSSION AND ANALYSIS**

As shown in Table II, SVG image format occupies less capacity in comparison with PNG image format, which evidences that SVG image format is more efficient than PNG image format in term of image’s capacity. For instance, the size of Lena image when saved in SVG image format using Inkscape software is 562,514 Bytes, while its size is 1,275,975 Bytes when saved in PNG image format. The difference in size of Lena image between SVG and PNG image formats is 713,461 Bytes, which acts 55.9% of PNG image’s size. Likewise, the size of Lena image when saved in SVG image format using Visio software is 288,569 Bytes, while its size is 841,419 Bytes when saved in PNG image format. The difference in size of Lena image between SVG and PNG image formats is thus, 552,850 Bytes, which is 65.7% of PNG image’s size. Similarly, the size of Lena image when saved in SVG image format using Dia software is 434 Bytes, while its size is 336,087 Bytes when saved in PNG image format. The difference in size of Lena image between SVG and PNG image formats is thus, 335,653 Bytes, which is 99.8% of PNG image’s size. Equally, the same calculations are applied to all other images as shown in Table II above. Overall, the difference in size between SVG and PNG image formats before steganography processing is about 18% - 99.9%. Correspondingly, Fig. 3, 4, and 5 below show the sizes of SVG and PNG image formats and the difference between them in Inkscape, Visio, and Dia image editing software respectively.

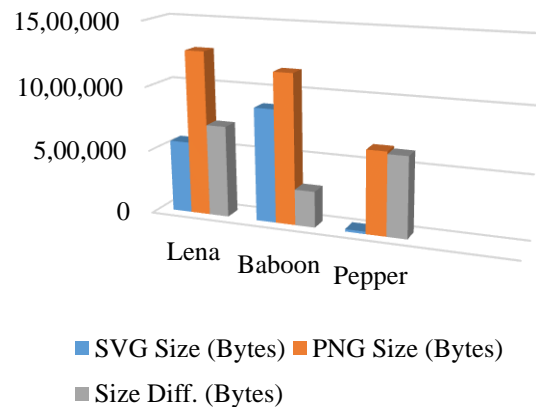


Fig. 3. SVG and PNG image sizes in Inkscape image editing software.

TABLE III. A COMPARISON BETWEEN SVG AND PNG IMAGES FORMAT SIZES AFTER STEGANOGRAPHY PROCESSING

Image Name	Inkscape Image Editing Software			Visio Image Editing Software			Dia Image Editing Software		
	SVG Size (Bytes)	PNG Size (Bytes)	Size Diff. (Bytes)	SVG Size (Bytes)	PNG Size (Bytes)	Size Diff. (Bytes)	SVG Size (Bytes)	PNG Size (Bytes)	Size Diff. (Bytes)
Lena	562,628	1,825,337	1,262,709	288,681	845,107	556,426	554	343,282	342,728
Baboon	881,562	1,372,420	490,858	1,083,371	1,323,995	240,624	556	753,170	752,614
Pepper	18,766	1,184,195	1,165,429	187,503	1,067,848	880,345	557	139,943	139,386

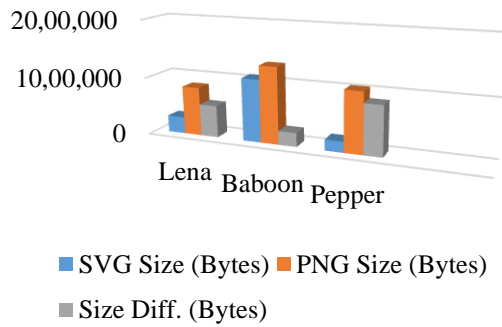


Fig. 4. SVG and PNG image sizes in Visio image editing software.

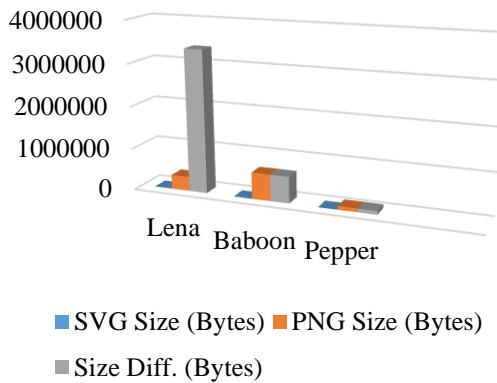


Fig. 5. SVG and PNG image sizes in Dia image editing software.

In addition, Table III above shows that SVG image format occupies less capacity than PNG image format after steganography processing, which evidences that SVG image format is more efficient than PNG image format as regards image's capacity. For example, the size of Lena image that is saved in SVG image format using Inkscape software after steganography processing is 562,628 Bytes, while its size is 1,825,337 Bytes when saved in PNG image format. The difference in size of Lena image between SVG and PNG image formats is 1,262,709 Bytes, which acts 69.2% of PNG image's size. Similarly, the size of Lena image that is saved in SVG image format using Visio software after steganography processing is 288,681 Bytes, while its size is 845,107 Bytes when saved in PNG image format. The difference in size of Lena image between SVG and PNG image formats is thus, 556,426 Bytes, which is 65.8% of PNG image's size. In the same way, the size of Lena image that is saved in SVG image format using Dia software after steganography processing is 554 Bytes, while its size is 343,282 Bytes when it is saved in PNG image format. The difference in size of Lena image between SVG and PNG image formats is thus 342,728 Bytes, which is 99.6% of PNG image's size. Likewise, identical calculations are applied to all other images as shown in Table III above. Overall, the difference in size between SVG and PNG image formats after steganography processing is about 18.2% - 99.9%. Correspondingly, Fig. 6, 7, and 8 show the sizes of SVG and PNG image formats and the difference between them in Inkscape, Visio, and Dia image editing software respectively after steganography processing.

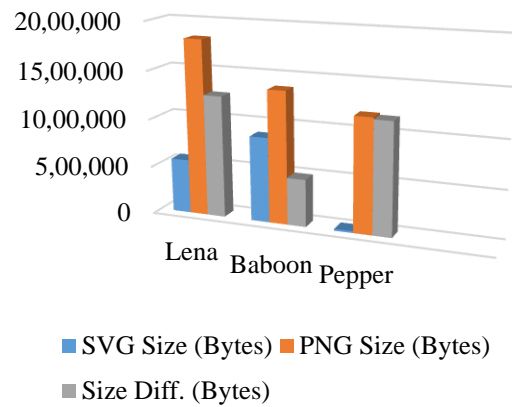


Fig. 6. SVG and PNG image sizes in Inkscape image editing software after steganography processing.

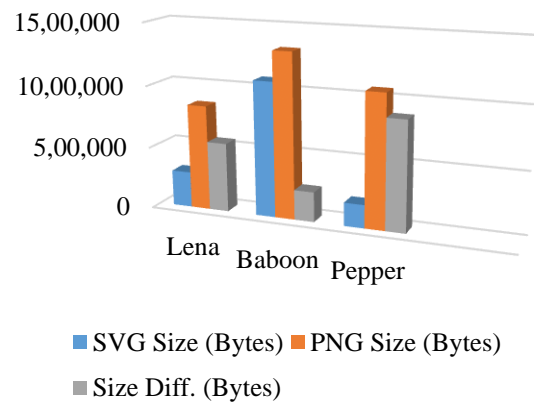


Fig. 7. SVG and PNG image sizes in Visio image editing software after steganography processing.

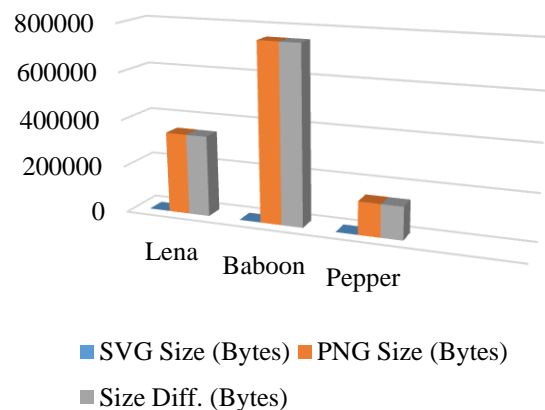


Fig. 8. SVG and PNG image sizes in Dia image editing software after steganography processing.

In addition, the following Tables IV, V, and VI evidence that SVG image format occupies less capacity in comparison to PNG image format after steganography processing. Thus, it can be inferred that SVG image format is more efficient than PNG image format in term of image's capacity.



TABLE IV. SVG AND PNG IMAGES SIZE DIFFERENCE BEFORE AND AFTER STEGANOGRAPHY PROCESSING

Image Name	Before/After Steganography Processing	Inkscape Image Editing Software			
		SVG Size (Bytes)	SVG Diff. (Bytes)	PNG Size (Bytes)	PNG Diff. (Bytes)
Lena	Before	562,514	114	1,275,975	549,362
	After	562,628		1,825,337	
Baboon	Before	881,448	114	1,161,746	210,674
	After	881,562		1,372,420	
Pepper	Before	18,652	114	645,490	538,705
	After	18,766		1,184,195	

TABLE V. SVG AND PNG IMAGES SIZE DIFFERENCE BEFORE AND AFTER STEGANOGRAPHY PROCESSING

Image Name	Before/After Steganography Processing	Visio Image Editing Software			
		SVG Size (Bytes)	SVG Diff. (Bytes)	PNG Size (Bytes)	PNG Diff. (Bytes)
Lena	Before	288,569	112	841,419	3,688
	After	288,681		845,107	
Baboon	Before	1,083,259	112	1,322,223	1,772
	After	1,083,371		1,323,995	
Pepper	Before	187,391	112	1,046,081	21,767
	After	187,503		1,067,848	

TABLE VI. SVG AND PNG IMAGES SIZE DIFFERENCE BEFORE AND AFTER STEGANOGRAPHY PROCESSING

Image Name	Before/After Steganography Processing	Dia Image Editing Software			
		SVG Size (Bytes)	SVG Diff. (Bytes)	PNG Size (Bytes)	PNG Diff. (Bytes)
Lena	Before	434	120	336,087	7,195
	After	554		343,282	
Baboon	Before	436	120	629,707	123,463
	After	556		753,170	
Pepper	Before	437	120	102,015	37,928
	After	557		139,943	

Table IV present the sizes of images that are saved in SVG and PNG formats using Inkscape image editing software before and after steganography processing. Thus, hiding the secret data “My Visa Credit Card number is 4539962848122779, CCV number is 483, and Expiration Date is 12/17” into SVG images format yields a fixed growth of image’s size, which is 114 Bytes for all SVG images format, while it yields a variable image’s size growth for PNG images format. This merit is achieved due to saving of the secret data outside SVG images format, and hence it provides a scalability. However, the secret data is when saved inside PNG images format, shows an increase in their sizes.

In addition, Table V present the sizes of images that are saved in SVG and PNG formats using Visio image editing software before and after steganography processing. Thus, hiding our secret data “My Visa Credit Card number is 4539962848122779, CCV number is 483, and Expiration Date

is 12/17” into SVG images format yields a fixed growth of image’s size, which is 112 Bytes for all SVG images format, while it yields a variable image’s size growth for PNG images format. This merit is achieved due to saving a secret data outside SVG images format, and hence it provides a scalability. However, when the secret data is saved inside PNG images format, it shows an increase in their sizes.

Finally, Table VI shows sizes of images that are saved in SVG and PNG formats using Dia image editing software before and after steganography processing. Thus, hiding our secret data “My Visa Credit Card number is 4539962848122779, CCV number is 483, and Expiration Date is 12/17” into SVG images format yields a fixed growth of image’s size, which is 120 Bytes for all SVG images format, while it yields a variable image’s size growth for PNG images format. This merit is achieved due to saving the secret data outside SVG images format, and hence it provides a

scalability. However, when the secret data is saved inside PNG images format, it shows an increase in their sizes.

## V. CONCLUSION

The current research has been conducted with the purpose to draw a comparative study on steganography digital images. It focuses on a nominated raster image, namely Portable Network Graphics (PNG), and a nominated vector image, namely Scalable Vector Graphics (SVG) Images Formats. The purpose of study has been to deduce optimal choices for image steganography, especially those that support web services. As shown above, SVG images format has been revealed in the study to be more efficient for image steganography than PNG images format in terms of capacity, scalability, simplicity, and performance. As demonstrated, SVG images format occupies fewer sizes before and after steganography processing, which makes it lighter and smoother for exchange through networks. As well, SVG images format provides scalability to a secret data, since it saves it externally rather than PNG images format, which saves it internally. Finally, processing steganography in SVG images format is simpler than PNG images format, since SVG images format are merely text files that provide the ease of processing using any direct text editor software, while being processed through RGB color model, and hence help to boost performance for processing steganography.

## REFERENCES

- [1] D. T. M. M.J.Thenmozhi, "A New Secure Image Steganography Using Lsb And Spiht Based Compression Method," *Int. J. Eng. Res. Sci.*, 2016.
- [2] N. K. M. A.- Anjali Tiwari, Seema Rani Yadav, "A Review on Comparison between Different Image Steganography Techniques," *vol. 3, no. 8*, pp. 355–358, 2014.
- [3] R. B. A. & O. M. A.-Q. Nagham Hamid, Abid Yahya, "Image Steganography Techniques: An Overview," *Int. J. Comput. Sci. Secur.*, no. 6, pp. 168–187, 2012.
- [4] A. Almutairi, "Optimal Specifications for a Secure Image Steganography Method," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 6, no. 2, pp. 198–202, 2017.
- [5] Sakshica and D. K. Gupta, "Various Raster and Vector Image File Formats," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 3, pp. 268–271, 2015.
- [6] P. M. Nishad and R. Manicka Chezian, "Various Colour Spaces and Colour Space Conversion Algorithms," *J. Glob. Res. Comput. Sci.*, vol. 4, no. 1, pp. 44–48, 2013.
- [7] H. K. Kelda and P. Kaur, "A Review: Color Models in Image Processing," *Int. J. Comput. Technol. Appl.*, vol. 5, no. 2, pp. 319–322, 2014.
- [8] M. J. Dahan, N. Chen, A. Shamir, and D. Cohen-Or, "Combining color and depth for enhanced image segmentation and retargeting," *Vis. Comput.*, vol. 28, no. 12, pp. 1181–1193, 2012.
- [9] A. C. Á, J. Condell, K. Curran, P. M. Kevitt, A. Cheddad, J. Condell, K. Curran, and P. Mc Kevitt, "Digital image steganography: Survey and analysis of current methods," *Signal Processing*, vol. 90, no. 3, pp. 727–752, 2010.
- [10] C. Seel-audom, W. Naiyapo, and V. Chouvatut, "A search for geometric-shape objects in a vector image: Scalable Vector Graphics (SVG) file format," in *2017 9th International Conference on Knowledge and Smart Technology (KST)*, 2017, pp. 305–310.
- [11] R. M. Mathis, "Constraint scalable vector graphics, accessibility and the semantic Web," in *Proceedings. IEEE SoutheastCon*, 2005., 2005, pp. 588–593.
- [12] D. M. Cabrita and W. Godoy, "PNG Optimization Techniques Applied to Lossless Web Images," *IEEE Lat. Am. Trans.*, vol. 10, no. 1, pp. 1398–1401, Jan. 2012.
- [13] H. Mao, Z. Hu, L. Zhu, and H. Qin, "PNG File Decoding Optimization Based Embedded System," in *2012 8th International Conference on Wireless Communications, Networking and Mobile Computing*, 2012, pp. 1–4.
- [14] Y. Jiazheng, H. Jinghua, W. Yujian, and B. Hong, "Converting real images to SVG based on XML," in *IET 2nd International Conference on Wireless, Mobile and Multimedia Networks (ICWMMN 2008)*, 2008, pp. 360–363.
- [15] Z. Liu, X. Cheng, C. Jia, and J. Yang, "Format Compliant Degradation for PNG Image," in *2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies*, 2013, pp. 720–723.
- [16] A. Jain and V. Jain, "PNG image copyright protection and authentication using SVD hash and AES," in *2014 International Conference on Advances in Engineering Technology Research (ICAETR - 2014)*, 2014, pp. 1–6.
- [17] T. Kumar and K. Verma, "A Theory Based on Conversion of RGB image to Gray image," *Int. J. Comput. Appl.*, vol. 7, no. 2, p. 9758887, 2010.
- [18] J. S. Kang, Y. You, M. Y. Sung, T. T. Jeong, and J. Park, "Mobile Mapping Service Using Scalable Vector Graphics on the Human Geographic," in *Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008)*, 2008, pp. 672–677.
- [19] C. Concolato, J. Le Feuvre, and J. C. Moissinac, "Design of an efficient scalable vector graphics player for constrained devices," *IEEE Trans. Consum. Electron.*, vol. 54, no. 2, pp. 895–903, May 2008.
- [20] I. Team, "Inkscape." 2017.
- [21] Microsoft, "Visio Professional 2016." 2017.
- [22] D. D. E. Team, "Dia Diagram Editor." 2017.
- [23] S. Vaidya, "OpenStego." 2017.
- [24] S. Singh and V. K. Attri, "Dual Layer Security of data using LSB Image Steganography Method and AES Encryption Algorithm," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 8, no. 5, p. 259266, 2015.
- [25] M. R. Shende; and A. Welekar, "Advanced Steganography for Hiding Data and Image using AudioVideo," *Int. J. Recent Innov. Trends Comput. Commun.*, 2016.
- [26] Sahar A. El\_Rahman, "A Comprehensive Image Steganography Tool using LSB Scheme," *I.J. Image, Graph. Signal Process.*, 2015.
- [27] D. Rawat and V. Bhandari, "Steganography Technique for Hiding Text Information in Color Image using Improved LSB Method," *Int. J. Comput. Appl.*, vol. 67, no. 1, p. 9758887, 2013.
- [28] Raju and Mohit Dhanda, "An Improved LSB based Image Steganography for Grayscale and Color Images," *Int. J. Curr. Eng. Technol.*, vol. 5, no. 5, p. 22774106, 2015.
- [29] Mohit, "An Enhanced Least Significant Bit Steganography Technique," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 5, no. 6, p. 22781323, 2016.
- [30] K. Muhammad;, J.Ahmad;, N. U. Rehman;, Z. Jan;, and R. J. Qureshi, "A Secure Cyclic Steganographic Technique for Color Images using Randomization," *Tech. Journal, Univ. Eng. Technol. Taxila*, 2014.

# Detection of Violations in Credit Cards of Banks and Financial Institutions based on Artificial Neural Network and Metaheuristic Optimization Algorithm

Zarrin Monirzadeh

Faculty of Computer Engineering,  
Department of Computer and  
Electronic Engineering, University  
of Eyvanekey, Semnan, Iran

Mehdi Habibzadeh

Faculty of Computer Engineering,  
Department of Computer and  
Electronic Engineering, University  
of Eyvanekey, Semnan

Nima Farajian

Faculty of Computer Engineering,  
Department of Computer and  
Electronic Engineering, University  
of Eyvanekey, Semnan, Iran

**Abstract**—Due to popularity of the World Wide Web and e-commerce, electronic communications between people and different organizations through virtual world of the Internet have provided a good basis for commercial and economic relations. These developments, although occurring for less than a century, electronic communications have always been subject to interference, cheating, fraud, and other acts of sabotage. Along with this increase in trading volume, there is a huge increase in the number of online fraud which results in billions of dollars of losses annually worldwide; this has a direct effect on customer service of banking systems, particularly electronic banking systems, and survival as a reliable financial service provider. Therefore, attention to fraud detection techniques is essential to prevent fraudulent acts and is the motive for many scientific researches. For this reason, business intelligence is used to identify financial violations in various economic, banking and other fields. Here, the focus is on algorithms and methods presented in data mining to deal with fraud by using neural networks. The main objective is to improve these methods or present new algorithms by studying the behavioral patterns of customers and the combined use of genetic algorithm to improve the performance of neural network and find the appropriate models for better decision making by implementing and testing the performance of the suggested algorithms. The results show that more strength was given to neural network by using genetic algorithm. In fact, genetic algorithm can raise our ability to control the training process. Moreover, it was concluded that criteria such as age, gender, marital status were not effective on detection; in fact, the most important effective criteria are information related to transaction.

**Keywords**—Financial fraud detection; neural networks; data mining; genetic algorithm

## I. INTRODUCTION

Although extensive research has been conducted on fraud detection, the need for these activities still persists due to the increasing number of financial and business activities and the increased use of modern technologies. Although there are still some up-to-date articles on this matter in prestigious journals, there is a lack of appropriate resources which include the latest research in this area and there is large-scale fraud in the financial and commercial areas. KPMG's 2003 research suggests an ever-increasing rate of fraud. The research indicates that 75% of the surveyed organizations experienced

instances of fraud. This figure is 13% higher than the 1998 figures. Hence, it is important to provide techniques for detecting fraud in e-commerce and research in this area. For a long time, traditional data analysis techniques have been used for detecting fraud. This requires complex and time-consuming research and requires the use of various fields of knowledge such as finance, economics, business methods, and legal debates [1]. Here, the focus is on algorithms and methods presented in data mining to deal with fraud by using neural networks. Artificial neural networks can work like a human brain and analyze information and correctly detect the problem when properly trained. Efficiency and function of a neural network is reasonable which it is properly designed [2]. For a correct design, parameters should be initialized and set reasonably. Typically, the method used to set input parameters is to use trial and error, through which various possible combinations are tested separately to select the best combination possible. There is always a lack of a regular approach to finding the best combination among different input parameters. Therefore, this study tends to introduce a method for determining the best combination among different input parameters. Neural network is considered as a very efficient functional approximation tool, which considers structure design, followed by optimal problem or network training. Instead of gradient-based methods, evolutionary optimization methods are used to determine the neural network weights (neural network learning or training) and genetic algorithm optimization code (the most well-known and most popular optimization algorithm in the field of evolutionary computing). In the field of fraud detection research, various techniques have been evaluated by various research communities and briefly investigated by numerous studies.

Carminati [5] suggested the BANKSEALER system which is a decision support system for analyzing online banking fraud. During the training phase of this system, easy models were developed to understand spending habits of customers based on past transactions. Halvaeie et al. [6] solved the credit card fraud detection problem using an artificial immune system. They developed a new model as artificial fraud detection model (AFDM) based on artificial immune system. This model used artificial immune and its improvement for fraud detection. Olszewski [7] suggested a fraud detection

method based on user account imaging and threshold type detection. The imaging method used in this approach was self-organizing mapping (SOM).

Artificial neural networks used for classification are widely used in many fields; one of their features is the unsupervised learning (Ghasemi & Asgharizadeh, 2016). Artificial neural networks are one of the methods used to identify fraud in bank cards. The advantage of neural networks to other methods is that it can learn from past transactions and improve the results over time [17].

Nagi et al. (2016) believed that fraud is one of the most common phenomena in business. According to Section 24 of the Iranian Standards of Audit, deceptive action of one or more directors, employees or third parties for an undue advantage refers to any intentional or unlawful act. Therefore, prevention or detection of important frauds in financial statements has always been the focus of investors, legislators, standardizers, managers and auditors [18]. This study examines the effectiveness of data mining techniques in detecting fraudulent behaviors of companies reporting fraudulent financial statements to identify effective factors on these behaviors. Data mining is a bridge between statistical science, computer science, artificial intelligence, modeling, machine learning and visual representation of data. In a process framework, it is possible to extract valid, previously unknown, intelligible and reliable information from a large database. It can be used in decision-making in important business activities such as improving the usefulness of information through identification of financial fraud [19].

To implement an effective neural network, genetic algorithm is used to detect financial violations to regulate the effective parameters on efficiency of neural network. The suggested genetic algorithm can be used to decide on topology of the network, number of hidden layers, number of nodes and other factors which are effective in design and efficiency of the neural network. The main objective is to improve these methods or develop new algorithms by studying the behavioral pattern of customers and integrating genetic algorithm to improve the performance of neural network and finding a suitable model for better decision making; performance of the suggested algorithms will be assessed to predict potential behavior of customers in the future [16].

## II. THEORETICAL FRAMEWORK

**E-banking:** E-banking is a set of services, technologies or processes used to remove time-consuming mechanisms and implement very in-house systems in banks. Electronic banking, as infrastructure of e-commerce, is one of the most important phenomena arising from information revolution and transformation of traditional ways of trading to replace it with e-commerce. Hence, electronic banking is considered as the main infrastructure of e-commerce due to the role of money and banking in commerce [3].

**Types of financial violations:** Financial violations are mainly carried out in two ways: direct and indirect. Directly, the physically lost or stolen card is used by other people. Indirectly, only the card number is stolen and used in phone purchases and other indirect purchasing methods, such as

Internet shopping. In the former, if the cardholder does not immediately find that the card is lost, it can only lead to financial loss. In the latter, the cardholder has no idea that he has shared his card with someone else and this may remain hidden for a long time. There is another type of indirect violation in which shared services of people are exploited; thus, the owners will have to charge their services sooner than the reasonable time or due date [4].

**Expert systems:** Expert systems refer to types of computational systems which are able to provide and reason in some rich areas of knowledge by solving problems and giving solutions [8]. Expert system detections encode knowledge in the form of rules; that is, they determine what should happen in what state by law. As an example, NIDES system, implemented by SRI, uses the approach of expert systems to identify attacks by using online monitoring of user activities [9]. NIDES include statistical analysis elements to detect abnormalities and rule analysis tools to detect abuses.

**Transition analysis:** This is an abuse detection technique in which attacks are displayed as a sequence of the monitored state transition. Activities which occur in an attack are defined as a transition between states. Attack scenarios are defined in the form of state transition diagrams. In these diagrams, nodes are system states and arcs are the related actions. In any case, if a final state is reached, it will mean the time of an attack. State Transition Analysis Tool (STAT) is a well-known regular expert system designed to search for known penetrations in an audit trail of multi-user computer systems [13]. Moreover, USTAT is also a prototype of STAT designed under the UNIX operating system [14].

**Clustering:** Data may contain complex structures from which even the best data mining techniques cannot extract meaningful patterns. Clustering provides a way to find the structure of complex data. Clustering refers to division of a heterogeneous population into a number of homogeneous subsets or clusters. Cluster refers to a set of information which is similar to other components of this set and is not similar to components of other sets. In clustering, there are no preset groups, and data is grouped simply by similarity, and the titles of each group are determined by the user [15].

**Neural networks:** Neural network is inspired by human brain; processing data is handled by many small processors which interact in parallel with each other to solve a problem. In these networks, a data structures is designed by programming methods, which can act as a neuron. This data structure is called a neuron. The network is trained by creating a network between these neurons and applying a training algorithm. By examining behavior of customers, their future behavior can be predicted. This requires a dataset consisting of characteristics of the former clients and their performance, whether they have a fraudulent and criminal function. By having this dataset and applying the right data mining techniques, one can predict the likelihood of fraud and criminal acts in new clients [15]. A neural network is a set of interconnected nodes designed by imitating human brain function. Each node has weighted communications to several other nodes on the adjacent layers [10]. In neural networks, the data structure designed by software can act as a neuron; this data structure is called a

node. Then the network is trained by creating a network between these nodes and applying a training algorithm to it. In this memory or neural network, the nodes have two active (on or 1) and inactive (off or 0) modes, and each edge (synapse or communication between nodes) has a weight. Positive weighted edges trigger or activate the next inactive node, and negative-weighted edges inactivate or inhibit the next connected node (if activated). An artificial neuron is a system with a large number of inputs and only one output. Neurons have two modes, training mode and operation mode. In training mode, the neuron learns to be triggered or fired against specific input patterns; however, the emerging trend of financial fraud is generally recognized through analyzing and extracting information (data mining) from the transaction database of financial institutions marked. This helps to formulate security policies and protocols and new authentication. In operation mode, when a detected input pattern is inserted, the corresponding output is provided. If the input is not part of the pre-identified inputs, the fire rules will decide for its triggering. Braves and Langdorff were the first to suggest an integration of continuous role-based systems and neural network-based approaches [11]. Falcon's fraud management system, a powerful tool for preventing fraudsters from abusing credit and debit cards, uses neural network algorithms. This system predicts the likelihood of fraud on an account by comparing the current transaction and past cardholder activities [12]. If this system detects a fraud-type transaction on a card, the cardholder will immediately be called by telephone; if the cardholder confirms fraud on the card, the card will be immediately blocked to prevent fraud. If the Falcon system detects any fraud, but it is not possible to call the cardholder, the card is temporarily blocked to ensure that the fraud is not committed and the cardholder must follow the situation by calling the bank; the card will remain blocked as long as the cardholder's contact is not recorded. The system is able to learn the cardholder's purchase habits by using neural networks and detect any irregularity in payment, and consider as a fraud. Machine learning techniques and technologies, adaptive pattern recognition, neural networks and statistical models have contributed in designing and developing the Falcon forecasting system.

### III. HYPOTHESES

**Hypothesis 1:** Genetic algorithm and neural networks used to discover violations in e-commerce systems provides better results.

**Hypothesis 2:** Auxiliary algorithms such as genetic algorithm used along with other algorithms such as neural networks provide a stronger configuration in detecting violations.

### IV. RESULTS

The data used includes various parameters such as name, gender, age, and address as personal specifications, and eventually information about transaction history, including transaction value, transaction time, and transaction status.

To test, parameters such as the number of primary population, the number of neurons, as well as the information

used can be changed to choose the best mode. All results are presented as mean. The test was iterated four times for each series of results and the mean of error and regression was presented for each iteration. This prevents random results. Moreover, variance was used to show the extent to which results of different tests were close to each other.

As it is clear, variance ranges from zero to one. The smaller and the closer numbers to zero indicate that the results of different tests are closer to each other and ultimately indicate the high reliability of results. The tables below also show regression (from the left, training regression, test evaluation regression and full regression, respectively).

First, the number of primary population was set at 30. All data available in dataset was used. The number of neurons was altered. Table I lists the results.

In Table II, only financial transaction information was used and demographic information such as name, age, and address were removed. The test mode is the same as the previous test mode and the results are presented in the Table II. Next, the number of primary population was set at 45; the results are shown in the Table III. First, all data available in dataset was used. In Table IV, a dataset of which personal data was removed was used and the results were presented. The number of primary population was set at 60. The results are presented for data with full specifications and data without personal specifications, respectively (Table V). In Table VI, data was used without personal information.

Obviously, the results of the neural network trained by genetic algorithm can be better than normal neural network. These results show that the neural network can be better trained by using optimization algorithms such as genetic algorithm (Table VII). In fact, this study indicates that optimization algorithms provide higher control on neural network training.

Genetic algorithm uses its unique features such as mutation and crossover which can be applied in a variety of ways and has a great potential in this regard. Thus, two-point crossover was used here to show capability of genetic algorithm. A finding of this study is that good results can be obtained without personal information; particularly in this method, since training time is longer than normal, the time will be longer when all data is used than the situation when this information is removed. Therefore, the latter is the final result of this study. For this purpose, these tests were run only on the data without personal information. Initially, the number of primary population was set at 30.

Table VIII shows that there was no need to increase the number of primary population, because the optimal result was obtained by this population. In fact, the need for greater population was reduced by using multiple crossovers. The greater population is required to increase overall search. However, this change can meet this need in shorter time. In fact, time is the weakness of this method, which was solved by using multiple crossovers. Better comparison is made below through diagrams. To clear the diagram, the scenarios 1 to 5 are described.

TABLE I. COMPARISON OF ERROR OF NEURAL NETWORK AND THE OPTIMIZED ERROR OF GENETIC ALGORITHM (POPULATION 30-ALL DATA)

Neuron No.	Suggested technique error	Error variance	Regression				NN error	Regression			
			0.86	0.85	0.87	0.86		0.81	0.81	0.81	0.81
[1,1,2]	0.022	0.0533	0.86	0.85	0.87	0.86	0.034	0.81	0.81	0.81	0.81
[1,1,3]	0.020	0.039	0.88	0.88	0.88	0.88	0.029	0.82	0.81	0.81	0.83
[1,2,3]	0.00884	0.109	0.947	0.94	0.9513	0.94649	0.0104	0.89	0.88	0.86	0.89
[2,2,3]	0.00839	0.038	0.952	0.95	0.952	0.951	0.012278	0.92	0.92	0.92	0.925

TABLE II. COMPARISON OF ERROR OF NEURAL NETWORK AND THE OPTIMIZED ERROR OF GENETIC ALGORITHM (POPULATION 30-FINANCIAL TRANSACTION)

Neuron No.	Suggested technique error	Error variance	Regression				NN error	Regression			
			0.875	0.84	0.88	0.86		0.82	0.815	0.83	0.82
[1,1,2]	0.0215	0.0263	0.875	0.84	0.88	0.86	0.0307	0.82	0.815	0.83	0.82
[1,1,3]	0.0198	0.0181	0.89	0.86	0.85	0.89	0.024	0.82	0.81	0.81	0.83
[1,2,3]	0.00787	0.0068	0.95	0.95	0.96	0.96	0.010	0.89	0.88	0.86	0.89
[2,2,3]	0.00809	0.0164	0.96	0.95	0.96	0.97	0.01107	0.93	0.91	0.91	0.94

TABLE III. COMPARISON OF ERROR OF NEURAL NETWORK AND THE OPTIMIZED ERROR OF GENETIC ALGORITHM (POPULATION 45-ALL DATA)

Neuron No.	Suggested technique error	Error variance	Regression				NN error	Regression			
			0.887	0.898	0.886	0.885		0.817	0.816	0.812	0.818
[1,1,2]	0.019	0.0071	0.887	0.898	0.886	0.885	0.057	0.817	0.816	0.812	0.818
[1,1,3]	0.00839	0.0214	0.875	0.95	0.952	0.951	0.0103	0.92	0.91	0.921	0.9396
[1,2,3]	0.00994	0.0381	0.97	0.967	0.97	0.97	0.0099	0.95	0.96	0.95	0.97
[2,2,3]	0.00717	0.037	0.98	0.98	0.978	0.98	0.00896	0.96	0.95	0.95	0.978

TABLE IV. COMPARISON OF ERROR OF NEURAL NETWORK AND THE OPTIMIZED ERROR OF GENETIC ALGORITHM (POPULATION 45-FINANCIAL TRANSACTION)

Neuron No.	Suggested technique error	Error variance	Regression				NN error	Regression			
			0.881	0.890	0.87	0.89		0.817	0.816	0.8122	0.818
[1,1,2]	0.017	0.046	0.881	0.890	0.87	0.89	0.051	0.817	0.816	0.8122	0.818
[1,1,3]	0.00804	0.0272	0.884	0.945	0.94	0.963	0.01	0.92	0.91	0.921	0.9396
[1,2,3]	0.00972	0.0189	0.94	0.97	0.969	0.98	0.00998	0.958	0.964	0.96	0.973
[2,2,3]	0.0067	0.0035	0.983	0.98	0.98	0.989	0.0075	0.978	0.96	0.968	0.979

TABLE V. COMPARISON OF ERROR OF NEURAL NETWORK AND THE OPTIMIZED ERROR OF GENETIC ALGORITHM (POPULATION 60-ALL DATA)

Neuron No.	Suggested technique error	Error variance	Regression				NN error	Regression			
			0.925	0.926	0.928	0.925		0.89	0.9	0.88	0.898
[1,1,2]	0.012278	0.0023	0.925	0.926	0.928	0.925	0.021	0.89	0.9	0.88	0.898
[1,1,3]	0.00884	0.077	0.945	0.948	0.951	0.946	0.0099	0.935	0.92	0.912	0.941
[1,2,3]	0.00387	0.024	0.968	0.969	0.9658	0.963	0.0064	0.961	0.957	0.95	0.96
[2,2,3]	0.00199	0.0041	0.89	0.982	0.981	0.993	0.0025	0.982	0.97	0.98	0.987

TABLE VI. COMPARISON OF ERROR OF NEURAL NETWORK AND THE OPTIMIZED ERROR OF GENETIC ALGORITHM (POPULATION 60-FINANCIAL TRANSACTION)

Neuron No.	Suggested technique error	Error variance	Regression				NN error	Regression			
			0.938	0.928	0.917	0.94		0.901	0.91	0.915	0.923
[1,1,2]	0.007278	0.0014	0.938	0.928	0.917	0.94	0.011	0.901	0.91	0.915	0.923
[1,1,3]	0.00684	0.08	0.957	0.949	0.951	0.96	0.0041	0.94	0.94	0.957	0.96
[1,2,3]	0.00497	0.042	0.978	0.971	0.972	0.981	0.0047	0.97	0.972	0.97	0.975
[2,2,3]	0.00169	0.006	0.986	0.982	0.98	0.993	0.0036	0.977	0.97	0.981	0.898

TABLE VII. COMPARISON OF ERROR OF NEURAL NETWORK AND THE OPTIMIZED ERROR OF TWO-POINT CROSSOVER GA (POPULATION 30)

Neuron No.	Suggested technique error	Error variance	Regression				NN error	Regression			
			0.88	0.871	0.87	0.886		0.84	0.82	0.827	0.83
[1,1,2]	0.0174	0.068	0.88	0.871	0.87	0.886	0.024	0.84	0.82	0.827	0.83
[1,1,3]	0.0082	0.024	0.894	0.88	0.879	0.89	0.021	0.868	0.854	0.858	0.87
[1,2,3]	0.00983	0.019	0.968	0.957	0.959	0.96	0.0097	0.938	0.94	0.94	0.94
[2,2,3]	0.00678	0.028	0.98	0.96	0.968	0.977	0.0091	0.97	0.95	0.954	0.962

TABLE VIII. COMPARISON OF ERROR OF NEURAL NETWORK AND THE OPTIMIZED ERROR OF TWO-POINT CROSSOVER GA (POPULATION 45)

Neuron No.	Suggested technique error	Error variance	Regression				NN error	Regression			
			0.927	0.9	0.89	0.91		0.878	0.87	0.869	0.88
[1,1,2]	0.00729	0.00796	0.927	0.9	0.89	0.91	0.0085	0.878	0.87	0.869	0.88
[1,1,3]	0.00692	0.078	0.948	0.945	0.994	0.963	0.01	0.92	0.91	0.921	0.9396
[1,2,3]	0.005	0.0178	0.978	0.96	0.957	0.97	0.00998	0.958	0.964	0.96	0.973
[2,2,3]	0.0017	0.0147	0.989	0.984	0.987	0.99	0.0075	0.98	0.974	0.97	0.98

Scenario 1: The neural network is trained by single-point crossover genetic algorithm; primary population is set at 30; the number of neurons is [2,2,3] and data lacks personal information.

Scenario 2: The neural network is trained by single-point crossover genetic algorithm; primary population is set at 45; the number of neurons is [2,2,3] and data lacks personal information.

Scenario 3: The neural network is trained by single-point crossover genetic algorithm; primary population is set at 60; the number of neurons is [2,2,3] and data lacks personal information.

Scenario 4: The neural network is trained by two-point crossover genetic algorithm; primary population is set at 30; the number of neurons is [2,2,3] and data lacks personal information.

Scenario 5: The neural network is trained by two-point crossover genetic algorithm; primary population is set at 45; the number of neurons is [2,2,3] and data lacks personal information.

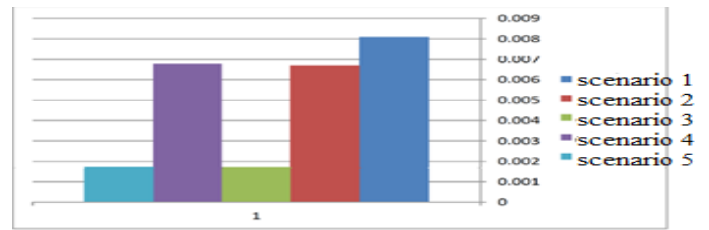


Fig. 1. Scenarios of primary population and single-point and two-point crossovers.

As shown in Fig. 1, two-point crossover could find results with smaller population to an optimal point. In this diagram, scenarios 3 and 5 are related to single-point and two-point crossovers with populations 60 and 45; the results are very close.

As shown in Fig. 2, factors such as population, bank growth, reward, salary, education are effective per unit time. Fig. 3 and 4 shows time-series prediction of multiple layer perceptron (MLP), neural networks and genetic algorithm. In these figures, the minimum time-series prediction is specified by red lines.

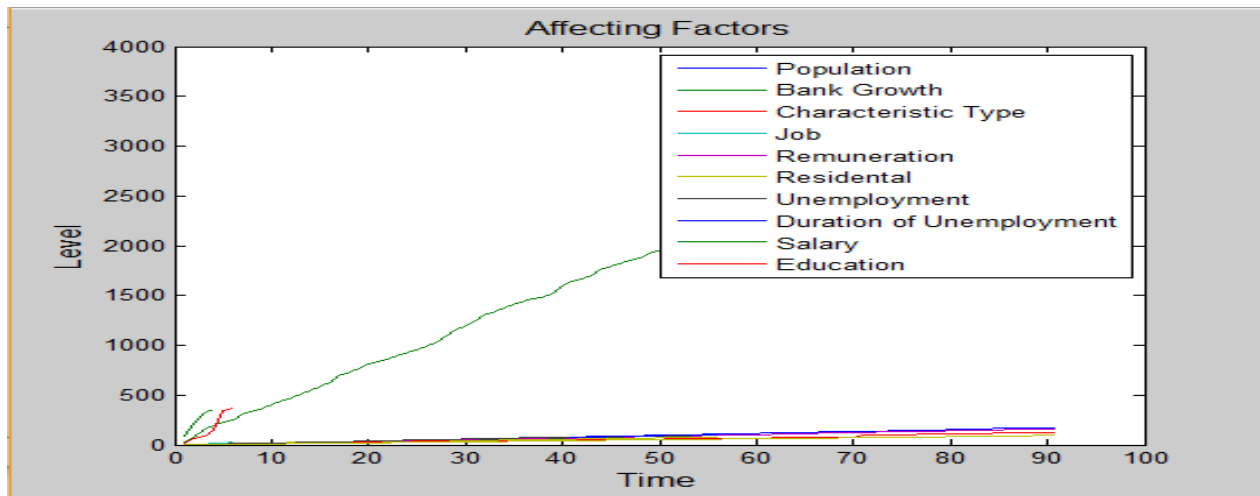


Fig. 2. Effective factors.

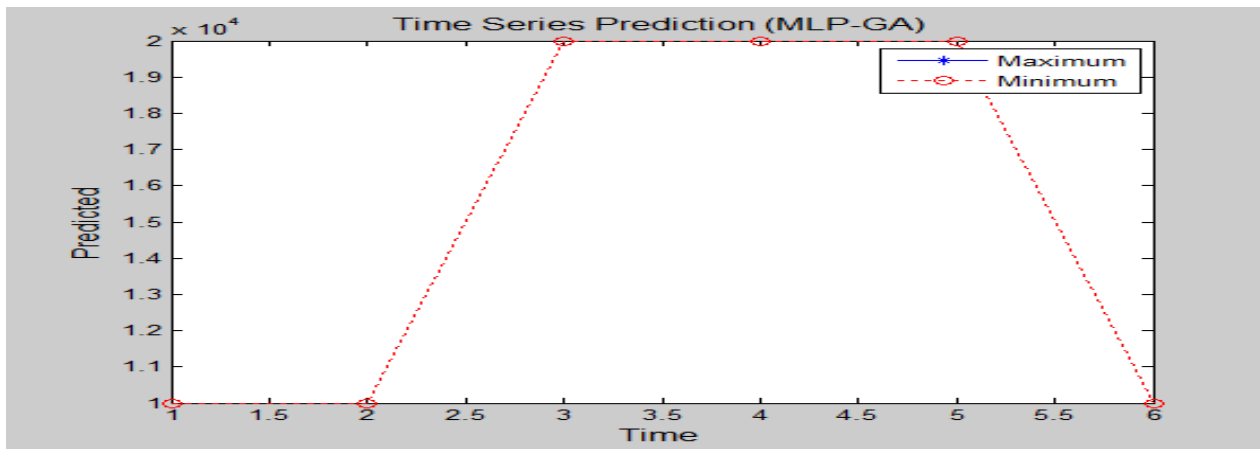


Fig. 3. Time-series prediction (MLP-GA).

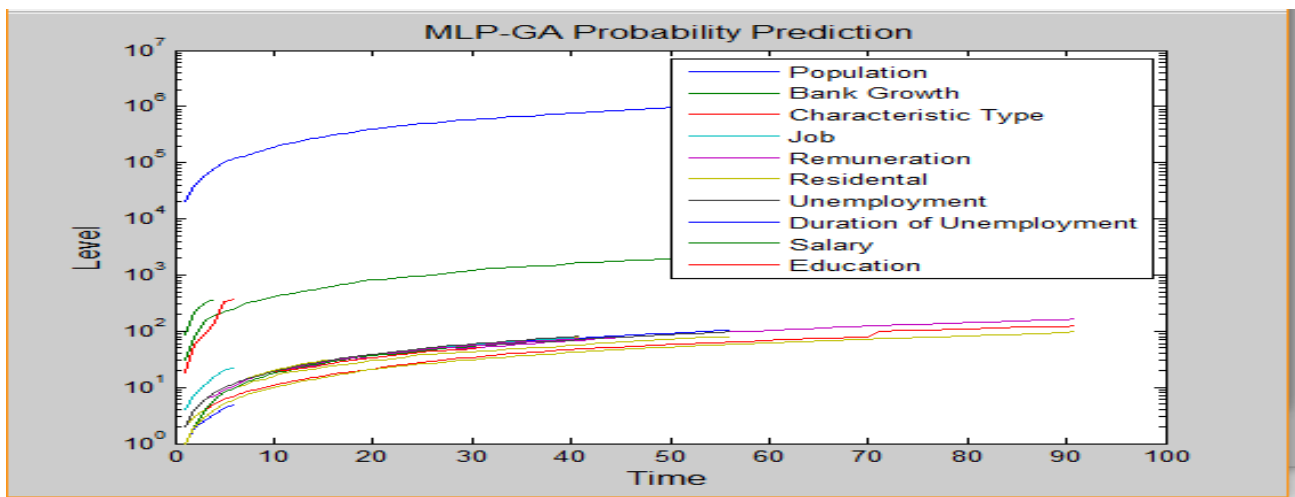


Fig. 4. MLP-GA probability prediction.

### V. CONCLUSION

By historical detection using library documents and literature review, this study provides the necessary evidence to answer the questions. The results showed that first, data mining techniques are useful for detection in fraudulent financial statements; second, data mining can be considered as focus of guiding thought in business management to detect fraud.

Hybrid use of fraud detection and fault detection approaches integrates the advantages of both methods and eliminates the weaknesses of each method. Using this approach in fraud detection techniques, it is very efficient to use techniques such as neural networks. Regardless of technical discussion, it is important to note that the expansion of e-commerce and increasing growth of financial services of banks and credit and financial institutions, the increase in the number of customers and penetration rate of users, and high volume of transactions have caused new problems and challenges, such as increased tendency of fraudsters to electronic banking, which require a careful examination of data; if there are no mechanisms for detecting and preventing fraud, there will be an increase in fraud in electronic banking. However, it is not possible to accurately examine this volume of data with routine methods. On the other hand, financial and credit institutions are looking for solutions which can quickly detect criminal acts. Hardware and software capacities provided in the present century with data mining techniques can be used to examine the high volume of this kind of information. This study discussed neural network training. Neural network is one of the smart and powerful tools for prediction. Therefore, it will be useful to focus on its improvement. The main element of neural network is training. Neural network training refers to determining weight parameters and its bias. The equation of input and output is obtained by determining these parameters. For this purpose, training data is observed to reach the proper value for these parameters by iteration. This process means finding the optimal value. That is why this study used genetic optimization algorithm instead of conventional neural network training techniques. The objective function for training neural network is the same as error rate between actual output and output of the neural network. On the other hand, fault detection

in financial transactions is another aspect of this study. For this purpose, special data was used. The results show that genetic algorithm could empower the neural network to achieve better results in e-commerce. In fact, genetic algorithm increases the ability to control training process. It can be concluded that criteria such as age, gender, marital status have no effect on detection; in fact, the most important effective criteria are transaction-related information.

### REFERENCES

- [1] A. Hatamirad, H. N. D. Shahriari, "fraud detection techniques in e-banking". economics new findings, vol. 134 , 2009.
- [2] S. N. D. Akbari, "financial fraud detection by data mining". AKSA IT Innovation Group, Iran, 2011.
- [3] H. Houshmand, "ecommerce and ebanking; challenges and solutions". 2010.
- [4] R. Varjini, M. Kani, "financial fraud detection techniques". Islamic Azad university of Gonabad. 2011.
- [5] M. Carminati., "BankSealer: A decision support system for online banking fraud analysis and investigation", Computers & Security., 2015
- [6] M. Halvaeie, N. Soltani, M. K. Akbari, "A novel model for credit card fraud detection using Artificial Immune Systems", Applied Soft Computing, pp. 2440-49, 2014.
- [7] D. Olszewski, "Fraud detection using self-organizing map visualizing the user profiles", Knowledge-Based Systems Vol.70, pp. 324-334, 2014.
- [8] T. F. Lunt, "A Real-Time intrusion Detection Expert System (IDES)-Final Technical Report". Technical Report. SRI Computer Science Laboratory, SRI International, from. <http://www.wenke.gtisc.gatech.edu>.1990
- [9] D. Anderson, T. Frivold, A. Tamaru , A. Valdes, "Next generation intrusion detection expert system (NIDES)", software user's manual,beta-update release, Technical Report SRIXSL-9547, Computer Science Laboratory, SRI International, from [www.thc.org/root/docs/intrusion-detection/...NIDES-summary.pdf](http://www.thc.org/root/docs/intrusion-detection/...NIDES-summary.pdf), 1994.
- [10] A. K. Ghosh, A. Schwartzbard, M. Schatz, "A Study in Using Neural Networks for Anomaly and Misuse Detection. 8th USENIX Security Symposium, from [www.portal.acm.org/citation.cfm?id=1251433](http://www.portal.acm.org/citation.cfm?id=1251433), 1999.
- [11] R. Brause, T. Langsdorf, M. Hepp, "Credit Card Fraud Detection by Adaptive Neural Data Mining", 11 th IEEE International Conference on Tools with Artificial Intelligence. Pp.103-106, 1999.
- [12] K. Hassibi, "Detecting Payment Card Fraud with Neural Networks", Singapore: World Scientific, ,2000.



- [13] K. Ilgun, R. A. Kemmerer, P. Porras, "A State transition analysis: A rule-based intrusion detection approach" *Software Engineering*, Vol. 21, pp. 181-199, 1995.
- [14] K. Ilgun, "USTAT A Real-time intrusion detection system for UNIX" *IEEE Symposium on Research in Security and Privacy*, pp.16-28, 2011.
- [15] I. C. Yeh, C. H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", *Expert Systems with Applications*, vol, 36, pp. 2473-2480., 2009.
- [16] A. Sharma, P. Kumar Panigrahi, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques", *International Journal of Computer Applications*, pp.0975 – 8887, Vol. 39, 2012.
- [17] Ghasemi, A. R. & Asgharizadeh, E. (2016). Presenting a hybrid ANN-MADM Method to Define Excellence Level of Iranian Petrochemical Companies. *Journal of Information Technology Management*, 6(2): 267-284.
- [18] Zakaryazad, Ashkan, and Ekrem Duman. "A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing." *Neurocomputing* 175 (2016): 121-131.
- [19] Semwal, Vijay Bhaskar, et al. "Design of Vector Field for Different Subphases of Gait and Regeneration of Gait Pattern." *IEEE Transactions on Automation Science and Engineering* (2016).

# Data Exfiltration from Air-Gapped Computers based on ARM CPU

Kenta Yamamoto, Miyuki Hirose, and Taiichi Saito  
Tokyo Denki University  
5 Senju-Asahi-Cho, Adachi-Ku, Tokyo 120-8551, Japan  
Tokyo, Japan

**Abstract**—Air-gapped Network is a network isolated from public networks. Several techniques of data exfiltration from air-gapped networks have been recently proposed. Air-gap malware is a malware that breaks the isolation of an air-gapped computer using air-gap covert channels, which extract information from air-gapped computers running on air-gap networks. Guri et al. presented an air-gap malware “GSMem”, which can exfiltrate data from air-gapped computers over GSM frequencies, 850 MHz to 900MHz. GSMem makes it possible to send data using the radio waves leaked out from the system bus between CPU and RAM. It generates binary amplitude shift keying (B-ASK) modulated waves with x86 SIMD instruction. In order to efficiently emit electromagnetic waves from the system-bus, it is necessary to access the RAM without being affected by the CPU caches. GSMem adopts an instruction that writes data without accessing CPU cache in Intel CPU. This paper proposes an air-gap covert channel for computers based on ARM CPU, which includes a software algorithm that can effectively cause cache misses. It is also a technique to use NEON instructions and transmit B-ASK modulated data by radio waves radiated from ARM based computer (e.g. Raspberry Pi 3). The experiment shows that the proposed program sends binary data using radio waves (about 1000kHz ~ 1700kHz) leaked out from system-bus between ARM CPU and RAM. The program can also run on Android machines based on ARM CPU (e.g. ASUS Zenpad 3S 10 and OnePlus 3).

**Keywords**—Air-Gapped Network; ARM CPU; data exfiltration; SIMD; NEON; GSMem

## I. INTRODUCTION

Air-gapped network is a network physically separated from other unsecured networks, and *air-gapped computer* is a computer in an air-gap network. Industrial control systems and security protection systems are often constructed in the air-gapped networks in which data leakage is prevented by restricting the use of Wi-Fi and Bluetooth and the access to removable storage such as USB flash drive.

Air-gap malware is a malware that breaks the isolation of an air-gapped computer using *air-gap covert channels*, which extract information from air-gapped computers running on air-gap networks.

EMSEC (Emission Security) [1] is an approach against attack using electromagnetic waves leaked from the computer. Usually, a computer emits various energy by data communication such as Wi-Fi or Bluetooth. However, more energy emissions are generated than what the user aware. For example, the fact that an electric current flows in a base board or a wiring may itself be an antenna. TEMPEST [2] is a specification of the defense technology against an attacker to exploit and techniques to steal information by using the emitted energy, such as electromagnetic waves or sound by the National Security Agency. In 1985, Van Eck [3] showed that electromagnetic radiation of monitor can be captured and image can be reconstructed using inexpensive devices as concrete exploit method of TEMPEST. Kuhn and Anderson demonstrated that the emission of electromagnetic radiation emitted from desktop computers can be controlled by software [4], [5]. Several methods of air-gap covert channels have been recently proposed. Mordechai et al. [6] proposed a method of transmitting data through an air gap using electromagnetic waves from a display cable. Hanspach et al. [7] proposed a method of transmitting data through an air gap using ultrasonic waves from a speaker.

Guri et al. a research team at Ben-Gurion University, presented an air-gap malware GSMem [8]. It generates electromagnetic waves from the memory bus between CPU and RAM to transmit B-ASK modulated signals in Intel architecture. GSMem executes an x86 SIMD instruction that exhausts the memory bus bandwidth to accelerate leakage of electromagnetic waves from the memory bus.

This paper presents an air-gapped covert channel on air-gapped computers based on ARM architecture CPU through which B-ASK modulated signals are transmitted over the AM frequency band (1,000 kHz - 1,600 kHz). GSMem adopts the particular instruction in x86 SIMD instruction set which can manipulate data without going through the CPU cache to efficiently access the memory bus, in order to transmit the modulated signal on Intel-based computer. On the other hand, since the CPU of the ARM architecture does not support instructions of application level to bypass the CPU cache, this paper proposes an algorithm that directly accesses memory avoiding CPU cache hit as much as possible and uses a NEON instruction that efficiently occupies the bandwidth of memory bus. The experiment executes the program that adopts the algorithm and shows the electromagnetic waves leaked from ARM computer with a spectrum analyzer.

---

A poster presentation and preliminary version of this paper were presented at IWSEC 2017 and CSS 2017, respectively.

The authors declare that there is no conflict of interest regarding the publication of this paper.

The remainder of this paper is organized as follows: Section 2 describes the basic technology to understand the proposed method in this paper. Next, Section 3 presents assorted related works. Section 4 describes a main method. Section 5 presents the result of measurement. Finally, Section 6 concludes this paper.

## II. BASIC TECHNICAL OUTLINE

This section provides a basic technical information to make it easier for readers to understand technical parts in the algorithm proposed by the previous researches and the proposed method.

### A. B-ASK Modulation

Binary amplitude shift keying (B-ASK) modulation is one of the modulation techniques for transmitting digital signals. It changes the amplitude corresponding to the transmitting binary data. B-ASK uses only amplitude modulation, and the frequency and phase are fixed. When the bit to be transmitted is "1", the amplitude of the carrier wave becomes large, and when the bit to be transmitted is "0", the amplitude of the carrier wave becomes small. Fig. 1 shows an example of an amplitude representation of bit string to be transmitted. Fig. 2 is an illustration of a signal to be sent (upper) and a signal modulated with B-ASK (lower).

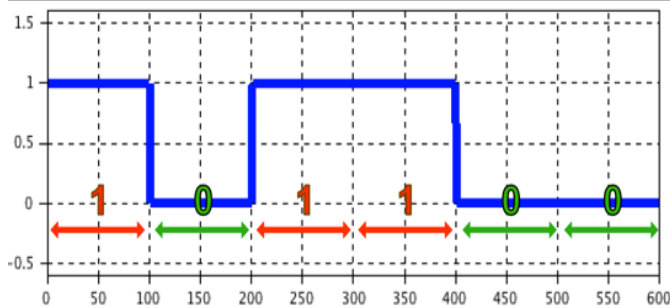


Fig. 1. Example of signal change according to bit string by B-ASK.

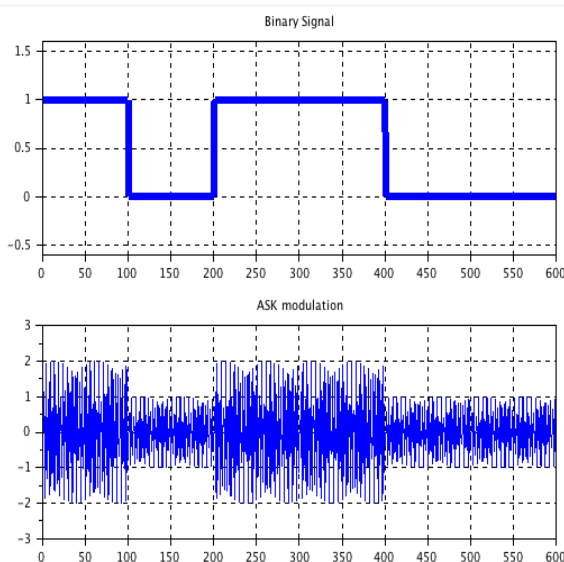


Fig. 2. The signal to be transmitted (upper) and the modulated signal (lower).

### B. CPU Cache

CPU cache is a small capacity memory installed in the CPU to hide the low-speed memory bus when the CPU transfers data to the memory. The bandwidth of the memory bus cannot catch up with the processing capability of the CPU, which becomes a bottleneck. CPU cache works as a very high-speed memory that holds the data and addresses of the memory accessed by the CPU in order to complement the performance difference between the CPU and RAM. In the case that the CPU accesses the data stored in the cache memory, it does not need to go through the memory bus, and therefore it is possible to avoid the bottleneck. CPU cache has various features: capacity, speed, data update method, etc. depending on the architecture. However, since the hardware automatically operates the cache, an application software needs not to control it. Although it is a high-speed memory, CPU cannot have a large-capacity cache memory like RAM. In many cases, the CPU cache has a multistage structure in order to support with increasing the capacity of the RAM and multi-core. They are called Level 1 (L1), Level 2 (L2), Level 3 (L3) in the order closest to the CPU core. Most of the L1 caches are installed for each CPU core, and the L1 cache is the fastest cache memory with the smallest capacity. L2 and L3 caches are cache memories with larger capacity, and they are shared with all CPU cores.

In Intel-based CPU for individual products, they have L1 and L2 caches about 32KB-256KB in each cores, and L3 as shared cache of up to 8MB. Cortex-A53, one of the ARM architectures, has up to 64KiB of L1 cache and 2MiB of L2 cache in each core.

In either architecture, in the L1 cache, it is divided into an instruction cache and a data cache. The instruction cache stores CPU instruction groups included in the program, and the data cache is an area for storing data processed by the program.

The CPU cache is stored by a unit called *line* and associated with the tag generated from the memory address. The line length in the ARM architecture is eight words [9], [23].

### C. SIMD Instructions

Single Instruction Multiple Data (SIMD) is an instruction set that can manipulate multiple data with a single instruction. In an algorithm capable of parallelization, it is possible to increase the effective speed of a program by processing a plurality of data with one clock by using SIMD. It is effective in an algorithm that parallelly calculates data equal to or larger than the data width supported by the processor in the general-purpose instruction set. Usually, the data width supported by the instruction set is divided and used for parallel calculation. For example, when the SIMD instruction set supports a calculation width of up to 128-bit, it is common to use a method of calculating four 32-bit floating point operations in parallel.

The Intel CPU can use the SSE instruction set, AVX, etc. the SSE-based instruction set supports to use up to 128-bit and AVX supports up to 256-bit.

Some ARM architectures support SIMD instructions called NEON. NEON supports up to 128-bit and is available in the Cortex-A family.

### III. RELATED WORK

#### A. GSMem

GSMem presented by Guri et al., is a malware that exfiltrates data from an air-gapped computer based on x86 CPU architecture. It utilizes the phenomenon that electromagnetic waves over the GSM frequency band are radiated when the CPU accesses RAM via the memory bus. GSMem malware infects a computer and performs it as a transmitter that sends a modulated signal by B-ASK over the GSM frequency band. The B-ASK modulation is a method of sending digital data by changing the amplitude of carrier wave according to binary symbols "1" and "0". GSMem involves much RAM access to generate large amplitude and less access to generate small amplitude. The amplitude of leaked wave becomes significantly large at GSM frequencies when memory access occurs. The receiver tries to demodulate leaked wave at GSM frequencies measuring the amplitude of it.

Guru et al. use the following technique for amplifying radiation of electromagnetic wave when accessing to RAM. They require the algorithm to involve direct RAM access avoiding CPU cache. Their technique uses an SSE2 instruction "MOVNTDQ" [10] on x86 architecture CPU in order to produce efficient access to RAM. MOVNTDQ is an instruction that stores data to RAM bypassing CPU caches. This is named `_mm_stream_si128` [11] in C++ library.

Algorithm 1 is a concept code which sends B-ASK modulated data by GSMem. This algorithm reads a binary data from an element of array *data* at index *bit\_index*, and if it is '1', the algorithm repeats executing MOVNTDQ for *tx\_time* nanoseconds. If the bit is '0', the algorithm sleeps for the same period.

---

**Algorithm 1.** Transmit data by the GSMem

---

```
1: buffer ← ALIGNED_ALLOCATE(16,4096)
2: tx_time ← 500000
3: for bit_index ← 0 to 32 do
4:   if (data[bit_index] == 1) then
5:     start_time ← CURRENT_TIME()
6:     while (tx_time > CURRENT_TIME() - start_time) do
7:       buffer_ptr ← buffer
8:       for i ← 0 to buffer_size do
9:         MOVNTDQ (buffer_ptr, 128bit_register)
10:        buffer_ptr ← buffer_ptr + 16
11:      end for
12:    end while
13:  else
14:    SLEEP (tx_time)
15:  end if
16: end for
```

#### B. System-Bus-Radio

William Entriken (github.com: fulldecent) presented the program "System-bus-radio"<sup>1</sup> that generates an AM frequency carrier wave amplification-modulated with sound data to be reproduced by a loudspeaker. System-bus-radio uses an algorithm similar to GSMem. It calculates "period" from audio frequency. It produces audio wave of frequency *f* by generating alternately "1" and "0". Fig. 3 shows an example of generating audio wave by calculated "period" from frequency *f*.

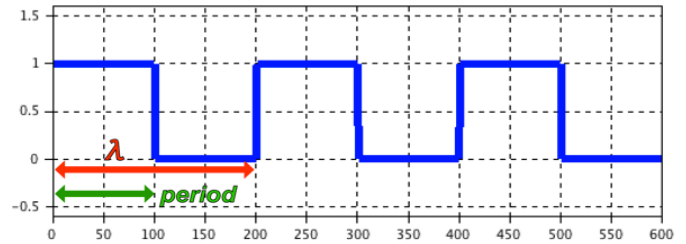


Fig. 3. System-bus-radio calculates "period" from  $f = 1/\lambda$  to obtain audio wave with frequency *f*. For example, if it wants sound of 2000 Hz, it calculates  $period = 1/2\lambda = f/2 = 0.00025$  seconds.

Even a common AM radio can be used to receive a signal as a receiver in System-bus-radio. Since the signal is amplitude-modulated, the AM radio can receive and demodulate it to reproduce the sound. System-bus-radio uses SSE2 instruction MOVNTDQ available on x86 architecture CPU.

### IV. PROPOSED METHOD

This section proposes an algorithm that transmits B-ASK modulated signal in ARM-architecture-based computers. In GSMem, Guri et al. adopted the SIMD instruction, MOVNTDQ in x86 architecture CPU. They showed that the instruction can efficiently and massively send data to the memory bus. Since ARM, however, has a different microarchitecture from x86, the instruction cannot be used. The proposed algorithm adopts a SIMD instruction that efficiently accesses the memory bus in ARM CPU based computer.

#### A. ARM Architecture

Low Power Double Data Rate (LPDDR) [12], [13] is a power saving standard of DDR memory. LPDDR memory is often adopted for many devices with ARM architecture CPU for power saving. Since the LPDDR memory has a 32-bit bandwidth bus, it is possible to occupy the memory bus even by using a 32-bit or 64-bit general-purpose instruction set. However, if the computer is in a multi-channel environment, any instruction in the general purpose instruction set cannot occupy the memory bus bandwidth.

In this paper, NEON instruction set is adopted for occupying the memory bus more certainly in a computer based on ARM CPU. NEON is a SIMD instruction set that is available on ARM architecture CPU. It can be used on ARMv7 and above, and operate 64-bit or 128-bit data.

---

<sup>1</sup> William Entriken (fulldecent). (2017) System-bus-radio. [Online]. [Accessed 25 October 2017]. <https://github.com/fulldecent/system-bus-radio/>.

Therefore, it is suitable for bandwidth occupation of the memory bus.

The instructions for operating data without using the CPU cache are equipped in x86 SIMD. However, in the NEON instruction set, there is no instruction to operate data explicitly without using the CPU cache. The ARM CPU has an instruction cache and a data cache, also has CPU modes disabling each cache. However, since the CPU's privileged mode is required to switch modes, invalidating the CPU cache is not a practical way for malware running in user application level. Therefore, it is needed to build an algorithm that generates access to RAM even if the CPU cache is valid, and select the optimal instruction from the NEON instruction set. The next section proposes an algorithm to avoid CPU data cache hit.

### B. Algorithm

In order to avoid reading cached data, it is only necessary that the data is not stored in data cache when the CPU refers to it. Here a concept code realizing an algorithm to avoid CPU data cache in ARM CPU is presented in **Algorithm 2**. This code modulates the electromagnetic waves from memory bus by B-ASK and transmits 4-bit data "1010". In order to transmit '1', it is necessary to repeat accessing the memory bus for a predetermined time to radiate electromagnetic waves. The algorithm repeats loading data on memory and releases electromagnetic radiation from the memory bus.

**Algorithm 2** is a simple algorithm for loading data from memory, but it changes the address of data to be loaded every time the load instruction is executed. CPU operates different data on different location each time. This trick makes it difficult to hit the CPU data cache.

---

#### Algorithm 2. A new algorithm for ARM computers

---

```
1: p = (int32_t *)malloc(size)
2: for (int i=0; i<=n; i++)
3:   p[i] = i;
4:
5: data_bits[] = {1, 0, 1, 0,}
6: period = 500000
7: for data in data_bits :
8:   if (data == 1):
9:     i = 0
10:    start = now()
11:    while (period > now() - start):
12:      va = vld1q_s32(p+i)
13:      i+=8
14:      if(i==limit) i=0
15:   if (data == 0):
16:     sleep(period)
```

As a SIMD instruction executed for loading data, the proposed algorithm adopts "VLD1.32" [14] which allows more data to be transferred than a general-purpose instruction set. It is implemented as the function "vld1q\_s32" in C++ language library. VLD1.32 is an instruction to load four 32-bit data stored in the RAM as a single vector into a register. It allows loading 128-bits data at a time.

This algorithm separates into a part of allocating a large array in memory and the other part of controlling the radiation electromagnetic waves from memory bus. When it executes RAM access, the CPU repeats loading the data allocated in the memory.

In Algorithm 2, the first line allocates the memory. It reserves "size" bytes of 32-bit data array when working on a 32-bit system. The size is needed to be at least larger than the L1 cache for the algorithm to effectively work. For example, the Cortex-A53 architecture, which is one of the ARM Cortex-A series installed in Raspberry pi 3, can have an L1 cache up to 64 KiB, so the size is needed to be larger than 64 KiB.

The lines 2-3 initialize each element of the array with each distinct value. "data\_bits" is an array of binary data to be transmitted. In this case, the data "1010" is sent.

The "period" means 500 microseconds. According to GSMem, if the period decreased, a higher bit rate is obtained, but the error rate is increased. The algorithm sets period to 500 microseconds in our algorithm like GSMem's one.

In the line 7, when a data in data\_bits is 1, the block starts while loop that performs memory operations and generates electromagnetic wave.

In the line 12, it loads the data into the registers as a single vector from the four addresses using the "vld1q\_s32" instruction. Then it adds 8 to the variable "i". It is the index to select memory address. That is, the index of the memory address next to be loaded is shifted by 8. This integer 8 means the size of the cache line in the ARM CPU. If "i" reaches the overflow value, set it to 0. The program repeatedly executes the code written on the lines 12-14 for the "period" time, where electromagnetic waves are generated from the memory bus.

In the last line, the algorithm sleeps for "period" microseconds when outputting "0". The amplitude of electromagnetic waves from memory bus becomes small.

### C. Details of Avoiding Cache

Once the CPU accesses a location in main memory, the data around the location is stored in the CPU cache in units of 8 words. Even when the CPU reads data of index 0 to 3 in an array, there is a possibility that data of index 0 to 7 is stored in the cache. Accordingly, on every memory access, the algorithm needs to read data at least 8 words far from the previously accessed data.

Fig. 4 shows the first action of loading four elements from an array, and Fig. 5 shows the next action of loading the four elements located 8 words far from the first ones in the array. Because the algorithm allocates an array larger than the total size of the CPU caches, the data in elements initialized in early stage have already been removed from the caches and then the actions success in causing cache miss and loading data directly from RAM. Also, by loading, CPU generates new data caches. To prevent loading from the same cache line, the algorithm shifts 8 words the position to be loaded each time.

The algorithm is described based on virtual memory terms and it is assumed that the virtual memory is almost not

fragmented. However, even though the virtual memory is fragmented and mapped to the physical memory, there is little influence for the purpose of outputting electromagnetic waves.

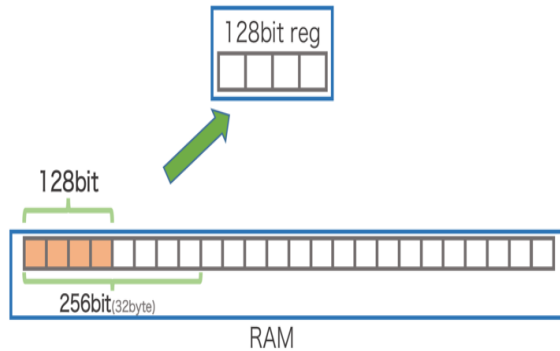


Fig. 4. The action of the first loading elements from memory to 128-bit register. Algorithm selects four elements in index 0-3.

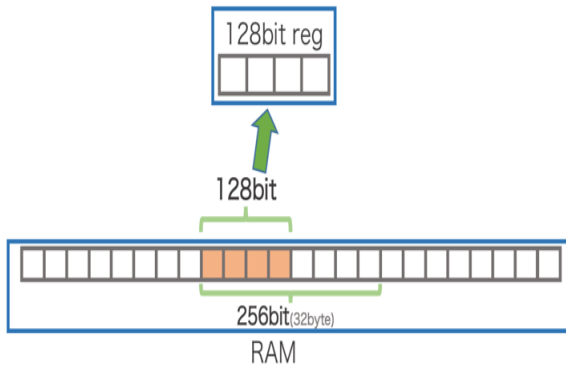


Fig. 5. The next action of loading elements. Algorithm selects four elements in index 8-11.

## V. MEASUREMENTS

In order to verify effects of the proposed algorithm in this section, frequency and time characteristics of the leaky electromagnetic wave from devices, based on ARM CPU, implemented the program described in the previous section, were measured by a spectrum analyzer and an oscilloscope.

### A. Frequency Domain

#### 1) Experimental arrangement

Based on the algorithm in the previous section, a data transmission program was created. It was forked from Systembus-radio made by William Entriken. This section shows the frequency characteristics when the program is executed on the computer based ARM CPU. Raspberry Pi 3 and ASUS Zenpad 3S 10 Z500M (Z500M) were used for the measurements.

Raspberry Pi 3 is running with Broadcom BCM2837<sup>2</sup>, which is one in a series of ARM architecture. The BCM2837

has four ARM Cortex-A53<sup>3</sup> CPU cores. Also Raspberry Pi 3 has 1 GB of LPDDR2 SDRAM. Raspbian Linux for experiment OS is used in this experiment.

Z500M uses Android 7.0 as operating system. Z500M is running on MediaTek MT8176<sup>4</sup>. Internally, the MT8176 has two cores of ARM Cortex-A72<sup>5</sup> and four Cortex-A53. Z500M is installed with 4 GB RAM. The RAM is connected to CPU by dual channel.

The measurement campaigns were conducted in a radio anechoic chamber. The receiving antenna was an omnidirectional, vertically polarized mono-pole antenna. Frequency-domain propagation gains were measured with a spectrum analyzer, as shown in Fig. 6. The DUT (Device Under Test) were Raspberry Pi3 and Z500M.

### 2) Results and analysis

The frequency-domain gains of Raspberry pi 3 and Z500M are shown in Fig. 7 and 8, respectively. The blue lines stand for ON state, the proposed program is executing. The black lines stand for OFF state, program is not executing. While a signal was found in ON state, no one was yielded practically in OFF state. From the measurement data, the peak (the highest amplitude value) of signals at approximately 1.5 MHz was observed in case of Raspberry Pi 3, and the peaks ranged 1.2 – 1.4 MHz in case of Z500M. In both cases, peak gains were approximately 4 dB.

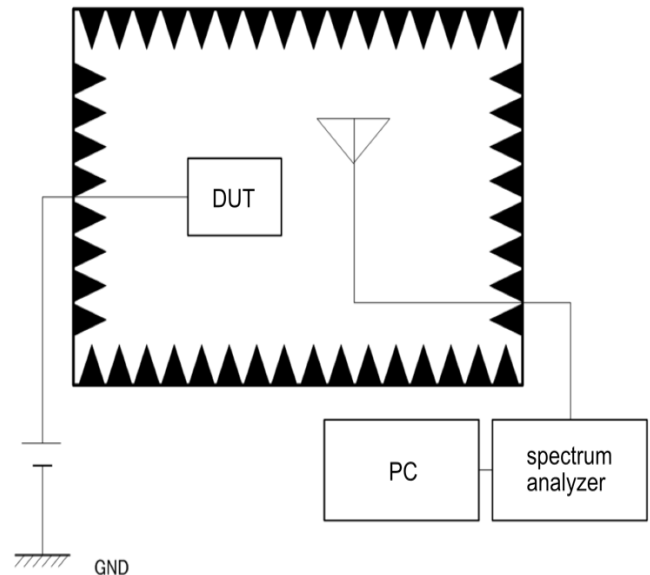


Fig. 6. Experiment setup.

<sup>3</sup> ARM Ltd., “ARM® Cortex®-A53 MPCore Processor Revision: r0p4 Technical Reference Manual”, [Online]. [Accessed 15 November 2017], [http://infocenter.arm.com/help/topic/com.arm.doc.ddi0500g/DDI0500G\\_cortex\\_a53\\_trm.pdf](http://infocenter.arm.com/help/topic/com.arm.doc.ddi0500g/DDI0500G_cortex_a53_trm.pdf)

<sup>4</sup> MediaTek Inc. (2016), MediaTek MT8176 for Tablets | MediaTek, [Online]. [Accessed 27 October 2017] <https://www.mediatek.com/products/tablets/mt8176>

<sup>5</sup> ARM Ltd., “ARM® Cortex®-A72 MPCore Processor Revision: r0p1 Technical Reference Manual”, [Online]. [Accessed 15 November 2017], [http://infocenter.arm.com/help/topic/com.arm.doc.100095\\_0001\\_02\\_en/cortex\\_a72\\_mpcore\\_trm\\_100095\\_0001\\_02\\_en.pdf](http://infocenter.arm.com/help/topic/com.arm.doc.100095_0001_02_en/cortex_a72_mpcore_trm_100095_0001_02_en.pdf)

<sup>2</sup> RASPBERRY PI FOUNDATION (2017), Raspberry Pi 3 Model B - Raspberry Pi, [Online]. [Accessed 25 October 2017] <https://www.raspberrypi.org/products/raspberry-pi-3-model-b/>

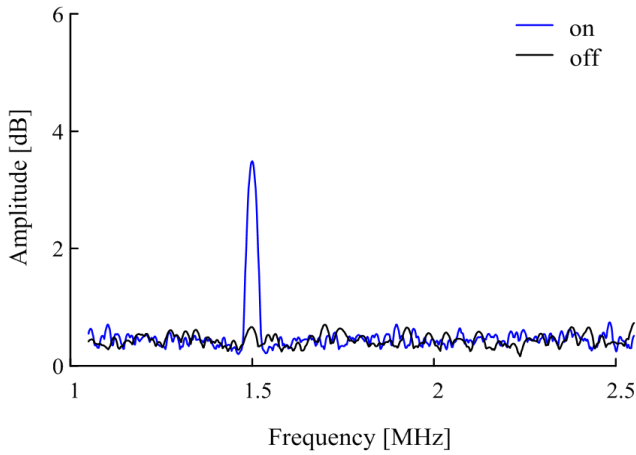


Fig. 7. An example of frequency-domain gains.(Raspberry Pi 3). “on” is in the program execution state. “off” is the state in which the program is not running.

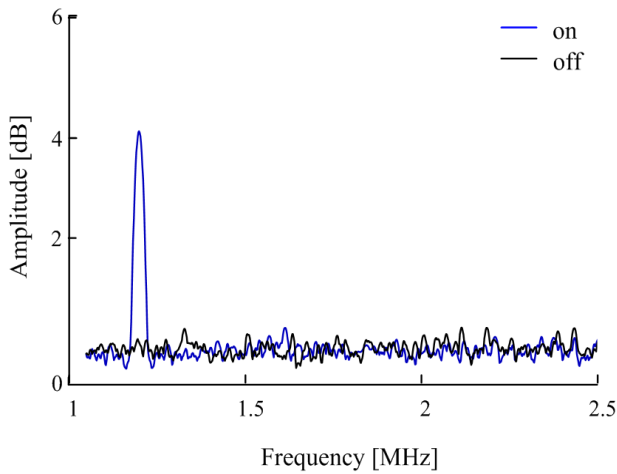


Fig. 8. An example of measured frequency spectrum.(ASUS Zenpad 3S 10 Z 500 M). “on” is in the program execution state. “off” is the state in which the program is not running.

## B. Time Domain

### 1) Experimental arrangement

The waveform in the time domain of the received signal with the same configuration was measured as in Fig. 6. The experiment was set up to analyze the time domain of the leaky waves, as show in Fig. 9. During the measurement, the oscilloscope was placed outside the radio anechoic chamber. The receiving antenna was an omnidirectional, vertically polarized mono-pole antenna. The DUT was Raspberry Pi3.

### 2) Results and analysis

Fig. 10 shows the waveform in the time domain of the received signal, and Fig. 11 shows the signal and the envelope of the received signal calculated from the measurement data. From the measured data, the received signal was performed the amplitude modulated. Also, the power of the envelope was distorted because of the noise from the DUT.

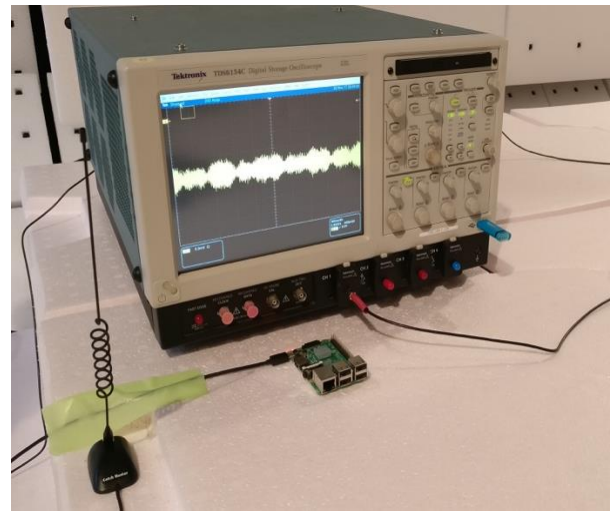


Fig. 9. The picture of measurement setup. (During measurement, the oscilloscope was placed outside the radio anechoic chamber.)

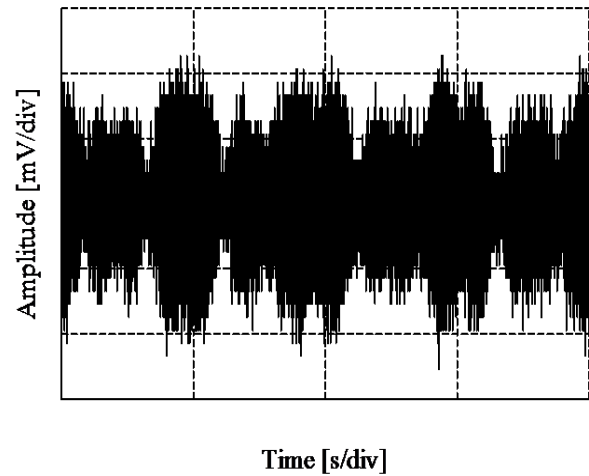


Fig. 10. An example of the waveform of the received signal (Raspberry pi 3).

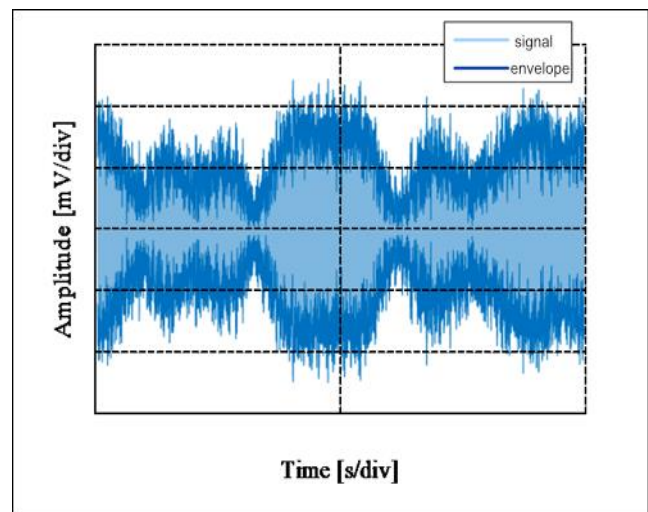


Fig. 11. The waveform and the envelope of the received signal (Raspberry Pi3).

TABLE I. A COMPARISON OF GSMEM AND PROPOSED ALGORITHM IN THIS PAPER

	Frequency	Platform	Memory usage	Languages
<i>GSMem</i>	GSM band (800MHz -)	Intel based PC	Low (~KB)	C++
<i>Proposed algorithm</i>	AM band (1.0 - 1.6MHz)	ARM and other CPUs	Medium (~MB)	C++, javaScript(asm.js)

## VI. DISCUSSION

The results shown in the previous section shows that the proposed algorithm is able to emit electromagnetic waves from the device (Raspberry Pi 3). However, since the ARM architecture is adopted for power saving devices, there is a possibility that the power of the emitted electromagnetic wave becomes smaller than GSMem.

Table I also shows a comparison between the proposed algorithm and GSMem. In GSMem, the frequency of electromagnetic waves emitted is the GSM frequency band, and the receiver is a modified mobile phone. In the proposed algorithm, the frequency of the emitted electromagnetic waves is AM frequency band, and ordinary AM radio can be used as the receiver.

The environment in which GSMem operates is limited to computers equipped with Intel CPU. CPU instructions used by GSMem are unique instructions not found in other architectures, and it is difficult to replace them. On the other hand, the proposed algorithm operates on a computer equipped with an ARM CPU. Furthermore, it can be used also on other platforms simply by replacing ARM specific instructions. This algorithm does not depend on the architecture, and it can be replaced with CPU instructions with memory access. The authors are doing work to operate this algorithm with a web browser.

The proposed algorithm requires assigning a memory greatly exceeding the capacity of the CPU cache. It involves more memory consumption than GSMem's algorithm.

## VII. CONCLUSION

This paper presented a new algorithm to exfiltrate data from ARM-based computers causing electromagnetic wave radiation, while Guri et al. presented the algorithm on Intel-based computers. Since the proposed algorithm accesses memory aggressively, the electromagnetic waves are radiated.

Frequency response of the electromagnetic wave is measured when the proposed algorithm is running with Raspberry Pi 3 and ASUS Zenpad 3S 10 Z500.

Guri et al. adopted an x86 CPU instruction that stores data in RAM explicitly bypassing the CPU cache. On the other hand, since SIMD of the ARM architecture has no instructions to bypass the CPU cache, the SIMD instruction (VLD 1.32) was adopted to load the data and proposed the algorithm that avoids CPU data cache hit by making data to be loaded different each time VLD 1.32 instruction is executed.

The measurement is concluded and it is confirmed that the program implementing the proposed algorithm is able to successfully radiate electromagnetic waves.

System administrator who manages ARM-based air-gap computers also needs to consider covert channels using electromagnetic radiation. Furthermore, the ARM computer is compatible with the mobile computer. Attackers will also be able to extract data without hacking the network.

Although GSMem used a specific SIMD instruction to bypass the CPU cache, this algorithm does not have to use such an instruction. The algorithm can also use generic instruction set instead of SIMD instruction. It has a disadvantage of requiring more memory than the capacity of the CPU cache, but it is effective in environments where the cache cannot be ignored. And I think that this algorithm can also be used in web browser and mobile OS, for example. It does not matter what kind of CPU these platforms are using. In future research, we will work on making programs applying the algorithm proposed in this paper work on web browsers and others.

## REFERENCES

- [1] R. J. Anderson, "Emission security," in Security Engineering, 2nd Edition, Wiley Publishing, Inc., 2008, pp. 523-546.
- [2] Rick Lehtinen, Howard Hecht, Deborah Russell, G. T. Gangemi, "Computer Security Basics: Computer Security", Oreilly & Associates Inc, pp.256, 2006.
- [3] W. van Eck, "Electromagnetic Radiation from Video Display Units: An Eavesdropping Risk? , " Computers and Security 4, pp. 269-286, 1985.
- [4] M. G. Kuhn and R. J. Anderson, "Soft tempest: Hidden data transmission using electromagnetic emanations," in Information Hiding, 1998, pp. 124--142.
- [5] M. G. Kuhn, "Compromising emanations: Eavesdropping risks of computer displays," University of Cambridge, Computer Laboratory, 2003.
- [6] G. Mordechai, G. Kedma, A. Kachlon and Y. Elovici, "AirHopper: Bridging the air-gap between isolated networks and mobile phones using radio frequencies," in Malicious and Unwanted Software: The Americas (MALWARE), 2014 9th International Conference on, IEEE, 2014, pp. 58-67.
- [7] M. Hanspach and M. Goetz, "On Covert Acoustical Mesh Networks in Air.," Journal of Communications, vol. 8, 2013.
- [8] Guri, M., Kachlon, A., Hasson, O., Kedma, G., Mirsky, Y. and Elovici, Y., (2015). GSMem: data exfiltration from air-gapped computers over GSM frequencies. In 24th USENIX Security Symposium, USENIX Security 15, pp. 849-864.
- [9] Sarah Harris, David Harris, "Digital Design and Computer Architecture: ARM Edition", Morgan Kaufmann, pp487-529, 2015.
- [10] Rajat Moona, "Assembly Language Programming in GNU/Linux for IA32 Architectures", Prentice-Hall of India Pvt.Ltd, pp.404, 2007.
- [11] Richard Gerber, "The Software Optimization Cookbook: High-Performance Recipes for the Intel Architecture (Engineer-To-Engineer)", Intel Press, pp.97, 2002.
- [12] JEDEC Solid State Technology Association, "JEDEC Standard: Low Power Double Data Rate 4 (LPDDR4)", JEDEC Standard JESD209-4, Aug 2014.
- [13] JEDEC Solid State Technology Association, "JEDEC Standard: DDR4 SDRAM", JEDEC Standard JESD79-4B, Nov 2013.
- [14] Bruce Smith, "ARM A32 Assembly Language: 32-Bit ARM, Neon, VFP, Thumb", Bruce Smith Books, 2017.

## APPENDIX

The proposed algorithm repeats trying to load four-word vector 8 words far from the previously loaded vector. It successfully generates electromagnetic radiation from memory bus between ARM CPU and RAM in the measurements. However, in order to estimate the efficiency that the algorithm avoids cache hit, this section presents an abstract data cache model and slightly modify the proposed algorithm into an inefficient version that randomly and independently chooses a position in the array and load four



words at the position each time. For simplicity, it is assumed that in mapping from virtual memory to physical memory, paging does not cause fragmentation of the array and page fault. Thus we identify an array in virtual memory with mapped contiguous data in physical memory.

This section considers an abstract data cache model and a modified algorithm as follows. The total amount of data cache is calculated in the form of (the cache line length) \* (the number of cache line) \* (the number of ways). Here let the cache line size be 32 words, the number of cache lines be 1024 and the number of ways be 2. Then the total amount of data cache is 16384 words. The modified algorithm allocates an array of the size three times as much as the total amount of data cache. For simplicity, it is also assumed that the data cache is used only for storing data in the array. Thus, each cache line stores only elements in the array. The algorithm randomly chooses an address of elements in the array and tries to load four-word data beginning at the address from the array.

If the four-word data does not include any array element stored in the data cache as illustrated in Fig. 12, the algorithm successfully loads the four-word data directly from memory through memory bus, without hitting cached data.

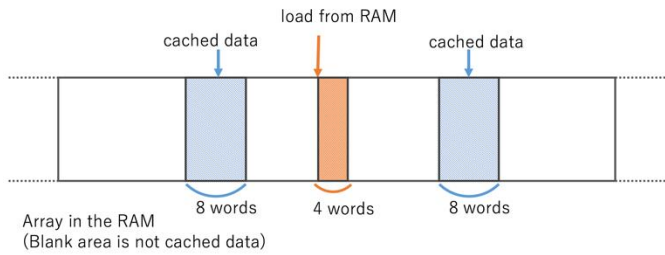


Fig. 12. An example of loading non-cached data from array in the RAM.

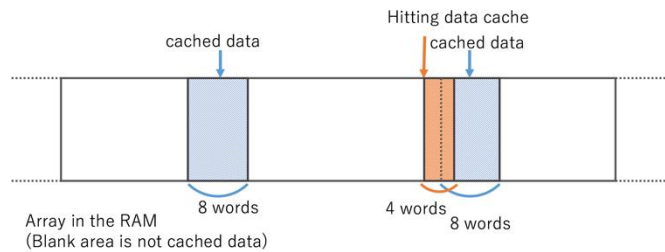


Fig. 13. An example of loading cached data from array in the RAM. Part of the data to be loaded is cached data.

If the four-word data at the randomly chosen address includes at least one array element stored in the data cache as illustrated in Fig. 13, this case is called "hitting data cache" since the algorithm may hit cached data and fail in loading four-word data directly from memory.

So, if the four-word area overlaps a part of the cached area, especially, if the chosen address is point to the position three words right shifted from the cached area, it is also the case of "hitting data cache". In the case that an eight word cached area and a four-word area that the algorithm tries to load overlap each other, the size of the combined area is at most 11 words. Since there exist at most 2048 cached areas of eight words exist in the array, there also may exist 2048 such combined areas of at most 11 words. If there are such combined areas at 2048 locations in the array, and the algorithm tries to access any element in these areas (2048\*11 words) in the array (49152 words), then the case of "hitting data cache" occurs.

Since the algorithm randomly chooses four-word area from the array, the probability that it can avoid the case of "hitting data cache" is (49152-22528)/49152. Consequently, the algorithm successfully causes electromagnetic radiation with the probability over 1/2.

Here, in order to estimate the efficiency of algorithm a random algorithm is used and a worst case such that whole data cache is occupied with data in the array is assumed.

However, in reality, since the operating system and device drivers may cause interrupts and many threads concurrently run, it is unlikely that the data cache holds only data from an array. Since electromagnetic waves are constantly radiated in the experiment, it is concluded that the algorithm rarely occurs cache hits in practice.

**Algorithm 3.** A random algorithm for ARM computers

```
1: p = (int32_t*)malloc(size)
2: for (int i=0; i<=n; i++)
3:   p[i] = i;
4:
5: data_bits[] = {1, 0, 1, 0,}
6: period = 500000
7: for data in data_bits :
8:   if (data == 1):
9:     i = 0
10:    start = now()
11:    while (period > now() - start):
12:      va = vld1q_s32( random(0 to n) ) //Random select
13:    if (data == 0):
14:      sleep(period)
```

# A Seamless Network Database Migration Tool for Insititutions in Zambia

Mutale Kasonde

Department of Electrical and Electronic Engineering  
University of Zambia  
Lusaka, Zambia

Simon Tembo

Department of Electrical and Electronic Engineering  
University of Zambia  
Lusaka, Zambia

**Abstract**—The objective of the research was to efficiently manage migration process between different Database Management Systems (DBMS) by automating the database migration process. The automation of the database migration process involved database cloning between different platforms, exchange of data between data center and different clients running non-identical DBMS and backing up the database in flexible format, such as eXtensible Markup Language (XML). This approach involved development of a “Database Migration Tool”. The tool was developed on a windows platform using Java Eclipse™ with four non-identical dummy Relational Databases (Microsoft Access, MySQL, SQL Server and Oracle). The tool was run in a controlled environment over the network and databases were successfully migrated from source to targeted destination option specified. The developed tool is more efficient, timely, as well as highly cost effective.

**Keywords**—Database management system; database migration; database structure; database migration toolkits and database cloning

## I. INTRODUCTION

Advancements in technology usually result in database migration, and an example would be for the National Pension Scheme Authority (NAPSA). A database (DB) is a persistent, logically coherent collection of inherently meaningful data, relevant to some aspects of the real world [1]. A collection of these databases is what forms the Database Management Systems (DBMS).

Database Management Systems perform a wide variety of roles such as allowing concurrency, controlling security,

maintaining data integrity, providing for backup and recovery, controlling redundancy, allowing data independence, providing a non-procedural query language as well as performing automatic query optimization.

Migrating a database involves migrating the tables and records from one database management system to another. The Transvive white paper, 2014 defines the term “Migration” as the movement of technology from older, or proprietary systems to newer, more versatile, feature-rich and cost-effective applications, and operating systems [2]. Data migration is usually undertaken for the purpose of replacing, upgrading server or storage equipment for a website consolidation, so as to conduct server maintenance or to relocate a data center.

However, because different database management systems have different formats for storing the database, the exchange of database tables and records between different database systems usually results in compromising the quality, or authenticity of the data in the transformation process. According to an Oracle White paper, 2011, up to 75% of new systems fail to meet expectations, often because of flaws in the database migration process, which in turn result in data that is not adequately validated for the intended task [3].

Some of the challenges associated with database migration processes include data loss particularly in a case of Poor Legacy Data Quality, having Wrong Data Migration Tools, inadequate knowledge in using the precise Data Migration Tools, failure to Test and validate Data Migration Process, and Absence of Data Governance Policies [4].

Currently, there is no comprehensive system to compartment the data migration process. The existing procedure for migration is semi manual, and involves fragmentary procedures, which entails using different tools in order to achieve a comprehensive process. As such, maintaining the structure of a database when migrating it from one database management system to another is quite a challenging task for most Organizations. Often times, organizations have to design a new database for the different database management system it wants to adopt. This current system of database migration is not only costly, but it's also ineffective, and may sometimes result in the loss of essential data in the migration process. This is because this system involves hiring a database designer every time the Organization has to switch to a different database management system. Therefore, by developing an automated database migration process, the challenges being experienced with the current migration process required to be addressed and eliminated.

This study intended to explore ways in which data migration process could be improved through the development of a new Seamless Database Migrator. This was expected to help overcome challenges associated with the network database, and deliver the data with such accuracy. Specifically, the new database migrator was expected to.

1) Eliminate the need for script writing when transferring tables in the database with the records.

2) Prompt the user to select the destination and source Database Management System, and specify what to migrate i.e. either the entire database with structure and data or just structure or data, thereby allowing different database management systems to exchange database tables and records in them, without any loss of data details in the transformation process.

The paper is organized as follows: Section 2 deals with literature review, which covers existing tools as well as the theoretical literature bordering on policy issues regarding database migration. The methodology is presented in Section 3. Section 4 brings out the results and the discussion of the baseline study conducted to identify challenges in the database migration process. System testing is presented in Section 5, and the last Section 6 contains the conclusion.

## II. LITERATURE REVIEW

In order to cope with a fast changing business environment, it is necessary to update the technological infrastructure constantly, and database migration is a routine part of this technology. Barron. C et al. suggested that the core reason for the need of database migration is mainly to upgrade the existing system into a developed system that conforms to the Industry requirements [5].

The tasks of a migration workflow are diverse and complicated, executing all these processes manually requires plenty of time, as well as a highly experienced migration team in both the source, as well as the target system. In a paper, reviewing database migration strategies, tools and techniques, Elamparithi, M and Anuratha, V singled out relational database migration (RDM) as an example. The authors stated that relational database migration was always a complex, time consuming, and magnified process due to heterogeneous structures and several data types of relational database [6]. This gives rise to certain risks and challenges in the data migration process.

The Arbutus Software Whitepaper summarized the risks of database migration as follows: Unrealistic estimates of data quality, inaccurate, missing or out of date source system documentation, as well as the inability to reconcile the target systems data to the source system [7].

To counter the above challenges of database migration, businesses have seen the need to develop effective methodologies of migrating databases. Several migration tools and strategies have since been developed in the software industry [8]. However, finding the effective methodologies for database migration still remains a challenge, and many current approaches to data migration suffer from a consistently low success rate. Arbutus Software White Paper approximates that between 70 and 90% of data migration projects either fail outrightly, or run over budget, with an average cost overrun of 10 times the original estimate [9]. This is mainly due to the unplanned issues that often occur at the later stages of a project.

Several researches have since been undertaken to address some of the problems associated with database migration, although no absolute solution has come forth. For example, Joseph R. Hudicka, provided a complete solution of data

migration methodology, which deals with row counts, column counts and related statistics to the source databases [10]. However, the problem with this methodology is that it does not migrate null and numeric values and error occurs for key constrain keys.

TABLE I. DEVELOPED DATA MIGRATION TOOL [15]

S. No	Name	Company	Source	From	To	Operating System
1	OSDM Toolkit	Apptility	Open	Oracle, SyBase, Informix, DB2, MS Access, MS SQL	PostgreSQL & MYSQL	Windows, Linux, Unix & Mac OS
2	DB Migration	Akcess	Closed	Oracle & MS SQL	PostgreSQL & MYSQL	Windows
3	Mssql2 Pgsq	OS Project	Open	MS SQL	PostgreSQL	Windows
4	MySQL Migration Toolkit	MySql AB	Open	MS Access & Oracle	MySQL	Windows
5	MySQL Migration Toolkit	Intelligent Convertors	Closed	MS Access, MS SQL, Dbase & Oracle	MySQL	Windows
6	Open DBcopy	Puzzle ITC	Open	Any RDB*	Any RDB*	OS Independent
7	Progression DB	Versora	Open	MS SQL	PostgreSQL, MySQL & Ingres	Linux & Windows
8	Shift2Ingres	OS Project	Open	Oracle & DB2	Ingres	OS Independent
9	SQLPorter	Real Soft Studio	Closed	Oracle, MS SQL, DB2 & Sybase	MySQL	Linux, Mac OS & Windows
10	SQLWays	Ispirer	Closed	All Relational Databases	PostgreSQL & MySQL	Windows
11	SwisSQL Data Migration Tool	AdventNet	Closed	Oracle, DB2, MS SQL, Sybase & MaxDB	MySQL	Windows
12	SwisSQL SQLOne Console	AdventNet	Closed	Oracle, MSSQL, DB2, Informix & Sybase	PostgreSQL & MySQL	Windows
13	MapForce	Altova	Closed	SQL Server, DB2, MS Access, MySQL & PostgreSQL	SQL Server, DB2, MS Access & Oracle	Windows, Linux & Mac OS
14	Centerprise Data Integrator	Astera	Closed	SQL Server, DB2, MS Access, MySQL & PostgreSQL	SQL Server, DB2, MS Access, MySQL & PostgreSQL	Windows
15	DBConvert	DB Convert	Closed	Oracle, DB2, SQLite, MySQL, PostgreSQL, MS Access & Foxpro	Oracle, DB2, SQLite, MySQL, PostgreSQL, MS Access & Foxpro	Windows

Ramaswamy, V.K. argued that effective and efficient migration of data is one of the cornerstones for the success of the process [11]. He further emphasized the fact that significant planning needed to be done before the actual process of data migration commences. He outlined a strategy for data migration in which he listed down the type of data to be migrated, timing of the data load, templates and tools for use in the migration process.

Sait S.A. et al. on behalf of Amazon Web Services (AWS) provided comprehensive strategies for migrating Oracle databases in which they stated that there is no absolute formula for migrating databases but that there are certain factors one needed to put into consideration before undertaking database migration [12]. These factors include the size of the database, the network connectivity between the source server and the target service, the version and edition of the oracle database software, the database options, tools and utilities that are available, as well as the time available for the migration process. Based on the above factors, the authors divided the migration process into two methods namely the One Step migration, which is ideal for small databases, and the two-step migration, which can be used for any size of the database.

Currently, a number of prototypes and tools have been developed to facilitate the migration of relational databases (RDBs) into target databases. Senior researchers Bin Wei, and Tennyson X. Chen, developed Data Migration Tool (DMT) for US National Oceanic and Atmospheric Administration (NOAA), outlining the criteria that need to be considered when evaluating a DMT [13]. However, while the criteria outlined by these authors may be adequate for the complex project of developing DMT, the complexity of a general extract, transform, and load (ETL) system may go beyond what these criteria can evaluate. Still the investigations are needed on dealing with complex files [14].

Jutta Hortsmann, J. suggested some examples of database migration tools as shown in Table I.

The data migration process under goes through several stages, which include planning, designing, cleansing, loading as well as verifying of the data. Fig. 1 shows the Data migration process.

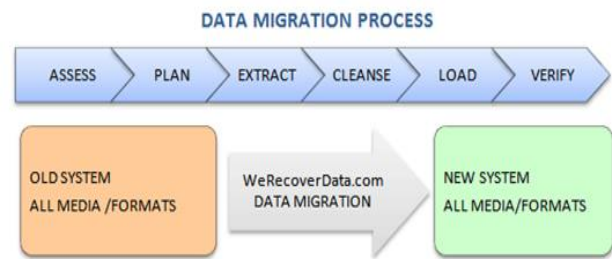


Fig. 1. Data migration process [16].

### III. METHODOLOGY

The method of analysis for this project was divided into two categories, guided mainly by the objectives of the study.

The first method of analysis involved getting expert opinion from IT personnel that were purposefully selected from the Applications and Database Administration sections of three institutions namely NAPSA, ZANACO and ZAMTEL. These IT experts were meant to capture as much qualitative data as possible regarding the existing database migration procedures, as well as identifying the challenges associated with it.

The second method of analysis involved designing the modules that made up the system. The overall purpose of this system design was to provide efficiency and effectiveness in data migration process. To achieve this process, the Java Programming Language and Database Management Systems technologies were used.

This approach involved developing the tool “Database Migrator Tool”. The tool was developed on a windows platform using Java Eclipse with four non-identical dummy databases (Microsoft Access, MySQL, SQL Server and Oracle). The automation of the database migration process involved database cloning between different platforms, exchange of data between data center and different clients, running non-identical DBMS and backing up the database in flexible format such as eXtensible Markup Language (XML). The tool was run in a controlled environment over the network. The following java support tools were used to support the technologies used in the system:

1) *MySQL connector (mysqlconnector.jar)*; which was used to connect MySQL database management system from java.

- 2) *SQL server (sqljdbc.jar)*; which was used to connect SQL server database management system from java, and
- 3) *Jackcess (jackcess.jar)*; which was used to connect SQL server database management system from java.
- 4) *Oracle connector (OJDB5.jar)*; this java library was used to connect Oracle database Management system from Java.

The Database Management Systems used included:

- 1) *Microsoft SQL Server* is a database management system whose primary function in this case was to store the database in SQL server format (.mdf) and retrieve data as requested by other software applications.
- 2) *MySQL* database management system was used to store the database in MySQL format (.frm). This database management system stored data in separate tables whose structures were organized into physical files.
- 3) *Microsoft Access Database Management System* combines the relational Microsoft Jet Database Engine with a graphical user interface and software-development tools. It is a member of the Microsoft Office suite of applications, included in the Professional and higher editions. Microsoft Access was used to store database data in its own format (.accdb) based on the Access Jet Database Engine.
- 4) *Oracle* database Management system was used to store and retrieve related information. Oracle database management system was used to store database data in its Oracle database format (.dbf)

IV. RESULTS AND IMPLEMENTATION OF NEW SYSTEM

This segment presents the results obtained from the baseline study as well as development and testing of system prototype. In order to confidently and significantly address the challenges associated with the old system, baseline study was conducted and proposed prototype application was developed.

The data collected from the baseline study was analyzed using descriptive statistics and the results were presented in form of charts. From the responses obtained from the respondents, 58% admitted that the tools that were currently being used for database migration were wrong tools. 42% said the tools were not wrong per ser. However, 71% of the responses indicated that the old system had no provision to test and validate the data migration process. Additionally, 87% had experienced data loss in the process of migrating the database when using the old system. 93% of the respondents admitted that because of the technical challenges associated with the old database migration process, such as constant data loss, failure to test and validate the migration process, engaging a consultancy every time the need arises, etc., the old system was

too expensive to manage. 88% further stated that there was inadequate knowledge among the users on the precise data migration tools to use. 55% of the respondents also stated that there were no existing data governance policies, a situation that made it difficult to manage this system. Fig. 2 shows the summary of the results from the baseline study.

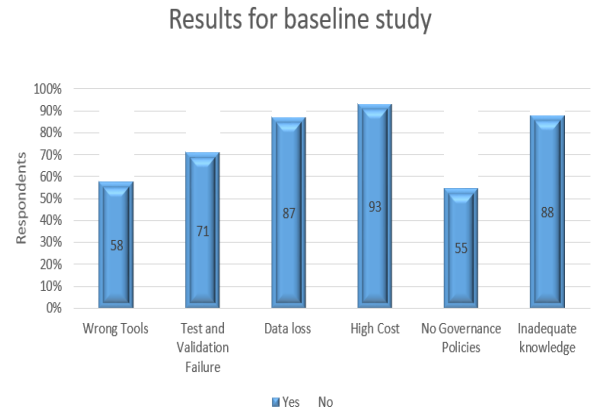


Fig. 2. Baseline study results.

A. The Architectural Design

This Database migrator was developed using a Top-Down approach. This approach involved decomposing the system into individual smaller modules, aimed at achieving the required detail. Fig. 3 shows the Internal Logic Design.

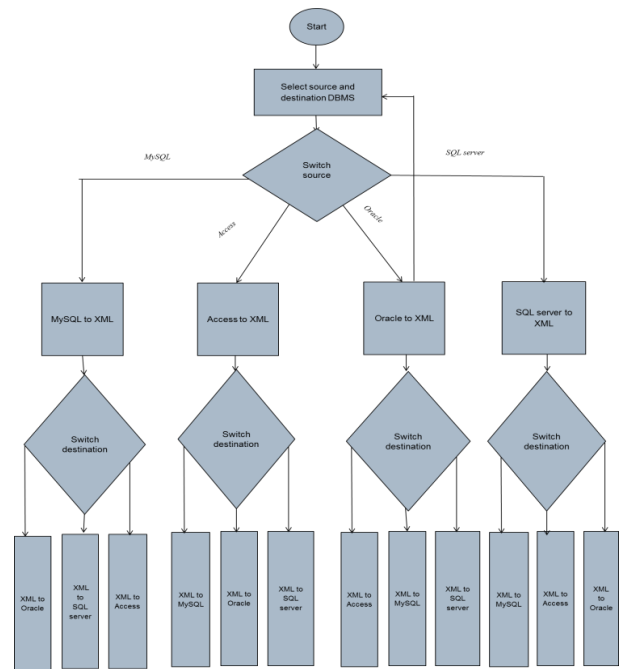


Fig. 3. Internal logic design.

**B. Conversion Modules**

A prototype was developed and a database migrator was effectively implemented using the technologies elaborated in the project methodology. Fig. 4 below shows the conversion module from XML to Oracle.

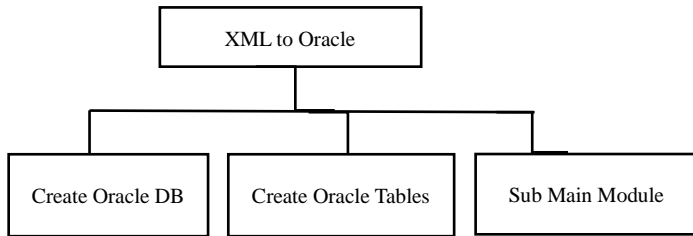


Fig. 4. XML to Oracle.

**C. The Interface Design**

Database migrator is a desktop application and the interface enables the user to use the system without any difficulties; Java Graphical User Interface (GUI) was used to design and develop the interface. The interface design comprises of the main interface, sub interface and the graphical interface.

**D. The Main Interface**

The main interface is the home interface for the system and appears when the system runs. It consists of buttons used for running the migration process where the user selects the source and destination Database management system. Fig. 5 show the Main Interface

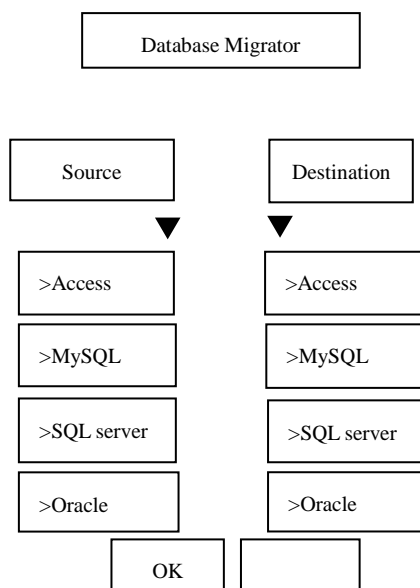


Fig. 5. The Main Interface.

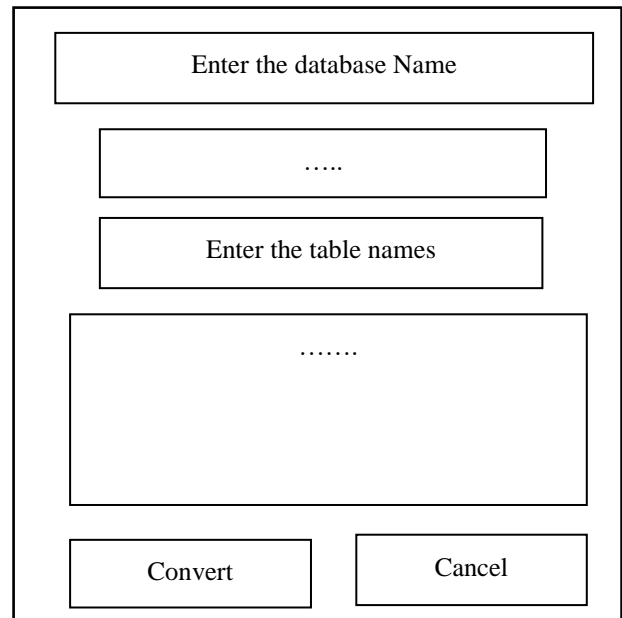


Fig. 6. Sub module.

**E. Sub Interface**

The sub interface is used for the actual conversion from one database system to the other. Upon specifying the details of the source database, the sub interface converts the source database system into the destination database system. In case of an error, the sub interface has a provision for cancelling the process.

All the four sub modules were converted to XML file, which in turn acts as a common ground for converting the database from one system to another. Once the design specification and project design was approved, system coding commenced. Fig. 6 shows the sub module.

**F. Graphical User Interface**

The graphical user interface allows the user to interact with the system; it was developed using Eclipse java (jdk 1.6.0). The user interface provided the user with a point and click interface which reduced user's errors because the user was not prompted to enter any information.

**G. The Migration Process**

The process involves selection of the source and destination Database Management System using the main interface by the system users. The sub interface then converts from one system to the other. Fig. 7 shows the source and destination interface.

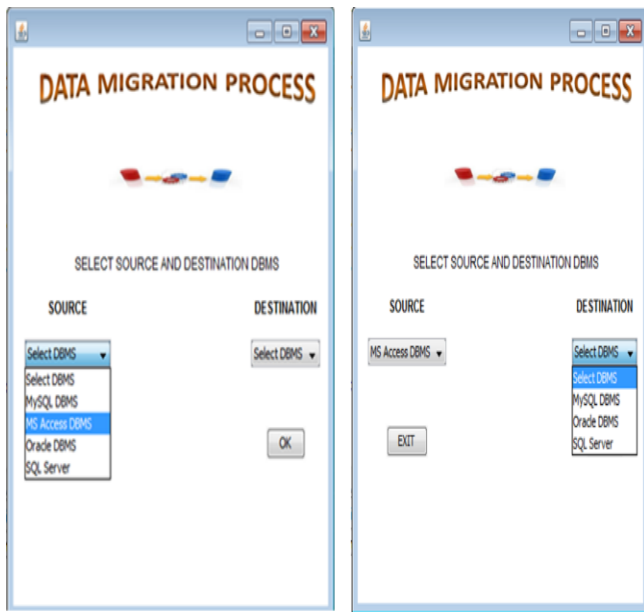


Fig. 7. Source and destination of DM.

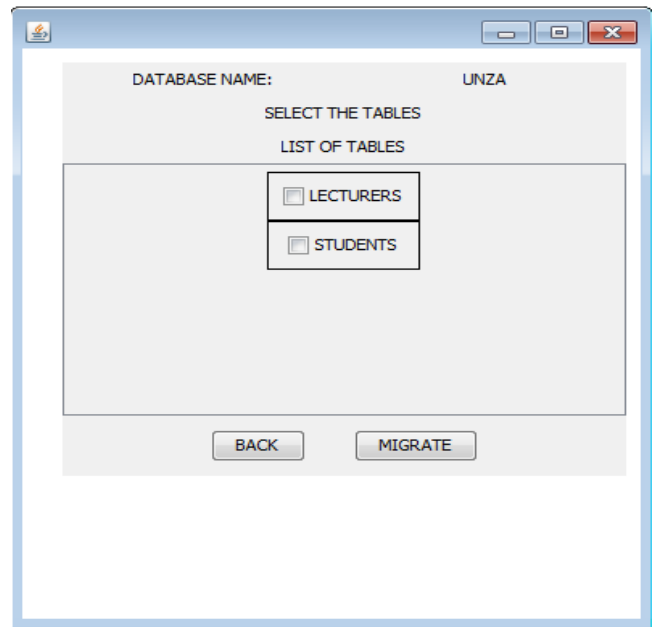


Fig. 9. Tables selected for migration process.

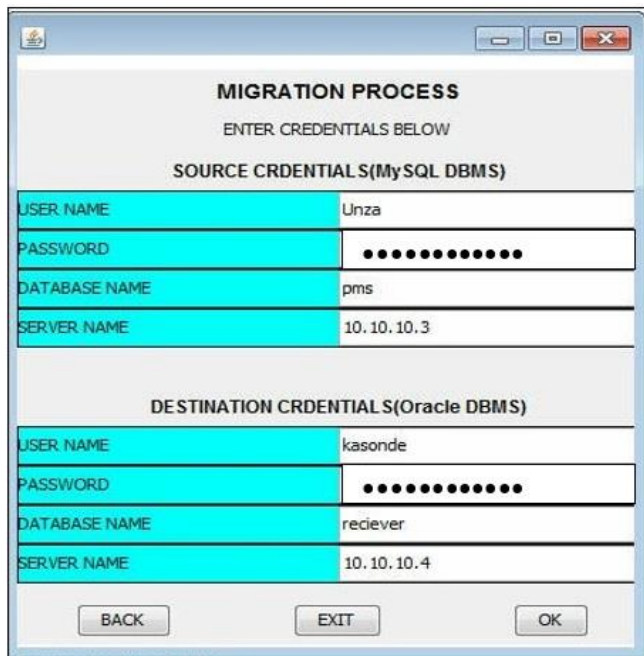


Fig. 8. Remote database credentials.

Once the user selects the source and destination databases, the tool prompts the user to specify credentials for the source and destination databases as shown in Fig. 8.

Upon authenticating the source database credentials, the tool prompts the users with options to specify what needs to be migrated, i.e. data, table or entire database. Fig. 9 shows the tables from source database.

TABLE II. TESTING OF CLASSES FROM XML FILE TO DATABASE

	Testing Scenarios	Expected Results	Actual Results
1	XML file to Microsoft Access	Access DB	XML file was converted to Access DB
2	XML file to MySQL	MySQL DB	XML file was converted to MySQL DB
3	XML file to SQL Sever	SQL Server DB	XML file was converted to SQL Server DB
4	XML file to Oracle	Oracle DB	XML file was converted to Oracle DB

## V. SOFTWARE TESTING

To ensure that the new system was working according to the intended objective, system testing was done both for the individual units as well as the integrated system. Unit testing was carried out by testing all the individual modules separately with connections to the database to see if module



performed all the expected functions. System testing was also carried out to determine how well the various components that comprise the system would interact in order to achieve the total system functionality. The main objective of testing was to detect and eliminate errors, as well as to check on whether it meets the requirements set out in the requirement specification document.

#### A. Integration Testing

After testing individual modules, integration testing was carried out and this involved putting the modules together and testing them to ensure that the object classes still work as expected even after being combined with other classes.

#### B. Testing Output

The classes for converting from database to XML file were tested separately to ensure that the selected source database management system was converted to XML. Table II shows testing of classes from Database to XML file whereas Table III shows testing of classes from XML file to Database.

TABLE III. TESTING OF CLASSES FROM DATABASE TO XML FILE

	Testing Scenarios	Expected Results	Actual Results
1	Microsoft Access to XML file	XML file	MS Database was converted to XML file
2	MySQL to XML file	XML file	MySQL Database was converted to XML file
3	SQL to XML file	XML file	SQL Database was converted to XML file
4	Oracle to XML file	XML file	Oracle Database was converted to XML file

The classes for converting from an XML file to a database were also tested separately to ensure that the xml files are converted to specified database management system. Table IV shows the system module testing.

TABLE IV. SYSTEM TESTING OF MODULES

	Testing Scenarios	Expected Results	Actual Results
1	Migrating from MS Access to MySQL.	MySQL DB	MS Access DB was successfully converted to MySQL DB
2	Migrating from MS Access to SQL Server.	SQL Server DB	MS Access DB was successfully converted to SQL Server DB
3	Migrating from MS Access to Oracle	Oracle DB	MS Access DB was successfully converted to Oracle DB
4	Migrating from MySQL to MS Access	SQL Server DB	MySQL DB was successfully converted to MS Access DB
5	Migrating from MySQL Server to SQL Server.	Access DB	MySQL Server DB was successfully converted to SQL Server DB
6	Migrating from MySQL Server to Oracle	Oracle DB	MySQL Server DB was successfully converted to Oracle DB
7	Migrating from SQL Server to MS Access.	Access DB	SQL Server DB was successfully converted to MS Access DB
8	Migrating from SQL Server to MySQL.	MySQL DB	SQL Server DB was successfully converted to MySQL DB
9	Migrating from SQL Server to Oracle.	Oracle DB	SQL Server DB was successfully converted to Oracle DB
10	Migrating from Oracle to MS Access	Access DB	Oracle DB was successfully converted to MS Access DB
11	Migrating from Oracle to MySQL Server	MySQL DB	Oracle DB was successfully converted to MySQL DB
12	Migrating from Oracle to SQL Server	SQL Server DB	Oracle DB was successfully converted to SQL Server DB

The system did the conversion successfully and the expected results were obtained. According to the minimum requirement of the system, the system performance was found to be fair. However, one important lesson learnt from this testing process was that modules which performed well when tested independently, do not always perform well when integrated with other modules. See Fig. 3, internal design.

In order to build a thorough contextual understanding of the issues at play, this project undertook an assessment of the old database migrator and challenges associated with it. Consequently, the project developed a new system that addresses the problems being experienced by the current database migration systems. The process was accomplished through planning and designing of a new system, cleansing, loading as well as verifying the new database migrator.

The new system does not support databases in Real Application Cluster (RAC) but only works for a single Database Node.

## VI. CONCLUSION AND RECOMMENDATION

The project objectives were achieved through the development of an automated database migrator. The results indicated that the new system was operating efficiently and effectively. It was therefore recommended that the new tool be adopted. The Seamless database migrator addresses the numerous challenges that are associated with the Database migration process. The implementation of the Seamless Database Migrator ensured that there was efficiency and effectiveness in the application and use of the database management system. With the use of a new DM tool, the challenges that were being experienced with the current migration process were addressed and eliminated. There was no data loss; no cost of hiring database designer to remodel

the new Database (DB), time taken to migrate was reduced tremendously.

## REFERENCES

- [1] Robbins, J.R, (1995, May), "Database Fundamentals" [Online]. pp-2. Available <http://www.esp.org/db-fund.pdf> [May. 16, 2017].
- [2] Transvive white paper (2014, Apr)"Migration Strategies & Methodologies" pp 4. Available <https://www.platformmodernization.org/transvive/Lists/ResearchPapers/Attachments/1/Transvive-MainframeMigrationStrategy-WP.pdf> [May. 16, 2017].
- [3] An Oracle White Paper. (2011,Oct.) "Successful Data Migration "[Online]. Pp.2-3. Available <http://www.ijssst.info/info/IEEE-Citation-StyleGuide.pdf> [May. 17, 2017].
- [4] An Oracle White Paper. (2011,Oct.) "Successful Data Migration "[Online]. Pp.4-11. Available <http://www.ijssst.info/info/IEEE-Citation-StyleGuide.pdf> [May. 17, 2017].
- [5] Barron C. Housel, Vincent Y. Lum, Nan Shu (1974), 'Architecture to an Interactive Migration System' in SIGFIDET 1974: Proceedings of the 1974 ACM SIGFIDET (now SIGMOD) workshop on Data description, Access and Control, New York, USA, pp. 157-169.
- [6] Elamarithi M, and Anuratha. V. "A Review on Database Migration Strategies, Techniques and Tools." World Journal of Computer Application and Technology 3 (3): 41/48, 2015.
- [7] Arbutus white paper. (2016,Oct.) "Top Three Data Migration risks "[Online]. Pp.2. Available <https://cdn2.hubspot.net/hubfs/3336879/ArbutusSoftware-June2017/Pdfs/arbutus-wp-data-migration.pdf?t=1503679160981> [May. 16, 2017].
- [8] Paul Dorsey, Joseph R. Hudicka) "Oracle Data Migration Handbook." McGraw-Hill Osborne Media, pp 23-27, 2000.
- [9] Arbutus white paper. (2016,Oct.) "Top Three Data Migration risks "[Online]. Pp.3-5. Available <https://cdn2.hubspot.net/hubfs/3336879/ArbutusSoftware-June2017/Pdfs/arbutus-wp-data-migration.pdf?t=1503679160981> [May. 16, 2017].
- [10] Joseph R. Hudicka , 'Oracle8 Database Design Using Uml Object Modeling' McGraw-Hill Professional , ISBN-13:9780078824746, pp 216-224, 1998.
- [11] Ramaswamy, V.K, Database Migration Strategy in ERP, London: United Kingdom, 2016
- [12] Sait A, S, Strategies for migrating Oracle Databases to AWS, Amazon Web Services, Inc: Amazon2014.
- [13] Bin Wei and Tennyson X. Chen (2012), 'Criterial for Evaluating General Database Migration Tools', [online] Research Report, RTI Press Publication No. OP-0009-1210. Research Triangle Park, NC:RTI Press. <http://www.rti.org/pubs/op-0009-1210-chen.pdf> (Accessed 17 July 2017
- [14] S. M. Abdelsalam Amaraga Maatuk, Migrating Relational Databases into object-based and XML databases., Published PhD thesis, Nortambria University, Newcastle, United Kingdom, 2009
- [15] Horstmann, J, Migration to Open Source Databases, Technical University, Berlin:Germany, 2005
- [16] Data recovery labs, "Data Migration Process" Internet: [www.werecoverdata.com](http://www.werecoverdata.com) [Mar. 10 2017]

# Software Engineering: Challenges and their Solution in Mobile App Development

Naila Kousar, Muhammad Sheraz  
Arshad Malik  
Department of Information  
Technology  
Government College University  
Faisalabad, Pakistan

Aramghan Sarwar  
Department of Information  
Technology  
Superior University  
Lahore, Pakistan

Burhan Mohy-ud-din, Ayesha  
Shahid  
Department of Software Engineering  
Government College University  
Faisalabad, Pakistan

**Abstract**—Mobile app development is increasing rapidly due to the popularity of smartphones. With billions of apps downloads, the Apple App Store and Google Play Store succeeded to overcome mobile devices. Throughout last 10 years, the amount of smartphones and mobile applications has been perpetually growing. Android and iOS are two mobile platforms that cowl most smartphones within the world in 2017. However, this success challenges app developers to publish high-quality apps to stay attracting and satisfying end-users. Developing a mobile app involves first to select the platforms the app can run, so to develop specific solutions (i.e., native apps). During application development a developer come across multiple challenges. In this paper, we have tried to find out challenges faced by developer during their development life cycle with their possible solution.

**Keywords**—Android; IOS; mobile apps; software quality; survey research; user requirements

## I. INTRODUCTION

The popularity of smart phone has gain the attention of developers. Smartphones are mobile devices that run software applications such as games, social network and banking apps. There are almost 4.77 billion mobile phone users in 2017. Mobile applications are software applications developed for use on mobile devices [9]. Total number of apps worldwide are hard to come by. There is no statistics on web apps – exist too many ways to develop and get them, so it's almost incredible to count these apps. But we can try to count mobile apps, which are represented on App Store (for apps that work on Apple devices) and Google Play (for apps that work on Android devices). Recent estimations indicate that by 2017 says that there are almost 2,800,000 apps on Google Play, 2.200,000 on Apple Store, 669,000 on window Store, 600,000 on Amazon Store and 234,500 on BlackBerry World [1].

However, programming languages and tools for developing mobile apps are platform-specific, like, Android applications are created in Java via the Android studio, whereas Apple iOS applications are developed either using Objective-C or Swift via the XCode tool. [7]As a result, to develop and maintain native apps for multiple platforms is a biggest challenge poignant the mobile development community.

Most of the work in the contest of mobile effort estimation has been concerned with the study of the issue that developer

face not on their solution. But in this paper we are also providing with proper solution.

## II. STUDY DESIGN

### A. Online Quotes Analysis

During the first phase, we found the online quotes made available by companies on the web, with the purpose of extracting an initial set of issue with solution. The context of the study consisted of every company having a website and providing an online form for requesting a quote about the development of a mobile apps. We used an automatic search tool, named GOOGLE-SCRAPER1, which is publicly available and open source.

### B. Survey with Experts

The goal of this step of the study was to conduct an interview and semi-structured survey by experts having a good knowledge of mobile apps development. The purpose was to exploit the involved experts in order to identify issues and their possible solution. The context of the study was composed by 20 developers with more than 4 years of experience in mobile development and effort estimation.

The selection of the types of participants involved in the study was not random. In fact, the selected project managers are responsible for leading the projects in their companies, in addition to managing the people, resources and the effort needed to complete the project. Some of them work for large companies, while the other work in local companies. The goal is bringing together the opinions of the participants and providing a joint solution.

Based on the idea rising from the interview part, we tend to design a semi-structured survey, as another supply of knowledge. Before publishing the survey, we tend to asked 3 mobile app developers to review the survey, so if there need any improvement or not.

### C. Participant Demographics

The participants involve in the interview and survey were 20 in number from different countries and companies .Some of them are IOS developer and remaining were android developer. We interviewed 10 developers from different companies and from remaining 10 we filled out the survey due to location issue. During each interview we write down their answer for

analysis. Table I represent each interviewer role with the mobile platforms in which they have expertise in and their work experience in mobile development.

Our survey was absolutely completed by 10 respondents. We conduct this survey on Aug 3, 2017 to a mobile development groups. They respondents belong to different countries and different age groups.

TABLE I. INTERVIEW PARTICIPANTS

ID	Role	Platform Experience	Location	Company	Dev Experience (Year)
P1	Android Developer	Android	Islamabad	Telenor Pakistan	5
P2	Android Developer	Android	Rawalpindi	iBrandify	4-5
P3	Android Team Lead	Android	Faisalabad		6
P4	IOS Team Lead	IOS	Islamabad	Mob Inspire (PVT) Ltd.	5
P5	Software Engineer	Android	Islamabad	Gazuntite	3-4
P6	Android + IOS Team Lead	Android, IOS, .Net	Faisalabad	Ingenious	5-6
P7	Sr. Android Developer	Android	Islamabad	Broadpeak Technologies	3-4
P8	Principal Software Engineer	IOS	Islamabad	TEO	4-5
P9	Mobile Developer	Xamarin, Phonegap And IOS	Dubai	App Emirates	5
P10	IOS Developer	IOS	Islamabad f- 10 market	SoftTech private Limited	3

### III. PROPOSED EXPERIMENTS

Our experiments included the following points:

- A. Creating Universal User Interfaces/ standard for GUI designing of app development
- B. Cross platform development issue
- C. Unclear/frequent changing requirements Issue
- D. Testing effects on development
- E. Technique used for supporting new API features in old API
- F. Maintenance

#### A. Creating Universal User Interfaces/ Standard for GUI Designing of App Development

Some research is already being done for creating a universal UI for mobile devices. Every mobile platform provide a distinct way for developer to address UI requirements. A noteworthy idea for mobile User Interface development identifies with screen size and its resolution for example Apple gadgets size are restricted based on the size of the iPhone and the iPad while Android give screens of dynamic sizes and resolutions.

Thus, UI design is tough and mobile application developers tough anticipate the targeted device [2]. Shneiderman's "8 [3] Golden Rules of Interface Design" are well received since their introduction [5]. However, these rules might not equally apply to mobile devices. Research by Gong and Tarasewich recommend that four of Shneiderman's tips promptly translate to mobile devices, including: enabling frequent users to use shortcuts, providing informative feedback, designing dialogs to yield closure, and supporting internal locus of control.

From interview and survey the developer provides the solution of designing Universal GUI. The IOS developers say that they follow Apple UI guidelines which involves constraints, size classes and ratio to support UI for all apple mobile devices. One IOS developer says there are different ways to designs your GUI in IOS i.e. XIB, MVC. But I mostly follow storyboard, because in storyboard there is easy to handle all views and they provide us same environment like device.

On the other hand Android developers says for GUI standard they are using Material Design. One said that I follow Material Design in most of the apps as it is recommended by Google as it gives better user experience. So if all developer follow apple guidelines and material design the can develop universal GUI for IOS and Android. Units

#### B. Cross Platform Development Issue

A current challenge for mobile developers is to decide which platform they have to choose for his or her mobile applications. To target large number of users, companies try to develop their app in all platforms i.e. in IOS and Android [4], [5]. The aim to target multiple platform is to target more user so that company can gain more profit and also increase its impact on the market. So Companies afford the charges of developer for developing cross platform app .It's also time consuming task. So company should hire expert developer for each platform.

A cross-platform mobile app development frame-works is Xamarin. It develop app for both Android and iOS using C#. Developers reuse their existing C# code, and share significant code across device platforms [10]. Xamarin or React-Native are cross platform Frameworks maintained by Microsoft and Facebook. In like manner, the home platform for Apportable is iOS. Developers build apps by utilizing Objective-C and the iOS SDK, and use Apportable [4].

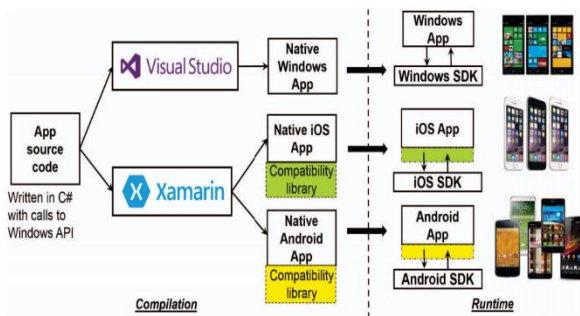


Fig. 1. Operation of a cross-platform mobile app development framework, using Xamarin[8].

In Fig. 1, overall operation of a cross-platform mobile app development framework, using Xamarin as a concrete example is shown. Developers build apps as they would for the Windows Phone, in C# using calls to the API of the Windows Phone SDK. This code can directly be compiled to Windows Phone apps using the Visual Studio toolchain. Xamarin allows developers to use the same code to build Android or iOS applications. Xamarin provides compatibility libraries that translate Windows SDK API calls in the code to the relevant API calls of the underlying Android and iOS SDKs.

Developers of both IOS and Android says cross platform development is a bigger issue. No doubt Xamarin and React-Native are available for cross platform development are available but their libraries and proper guidelines support are not available. So if we stuck in some point we face a lot of problem that cross platform technologies do not have much help on the internet and they have limited access to the mobiles functionality. Generally they have compatibility issue, memory leakage suitable for small apps.

*We ask why software houses not develop apps on cross platform tools ?*

They told us Cross platform developers are not found easily with good expertise. Moreover follow cons exist in cross platform like platform limitations, User Experience, Integration challenges so companies are afraid to try new things in terms that they would earn from that in long run or not so they avoid cross platforms.

One said us that native apps have large community and easy to develop other than cross platform. Cross Development requires additional time and effort in order to mimic the native look and feel. Then

*We ask developers why they not learn and adopt new development platforms like Xamarin?*

They said Some developers want to learn new things and some are not, but if their respective companies encourages and gives time to learn new technologies so they should prefer to learn new things, because time required for learning new technologies and due to burden of pending task project developer have lack of time.

One Said learning new tools depends on the requirement and work required from a developer. A good developer can learn any tool if it's required from him/her with some time but

Xamarin is highly priced and there is not much help available online. But they also said with the passage of time these technologies will mostly use for development. One developer also said I don't think cross platform development should be an issue. It's all about basic concepts. Object Oriented Concepts remains the same no matter what language (OOP based) we use. I have tried my hand on Unity 3D engine as well and so far, I don't seem to encounter any problem.

### C. Unclear/Frequent Changing Requirements Issue

In Software Application Development if the user change its requirements in the initial stage then it's still alright to adopt the new modification in requirements. If the requirements are changing frequently however the changes are communicated well ahead and enough time is given to testing the application then it is not a problem at all. However, if the requirement change in later stage of SDLC then cost to fix is incredibly high [6].

When we ask the same question from developers: Do unclear/frequent changing requirements affect your development? They said obviously frequent changing in requirements effects the development time and the quality too because when you start developing the app and then the requirement changes you have to reconsider everything as a developer and sometimes it needs to start from the very beginning. They said it is one of the major hurdle in a developer's task. Requirements are core for any development. If requirements are not clear or vague, it affects badly.

One says For example, if I am developing the app with particular requirements and during development the client changes his mind then I have to change the previously developed classes or modules it may be big change or small, it depends on new change. Unclear changes also waste the time of the both parties. Some developers also provide the solution of this issue they said Agile processes have slots to cope with these situations.

### D. Testing Effects on Development

Testing your code is very important. Testing aim is to find more error to make your app more effect. Testing improve the quality of your project/App [5]. Basically small scale organization don't focus too much on testing, Developer is also a tester means developer test the app which they build, as we know it's difficult to find our mistakes so a developer can't test its own app. One developer during interview said that their organization has little concept of testing. Some organization test application only on emulator which has limited features, lack of mobility, location services, sensors, or different gestures, so when application is launch its failed sometime and loss its popularity .

When we ask developer is testing effects your development their answer was yes its effect but in positive way, because we come to know either our app is fulfilling its requirement or not. Major focus of every testing should be functionality, content-based and exceptions/crashed/user-behavior.

One Developer says testing aim is to find issue and through these issues I learned new thing and make application perfect. Some says Testing is mostly automated in our company. Some

follow manually testing method. The other says their company follow smoke, Unit, regressive load, alpha and beta testing.

One says mostly we followed scrum system for app but it's not good for small company because our project is not too big like Facebook or WhatsApp. One of them tell us that we have a QA department in our office. They manually test the apps developed. Apps are tested by running it on different devices.

Testing is compulsory to deliver fine and bug free product. Mostly we don't get time for fixing the issues because side by side testing is not performed. They said side by side testing must need to make app more stable and after testing we may need some more changes according with better flow and by side by side testing we get time to fix bugs.

#### E. Technique Used for Supporting New API Features in Old API

Some new APIs have some hardware compatibility, new feature is being introduce in new API which doesn't being supported in old, which is a big issue for developer. So we ask how they solve this, they said we Use support libraries to use new features of new API in old API. Also if some new feature is not supported by support library then find some custom library to achieve required functionality.

#### F. Maintenance

One issue developer face is also to maintenance of app during its life cycle which is difficult and challenging task, because sometime it's very difficult to analyze the reason of app crash.

When app crash, some log should be mail to developer, so that developer can find the reason of crash to fix it.

### IV. CONCLUSIONS

Our study has given us a better, more objective understanding of the real challenges faced by the mobile application developers today, past verbose stories with their appropriate solution.

Our outcomes uncover that dealing with various mobile platforms is a standout amongst the most difficult parts of mobile development. Since mobile platforms are moving toward fragmentation as opposed to unification, the development procedure cannot use data and information from a platform to another platform. When the 'same' application is developed for multiple platforms, developers now treat the mobile app for each platform independently and manually check that the functionality is preserved across multiple platforms. We also provide solution of cross platform app development through Xamarin. Creating a reusable user-interface design for the app is also a trade-off between consistency and adhering to each platform's standards. Our

study also shows that mobile developers need better analysis tools to measure and monitor their apps. Also, testing is a huge challenge currently. Most developers test their mobile apps manually. Unit testing is not common inside the mobile community and current testing structures don't give a similar level of help for different platforms.

Also, most developers feel that present testing tools/devices are powerless and temperamental and don't reinforce fundamental features for mobile testing such as mobility (e.g., changing network connectivity), area administrations, location services, sensors, or different gestures and inputs. Finally, emulators appear to lack several real features of cell phones, which makes analysis and testing considerably additionally difficult.

There remain a large number of complex issues where further work is needed. In addition, there is a mobile "angle" to almost every aspect of software engineering research, where the characteristics of mobile applications and their operating environments present a new or different set of research issues.

### ACKNOWLEDGMENT

This research paper was supported by the Govt. College University Faisalabad, Pakistan, under department of Information Technology supervised by Dr Muhammad Sheraz Arshad Malik.

### REFERENCES

- [1] Malavolta, I.: 'Web-based hybrid mobile apps: state of the practice and research opportunities' (2016. 2016).
- [2] <https://www.nngroup.com/articles/do-interface-standards-stifle-design-creativity/>
- [3] Dehlinger, J., and Dixon, J.: 'Mobile Application Software Engineering: Challenges and Research Directions' (2017. 2017).
- [4] Martinez, M., and Lecomte, S.: 'Towards the quality improvement of cross-platform mobile applications' (2017. 2017).
- [5] Amatya, S., and Kurti, A.: 'Cross-Platform Mobile Development: Challenges and Opportunities', in Trajkovik, V., and Anastas, M. (Eds.): 'ICT Innovations 2013: ICT Innovations and Education' (Springer International Publishing, 2014), pp. 219-229.
- [6] Wasserman, A.I.: 'Software engineering issues for mobile application development'. Proc. Proceedings of the FSE/SDP workshop on Future of software engineering research, Santa Fe, New Mexico, USA2010 pp. Pages.
- [7] [https://developer.xamarin.com/guides/cross-platform/application\\_fundamentals/building\\_cross\\_platform\\_application\\_s/part\\_0\\_-\\_overview/#](https://developer.xamarin.com/guides/cross-platform/application_fundamentals/building_cross_platform_application_s/part_0_-_overview/#), accessed Dec,13 2017.
- [8] Boushehrinejadmoradi, N., Ganapathy, V., Nagarakatte, S., and Iftode, L.: 'Testing Cross-Platform Mobile App Development Frameworks (T)', in Editor (Ed.)(Eds.): 'Book Testing Cross-Platform Mobile App Development Frameworks (T)' (2015, edn.), pp. 441-451.
- [9] <https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/>, accessed Feb ,12 2017
- [10] <http://www.binariiks.com/blog/tips/mobile-app-development-react-native-vs-ionic-vs-xamarin/>, accessed Nov,20 2017

# Analysis of Valuable Clustering Techniques for Deep Web Access and Navigation

Qurat-ul-ain, Asma Sajid, Uzma Jamil  
Department of Computer Science  
Government College University Faisalabad  
Faisalabad, Pakistan

**Abstract**—A massive amount of content is available on web but huge portion of it is still invisible. User can only access this hidden web, also called Deep web, by entering a directed query in a web search form and thus accessing the data from database which is not indexed with hyperlinks. Inability to index particular type of content and restricted storage capacity is significant factor behind the invisibility of web content. Different clustering techniques offer a simple way to analyze large volume of non-indexed content. The major focus of research is to analyze the different clustering techniques to find more accurate and efficient method for accessing and navigating the deep web content. Analysis and comparison of Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Hierarchical and K-means method have been carried out and valuable factors for clustering in deep web have been identified.

**Keywords**—Deep web; clustering; Latent Dirichlet Allocation; Latent Semantic Analysis; hierarchical methods; K-means methods

## I. INTRODUCTION

The complicated structure of deep web requires sophisticated methods to access and navigate the content and data on deep web databases. Unlike the indexed surface web, deep web has no hyperlink web crawling. The complexity of deep web doesn't meet up the simple navigational access methods and techniques of surface web. Thus it requires different techniques for data extraction from deep web databases.

For the enhancement of the productivity of these search engines, the programmers are trying hard to bring the content of deep web to the surface. They not only try to search valid data but also search in a way without flooding out the users with irrelevant information. Researchers and programmers of famous search engines like Google are trying to provide data which is richer in content and fulfills user demands. Google's researchers are working on algorithm for Google's Deep web crawl [17].

The main focus of research is to analyze the different clustering techniques to find more precise and better clustering technique for access and navigation of deep web. It may be truly useful to understand the minute detail of clustering techniques and algorithms and helps to put the foundation for developing more refined techniques for the data access and navigation from Deep web.

The remainder of paper is organized as follows. Section II elaborates on previous work, Section III presents the

attempted dataset and proposed methodology, Section IV discusses our experimental results and the last Section V contains concluding remarks and demonstrates future work.

## II. LITERATURE REVIEW

Various techniques of deep web clustering and classification have been presented before the comparative study of which can be found in [10]. Several researchers contributed various approaches regarding web clustering and data extraction. Here we thoroughly discuss presented clustering techniques and algorithms regarding the inspiration towards our work.

HTML structure of web documents is becoming more complex and diverse now days thus making it complicated to extract information from web pages [1]. Dr Jill Ellsworth in 1994 initially named the term "invisible web" to denote the data which was hidden from traditional search engines [2]. Google, in 2005, provides a mechanism that allows search engines and other interested users to access deep web resources and content on certain web server and database [3].

Commercial search engines have started discovering alternative method to access deep web. BrightPlanet presented the study about deep web in 2000(a massive depository of databases and data which was hidden from search engines) declaring about deep web which was 500 times greater than surface Web having indexes available at search engines [4]. In deep web harvested search engines like Deep Web Harvester of Bright Planet Extract each individual word each time it access a web page [5]. U.S Naval Research Laboratory developed TOR network in 2002. Tor browser permits user to access deep web content anonymously and routing the encrypted requests so that traffic can be hidden from network surveillance tools [6].

Clustering is a nucleus task in data mining. Clustering is defined as "the objects are clustered or grouped based on the principle of maximizing the inter-class similarity and minimizing the intra-class similarity". Famously used clustering methods can be categorized as hierarchical methods and partitioning methods. Hierarchical decomposition can be categorized as agglomerative or divisive. Partition clustering technique generates primary partitioning and employs iterative rearrangement technique that tries to upgrade partition through proceeding clusters from a group of cluster to another. Hierarchical modeling, clustering, complex mapping and parameter knowledge gain from user connectivity based

approach is need to be developed to meet the requirements [7], [8].

Clustering requires comprehensiveness, usability, ability to deal with different kinds of elements, finding of clusters with uninformed shape and ability to handle noisy data [9].

Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation are widely used clustering techniques to access and navigate the content of deep web. Clustering is the partitioning of data in similar objects. Images, words, patterns and documents can be clustered. Clustering techniques for deep web pages are Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Both are clustering algorithms that work on text [10].

LDA semantic based technique for heterogeneous of deep web data sources is presented. LDA is a generative probabilistic model for forming content illustration of deep web database. The document consists of topics; the core work of LDA is clustering the words in document and in topics. The DWSemClust semantics based technique is developed. CAFC-C, CAFC-CH are compared to DWsemClust. CAFC employs random document selection and CAFC-CH employs the hub based clustering of induced similarity. Both of them lack the semantic based similarity of vocabulary. DWSemClust is more suitable for sparsely distributed web sources [11], [12].

Bayesian networks are the root of many successful probabilistic topic models. But the issue of the models based on Bayesian network is the complexity of structure of model as the deduction of latent topic model distribution is frequently undetectable. LDA, hierarchical Bayesian model for the inference of topic models is much time consuming. The deep neural network (DNN) approach helps the topic model inference with low computation [21].

The classification techniques use more complex symbolic representation of instances. It does not work better on large dataset. In classification algorithm like KNN, the larger the dataset, the less accurate classification is done. Whereas Clustering techniques and algorithms are used to accelerate resource retrieval process on large dataset. It also enhances the efficiency of decision making task.

Zhengyu Yang with other fellows presented a new approach of automatic replication in SSD-HDD data processing and caching process. The exchange between Input/output performance and fault tolerance is balance efficiently through auto replica (a manager) in distributed caching and data processing systems with SSD-HDD tier storage systems [19]. Approximation algorithm is presented for automatic data placement in datacenters of all-flash multi-tier. Auto-tiering provides solution regarding allocation and migration over multiple SSD. It helps in optimizing the performance and decreasing migration operating cost. It makes the issue of polynomial time simpler and resolvable [20]. The auto-tier and auto-replica [19], [20] approach are machine learning approaches which can be further enhanced in future to operate on deep web databases.

### III. METHODOLOGY

To elaborate what research is conducted it is helpful to demonstrate wholly the materials and method of research. The methodology includes the overall description of research. It provides a compact view of work which is done in this research.

- *Data Gathering*: First step was the collection of data.
- *Implementation and Comparison*: Different clustering techniques of deep web are implemented and analyzed.
- *Identification of Valuable Factors of clustering*: Valuable factors of various clustering techniques for deep web access are identified and compared. These factors include flexibility, usability, complexity, sensitivity, adaptability and scope. On the basis of these factors the comparison between differences is performed.

#### A. Data Collection

Data is collected from various sources. The collected data is then used for existing algorithm's implementation

1) *Data Set Information (bag of words)*: The dataset consists of text collection in shape of bag of words. NIPS conference paper 1987 and review of Psychological articles is gathered in the dataset. After tokenization, removal of words vocabulary was diminished by merely keeping the words coming more than 10 times. No class labels are assigned to the datasets due to copyrights factors.

TABLE I. INFORMATION ABOUT DATASET USE IN RESEARCH

Data set Information				
Sr #	Dataset	Source	Characteristics	Format/Pattern
1.	Bag of words dataset (Nips and Psychreview)	UCI Machine learning Repository	Dataset Characteristics	Text
			Attribute Characteristics	Integer
	NIPS proceeding papers		Bagofwords_nips	Document word count
			Words_nips	Vocabulary
			Authors_nips	
	Authordoc_nips	Author word count		
Psych review Abstract		Bagofwords_psychreview	Document word count	
		Words_psychreview	Vocabulary	
2.	Temprature Sheet	National Center for Environment Information	Dataset Characteristics	Numeric
			Attribute Characteristics	Integer
3.	Movie Space	MovieLen	User movie ratings	Numeric



Table I discusses the datasets that are appropriate for topic modeling and clustering. bagofwords\_nips.mat and bagofwords\_psychreview file (numeric values) and authors.nips.mat and authordoc.nips.mat (vocabulary file) are provided for every text collection. The dataset is implemented on LDA and LSA clustering techniques. Document word Count means the total number of words in document.

Vocabulary in dataset means the words or letters used. Author word count includes counting of words and letter of author names.

2) *Data Set Information (Temperature Sheet)*: National Center for Environment Information contains daily, weekly, monthly and yearly temperature forecast. The dataset includes 1981 to 2010 normal temperature consisting of 30 year temperature of all stations. The stations are represented by station numbers.

The dataset consists of daily normal weather condition and climate records. It contains most numeric values. It is very small dataset to test the efficiency of Hierarchical clustering for smaller dataset.

3) *Data Set Information (movie Space)*: The dataset consists of user's movie rating of different years and types. The data set contains numeric values and is employed in Hierarchical Clustering, k-means clustering and pLSA.

### B. Analysis and Comparison

Analysis and comparison of different clustering techniques to access and navigate the deep web are conducted to find and evaluate the functionality and performances of these techniques.

Clustering techniques are implemented in Matlab (MATLAB R2014a) for analysis. Matlab has momentous support for fast prototyping algorithms, graphing and matrix operations.

#### 1) Techniques of clustering

a) *Latent Dirichlet Allocation*: Latent Dirichlet Allocation is a generative model which affirms that documents have multiple topics. Topic is a distribution over a fixed vocabulary. All documents of homogenous set contribute to the similar combination of topics but every document demonstrates these topics with distinct ratio [13].

LDA is mostly used for the modeling of text corpora. The notion of "bag of words" is implemented in these models. The topic in this model has discrete distribution of words from some finite lexicons. LDA moulds each documents as combination of clusters. Psychreview and NIPS dataset is implemented with LDA algorithm.

b) *LDA Gibbs*: LDA Gibbs sampling is an approach now in use to solve the good probabilities of LDA methods. Steyvers and Griffiths introduced the approach of Gibbs sampling which contains the Markov Chain Monte Carlo procedure [22]. It is generally used for statistical inference. The algorithm makes use of random numbers and produces different results when executed. Execution of LDA Gibbs

(basic Topic Model dataset of psychreview and nips' bagofwords is performed to extort the set of topics and presents the most liable words per topic.

### Algorithm

1. Input: bag of words (consisting of number of times each word occur)
2. Output: Topic assignment to each word token
3. Calculate the number of times each word is given the topic.
4. Number of times the topic is allocated to document.

c) *Latent Semantic Analysis*: LSA attempts to map words and documents in concept space or clusters for comparing by implementing centroid-based clustering. It compares the meanings and concepts behind the words. It analyzes and examines the documents for finding original concepts and notions of these documents [14], [15]. Some suitable conditions to apply LSA techniques are as follows:

- 1) When documents contain same writing style.
- 2) When each and every document has focuses on particular topic.
- 3) When a word has higher probability of belonging to a topic than another topic and lower probability with other topics.

d) *pLSA*: Probabilistic Latent Semantic Analysis is an upgrade to LSA technique. Words in topics from pLSA are closely related than words in LSA. Topics are multinomial random variables in pLSA, and a particular topic produce each word and thus various words are originated by various topics. The larger the number of documents the larger the pLSA model, is the limitation of pLSA model.

Table II highlights the differences between pLSA and LSA. The LSA method originates from Linear Algebra and acts upon the Singular Value Decomposition (SVD). The pLSA method has the foundation on mixture decomposition. The advantage of using pLSA statistical model over SVD is that it permits to join diverse models methodologically.

TABLE II. DIFFERENCE BETWEEN LSA AND PLSA

Sr#	Latent Semantic Analysis	Probabilistic Latent Semantic Analysis
1	Highest Gaussian Error	Highest Likelihood Function
2	No apparent explanation of parameters	Polynomial Distribution of Parameters
3	Singular Value Decomposition is precise.	pLSA EM congregate to confined best possible

2) Types of clustering

a) Hierarchical Clustering: Hierarchical algorithmic methods use similarity or distance matrix. Splitting or merging of one cluster is performed at one time. Dendrograms are used to represent hierarchical clustering.

Hierarchical Methods

Divisive: Divisive is the top bottom approach for clustering. Divisive clustering is less blind to the global structure of data. The idea of divisive clustering is that all objects are in one cluster. The cluster is divided into sub-clusters which are sequentially separated into more sub-clusters. This process persists unless the preferred cluster is acquired. The divisive clustering follows the top-down approach of hierarchical structure.

Agglomerative: Every object embodies its own cluster. The clusters are sequentially merged unless the desired pattern of cluster is achieved. The fundamental function of agglomerative clustering is the calculation of proximity among two groups of clusters. Agglomerative clustering follows the bottom up hierarchy [16].

- 1) Initiate with point as single clusters
- 2) At every step, merge the closest pair of clusters until only a cluster left.

Algorithm

1. Calculate the proximity/similarity matrix
2. Let each data point be a cluster
3. Merge the two nearest and most similar closest clusters. Update the proximity/similarity matrix
4. Repeat 3 & 4 until all patterns are in individual Cluster.

b) K-means Clustering: K-means is a heuristic approach of partitioning clustering. Each cluster is connected with a

central point called centroid. Each point in cluster is linked to cluster with nearest and closest cluster. Number of clusters must be identified and denoted as  $k$ . The aim is to reduce the summation of distances of the points to their relevant centroid. Mixture model (EM algorithm: dealing with clusters having uncertainty), k-medoids(better for noise and outlier), k-median and k-models are the variations of k-means method.

Algorithm

1. Choose  $K$  points as early centroids.(initial centroids are selected randomly)
2. From  $K$  clusters allocate all points to the nearest and closest centroids.
3. Recomputed the centroid of every cluster
4. Reiterate step 2& 3 unless the centroids don't change.
5. Selection of  $K$  points may be performed by using some method.

IV. RESULTS AND DISCUSSION

Results are presented in this section by implementing various algorithm and different dataset. Major outcome of this research is describes below.

A. Latent Dirichlet Allocation

LDA Gibbs algorithm is implemented on the bag of words (NIPS and psychreview) dataset. It proves better than LSA traditional algorithm by accessing and extracting desired information. It is proved as time efficient techniques with a large dataset. As the number of iterations increase the time efficiency of LDA is disturbed. Words extraction from different topic models is depicted as below. Table III shows the detailed comparison of LDA and LSA.

Fig. 1 shows the most occurred words in first ten topics with ten iterations. It offers more compact view of data extraction from documents. LDA (LDA-Gibbs) technique is more accurate to present possible desired results.

TABLE III. COMPARISON BETWEEN LDA AND LSA

Comparison between LDA and LSA Latent Dirichlet Allocation (LDA) vs. Latent Semantic Analysis (LSA)		
Factors	Latent Dirichlet Allocation (LDA)	Latent Semantic Analysis (LSA)
Usability	More effective at finding word-level topics with large dataset.	Less effective as compared to LDA
Time Complexity	Less Time consuming	Much Time consuming
Suitability	Suitable for large dataset as well as smaller data set	It performs efficiently with smaller data set but it is not suitable for large dataset
Flexibility	Gibbs LDA sampling is easier to compute	Singular Value Decomposition is difficult to compute
Capacity	Provide a probabilistic model at document level	Offers no probabilistic model at document level
Usage	It assigns Probability for document/topic/word in each cluster	Probabilistic LSA defines the probability of /topic/word in each cluster.

```

Iteration 0 of 10
Elapsed time is 99.879548 seconds.

Most likely words in the first ten topics:

ans = |

'units network hidden input networks net output'
'variables learning algorithm belief probability distribution problem'
'spike noise information neuron signal neurons code'
'state learning states policy action optimal time'
'model memory neural network capacity figure results'
'data clustering cluster algorithm estimate tree model'
'representation field feature image recognition level top'
'neuron network neurons neural activation input time'
'recognition classification training class classes table data'
'class vector network neural function risk networks'
    
```

Fig. 1. LDA Word extraction in different Topics.

```

Example topics of chain 1 sample 1
ans =

'perceptual conditions patterns result organization'
'theories similarity proposed psychological dimensions'
'word words network semantic model'
'problems research strategies empirical theoretical'
'models data based simple rules'

Example topics of chain 1 sample 2
ans =

'perceptual conditions result patterns psychological'
'theories similarity proposed shown dimensions'
'word model words network semantic'
'research problems theoretical strategies methods'
'models data based rules simple'

Example topics of chain 2 sample 1
ans =

'account spatial defined objects series'
'problem problems related variables psychological'
'information processing stage motion rt'
'visual perception perceptual target masking'
'social approach levels system principles'

Example topics of chain 2 sample 2
ans =

'account alternative objects terms spatial'
'problem problems variables independent solving'
'information processing motion stage stages'
'visual perception perceptual target masking'
'social approach levels specific empirical'
    
```

Fig. 2. LDA topic generation.

Fig. 2 depicts that Topics generated by LDA algorithm in 100 iterations from 2 samples. The lesser the iterations the lesser the time consumption is observed.

### B. Latent Semantic Analysis

LSA algorithm is implemented on Bag of Word (NIPS & psychreview) dataset. The LSA measures the likelihood of every word in a topic model. The word extraction from topics is a time consuming process in LSA as compared to LDA. A small dataset is easily accessed in lesser time than large dataset.

A pLSA code with a large data set with different iterations has run and produced different elapsed time. The execution time on large dataset with 100 iterations is 2331.868618 and 121.118317 sec with 50 iterations. Table III compares LDA and LSA on the basis of various factors.

The pLSA working of algorithm with minimum iteration produce fast execution, whether the execution on large dataset with much iteration is time-consuming.

Fig. 3 shows the likelihood of occurrences of words in topics through pLSA EM (Expectation-Maximization) steps with 10 iterations. It generates results with top 20 words in top 10 topics.

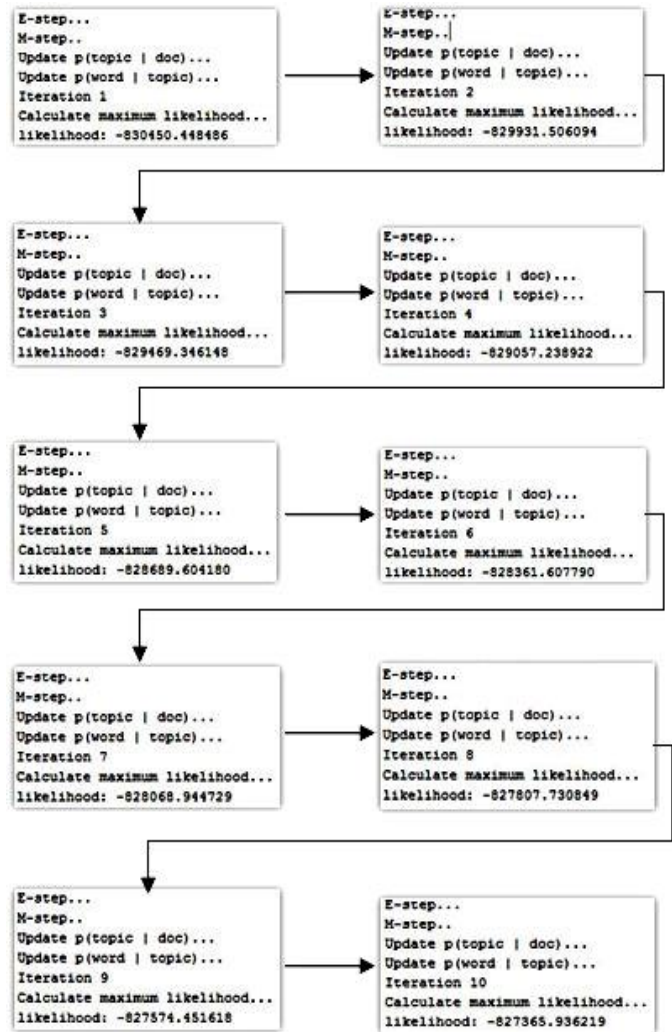


Fig. 3. Likelihood of occurrences of words in a topic.

TopN(20) keywords for topic 1	TopN(20) keywords for topic 2
middai (0.000146)	shekel (0.000146)
minim (0.000146)	legran (0.000146)
atrit (0.000146)	bark (0.000146)
none (0.000146)	hizb (0.000146)
fail (0.000146)	marguli (0.000146)
oil (0.000146)	aeronaut (0.000146)
2003 (0.000146)	hawaii (0.000146)
norfolk (0.000146)	handler (0.000146)
violin (0.000146)	kordofan (0.000146)
disgrac (0.000146)	carden (0.000146)
nanga (0.000146)	soire (0.000146)
promptli (0.000146)	bassist (0.000146)
faithless (0.000146)	candlelight (0.000146)
ettiquit (0.000146)	assi (0.000146)
muscat (0.000146)	tab (0.000146)
merger (0.000146)	chao (0.000146)
deploy (0.000146)	alexi (0.000146)
swimmer (0.000146)	cegypt (0.000146)
farouq (0.000146)	hold (0.000146)
200000 (0.000146)	unnatur (0.000146)

Fig. 4. pLSA's words extraction from topics with possible likelihood.

Fig. 4 shows pLSA topic extraction with word's occurrence likelihood that these words and keywords are extracted from each topic with maximum likelihood of word occurrence in these topics. The 10 iterations create 10 topics with each topic consisting of 20 top keywords in those topics.

### C. Hierarchical clustering

Hierarchical clustering algorithm is implemented on Movies and Temperature datasets [18].

#### 1) Types of Linkage Function

a) **Single Linkage:** Merging of two clusters where two nearest elements have the minimum distance. It produces the minimum spanning tree. It promotes the expansion of extended clusters. It is highly sensitive to the noise.

b) **Complete Linkage:** Merging of two clusters in every step that merging has the maximum distance. It promotes dense clusters. It doesn't work efficiently if extended clusters are presented.

c) **Average Linkage:** Keeping in view the sensitivity of complete linkage clustering to outliers and the predisposition of single linkage clustering to create big chains that don't match up the discerning idea of clusters as solid objects is observed. Agglomerative clustering is very strong and vigorous with average cluster distance and linkage. Fig. 5 shows the comparisons of types of linkages.

The algorithm is implemented on MovieSpace dataset and Temperature Sheet dataset. As the hierarchical clustering is implemented on large dataset of MovieSpace, it consumes more CPU time and memory space and affects the cost of Input/output. It is slower than k-means algorithm on same dataset and takes longer time to generate result.

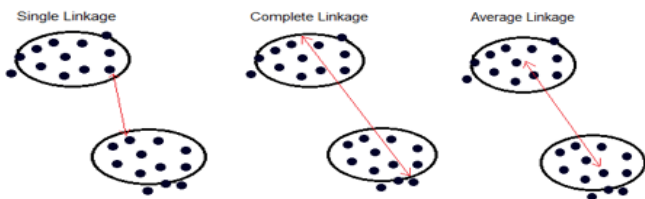


Fig. 5. Single, complete and average linkages.

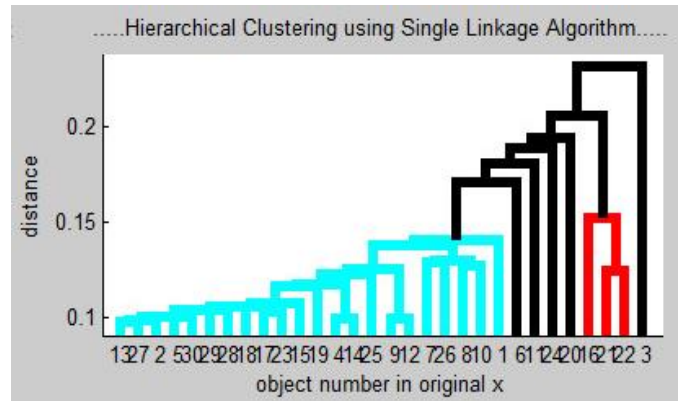


Fig. 6. Hierarchical clustering using single linkage algorithm on large dataset.

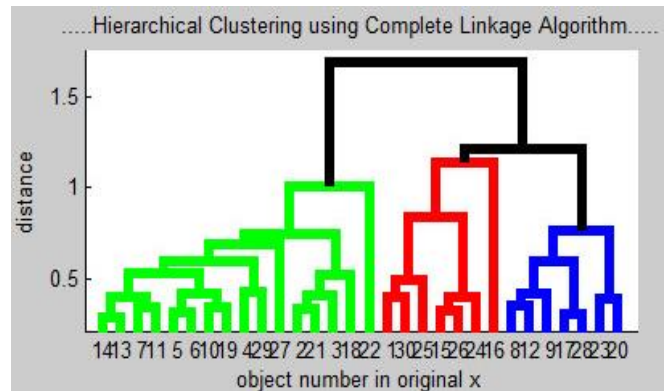


Fig. 7. Hierarchical clustering using average linkage algorithm on large dataset.

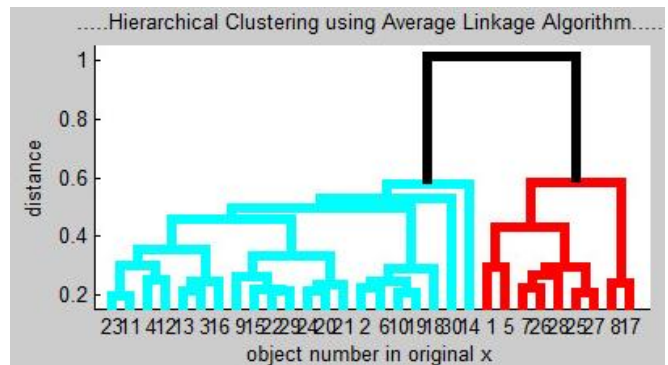


Fig. 8. Hierarchical clustering using complete linkage algorithm on large dataset.

The implementation of hierarchical algorithm on smaller dataset of Temperature sheet produces different results. The small number of instances in dataset results in low Input/output cost and less execution time. It produces following results on movieSpace dataset.

In Fig. 6, 7 and 8 the pairs of object forming cluster are depicted in object number in original X (Y label). These figures show the hierarchical tree of Single, complete and average linkage function which was performed for hierarchical clustering on MovieSpace dataset. After analysis of these figures, the execution time for the movieSpace dataset was obtained is 282 sec.

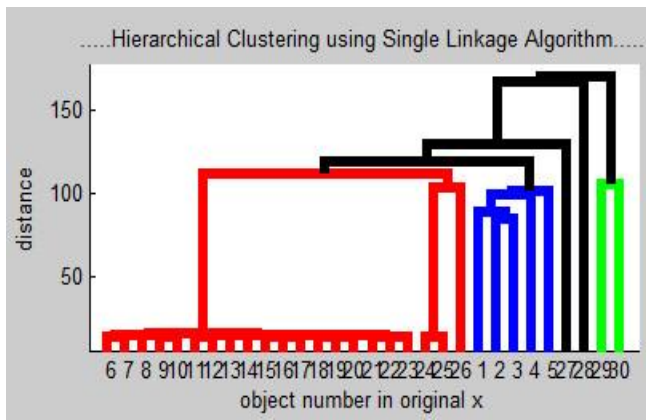


Fig. 9. Hierarchical clustering using single linkage algorithm on small dataset.

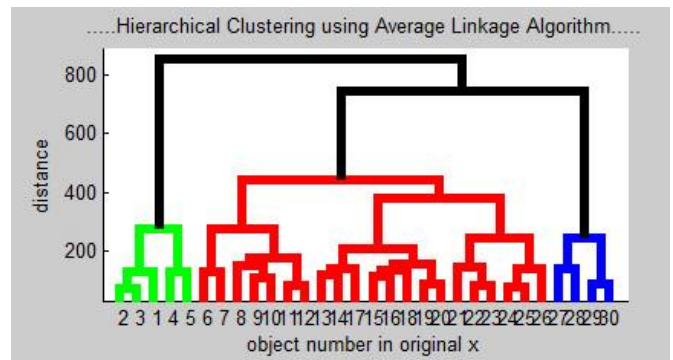


Fig. 11. Hierarchical clustering using complete linkage algorithm on small dataset.

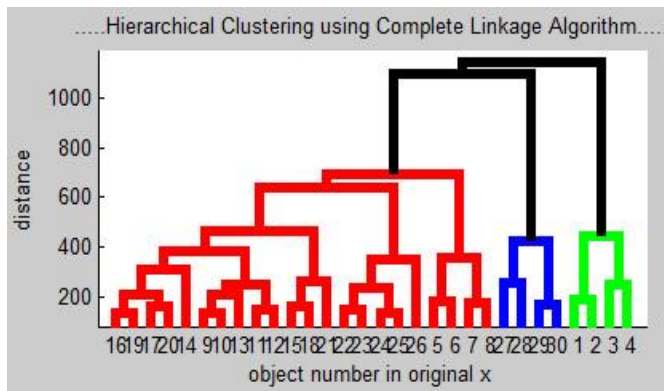


Fig. 10. Hierarchical clustering using average linkage algorithm on small dataset.

Fig. 9, 10 and 11 shows single complete and average linkage algorithm implementation for small dataset of temperature respectively. Execution time for Temperature dataset is 41 sec.

#### D. K-means Clustering

K-means algorithm is implemented on MovieSpace dataset. On data set of MovieSpace the K-means algorithm is implemented which produce faster results and low execution time than the same dataset's implementation in Hierarchical algorithm. The algorithm produces random cluster each time when executed. It generates efficient result with sparse (not dense) data with no noise. It utilizes less Input/output and memory storage for execution. K-means performs better with large data set as compared to hierarchical clustering.

Table IV displays the complete comparison of hierarchical and K-means clustering.

TABLE IV. COMPARISON BETWEEN HIERARCHICAL CLUSTERING AND K-MEANS CLUSTERING

Hierarchical Clustering vs. K-means Clustering		
Factors	Hierarchical Clustering	K-means Clustering
<b>Nesting</b>	Combination of nested clusters, which are arranged as tree.	Combination of objects in related clusters such that every object is in just a single cluster.
<b>Complexity</b>	Time and Space complexity(Non-Linear) Time complexity is at least $O(m^2)$ m is the total number of instances that is not linear with number of objects in cluster. Clustering a large dataset may have immense I/O cost.	Linear Complexity The algorithm works well with very large number of instances. Better of Large dataset.
<b>Sensitivity</b>	Problem with noise and outliers in data	Problem with noise and data that has outliers. Appropriate only when mean is characterized. It needs the number of clusters in advance.
<b>Result Demonstration</b>	The result of hierarchical clustering is presented in the form of dendrogram.	The results of K-means clustering are presented mostly in cluster points and plots.
<b>Back-tracking</b>	No back-tracking is observed as hierarchical clustering can never go back to previous step	K-means is a randomized algorithm , it always select clusters randomly each time
<b>Usability</b>	The Use of Hierarchical clustering is normally constrained with numeric attributes. A hierarchy of documents in deep web database can also be maintained	The Use of K-means is frequently constrained with numeric attributes.
<b>Adaptability</b>	Show better performance on data set consisting of non-identical clusters containing string and having a common centers or clusters	Show better performance on data set which has isotropic clusters and not as adaptable as hierarchal single link method.

TABLE V. K-MEANS IMPLEMENTATION OVER DATASET

Sr#	K	D	N	X	INIT	answer
1.	2	2	3000	5	19	(3 2 3 1)
2	3	3	3000	5	19	(1 2 1 3)

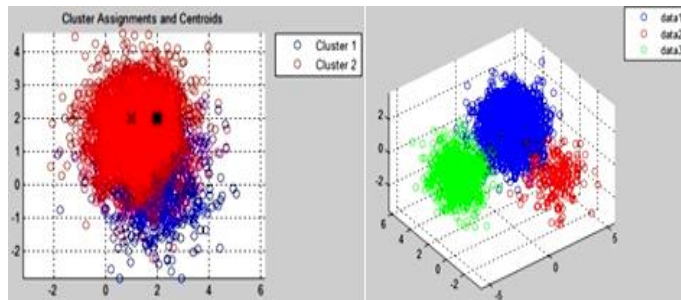


Fig. 12. Views of clusters assignment and centroids.

In Table V, K-means takes required number of clusters and the starting means as inputs and generates final means as output. Means of cluster are described first and last means. K is the number of clusters,  $d$  is dimensions (2nd dimensional or 3<sup>rd</sup> dimensional), X and INIT are the required numbers of clusters and initial means.

Fig. 12, Sensitivity of K-means can be seen in center initialization of a cluster. When the center initialization is done poorly it may lead to bad intersection speed and on the whole bad clustering. K-means clustering method groups the same the type of objects in similar cluster.

## V. SUMMARY AND CONCLUSION

The complicated structure of deep web requires sophisticated methods to access and navigate the content and data on deep web databases. The comparative analysis of clustering techniques demonstrates that to extract information from deep web databases is the complex task. It deduces the weaknesses of these techniques to overcome for better performance. LSA is beneficial to work with small datasets; it is much time-consuming working with large datasets. The LDA (LDA-Gibbs) technique is far better than LSA to present possible desired results. Hierarchical and partitioning methods are beneficial for structuring the content on deep web databases. The random allocation of documents in cluster having similar or dissimilar documents produces time efficient method, whereas, hierarchical structure of documents of similar category produces time consuming methods. The combination of both methods may enhance some features for structuring document.

The analysis provides new directions to refine these techniques. The future work focuses on designing, modification and amalgamation of existing techniques for better performance and functionality. Genetic algorithm based clustering techniques are taken into consideration for future

technological enhancements. The Combination of clustering technique with other data mining or machine learning technique like Deep Neural networks and artificial neural networks may provide a more optimized and refined technique of data access on deep web. For maximum desired search outputs semantic as well as syntactic accuracy of search prediction can be devised through discovering unique techniques to deal with certain semantic and linguistic properties of deep web sites.

## REFERENCES

- [1] Lavanya M. & Dr. Usha Rani "A framework for vision-based Deep Web Data Extraction for web", September 2012.
- [2] Michael K. Bergram "The Deep Web: Surfacing Hidden values". White Paper: *Deep Content*, September 2001.
- [3] Chertoff M. & Simon T.(2015,February)The Impact of the Dark Web on Internet Governance and cyber security. *Paper Series No 6.Global Commission on Internet Governance*
- [4] Steve Pederosen(2013,March) Understanding the Deep Web in 10 Minutes. White Paper.
- [5] BrightPlanet (2012), What is Deep Web Harvest? <https://brightplanet.com/2012/07/what-is-a-deep-web-harvest/>
- [6] Dinglein R., Mathewson N. & Syerson P.( 2004,August) "Tor: Second Generation Onion Router". In *proceeding of 13<sup>th</sup> conference on USENIX Security Symposium*. Volume 13.
- [7] Han J. et al., "Data Mining: concepts and techniques". *Third Edition*, 2012
- [8] Wensheng Wu et al.,(2004,June) "An Interactive Clustering-based Approach to Integrating Source Query Interfaces on the Deep Web". In *proceedings of SIGMOD June 2004*.
- [9] Estivill Castro V.,(2002, June) "Why so many clustering algorithm- A position Paper". *SIGKDD*. Volume 4, Issue 1.
- [10] Muhunthaadithya C & Rohit J.V et al., "Clustering of Deep WebPages: A comparative study".*International Journal of Computer Science Information Technology(IJCSIT)*. Volume 7, No 5, October 2015
- [11] Umara Noor & Ali Daud et al., "Latent Dirichlet Allocation based Semantic Clustering of Heterogeneous Deep Web Sources", September 2013 [ *In proceedings of 5<sup>th</sup> International Conference on Intelligent Networking and collaborative Systems*]
- [12] Ben Eysenbach( 2016,May) Latent Dirichlet Allocation and Application to DSPACE.
- [13] David M.Blei , Andrew Y.Ng et al., (2003, March) Latent Dirichlet Allocation. *Journal of Machine Learning Research*
- [14] Katzman D. (2010,June) "Clusters that Think". *Article*. Link <http://deepwebtechblog.com/clusters-that-think/>
- [15] Landauer T. K. & Peter W. Foltz et al., (1998) "An Introduction to Latent Semantic Analysis". *Discourse Processes*, 25,259-284
- [16] Manpreet Kuar and Usvir Kuar."Comparison Between K-Mean and Hierarchical Algorithm Using Query Redirection". *International Journal of Advanced Research in Computer Science and software Engineering*. Volume 3, Issue 7, July 2013
- [17] Jayant Madhavan, David Ko et al., "Google's Deep-Web Crawl"
- [18] Sivtalana Volkova, "Latent Dirichlet Allocation".*Final Year Project Report*"Hierarchical clustering Algorithm" <https://github.com/meskatjahan/Hierarchical-clustering-Algorithm>
- [19] Zhengyu Yang et al., "AutoReplica: Automatic data replica manager in distributed caching and data processing systems".
- [20] Zhengyu Yang et al., "AutoTiering: Automatic Data Placement Manager in Multi-Tier All-Flash Datacenter"
- [21] Dongxu Zhang et al., "Learning from LDA using Deep NeuralNetworks"
- [22] "LDA Gibbs". *TopicToolbox* <https://github.com/huashiyiqike/topictoolbox>

# Pre-Trained Convolutional Neural Network for Classification of Tanning Leather Image

Sri Winiarti, Adhi Prahara, Murinto, Dewi Pramudi Ismi

Informatics Department  
Universitas Ahmad Dahlan  
Yogyakarta, Indonesia

**Abstract**—Leather craft products, such as belt, gloves, shoes, bag, and wallet are mainly originated from cow, crocodile, lizard, goat, sheep, buffalo, and stingray skin. Before the skins are used as leather craft materials, they go through a tanning process. With the rapid development of leather craft industry, an automation system for leather tanning factories is important to achieve large scale production in order to meet the demand of leather craft materials. The challenges in automatic leather grading system based on type and quality of leather are the skin color and texture after tanning process will have a large variety within the same skin category and have high similarity with the other skin categories. Furthermore, skin from different part of animal body may have different color and texture. Therefore, a leather classification method on tanning leather image is proposed. The method uses pre-trained deep convolution neural network (CNN) to extract rich features from tanning leather image and Support Vector Machine (SVM) to classify the features into several types of leather. Performance evaluation shows that the proposed method can classify various types of leather with good accuracy and superior to other state-of-the-art leather classification method in terms of accuracy and computational time.

**Keywords**—Leather classification; tanning leather; convolution neural network (CNN); deep learning; support vector machine (SVM)

## I. INTRODUCTION

Small and medium sized industries lately have experienced a rapid growth. One of them is leather craft industry which produces various kinds of leather craft item, such as gloves, wallet, belt, sandals, shoes, jacket, and bag. The growth of leather craft industry also affects leather tanning factories to increase their production in order to meet the demand of leather craft materials. To achieve large scale production, an automation system should be implemented in the leather tanning factories.

The automation system involves automatic grading based on type and quality of leather. Type of leather usually distinguished based on color and texture using global and/or local statistical geometrical features with machine learning approach [1], [2]. Quality of leather is mainly determined by the size and location of leather defect. Leather defect can be categorized into five types: lines, holes, stains, wears, and knots [3]. To locate the defects, researchers use morphological operation [4], [5], clustering [6], [7], or machine learning approach [3], [8]-[10].

Leather craft materials usually come from cow, crocodile, lizard, buffalo, goat, sheep, and stingray skin. The animal skin will go through tanning process before it can be used as crafting materials. In every level of tanning process, tanning agent will alter the physical properties and chemical compositions of the skin. The skin will become durable, pliant, and may have different color and texture from the original. Therefore, after tanning process, animal skins will have a large variety of color and texture within the same skin category and have high similarity with other skin categories. This make them difficult to be distinguished. Furthermore, skin from different part of animal body may have different color and texture.

The result of classification determines the grade of tanned leather. The grade of leather will affect the price of leather. Therefore, classification procedure is the most important in automation system of tanning leather production because it is directly affects the price of final tanning leather products. Furthermore, a high return rate and disputes between customer and manufacturing industry which caused by failure in classification of leather usually cause additional costs [4].

From leather craftsmen point of view, the correct result of leather grading is important because it will be used as consideration to determine the type of leather craft product that will be made. Mistakes in determining the type of leather for leather craft products can make the resulting leather craft products become unfavorable and impact on loss in sales. Leather craftsmen in some area do not yet have the knowledge about feasible standard of leather crafting. The type and quality of leather that will be used as leather craft materials are known based on experience and tradition which inherited from generation to generation.

Therefore, this research proposes a method to classify type of leather on tanning leather image and performs performance comparison to evaluate the method. In summary, the contributions of this work are given as follows:

- 1) The proposed method uses tanning leather images as input because tanning leathers are the final product of leather tanning factory and will become leather craft materials for leather craft industry. Therefore, the proposed method can be used and will benefit both of tanning leather factory and leather craft industry.
- 2) The features of tanning leather images are extracted using specific layer from pre-trained deep convolutional neural

network (CNN). As the layer goes deeper, the richer features will be extracted from the tanning leather images.

3) Classification is done using linear Support Vector Machine (SVM). SVM performs well in few training data, easy to configure the parameters, and has potential to perform real time classification.

4) Finally, performance of the proposed method is compared with the other state-of-the-art leather classification method.

The rest of this paper is organized as follow: Section 2 presents the related works, Section 3 presents the proposed leather classification method, Section 4 presents the results and discussion, and the conclusion of this work is described in Section 5.

## II. RELATED WORKS

Leather classification method has been proposed and developed by many researchers. Most of them focus on making an automation system to detect lather defects [3]-[10]. Leather defects inspection system that has purpose to detect the size and location of defects on leather surface is one of the characteristic to determine the quality of leather along with other characteristics such as the type of leather and the correlation between usable and unusable areas on leather. In this research, authors focus on classifying type of leather as it is also an important aspect that affect the quality of leather.

In the tanning leather classification, the difficulties come from large variety and similarity of color and texture of tanning leather. Researchers use statistical geometrical features with machine learning approach to classify the type of leather. In [1], an improved statistical geometrical features (ISGF) is proposed to classify chrome tanning and vegetable tanning leather. They use two classifiers based on Fisher criterion and Learning Vector Quantization (LVQ) network to compare the chrome tanning and vegetable tanning leather. The result shows that ISGF outperforms the performance of SGF.

In their research, [2] use both global and local features to classify leather. The global features are extracted from the two dimensional power spectrum of enhanced leather trench image. The local features are based on mathematical morphology operation of segmented leather image. The features then compared with set of training image that already classified by the experts.

In this research, authors use rich features obtained from pre-trained CNN. The rich features obtained from pre-trained CNN also have been used in some classification problems [11]-[14]. The features are classified using SVM into several types of leather. Authors perform the same procedure and use the same training data for the other leather classification and compared its performance. The other leather classification uses hand-crafted features which are color moments and some statistical measurements from Gray Level Co-Occurrence Matrix (GLCM). GLCM is well known texture features extraction method and has been used in many texture based classification in wide range of applications [15]-[18]. The features are also classified using SVM.

## III. METHODOLOGY

This research proposes a method to classify the type of leather to be used in the quality measurement of leather craft materials. General procedure of the proposed method is shown in Fig. 1. From Fig. 1, tanning leather image is resized to fit the input of pre-trained CNN. Authors use the seventh layer of AlexNet [19] which originally trained for ImageNet challenge [20] to extract the features. The features will be classified using linear SVM into five types of leather: monitor lizard, crocodile, sheep, goat, and cow.

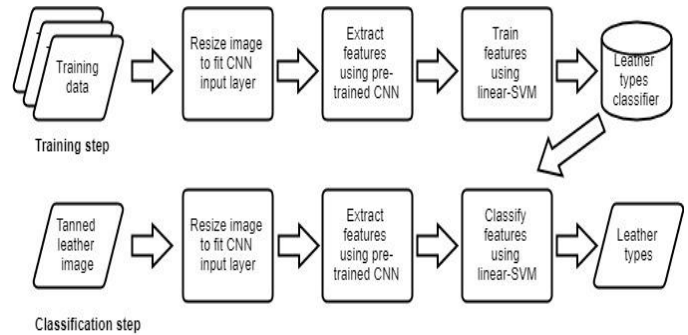


Fig. 1. General procedure of the proposed leather classification method.

### A. Tanning Leather Image

Tanning is a process of making leather from the skins of animals. Tanning agents can alter the physical properties and chemical compositions of the leather. The result of tanning is more durable, pliant, and texturally practical material. There are three methods of tanning leather: chroming, vegetable, and combination. Each tanning method produces different materials, aesthetically and texturally e.g. chrome tanned leather is soft and oiled type while vegetable tanned leather is sturdy and hefty. Because tanning agents can change the color and texture of leather surface, the leather images acquired from the industry have large variety even from the same animal. Additionally, the image can be damaged by the external noises.

### B. Hand-Crafted Feature Extraction

The hand-crafted leather classification usually based on color and texture. The most common statistical color features are color moments and for statistical texture features are Gray Level Co-Occurrence Matrix (GLCM).

1) *Color features extraction*: Color features can be extracted using color moments. Color moments are characteristic measurement of color distribution from image. Color moments consist of mean, standard deviation, skewness, kurtosis, and other higher order moments. Color moments are scale and rotation invariant. Color moments that used in this research are explained as follows:

a) Mean is the average value of color on image. If  $N$  is the number of pixel on image and  $p_i$  is the  $i^{th}$  pixel on image then mean ( $\mu$ ) can be calculated using (1).

$$\mu = \frac{1}{N} \sum_{i=1}^N p_i \quad (1)$$

b) Standard deviation is the square root of the variance. The standard deviation ( $\sigma$ ) can be calculated using (2).



$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N |p_i - \mu|^2} \quad (2)$$

c) Skewness is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right. The skewness ( $s$ ) can be calculated using (3).

$$s = \sqrt[3]{\frac{1}{N-1} \sum_{i=1}^N |p_i - \mu|^3} \quad (3)$$

d) Kurtosis is a measure of how the outlier-prone a distribution. Kurtosis  $k$  can be calculated using (4).

$$k = \sqrt[4]{\frac{1}{N-1} \sum_{i=1}^N |p_i - \mu|^4} \quad (4)$$

2) *Texture features extraction*: Texture features can be extracted from Gray Level Co-Occurrence Matrix (GLCM). GLCM uses spatial correlation between pixels. Feature extraction using GLCM is done by measuring the occurrence level of paired pixel with specific value on the image then calculate the statistical texture features [21]. The features can be measured from four different orientations or offsets which are  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$ , and  $135^\circ$ . Some of the statistical measurements of GLCM that used in this research are explained as follows:

a) Contrast is a measure of the intensity between a pixel and its neighbor over the whole image. Contrast is used to measure the local variance level in GLCM matrix. If  $p(i, j)$  is the GLCM in coordinate  $(i, j)$ , then contrast can be calculated using (5).

$$contrast = \sum_{i,j} |i - j|^2 p(i, j) \quad (5)$$

b) Correlation is a measure of how correlated a pixel is to its neighbor over the whole image. Correlation is used to measure the occurrence of paired pixels in GLCM. If  $\mu_i = \sum_{i,j} i \cdot p(i, j)$  and  $\mu_j = \sum_{i,j} j \cdot p(i, j)$  is the mean of GLCM  $i$  and  $j$ , and  $\sigma_i = \sqrt{\sum_{i,j} p(i, j) (i - \mu_i)^2}$  and  $\sigma_j = \sqrt{\sum_{i,j} p(i, j) (j - \mu_j)^2}$  is the standard deviation of GLCM  $i$  and  $j$  then correlation can be calculated using (6).

$$correlation = \sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)p(i, j)}{\sigma_i \sigma_j} \quad (6)$$

c) Energy is the sum of squared elements in the GLCM and can be calculated using (7).

$$energy = \sum_{i,j} p(i, j)^2 \quad (7)$$

d) Homogeneity is a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal and can be calculated using (8).

$$homogeneity = \sum_{i,j} \frac{p(i, j)}{1 + |i - j|} \quad (8)$$

e) Entropy is a measure of non-uniformity and texture complexity of image. Entropy can be calculated using (9).

$$entropy = - \sum_{i,j} p(i, j) \log p(i, j) \quad (9)$$

### C. Pre-Trained CNN for Feature Extractor

Convolutional Neural Network (CNN) is a powerful machine learning technique. In the deep learning field, CNNs are trained using large collections of diverse images. From these large collections, CNNs can learn rich feature representations for a wide range of images. CNNs have many layers namely input layer, convolutional layers, Rectified Linear Unit (ReLU) layers, cross channels normalization layers, average pooling layers, max pooling layers, fully connected layers, dropout layers, softmax layers, and output classification layers. Each layer in the networks takes in data from the previous layer, transforms the data, and passes the data on the next layer. The network will learn directly from the data and increases the complexity and detail of what it is learning from layer to layer. The function of some layers are explained as follows:

a) Convolutional layer puts the input images through a set of convolutional filters, each of which activates certain features from the images.

b) Pooling layer simplifies the output by performing nonlinear downsampling, reducing the number of parameters that the network needs to learn.

c) ReLU layer allows faster and more effective training by mapping negative values to zero and maintaining positive values.

d) Fully connected layer has outputs a vector of  $K$  dimensions where  $K$  is the number of classes that the network will be able to predict. This vector contains the probabilities for each class of any image being classified.

e) The final layer of the CNN architecture uses a softmax function to provide the classification output.

In order to leverage the power of CNNs without investing time and effort into training is to use a pre-trained CNN as a feature extractor. Layer that will be used as feature extractor is the fully connected layer that extract richer features compared to the lower layer. Most of the models have been trained on the ImageNet dataset [20], which has 1000 object categories and 1.2 million training images.

One of the model is AlexNet [19] which published in 2012. It can classify images into 1000 different categories, including keyboards, computer mice, pencils, and other office equipment, as well as various breeds of dogs, cats, horses, and other animals. The architecture of AlexNet is shown in Fig. 2 and the sequence of layer that used in the model is listed below:

- Input image layer: 227x227x3
- Convolutional layer 1: 96 11x11 filters at stride 4
- ReLU layer 1
- Cross channel normalization layer 1
- Max pooling layer 1: 3x3 filters at stride 2
- Convolutional layer 2: 256 5x5 filters at stride 1, pad 2
- ReLU layer 2
- Cross channel normalization layer 2
- Max pooling layer 2: 3x3 filters at stride 2
- Convolutional layer 3: 384 3x3 filters at stride 1, pad 1
- ReLU layer 3
- Convolutional layer 4: 384 3x3 filters at stride 1, pad 1

- ReLU layer 4
- Convolutional layer 5: 256 3x3 filters at stride 1, pad 1
- ReLU layer 5
- Max pooling 3: 3x3 filters at stride 2
- Fully connected layer 6: 4096 neurons
- ReLU layer 6
- Fully connected layer 7: 4096 neurons
- Fully connected layer 8 (softmax layer): 1000 neurons
- Classification output layer

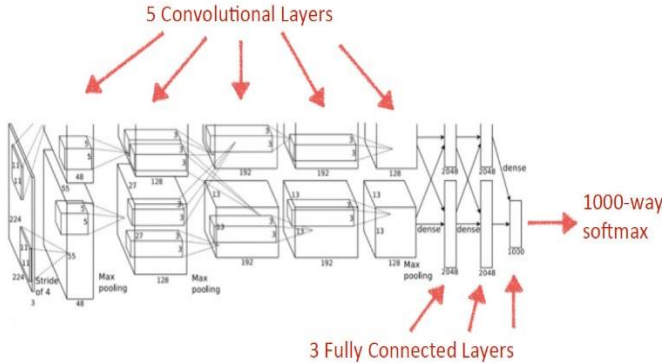


Fig. 2. The architecture of AlexNet [19].

#### D. Tanning Leather Image Classification

Support vector machine (SVM) is used to classify tanning leather image from set of features extracted using pre-trained CNN. SVM is a classifier that determined by separation which called hyperplane. Hyperplane can be calculated by maximizing margin or distance from two set of object from two different class. Classification with SVM consist of training and classification. In the training step [22], if there are given training data  $x_i \in R^n, i = 1, \dots, l$  in two classes and label  $y \in R^l$  such that  $y_i \in \{1, -1\}$ , it can be solved with (10).

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^l \xi_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad (10)$$

$$\xi_i \geq 0, i = 1, \dots, l$$

where  $w$  is weight vector,  $b$  is bias,  $\xi_i$  is slack variables,  $\phi(x_i)$  maps  $x_i$  into higher dimensional space and  $C > 0$  is the regularization parameter and the decision function is shown in (11).

$$\text{sgn}(w^T \phi(x) + b) = \text{sgn}(\sum_{i=1}^l y_i \alpha_i K(x_i, x) + b) \quad (11)$$

where  $K(x_i, x_j)$  is the kernel function. After training process, parameter  $y_i \alpha_i \forall i$ ,  $b$ , label names, support vectors, and kernel parameter saved as output model from training SVM. In classification, voting strategy is performed for each data  $x$  which will be designated to be in a class with the maximum votes [22]. Optimal parameter is selected using  $k$ -fold cross validation which is a method to do cross validation by dividing training data into  $k$  set which has  $(k-1)$  as training data and the rest will be the test data. After training process we will get variable  $w$ ,  $x$ , and  $b$  for each class, then the classification process can be done with these steps:

- 1) Calculate kernel.
- 2) Calculate decision function using (11).
- 3) Repeat step 1 and 2 for other classes.
- 4) Determine the class by function which gives the most maximum result.

#### E. Performance Evaluation

The evaluation procedure uses confusion matrix then calculates the accuracy, specificity, sensitivity, and precision using (12), (13), (14), and (15) respectively where TP is true positive, FP is false positive, TN is true negative, and FN is false negative. In order to handle imbalance dataset between positive and negative test data, authors use ratio as weight in the calculation of statistical performance measurement to balance the influence of the dataset.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} \quad (13)$$

$$\text{Sensitivity} = \frac{TP}{(TP+FN)} \quad (14)$$

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (15)$$

### IV. RESULT AND DISCUSSION

The proposed method runs on Laptop with Intel i7 processor, 16GB of RAM, and graphics card NVIDIA GTX 1050 4GB. Performance evaluation is done using MATLAB. The proposed method will be tested to classify leather into five types: monitor lizard, crocodile, sheep, goat, and cow skin.

Training data consist of 1000 leather images i.e. 200 leather images for each category. Authors conduct performance evaluation test on 3157 leather images. Input tanning leather images are taken using camera with 15-50 cm distances from leather objects and saved into images with size 512x512 pixels. Fig. 3 shows the samples of leather image used in this research. From Fig. 3, tanning leathers may have different texture within the same category which caused by tanning process or the skins are taken from different part of the animal body e.g. reptiles have different color and texture of skin in their belly and back.

#### A. Result of Hand-Crafted Feature Extraction

For the hand-crafted feature extraction, input images are converted into grayscale. Authors extract 24 global features from input image, consist of 4 features from color moments namely mean, standard deviation, skewness, and kurtosis and the next 20 features come from statistical texture measurement of GLCM in 4 offsets, namely, contrast, energy, correlation, homogeneity, and entropy. Features are standardized then classified using SVM. Confusion matrix and statistical measurement per category for this hand-crafted leather classification scheme is shown in Table I. From Table I, all samples of crocodile, sheep, and goat tanning leather images can be recognized perfectly except in the monitor lizard and cow categories only 924 of 1000 and 987 of 1000 samples are recognized.

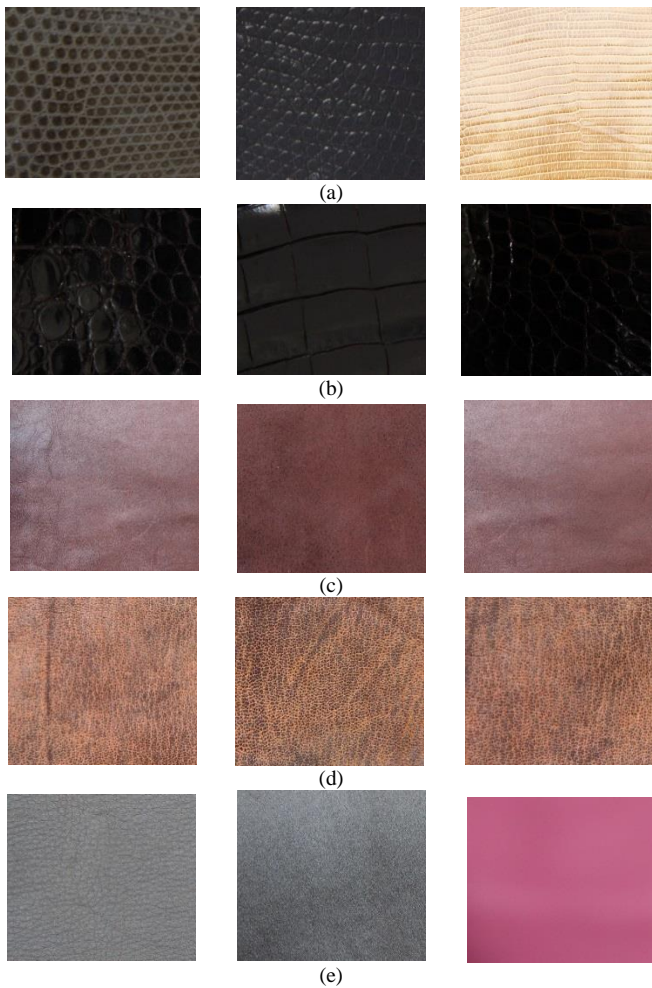


Fig. 3. Sample of leather images (a) monitor lizard (b) crocodile (c) sheep (d) goat (e) cow.

In the case of monitor lizard category, 74 samples falsely classified as crocodile, 1 samples falsely classified as sheep, 1 samples falsely classified as cow, and 12 of cow samples are falsely classified as monitor lizard skins. Because the false positive of monitor lizard samples which classified as crocodile skins, the precision of crocodile category becomes 0.9668 and the sensitivity of monitor lizard category becomes 0.9240. The accuracy of monitor lizard category also drops to 0.9592.

**B. Result of Feature Extraction using Pre-Trained CNN**

For the feature extraction using pre-trained CNN, input image is resized to 227x227 to fit CNN input layer. Authors use the eighth layer of AlexNet which is the last fully connected layers before output as activation function to extract features in order to gain rich features for classification. Another reason is to make the features compact because the eighth layer produces 1000 features and the other fully connected layers produce 4096 features. Features are classified using SVM. Confusion matrix and statistical measurement per category for leather classification scheme using pre-trained CNN is shown in Table II.

TABLE I. CONFUSION MATRIX AND STATISTICAL MEASUREMENT OF LEATHER CLASSIFICATION USING HAND-CRAFTED FEATURE EXTRACTION

Confusion Matrix					
Actual	Predicted				
	Monitor Lizard	Crocodile	Sheep	Goat	Cow
Monitor Lizard	924	74	1	0	1
Crocodile	0	1000	0	0	0
Sheep	0	0	150	0	0
Goat	0	0	0	7	0
Cow	12	0	1	0	987
Statistical Measurement					
Accuracy	0.9592	0.9828	0.9998	1.0000	0.9933
Specificity	0.9944	0.9657	0.9997	1.0000	0.9995
Sensitivity	0.9240	1.0000	1.0000	1.0000	0.9870
Precision	0.9940	0.9668	0.9993	1.0000	0.9995

TABLE II. CONFUSION MATRIX AND STATISTICAL MEASUREMENT OF LEATHER CLASSIFICATION USING PRE-TRAINED CNN

Confusion Matrix					
Actual	Predicted				
	Monitor Lizard	Crocodile	Sheep	Goat	Cow
Monitor Lizard	998	0	0	0	2
Crocodile	0	1000	0	0	0
Sheep	0	0	150	0	0
Goat	0	0	0	7	0
Cow	0	0	0	0	1000
Statistical Measurement					
Accuracy	0.9990	1.0000	1.0000	1.0000	0.9995
Specificity	1.0000	1.0000	1.0000	1.0000	0.9991
Sensitivity	0.9980	1.0000	1.0000	1.0000	1.0000
Precision	1.0000	1.0000	1.0000	1.0000	0.9991

TABLE III. LEATHER CLASSIFICATION PERFORMANCE COMPARISON

Datasets	Leather Classification Method	
	Hand-crafted features (color moments + GLCM)	Features from pre-trained CNN (8 <sup>th</sup> fully connected layer)
Average Accuracy	0.9870	0.9997
Average Specificity	0.9919	0.9998
Average Sensitivity	0.9822	0.9996
Average Precision	0.9919	0.9998
Average Computational Time	0.3793 seconds	0.2015 seconds

From Table II, all test samples are perfectly classified except in the monitor lizard category. In the case of monitor lizard category, 998 of 1000 samples are recognized and only two samples are falsely classified as cow skins. This small mistakes means, the features extracted from pre-trained CNN can exploit the characteristic of each leather category. Furthermore, the feature selection procedure already handled by pre-trained CNN which learned from large categories of image.

The statistical measurements per category in Tables I and II are averaged to calculate the whole performance of the proposed method. Authors also measure the computational time and compare the result from both classification schemes. From Table III, the hand-crafted features extraction scheme produces lower performance than pre-trained CNN. The computational time for hand-crafted features is also longer than pre-trained CNN despite only use small number of features. In the hand-crafted feature extraction, input image needs to be transformed into GLCM matrix before calculating the statistical texture features while in the feature extraction using pre-trained CNN, the features are directly extracted from input image with activation function which are the weights from specific fully connected layer.

The outstanding classification performance of the proposed method mainly caused by utilizing pre-trained CNN in this case AlexNet that designed for harder classification problem with 1000 categories of objects and trained using 1.2 millions of images. Therefore, for smaller and simpler five categories classification task can produce good performance when using pre-trained CNN as feature extractor.

## V. CONCLUSION

In this research, a method to classify tanning leather image is proposed. The classification process is used to ensure the quality of leather craft materials produced by leather tanning factories and used by leather craft industries by recognizing the type of animal skin. The method uses tanning leather image as input then perform feature extraction using pre-trained CNN. The classification is done using SVM into five leather categories. The performance evaluation shows that the proposed method can classify leather with good accuracy and precision. The proposed method also superior to hand-crafted classification scheme. For future works, authors aim to add more types of leather to be classified, improve the classification performance, speed-up the computational time, and implement the method in mobile application.

## ACKNOWLEDGMENT

This research is supported by Institute of Research and Development (LPP) Universitas Ahmad Dahlan research grant no. PP-001/SP3/LPP-UAD/IV/2017.

## REFERENCES

- [1] L. Wang and C. Liu, "Tanning Leather Classification using an Improved Statistical Geometrical Feature Method," in *2007 International Conference on Machine Learning and Cybernetics*, 2007, pp. 1765–1768.
- [2] Q. Wang, H. Liu, J. Liu, and T. Wu, "A new method for leather texture image classification," in *[1992] Proceedings of the IEEE International Symposium on Industrial Electronics*, pp. 304–307.
- [3] C. Kwak, J. A. Ventura, and K. Tofang-Sazi, "A neural network approach for defect identification and classification on leather fabric," *J. Intell. Manuf.*, vol. 11, no. 5, pp. 485–499, 2000.
- [4] C. Yeh and D.-B. Perng, "Establishing a Demerit Count Reference Standard for the Classification and Grading of Leather Hides," *Int. J. Adv. Manuf. Technol.*, vol. 18, no. 10, pp. 731–738, Nov. 2001.
- [5] F. P. Lovregine, A. Branca, G. Attolico, and A. Distanto, "Leather inspection by oriented texture analysis with a morphological approach," in *Proceedings of International Conference on Image Processing*, pp. 669–671.
- [6] K. Hoang, W. Wen, A. Nachimuthu, and X. L. Jiang, "Achieving automation in leather surface inspection," *Comput. Ind.*, vol. 34, no. 1, pp. 43–54, Oct. 1997.
- [7] H. Chen, "The Research of Leather Image Segmentation Using Texture Analysis Techniques," *Adv. Mater. Res.*, vol. 1030–1032, pp. 1846–1850, Sep. 2014.
- [8] R. Viana, R. B. Rodrigues, M. A. Alvarez, and H. Pistori, "SVM with Stochastic Parameter Selection for Bovine Leather Defect Classification," in *Advances in Image and Video Technology*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 600–612.
- [9] M. Jawahar, N. K. C. Babu, and K. Vani, "Leather texture classification using wavelet feature extraction technique," in *2014 IEEE International Conference on Computational Intelligence and Computing Research*, 2014, pp. 1–4.
- [10] A. Branca, M. Tafuri, G. Attolico, and A. Distanto, "Automated system for detection and classification of leather defects," *Opt. Eng.*, vol. 35, no. 12, p. 3485, Dec. 1996.
- [11] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *Proc. 2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 580–587, Jun. 2014.
- [13] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2015, pp. 1–6.
- [14] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [15] R. Maurya, Surya Kant Singh, A. K. Maurya, and A. Kumar, "GLCM and Multi Class Support vector machine based automated skin cancer classification," in *2014 International Conference on Computing for Sustainable Global Development (INDIACom)*, 2014, pp. 444–447.
- [16] G. Preethi and V. Sornagopal, "MRI image classification using GLCM texture features," in *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)*, 2014, pp. 1–6.
- [17] M. Hall-Beyer, "Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales," *Int. J. Remote Sens.*, vol. 38, no. 5, pp. 1312–1338, Mar. 2017.
- [18] S. Mukherjee and S. Pandey, "Road Surface Classification Using Texture Synthesis Based on Gray-Level Co-occurrence Matrix," Springer, Singapore, 2017, pp. 323–333.
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Curran Associates Inc., pp. 1097–1105, 2012.
- [20] ImageNet. Large-scale hierarchical image database. Available online at <http://www.image-net.org>. Accessed on November 22<sup>nd</sup>, 2017.
- [21] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.
- [22] C.-C. Chang and C.-J. Lin, "LIBSVM," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

# Iteration Method for Simultaneous Estimation of Vertical Profiles of Air Temperature and Water Vapor with AQUA/AIRS Data

Kohei Arai

Department of Information Science  
Saga University  
Saga City, Japan

**Abstract**—Iteration method for simultaneous estimation of vertical profiles of air temperature and water vapor with the high spectral resolution of sounder of AQUA/AIRS data is proposed. Through a sensitivity analysis based on the proposed method for the several atmospheric models simulated by MODTRAN, it is found that the proposed method is superior to the conventional method by 41.4% for air temperature profile and by 88.9% for relative humidity profile.

**Keywords**—Inversion; tropopause; AQUA; AIRS; Air temperature; sounder; MODTRAN

## I. INTRODUCTION

Atmospheric sounding can be improved by using the high spectral resolution of sounder, such as AQUA<sup>1</sup>/AIRS<sup>2</sup> onboard AQUA satellite, IASI<sup>3</sup>, instead of HIRS<sup>4</sup>, TOVS<sup>5</sup> used in the past [1]. Estimation of air-temperature profile with AQUA/AIRS data on the tropospheric boundary is tried [2]. Meanwhile, water vapor and air-temperature profile estimation with AIRS data based on Levenberg-Marquadt is proposed [3]. On the other hand, a sensitivity analysis for air temperature profile estimation method around the tropopause using simulated AQUA/AIRS data is carried out [4]. Also, a method for water vapor profile retrievals by means of minimizing difference between estimated and brightness temperature derived from AIRS data and radiative transfer model is proposed [5].

These sensors have large number of channels, and have large amount of atmospheric sounding information in the measurement data. However, for the retrieval of air temperature profile, it is not practical nor an advantage to use all spectral points. Therefore, it is important for this work to eliminate those redundant channels whose information does not add to the final retrieval accuracy and even before for the sake of efficiency, those channels potentially contaminated by solar radiation or significantly affected by other gases (not required for temperature profiling). Because of the influence due to the

thermal radiation from earth's surface and the sharp variation of air temperature at around tropopause<sup>6</sup>, on the other hand, the retrieval accuracies of air temperature at the surface and around the tropopause are not so high (~4 K) in the previous works<sup>7</sup> even using the high spectral resolution of sounder. One of the factors for this result is caught by insufficient channels selection. Some of redundant channels have to be removed from the air temperature profile estimation.

In this paper, a simultaneous estimation of vertical profiles of air temperature and water vapor with AQUA/AIRS data is proposed. There is a relation between vertical files of air temperature and water vapor. For instance, when sea surface temperature is raised then water vapor in the atmosphere is increased accordingly. Vertical profiles estimation accuracies can be improved if the relation is used in the estimation method. In order to confirm the effect of simultaneous estimation, MODTRAN<sup>8</sup> (Moderate Resolution Transmittance Code) is used for simulation.

The following section describes the theoretical background and related research works followed by the proposed method. Then, sensitivity analysis is described followed by simulation result. Finally, conclusion is described with some discussions and future research work.

## II. PROPOSED METHOD AND RELATED STUDIES

### A. Theoretical Background and Related Studies

The general forward model of the equation mapping the state (atmospheric profile) into measurement space (satellite-measured radiance or brightness temperature spectrum) is expressed as follows [6]:

$$y = F(x) + \varepsilon \quad (1)$$

Where,  $y$  is the measurement vector,  $F(x)$  is the forward model operator for a given state  $x$ , and  $\varepsilon$  is the measurement error.

The measurement error characteristics should be known, and the measurements  $y$  should be corrected before using them

<sup>1</sup> <https://aqua.nasa.gov/>

<sup>2</sup> [https://ja.wikipedia.org/wiki/Aqua\\_\(%E4%BA%BA%E5%B7%A5%E8%A1%9B%E6%98%9F\)](https://ja.wikipedia.org/wiki/Aqua_(%E4%BA%BA%E5%B7%A5%E8%A1%9B%E6%98%9F))

<sup>3</sup> <https://www.eumetsat.int/website/home/Satellites/CurrentSatellites/Metop/MetopDesign/IASI/index.html>

<sup>4</sup> <https://www.eumetsat.int/website/home/Satellites/CurrentSatellites/Metop/MetopDesign/HIRS/index.html>

<sup>5</sup> <https://eosweb.larc.nasa.gov/content/tovs>

<sup>6</sup> <https://en.wikipedia.org/wiki/Tropopause>

<sup>7</sup> <https://climatedataguide.ucar.edu/climate-data/airs-and-amsu-tropospheric-air-temperature-and-specific-humidity>

<sup>8</sup> <http://modtran.spectral.com/>

in the retrieval. Given a reasonable air temperature profile for  $x$ , equation (1) can be approximately linearized as follows:

$$y - y_0 = K_0(x - x_0) \quad (2)$$

Where,  $K_0 = \partial F(x) / \partial x$  is the weight function matrix with respect to  $x_0$ , and  $x_0$  is suitable reference state.

For the retrieval of air temperature profile from brightness temperature measurements  $y$ , the inverse problem associated with (1) is proposed by the concept of Bayesian optimal estimation<sup>9</sup> described by Rodgers [6] as follows:

$$x = x_a + (K^T S_\epsilon^{-1} K + S_a^{-1})^{-1} K^T S_\epsilon^{-1} (y - F(x_a)) \quad (3)$$

Where,  $x_a$  is a priori profile for air temperature,  $S_a$  is the a priori error covariance matrix and  $S_\epsilon$  is the measurement error covariance matrix.  $K$  is the weighting matrix.  $F(x_a)$  are the brightness temperatures from simulation with respect to  $x_a$ . Considered the problem is nonlinear, an iterative optimal estimation is selected as follows:

$$x_{i+1} = x_a + (K_i^T S_\epsilon^{-1} K_i + S_a^{-1})^{-1} K_i^T S_\epsilon^{-1} ((y - F(x_i)) + K_i(x_i - x_a)) \quad (4)$$

Where, subscript  $i$  is the iteration index. The optimum can be obtained only by 2 ~ 4 iterations because the problem is moderately nonlinear. This is the conventional method for vertical profile estimation with thermal infrared sounder data. Namely, vertical profile of air temperature is estimated separately with that of water vapor, independently.

With respect to high spectral resolution of sounder, thousands of measurements data are obtained in accordance with the channels. Air temperature profile ( $x$ ) to only a few dozens of air temperatures at different altitude levels is usually dispersed. It is not practical nor an advantage to use all spectral points. Some of them, which information are not required for temperature profiling, even decrease the retrieval accuracies. Thus, it is important to sufficiently select channels suitable for the retrieval of air temperature profile.

### B. Principle of Vertical Profile of Air Temperature and Water Vapor Estimations

Fig. 1 is spectral absorption characteristic in thermal IR region. The region 620 ~ 740  $\text{cm}^{-1}$   $\text{CO}_2$  absorption bands can be used for the retrieval of air temperature profile. More accurate reduction of the abundant channels is adopted the Information Content (IC) measure<sup>10</sup> as described by Rodger [7].

It, however, generally used in pre-process for the inverse scheme. It also means that the channels cannot be changed as the inverse scheme described as above to retrieve the air temperature profile after the pre-process for channels selection is performed. From the line-by-line computation for absorbance coefficient, on the other hand, absorbance

coefficient is not only depended on the spectrum, but also associated with the air temperature [8]. Meanwhile, it is nonlinear between the absorbance coefficient and the air temperature. One channel which information can add best to the retrieval accuracies at the certain temperature may not be at the other temperature. This relation becomes more significant at around the tropopause of which the sharp equatorial structure with the sharp temperature changes. One of the factors, thus, that result in the difficulty to improve the retrieval accuracies of air temperature at around tropopause is that the channels cannot change in the inverse scheme with respect to the different retrieval temperature profile.

As mentioned before, there is cross link between air temperature and water vapor. Therefore, it may be possible to improve estimation accuracy by using the relation. Iteration method for simultaneous estimation of air temperature and water vapor profiles is then proposed.

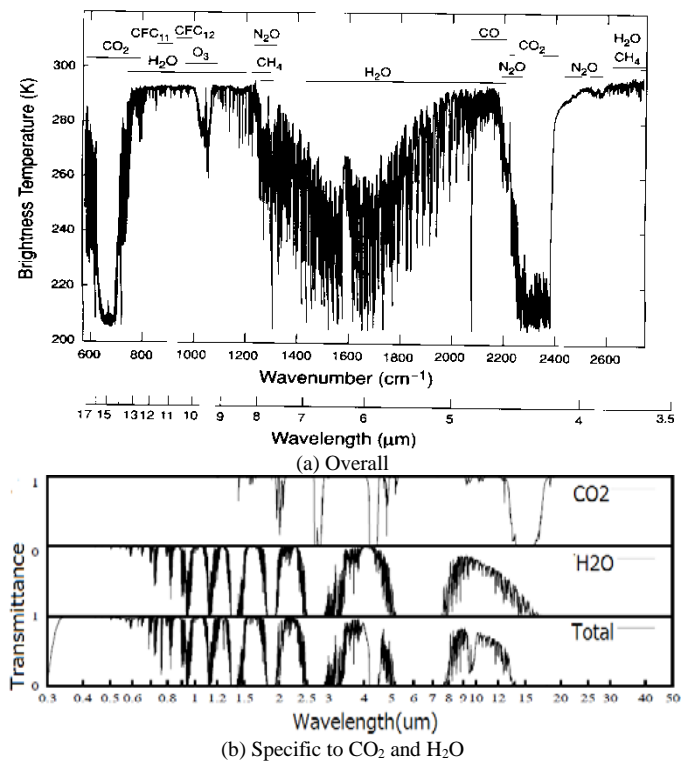


Fig. 1. Spectral absorption characteristic in thermal IR region.

## III. PROPOSED METHOD

### A. Weighting Function

The weighting function is expressed as follows:

$$K = \partial T_b / \partial T \quad (5)$$

Where,  $T_b$  denotes brightness temperature while  $T$  denotes air temperature. At sensor brightness temperature,  $T_b$  and air temperature,  $T$  can be estimated with MODTRAN. Fig. 2(a) shows an example of the input parameters of MODTRAN while Fig. 2(b) shows the output of  $T_b$  and  $T$  come out from MODTRAN.

<sup>9</sup> [https://en.wikipedia.org/wiki/Bayesian\\_optimization](https://en.wikipedia.org/wiki/Bayesian_optimization)

<sup>10</sup> [https://en.wikipedia.org/wiki/Information\\_criterion](https://en.wikipedia.org/wiki/Information_criterion)

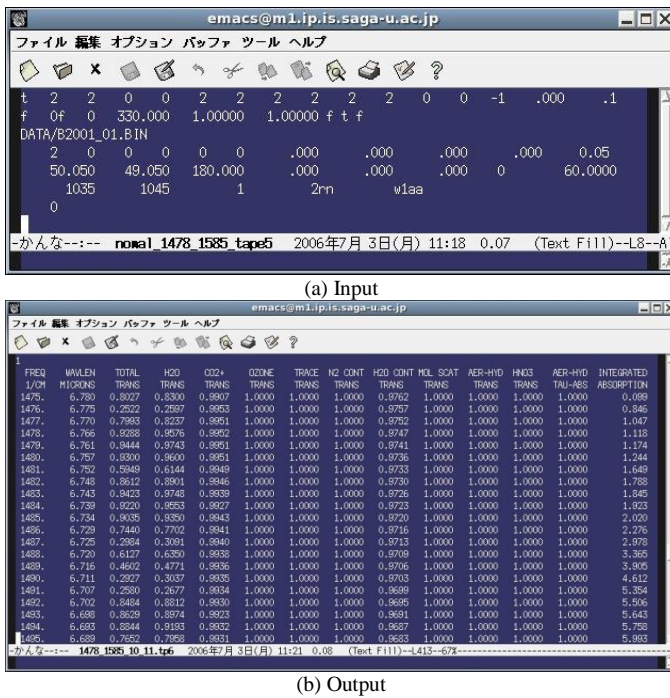


Fig. 2. Example of input parameters and output of MODTRAN.

Thus, atmospheric transparency can be estimated with MODTRAN followed by weighting function because the derivative of the transparency is weighting function. Fig. 3 shows an example of estimated atmospheric transparency and weighting function.

If the input parameter of certain wave number is selected for MODTRAN, then weighting function can be estimated accordingly as shown in Fig. 4.

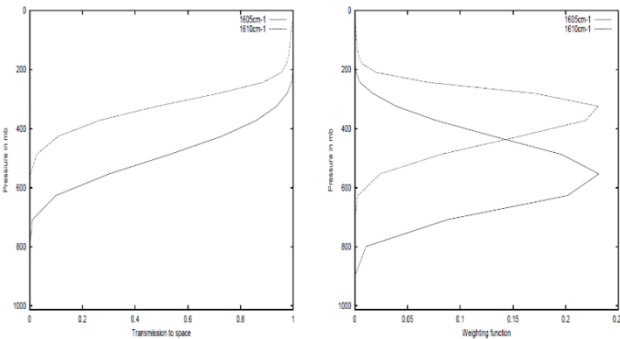


Fig. 3. Example of estimated atmospheric transparency and weighting function.

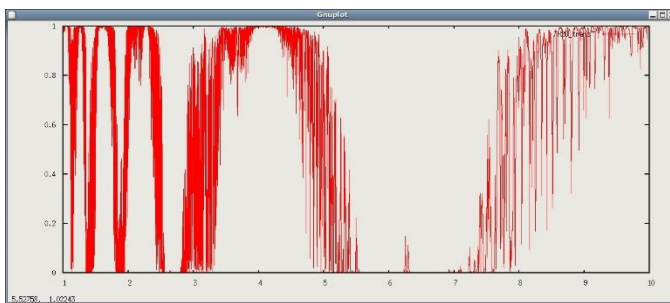


Fig. 4. Example of atmospheric transparency.

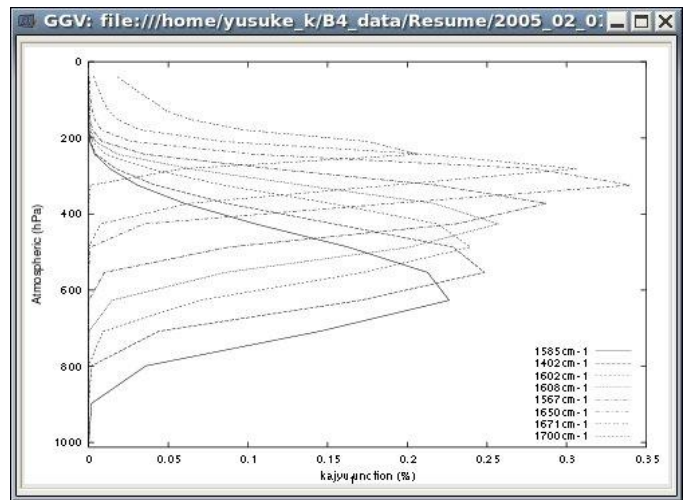
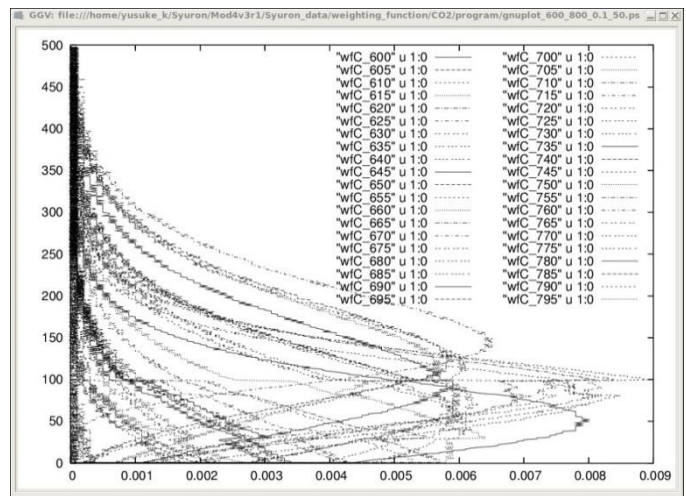
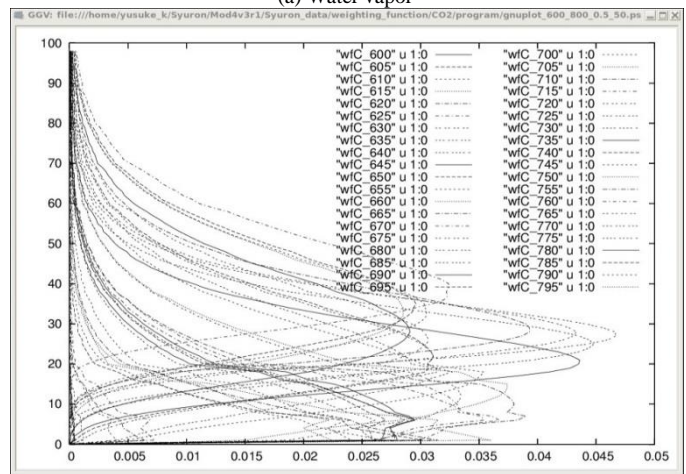


Fig. 5. Weighting functions for the wave number, 1478, 1483, 1508, 1514, 1519, 1541, 1544, 1558, 1585  $\text{cm}^{-1}$ .

When the wave number of 1478, 1483, 1508, 1514, 1519, 1541, 1544, 1558, 1585  $\text{cm}^{-1}$  is selected, then the weighting functions are estimated as shown in Fig. 5.



(a) Water vapor



(b) Air temperature

Fig. 6. Finding most appropriate wave number for estimation of water vapor and air temperature profiles.

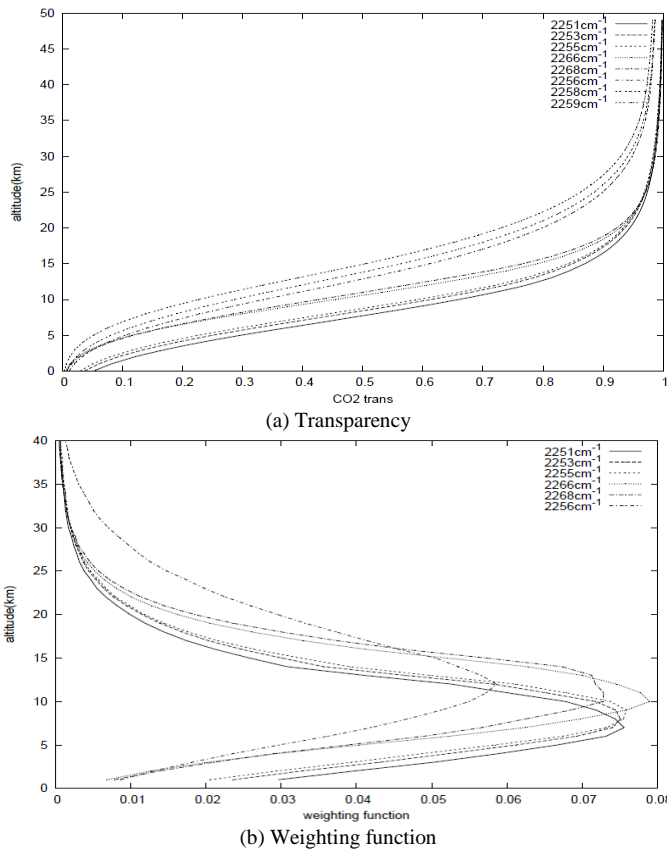


Fig. 7. Relation between CO<sub>2</sub> transparency and altitude while weighting function of CO<sub>2</sub>.

This is the same thing for estimation of weighting functions at arbitrary wave numbers as shown in Fig. 6. When the wave number is selected at the water vapor absorption bands, then weighting function for water vapor profile estimation can be selected as shown in Fig. 6(a). That is the same thing for air temperature profile estimation. When the wave number is selected at the air temperature absorption bands, then weighting function for air temperature profile estimation can be selected as shown in Fig. 6(b).

As shown in Fig. 4, atmospheric transparency due to carbon dioxide of air temperature is calculated with MODTRAN. Fig. 7(a) shows the relation between CO<sub>2</sub> transparency and altitude while weighting function of CO<sub>2</sub> is shown in Fig. 7(b).

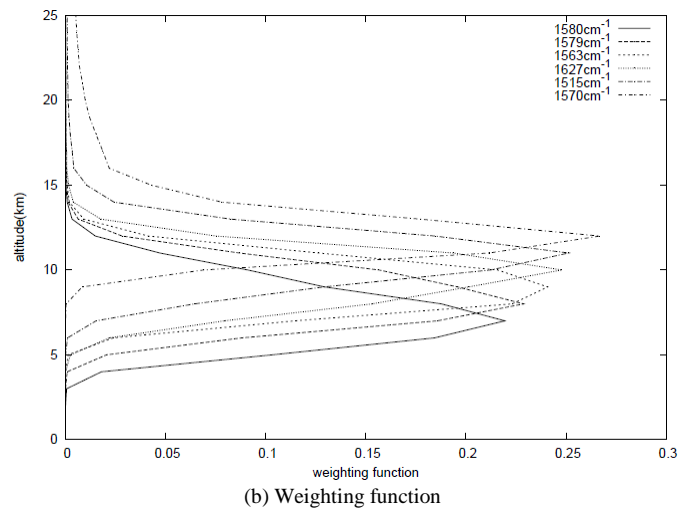
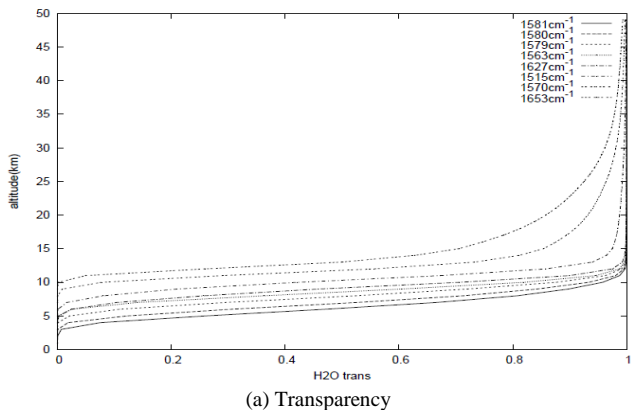


Fig. 8. Relation between H<sub>2</sub>O transparency and altitude while weighting function of H<sub>2</sub>O.

On the other hand, Fig. 8(a) shows the relation between H<sub>2</sub>O transparency and altitude while weighting function of H<sub>2</sub>O is shown in Fig. 8(b).

### B. Relation between Water Vapor and Air Temperature

It is possible to calculate air temperature and precipitable water, or water vapor content in the atmosphere by using MODTRAN. Fig. 9(a) shows the relation for Mid-Latitude Summer in day time model while Fig. 9(b) shows the relation for Mid-Latitude Winter in night time model, respectively.

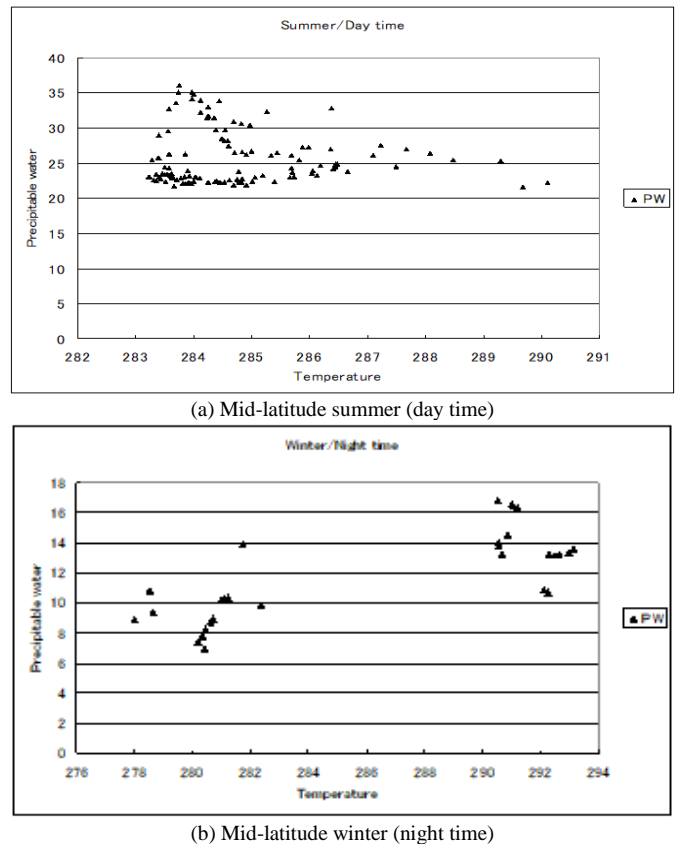


Fig. 9. Relation between air temperature and water vapor.



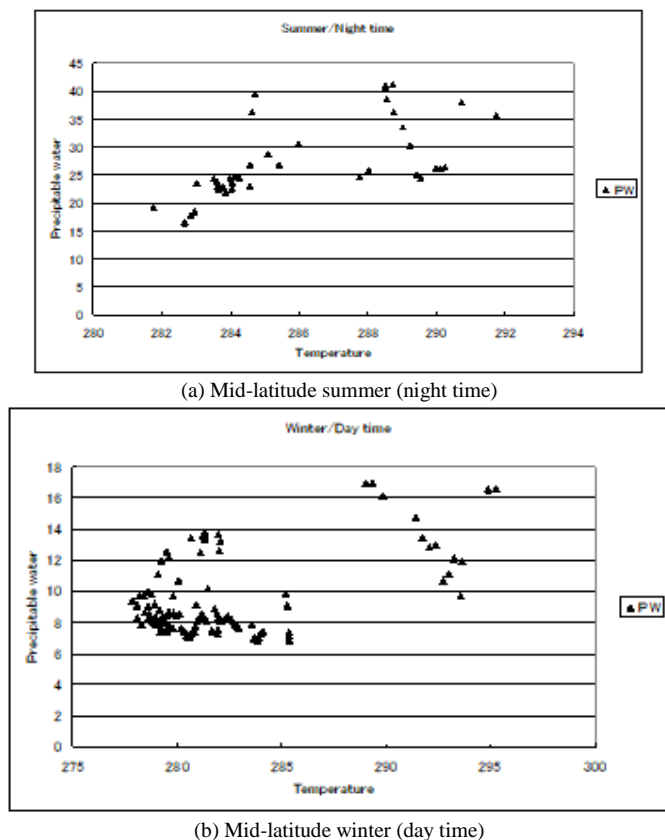


Fig. 10. Relation between air temperature and water vapor.

Meanwhile, Fig. 10(a) shows the relation for Mid-Latitude Summer in night time model while Fig. 10(b) shows the relation for Mid-Latitude Winter in day time model, respectively.

As shown in Fig. 9 and 10, it is obvious that there is relation between air temperature and water vapor. Therefore, it may be possible to improve estimation accuracy by using the relation. Iteration method for simultaneous estimation of air temperature and water vapor profiles is then proposed.

The proposed method is an iteration method. Firstly, initial value of air temperature and relative humidity is selected. Then, air temperature profile is estimated with the initial value of relative humidity by the conventional method. After that, relative humidity is estimated with the estimated air temperature profile just mentioned above by the conventional method. These processes are repeated iteratively until the residual error reaches to the convergence radius.

#### IV. ACCURACY EVALUATION

##### A. Preliminary Results

When the wave number of 1478, 1483, 1508, 1514, 1519, 1541, 1544, 1558, 1585  $\text{cm}^{-1}$  is selected, then the weighting functions are estimated as shown in Fig. 5. Fig. 11 shows water vapor (Relative Humidity) profile estimated with the conventional method. In the figure, solid line shows MODTRAN derived profile while dotted line shows the estimated profile with the conventional method. There are some differences as shown in Fig. 11 clearly.

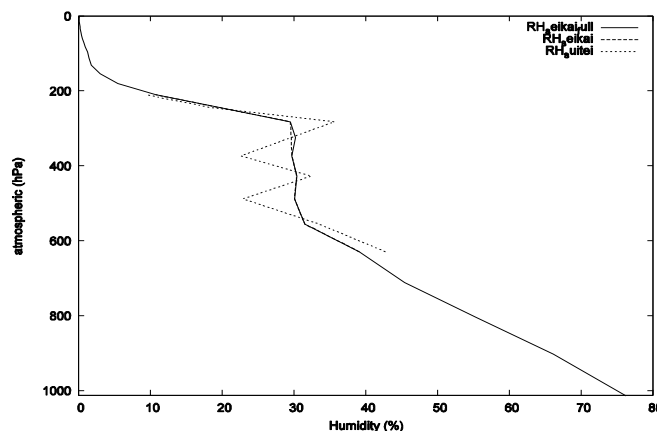


Fig. 11. Estimated water vapor profile.

TABLE I. EXAMPLE OF THE ESTIMATION ACCURACY EVALUATION RESULT FOR RELATIVE HUMIDITY PROFILE FOR MID-LATITUDE SUMMER MODEL

Wave number $\text{cm}^{-1}$	Peak (km)	Correct Rh (%)	Estimated R h (%)	Diff.
1585	4	39.049	42.640	3.591
1402	5	31.417	33.380	1.963
1602	6	29.980	22.828	7.152
1608	7	30.310	32.359	2.049
1567	8	29.630	22.573	7.057
1650	10	29.440	35.553	6.113
1671	11	19.480	18.332	1.148
1700	12	10.694	9.329	1.365

Estimation accuracy is calculated through comparisons between the estimated relative humidity and the MODTRAN derived humidity. An example of the estimation accuracy evaluation result is shown in Table I.

##### B. Results for the Proposed Method

Fig. 12(a) and (b) shows air temperature (a) and relative humidity (b) profiles of MODTRAN derived and the estimated by the conventional method as well as the estimated by the proposed method, respectively.

TABLE II. RMSE: ROOT MEAN SQUARE ERROR OF THE CONVENTIONAL AND THE PROPOSED METHODS FOR ESTIMATION OF AIR TEMPERATURE (T) AND RELATIVE HUMIDITY (RH) PROFILES

Profile	Conventional	Proposed
(T)	1.031(K)	0.607(K)
(RH)	1.779(%)	0.197(%)

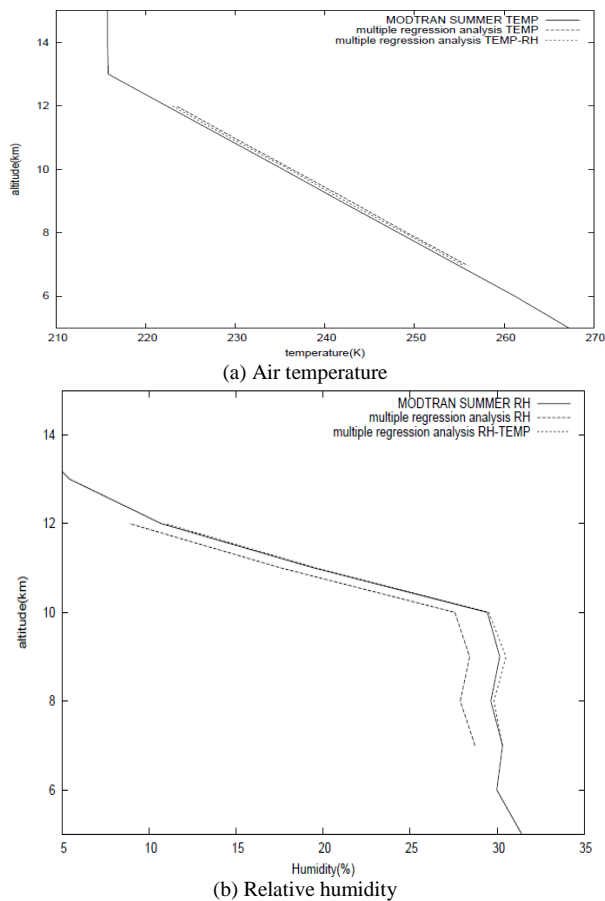


Fig. 12. Air temperature (a) and relative humidity (b) profiles of MODTRAN derived and the estimated by the conventional method as well as the estimated by the proposed method.

Table II shows RMSE: Root Mean Square Error of the conventional and the proposed methods for estimation of air temperature (T) and relative humidity (RH) profiles. It is obvious that the proposed method is superior to the conventional method by 41.4 % for air temperature profile and by 88.9 % for relative humidity profile.

## V. CONCLUSION

Iteration method for simultaneous estimation of vertical profiles of air temperature and water vapor with the high spectral resolution of sounder of AQUA/AIRS data is proposed. Through a sensitivity analysis based on the proposed method for the several atmospheric models simulated by MODTRAN, it is found that the proposed method is superior to the conventional method by 41.4 % for air temperature profile and by 88.9 % for relative humidity profile.

Further study is highly required for another experiment with a variety of dataset and atmospheric conditions.

## ACKNOWLEDGMENT

The author would like to thank Dr. Xing Ming Liang of NOAA/NESDIS and Mr. Naohisa Nakamizo of Saga University for their effort to conduct simulations and the experiments. Also, the author would like to thank all members of the fourth research group of Information Science Department of the Science and Engineering Faculty, Saga University for their valuable comments and suggestions as well as discussions.

## REFERENCES

- [1] Jeffrey, A.L., Elisabeth, W., and Gottfried, K., Temperature and humidity retrieval from simulated Infrared Atmospheric Sounding Interferometer(IASI) measurements, *J. Geop. Res.*, 107, 1-11, 2002.
- [2] Kohei Arai and Liang XM. Estimation of air-temperature profile with AQUA/AIRS data on the tropospheric boundary, Abstract, COSPAR A1.1, A-00715, 2006.
- [3] Kohei Arai and Naohisa Nakamizo, Water vapor and air-temperature profile estimation with AIRS data based on Levenberg -Marquadt, Abstract of the 50th COSPAR(Committee on Space Research/ICSU) Congress, A 3.1-0086-08,995, Montreal, Canada, July 2008
- [4] Kohei Arai and XingMing Liang, sensitivity analysis for air temperature profile estimation method around the tropopause using simulated AQUA/AIRS data, *Advances in Space Research*, 43, 3, 845-851, 2009.
- [5] Kohei Arai, Method for water vapor profile retrievals by means of minimizing difference between estimated and brightness temperature derived from AIRS data and radiative transfer model, *International Journal of Advanced Computer Science and Applications*, 3, 12, 145-148, 2012.
- [6] Rodgers, C.D., *Inverse method for atmospheres: theory and practice*, World Sci., Singapore, 2000.
- [7] Rodgers, C.D., Information content and optimization of high spectral resolution measurements. In *Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research II*, vol. 2830, pp. 136-147, Int. Soc. For Optical Eng., Bellingham, Wash.,1996.
- [8] Liou, K.N., *An introduction to atmospheric radiation*. Elsevier Science, USA, 2002.

## AUTHORS PROFILE

**Kohei Arai**, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post-Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission-A of ICSU/COSPAR since 2008. He received Science and Engineering Award of the year 2014 from the minister of the ministry of Science Education of Japan and also received the Best Paper Award of the year 2012 of IJACSA from Science and Information Organization: SAI. In 2016, he also received Vikram Sarabhai Medal of ICSU/COSPAR and also received 37 awards. He wrote 37 books and published 570 journal papers as well as 370 conference papers. He is Editor-in-Chief of International Journal of Advanced Computer Science and Applications as well as International Journal of Intelligent Systems and Applications. <http://teagis.ip.is.saga-u.ac.jp/>

# A Robust System for Noisy Image Classification Combining Denoising Autoencoder and Convolutional Neural Network

Sudipta Singha Roy

Institute of Info. and Comm.  
Technology  
Khulna University of Engineering &  
Technology  
Khulna, Bangladesh

Sk. Imran Hossain, M. A. H.  
Akhand

Dept. of Computer Science and  
Engineering  
Khulna University of Engineering &  
Technology  
Khulna 9203, Bangladesh

Kazuyuki Murase

Graduate School of Engineering  
University of Fukui  
Fukui 910-8507,  
Japan

**Abstract**—Image classification, a complex perceptual task with many real life important applications, faces a major challenge in presence of noise. Noise degrades the performance of the classifiers and makes them less suitable in real life scenarios. To solve this issue, several researches have been conducted utilizing denoising autoencoder (DAE) to restore original images from noisy images and then Convolutional Neural Network (CNN) is used for classification. The existing models perform well only when the noise level present in the training set and test set are same or differs only a little. To fit a model in real life applications, it should be independent to level of noise. The aim of this study is to develop a robust image classification system which performs well at regular to massive noise levels. The proposed method first trains a DAE with low-level noise-injected images and a CNN with noiseless native images independently. Then it arranges these two trained models in three different combinational structures: CNN, DAE-CNN, and DAE-DAE-CNN to classify images corrupted with zero, regular and massive noises, accordingly. Final system outcome is chosen by applying the winner-takes-all combination on individual outcomes of the three structures. Although proposed system consists of three DAEs and three CNNs in different structure layers, the DAEs and CNNs are the copy of same DAE and CNN trained initially which makes it computationally efficient as well. In DAE-DAE-CNN, two identical DAEs are arranged in a cascaded structure to make the structure well suited for classifying massive noisy data while the DAE is trained with low noisy image data. The proposed method is tested with MNIST handwritten numeral dataset with different noise levels. Experimental results revealed the effectiveness of the proposed method showing better results than individual structures as well as the other related methods.

**Keywords**—Image denoising; denoising autoencoder; cascaded denoising autoencoder; convolutional neural network

## I. INTRODUCTION

Categorization of objects from images is a complex perceptual task and is termed as image classification. Classification of images utilizes multispectral data. The underlying multispectral pattern of the data of each individual pixel is utilized as the quantitative basis for classification [1]. In the past decade, image classification has shown major advances in terms of classification accuracy. In recent times,

image classification models are rapidly being used in various application fields, such as handwritten numeral recognition [2], recognition of traffic signs from roadside boards [3]-[5], segmentation of Magnetic Resonance Image (MRI) [6], identification of chest pathology [7], face detection from images [8] and so on. Existing models are categorized into unsupervised and supervised modes.

Unsupervised classification based models try to find out the underlying representation in the input images without considering whether the images are labeled or not. One conventional model of this genre is stacked autoencoders (SAE) [9]-[11]. With an intention to learn features, SAE stacks shallow autoencoders which at first encodes the original input image to a vector of lower dimension and then decodes this vector to the original representation of the image. Shin et al. [12] showed the application of stacked sparse autoencoders (SSAEs) to classify medical images which made a noteworthy promotion in terms of classification accuracy. Norouzi et al. [13] inaugurated stacked convolutional restricted Boltzmann machine (SCRBM) where they applied a modified training process rather than the conventional one for individual restricted Boltzmann machine (RBM) and finally combined them in a stacked manner to implement the deep architecture. Later, Lee et al. [14] instigated another variant of deep belief network (DBN) called convolutional DBN (CDBN) by placing convolutional RBMs (CRBM) instead of traditional RBMs at each layer and then joined the layers in a convolutional structure to ensure the construction of a hierarchical model and it produced better feature representation [15]. With the practically identical considerations, Zeiler et al. [16], [17] modified traditional sparse coding technique [18] to build deconvolutional model that decomposes the input data in a convolutional way, at the same time, maintains a sparsity constraint. In contrast to conventional sparse coding technique, this approach produces mid-level delineations of data with more affluent learned features.

Unlike unsupervised classifiers, supervised classification based models require labeled data to complete their training process. In this category, deep neural network (DNN) does the task efficiently implementing the idea of human visual system.

Layer-wise pre-training and fine-tuning makes DNN successful in image processing tasks such as classification, feature extraction etc. Convolutional neural network (CNN) [19]-[22] is the most successful hierarchical deep neural network structure. Shared weights, three-dimensionally arranged and locally connected neurons make the architecture of CNN distinctive to ordinary neural networks and contribute to its superior performance to most of the image classification algorithms [23]. The unique characteristics of CNN such as weight sharing and preservation of the corresponding locality, which make the deep architecture the most suitable for 2D images to conserve a better epitome, are the outcome of using convolution and following subsampling layer. Right now, CNN based models are being used vastly in 2D material identification and various cases [3]-[8].

One major challenge in image classification tasks is the presence of noise that corrupts the original shape of the objects in the image and makes it difficult for the classifiers to be used in real life scenarios. Unlike human visual system, which is capable of classifying objects ignoring a certain amount of perturbation present in the image, these classifiers suffer in quality if the test image contains noise. Although DNN based methods outperformed others in image classification, their performances are deteriorated during classification of noisy images. However, it is quite impossible to work with noiseless images in practical cases. During acquisition and transition phases, corruption of digital images due to noise is common. As DNN based models are trained to work with noiseless images, their accuracy noticeably drops when they are applied in real life applications. The main reason that works behind the occurrence of this incident is the affection of the DNN based models to the training data. Because of this sensitive behavior towards the training data, often these models perform misclassification if the test data is subject to a significant amount of noise and distortion [24].

It is an open challenge to develop image classification systems for the real-life noisy environment. Lu and Weng [25] investigated different image classification models and finally came to a conclusion that denoising images prior classification is the best possible way to make the DNN based models more compatible with practical cases. Their survey gives the evidence of the fact that training classifiers with noisy images may enhance the precision a tad; however, it is not satisfiable. So, applying image denoising techniques before feeding the image to the classifier has become a compulsory to fit the DNN based classifiers in real life scenario.

A number of researches have been conducted to recover the true image from the noisy form by applying both spatial and transform domain [26]. Several pioneer image denoising researches used wavelet transformation techniques [27], partial differential equation based approaches [28]-[30], and conveyed scant coding approaches [18], [31], [32]. Singh et al. [33] introduced a multi-class classifier for images which are adulterated with Gaussian noise. To accomplish image denoising they utilized NeighShrink thresholding over the wavelet coefficients to wipe out wavelet coefficients which are responsible for the noise present in the image and find out just the useful ones. However, these denoising approaches face problems in case of heavily noised image and are

computationally complex. In the process of image denoising using spatial filtering techniques images gets blurred, where transfer domain filtering models are time-consuming as well as computationally complex.

Recently, artificial neural network (ANN) based models are being adopted in image denoising tasks. A variant of autoencoder (AE) named denoising autoencoder (DAE) [34], [35] has been introduced to serve the purpose of image denoising and shows a better performance compared to the traditional ones. In DAE the initial input gets corrupted by arbitrary noise then it is trained to restore the original image from its' corrupted version. In [36] Vincent et al. stacked a number of denoising autoencoders and established a deep network named stacked denoising autoencoder (SDAE) which is widely implemented for unsupervised learning. Agostinelli et al. [37] developed an adaptive multi-column DNN combining multiple stacked sparse DAEs (SSDAE), where the multi-column architecture empowers the model to deal with images corrupted by not one type but three different types of noises. Utilizing non-linear optimization technique, they figured out the most favorable column weights at first and then individual models were trained to make them anticipate those optimal weights. Incorporating the idea of AEs and convolutional operation Masci et al. [38] introduced convolutional autoencoder (CAE) which can preserve better spatial locality. CAE is based on CNN and it learns to reconstruct the images at the output end from the input image set applying convolutional approach so that the kernels convolve over the 2D images and at each layer generates more abstract feature maps. In order to use this convolutional structure of AEs for image denoising task, Gondara [39] deployed DAEs along with CAEs. She utilized the DAE, at first, to denoise medical images and then CAEs to generate a better representation of the images. Xu et al. [40] implemented a deep CNN architecture that can find out the features of blur degradation present in an image.

Du et al. [41] introduced stacked convolutional denoising autoencoder (SCDAE). To maintain a hierarchical structure, they arranged a stack of DAEs in a convolutional manner. Additionally, they embedded a whitening layer in front of each and every convolutional layer to enclose the input feature maps. Most recently, Roy et al. [42] applied convolutional denoising autoencoder (CDAE) followed by a DAE and arranged them in a cascaded manner, rather than in a stacked way to deal with data subject to massive noise. They showed that if two AE based models are individually trained to denoise images subject to regular level of noise, the cascaded architecture of them can show a great performance in case of denoising massive noisy images. Still, these models suffer from one limitation: their performances require the presence of a quite same proportion of noise in both training and testing dataset. To fit the DNN based image classifiers in real life scenario, models should be able to work with variable level of noise i.e., regular to massive level of noise.

The aim of this study is to develop a robust image classification system which performs well in any noise level with a minimized computational cost, at the same time, omits the requirement of arranging multiple training of the system with images containing variable proportion of noise separately

to deal with images subject to variable level of noise. The proposed method first trains a DAE with low-level noise injected images and a CNN with noiseless native images independently. Then it arranges the two trained models in three different combinational ways: CNN, DAE-CNN, and DAE-DAE-CNN. Finally, it combines the outcomes of these three combinations for system outcome. The motivation of such arrangement is the adaptation of noise in DAE and image classification ability of CNN. Since CNN is trained with native images without noise it is well for noiseless image classification. On the other hand, DAE-CNN and DAE-DAE-CNN structures perform well for low-level and high-level noisy cases, respectively [42]. In DAE-CNN, DAE first removes noises from noisy input images and then CNN is fed with these restored images for classification purpose. In DAE-DAE-CNN, two pre-trained DAEs are cascaded together and followed by a CNN. First DAE denoises the input images which are further filtered by the next DAE and therefore CNN gets the restored images, which is better suited to classify the test images in case they are adulterated with massive noise even though both DAEs are same and trained with the low noise level. The winner-take-all combination gives system output emphasizing confidence of individual structure and thus the proposed model performs well to classify images for noiseless to high-level noise cases. In this study, the proposed method is tested with MNIST handwritten numeral dataset and its performances are compared with other related methods.

The rest of the paper is designed as follows. Section II describes the proposed robust system for noisy image classification along with some preliminaries for better understanding. Section III shows the result of the proposed method as well as performance comparison with some other existing related research works. Finally, a brief conclusion of this work is presented in Section IV.

## II. ROBUST SYSTEM FOR NOISY IMAGE CLASSIFICATION

In practical life, image classification models suffer from noise, injected in the image while acquiring and transmitting, as well as other imperfections existing in the image. Existing systems are found to be effective for a fixed level of noise on which they are trained. On the other hand, this study investigated a robust system which performs well in classification of images in spite of the varying level of noise present in the image. The proposed method first trains a DAE with low noise level and a CNN independently. The main novelty of the proposed system is the innovative combinational arrangements of the trained DAE and CNN for three different structures. This section first describes the training of individual DAE and CNN briefly; and then explains the proposed system.

### A. Review of DAE and CNN

The main computational components of the proposed system are DAE and CNN. Well studied standard DAE and CNN architectures are considered in this study. For a better understanding as well as to make the paper self-contained, DAE and CNN are presented briefly.

1) *Denoising Autoencoder (DAE)*: Autoencoder (AE) is a three-layered neural network, which is unsupervised and deterministic in nature. It maps the input into a hidden

representation through encoder and then decoder maps it back to a reconstruction, which is of the same shape of the input. DAE, unlike basic AE, forces the hidden layer to capture the information about how the inputs are statistically dependent on each other instead of learning trivial features [14]. This is done by the corruption of input dataset  $x$  into  $\tilde{x}$  stochastically ( $q_D(\tilde{x}|x)$ ) which is mapped into a hidden representation  $y$ .

$$y = \mu(W(q_D(\tilde{x}|x) + b)), \quad (1)$$

where  $W$  is the weight of the input-hidden layer and  $q_D$  represents the type of distribution with a certain probability  $D$ .  $q_D$  depends on two parameters: one, the distribution of the original input  $x$  and two, the type of noise corrupting the images. Clinched alongside practical scenarios, binomial noise is utilized while working with black and white pictures whereas, to color pictures uncorrelated Gaussian noise is superior suiting. At that point,  $\tilde{x}$  is mapped to a low dimensional hidden depiction  $y$  using nonlinear deterministic function  $\mu$ . Finally, this hidden representation gets mapped into a reconstruction  $z$  which has as close as possible resemblance to input  $x$ . This process also passes through another nonlinear deterministic function  $\Phi$ .

$$z = \Phi(W'y + b') \quad (2)$$

where  $W'$  is the weight of the hidden-output layer. Thus, DAE is capable to generate representations of features, which is suitable for the classification task. The architecture of DAE is shown in Fig. 1. The training of DAE requires it to be fed with noisy images putting the corresponding native images at the output layer. During backpropagation, this model learns to filter out the underlying noise from the input image and reconstruct a noiseless one. The detailed description, as well as training of DAE, is available in the previous studies [42].

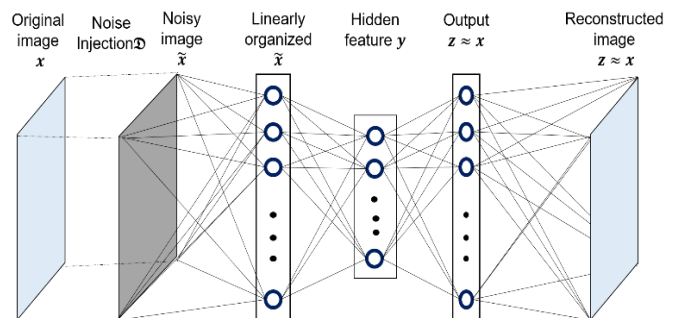


Fig. 1. Denoising Autoencoder (DAE) architecture.

2) *Convolutional Neural Network (CNN)*: CNN [19] is a variant of neural network popular for object detection and segmentation task. Regular neural network or multi-layer perceptron has some limitations: it suffers from overfitting; it ignores the fact that there is a strong correlation among neighborhood pixels and it is sensitive to any kind of transformation of the image. CNN overcomes these problems by ensuring spatial local connection, weight sharing, and subsampling. The operation of a CNN is done on the premises of two basic operations: convolution and subsequent subsampling. Convolution operation forces a kernel, which is an organization of weights and bias, to convolve over input

feature map (IFM) which in the end results in a convolved feature map (CFM). Throughout the convolution operation, the same kernel is applied to each and every small segment of each IFM, which is called the local receptive field (LRF), to acquire every specific point of the CFM. Throughout this process, both weights sharing among each and every position as well as the preservation of special locality are done simultaneously. From an IFM the CFM can be calculated by.

$$CFM_{(x,y)} = \tau(\sum_{i=1}^{K_h} \sum_{j=1}^{K_w} K_{(i,j)} * IFM_{(x+i,y+j)} + \beta) \quad (3)$$

where  $\tau$  and  $*$  symbolize the activation function and the 2-D convolution operation accordingly. The bias of the applied kernel  $K_h \times K_w$  is symbolized by  $\beta$ . To conduct the experiment here, relu is used as the activation function, whereas for every latent map single bias is used.

The feature map, obtained from convolution operation, is processed by applying the following subsampling layer in order to gain a simplified form. This procedure of simplification is accomplished by choosing important features from a locale and discarding whatever is left of the ones [41]. Having different sub-sampling methods available, throughout

the experiments here, max-pooling [21] has been utilized. Max-pooling operation picks the most important feature over non-covering sub-regions and this process can be defined as:

$$FM(x, y) = \mathcal{s}(\sum_{i=0}^{R-1} \sum_{j=0}^{C-1} CFM_{(xR-1+i,yC-1+j)}) \quad (4)$$

where  $\mathcal{s}$  symbolizes the max-pooling operation over the pooling locale and the size of the pooling area is represented as  $R \times C$  matrix.

Fig. 2 shows the most studied CNN architecture which is considered in this study. The CNN has two convolutional layers of filter size of  $5 \times 5$  and the subsampling layer with a pool size of  $2 \times 2$ . A subsampling layer follows each convolutional layer. The convolutional and pooling layers together extract the features of the image. There is a fully connected layer and the input of which is the output of the second subsampling layer. This layer uses the extracted feature to classify the image depending upon the training dataset. The parameters of the network, as well as the kernel, get updated during the training process until the desired accuracy is achieved. The detail description of CNN training is also available in the previous studies [2], [42].

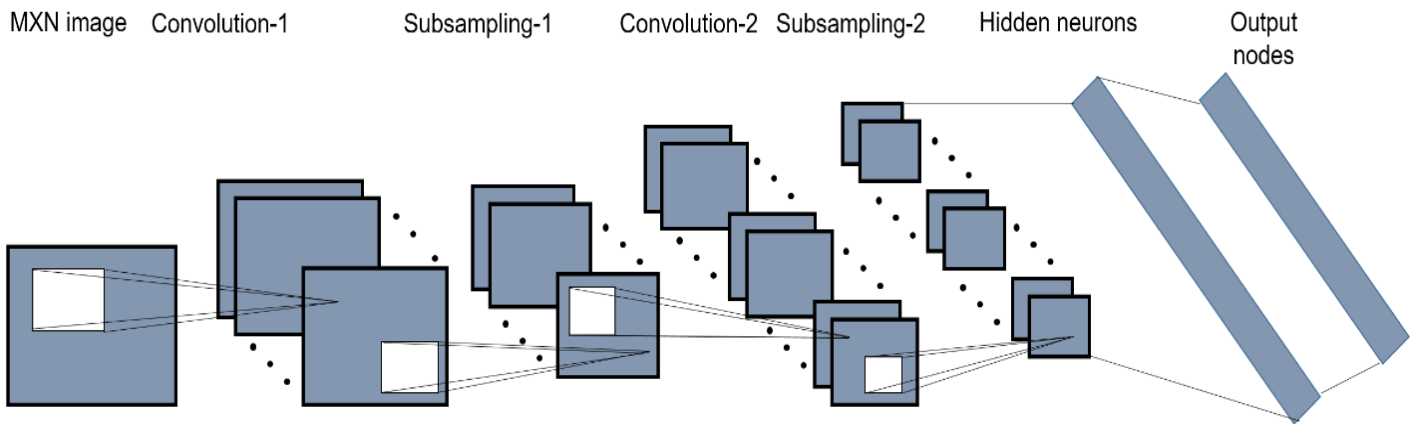


Fig. 2. CNN architecture for classification.

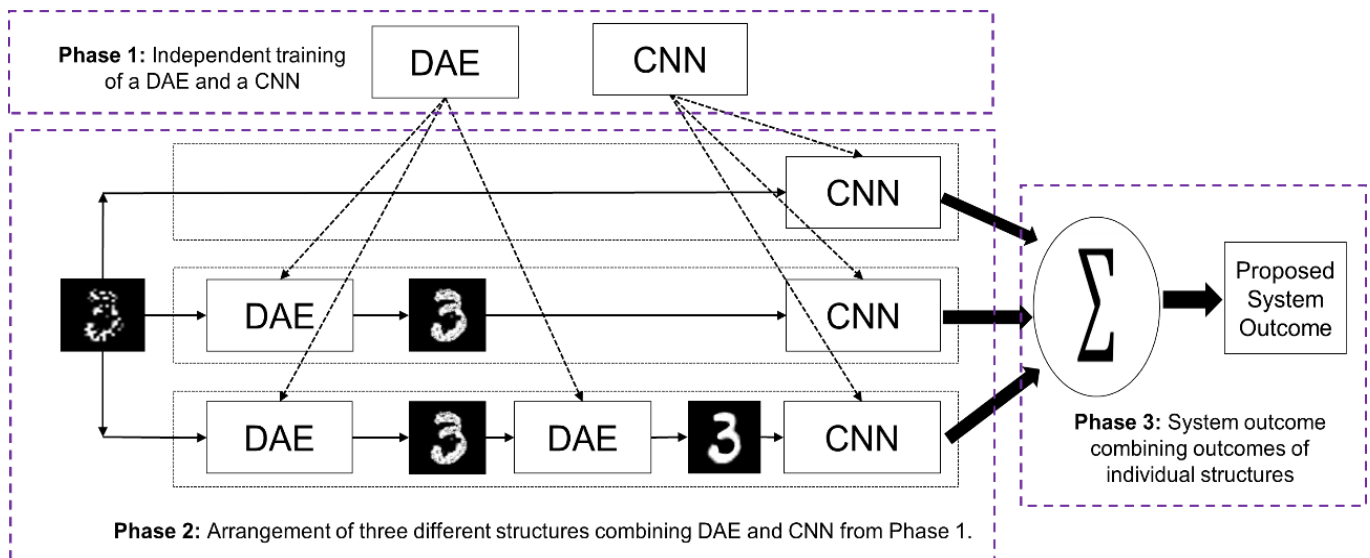


Fig. 3. Proposed robust system for noisy image classification combining three different structures with DAE and CNN.

### B. Proposed Robust System Combining Different Structures with DAE and CNN

Fig. 3 is the topological structure of the proposed Robust System based on DAE and CNN (RSDAECNN). Proposed RSDAECNN has three different functional phases. It trains a single DAE and a single CNN in Phase 1. In Phase 2, three different structures are arranged by copying the same DAE and CNN from Phase 1. In Phase 3, system outcome is prepared by combining outcomes of individual structures. Although proposed system consists of three DAEs and three CNNs in different structure layers, the DAEs and CNNs are the copy of same DAE and CNN trained in Phase 1. Thus, trainings of the DAE and the CNN in Phase 1 are the main computational elements in the proposed system.

Training of a single DAE and a single CNN only in Phase 1 makes the proposed system computationally efficient. The CNN is trained with native images and is used for classification purpose. The DAE is trained to restore the native image from the noisy image. In this study, the DAE is trained for regular level noise. Classification with CNN after restoring images through DAE might be helpful for noisy image classification. Since DAE is trained with regular level noise, different structures with different organizations of this DAE are managed to make the system adaptable to real-life environment where the noise level is not defined.

The main novelty of the proposed model is the innovative arrangements of pre-trained DAE and CNN to produce three different structures where each of the structures is responsible for dealing with images corrupted by a specific noise level. Each individual structure shown in Phase 2 has significant motivation to use in the proposed system. The CNN, common in all three structures, is used for classification purpose as CNN outperforms all other models in case of image classification [23]. To classify noiseless images CNN alone is good enough as CNN is trained with native images. This is the motivation of first structure with a CNN only in the system as seen in Phase 2 of Fig. 3. However, images corrupted by noise require a prior denoising step to improve the classification accuracy of the image classifier. For this purpose in the proposed system, DAE is used as the image denoiser. This DAE is trained only once with images corrupted by regular level noise (such as 20%). So, only a single DAE is sufficient enough to reconstruct the corresponding noiseless native form by filtering the images subject to regular level of noise. From this point of view, a DAE-CNN is placed as the second structure to emphasize the classification of images which are corrupted by regular level of noise. As the DAE and CNN are already trained separately, the DAE-CNN needs no further training. In DAE-CNN, DAE filters images subject to noise and then CNN classify the restored images.

A different structure DAE-DAE-CNN is developed to emphasize classification of images with massive noise because DAE-CNN structure is not sufficient enough to classify images in case of massive noise present in them. The DAE is trained with regular noise only and can't reconstruct native images which are corrupted with massive level of noises, such as if the percentage of noise in the images are around 50%. Roy et al. [42] showed that cascaded architectures of DAEs, where each

of the DAEs is trained with 20% noisy images, can reconstruct images of good quality even if the noise level present in the images is 50%. Following this idea, a cascaded DAE-DAE is arranged as the image denoiser in the third structure DAE-DAE-CNN. In DAE-DAE-CNN, both the DAEs are the same in terms of architecture as well as all the corresponding parameters as they are copies of trained DAE in Phase 1. CNN is also the duplicate of the trained CNN in Phase 1 as like other two structures. Therefore, no additional training is required for this structure. In DAE-DAE-CNN operation, at first the image is filtered by the first DAE, then the intermediate image is further filtered by the following DAE. So, the pre-trained CNN is sufficient enough to classify the reconstructed image from massive noisy images after they are filtered by DAE-DAE. However, this model doesn't suit in case the level of noise in the image is not that much because restoration through DAE-DAE might overshoot to different images.

The proposed robust system combines the outcomes of the three structures, which are specialized to different noise levels while classifying an image subject to unknown level of noise. Among the three structures, structure with CNN alone is best suited for noiseless image classification. With DAE, DAE-CNN and DAE-DAE-CNN structures are suitable for images with comparatively less and heavier noise levels, respectively. An image with unknown level of noise is fed to all three structures at the same time and generates different outcomes. In Phase 3, winner-take-all combination is employed to generate outcome of proposed RSDAECNN system. Combination of outcomes from several individual systems is generally used in ensemble of classifiers and winner-take-all combination emphasizes the individual best confident system [43]-[44]. Therefore, the outcome of the proposed system will be correct classification selecting the outcome of the most confident structure. As an example, if the input image is noiseless, system outcome might come from the structure with CNN alone. On the other hand, system outcome might come from DAE-CNN and DAE-DAE-CNN for input image with less and heavier noise levels, respectively. As an example, if an image of '3' with massive noise is placed to the system, DAE-DAE will restore the original image as shown in Fig. 3 and CNN will classify it correctly.

### C. Significance of the Proposed Model

There are several notable differences between the existing models and the proposed one on the premises of noisy image classification. Existing noisy image classification methods are found suitable for defined noise level. To work with less noisy data these models need to be trained with less noisy data whereas to classify massive noisy data the training data set should be corrupted by similar proportionate of noise. However, the proposed model can work with zero to massive level of noise due to the innovative arrangement of trained DAE and CNN for three different structures.

This model also omits the necessity of the system to be trained for images with different noise levels. It requires a single DAE to be trained with images containing regular level of noise and a CNN with noiseless images. Instead of using multiple training it places different arrangements of this trained DAE and CNN to deal with images carrying different

proportion of noise. Thus, it reduces the pre-processing time for preparing the training dataset.

One more significant contribution of this work is the computational efficiency. To develop the proposed model with three DAEs and three CNNs, only one DAE and one CNN are trained independently. The cascaded DAEs in DAE-DAE-CNN also contains same trained DAE. Innovative arrangements of a trained DAE and a trained CNN makes the system computationally efficient.

### III. PERFORMANCE EVALUATION

This section investigates the performances of proposed RSDAECNN on the benchmark image dataset MNIST numeral images [19]. This section first gives the description of the dataset and the experimental setup used to work over this dataset and afterward the results of the experiments conducted on images of different noise levels and lastly looks at the capability of the proposed model against existing ones. This model is implemented in Matlab R2017a. The performance analysis has been conducted on MacBook Pro Laptop (CPU: Intel Core i5 @ 2.70 GHz and RAM: 8.00 GB) in OS-X Yosemite environment.

#### A. Dataset Description and Experimental Setup

MNIST database [19] consists of 70000 sample gray-scaled images of handwritten digits collected from individuals having different writing styles. There are two sets of data: training set, which consists of 60000 images, and testing set of 10000 images. For each of 10 digits there are 6000 training samples and 1000 testing samples. Images in this dataset are of size  $28 \times 28$ .

In order to conduct a fair analysis of the proposed model's performance against the existing ones, a uniform experimental environment is required. The DAE used here has 784 input nodes as the images in the MNIST dataset are of  $28 \times 28$  size and DAE can be fed with linearized data only. DAE includes 784 input neurons, 500 hidden neurons and 784 output neurons. The input of DAE is a linearly oriented noisy image of size of  $28 \times 28$  whereas the output is the linearly oriented raw image of size  $28 \times 28$ .

CNN, the only classifier used here, is trained with clean images of  $28 \times 28$  size. The CNN used here is two layered where each layer contains a convolution layer and a following subsampling layer. The kernels and other parameters are initialized randomly. The filter used for the convolution task in both layer is a  $5 \times 5$  matrix. This filter slides over the original image and for every position the dot product is calculated which results in the feature map. The size of the feature map is  $24 \times 24$  and the depth is 6 as the number of filters used is 6. Afterwards, max pooling is applied separately on each feature map with a spatial neighborhood of  $2 \times 2$  window and the size of the feature map becomes  $12 \times 12$ . It is followed by another convolution and pooling operation with the same sized kernels and pooling region as before, which further reduces the size of the feature map to 192 as the depth of convolution layer used here is 12. The output of the second pooling layer acts as the input of the fully connected layer, which calculates the output probabilities for each class. So, there would be 192 nodes in the hidden layer. The data in this benchmark dataset is

distributed among 10 classes. That's why the CNN used here contains 10 nodes in the output layer.

In order to deal with noises in the images, all the images in the training set are corrupted with 20% random noise. For, the testing purpose, the test dataset is used once as it is, then corrupted with 10% noise, afterward, they are adulterated with 20% noise and finally to check the performance of our model with massive noisy images, we increased the level of noise included in the images to 50%. To add noise in the training and testing image samples zero masking noise has been used where a random matrix is initialized with the same size of training data with some of the pixels within the data being randomly OFF having probability of 20%. For testing purpose, another three random matrices of same size are initialized where 10%, 20% and 50% data are randomly turned OFF. These matrices are multiplied with the raw images to generate the noisy images.

#### B. Experimental Result and Analysis

This section evaluates the classification performance of the proposed system against MNIST dataset on the premises of various proportionate of noise present in the image to validate its performance in case of dealing with variable level of noise. To simulate the performance of the proposed system for real world scenario where images can be noisy but prior knowledge about the level of noises is not possible, different level of noises has been added to the dataset because MNIST does not carry noises.

In this study, we implemented masking noise where fraction of the pixels of input image is forced to be zero having probability of 0%, 10%, 20% and 50%. At first a detailed presentation has been given for a sample image containing different noise levels as well as the reconstructed ones from DAE and DAE-DAE and finally the classification results. Experimental results for the dataset are collected for individual structures as well as proposed RSDAECNN system and are compared with other prominent methods. The performance of the system is analyzed on the basis of image reconstruction as well as classification accuracies represented by both confusion matrices and accuracy graphs.

Table I delicates the outputs of image denoising step applying DAE and DAE-DAE architectures as well as the obtained classification results from three different structures (i.e., CNN, DAE-CNN, DAE-DAE-CNN) as well as the classification result of the proposed PSDAECNN. Images of '3' with different noise levels are considered as inputs of the proposed system those are classified with individual structures and generate system outcome. In case of noiseless image, it is seen that first structure (i.e., only CNN) correctly classified the image as "3". However, the reconstructed image obtained from a DAE seems more like numeral "8" and whenever it is filtered by DAE-DAE the image turned into the image of numeral "8". There remain two reasons behind the occurrence. Firstly, the DAEs being used here is the pre-trained DAE which is learned to reconstruct noiseless image from a noisy version of it. In case it is fed with a noiseless image it is not possible for it to know whether the image contains no noise and tries to reconstruct an image taking the input image as a noisy image. Secondly, the structures of numeral "3" and "8" are quite same.



So, DAE takes the input image and reconstruct an image like the numeral “8”. The DAE-DAE architecture is a two-layered cascaded form of the same DAE. The same scenario happens with it also. So, both the DAE-CNN and DAE-DAE-CNN misclassify the image as numeral “8”. Still, the proposed model classifies this image accurately as “3” because single CNN classified it correctly with more confident level. In case of both 10% and 20% noisy form of the very same image, only DAE-DAE-CNN misclassifies it as numeral “8”. In case of DAE-DAE-CNN structure, the image is first filtered by the frontier DAE. As the proportion of noise present in the image is less, this frontier DAE is sufficient enough to output the noiseless and good quality image. This reconstructed image is then again fed to the following DAE which also takes it as noisy image and tries to denoise it which in the end outputs a disordered image which looks like numeral “8”. The scenario is different in case of 50% noisy images. This time without any additional denoising technique CNN classifies it as numeral “5” whereas, both the DAE-CNN and the DAE-DAE-CNN classify it correctly. Though DAE-CNN classifies it correctly, from the figure it is clearly observable that the quality of the image reconstructed by DAE-DAE is far better and more like the original one as it is in case of the reconstructed one from the single DAE. Finally, the proposed RSDAECNN classified correctly all four cases although individual structures generate different outcomes.

Fig. 4 shows test set image classification accuracy of the proposed RSDAECNN system along with individual structures (i.e., CNN, DAE-CNN and DAE-DAE-CNN) for 0%, 10%, 20% and 50% noisy images up to 400 epochs. For 0% noise (Fig. 4(a)), structure with CNN alone achieved the highest

classification accuracy among three individual structures and showed classification accuracy of 99.31%. On the other hand, classification accuracies of DAE-CNN and DAE-DAE-CNN are 97.83% and 95.99% accordingly. The reason behind these two models poor performance compared to single CNN is that CNN is trained with noiseless native images. Whenever any noiseless image is fed to a DAE trained with 20% noisy images, the DAE tries to convert the shape of the image to some other form assuming that the image is corrupted by 20% noise and results in producing a deformed image. The scenario is worse in case cascaded DAE is used. So, logically DAE-CNN and DAE-DAE-CNN perform worse compared to single CNN. However, because of using winner-takes-all model for final class label selection, the proposed model shows a better classification accuracy than these two models and same as the single CNN. For 10% noise (Fig. 4(b)) DAE-CNN is shown best suited individual structure because DAE is trained with 20% noise level. For this case performance of the proposed method is same as DAE-CNN. DAE-CNN is showed as best individual structure for 20% noise (Fig. 4(c)), but interestingly proposed model performed better than DAE-CNN for this case. On the other hand, for 50% noise case, DAE-DAE-CNN outperformed CNN and DAE-CNN. The reason behind is already explained that cascaded DAEs perform well than single DAE in case of image with massive noise as they are both trained at 20% noise level; after the first DAE works on a noisy image the second one gets an image with relatively less noise which gets further denoised. In such heavy noise case, proposed method showed the similar performance of DAE-DAE-CNN. Finally, considering all the scenarios the proposed model performs the best for noiseless to heavy noise cases.

TABLE I. SAMPLE OF ORIGINAL IMAGES WITH AND WITHOUT NOISE AND THEIR RECONSTRUCTION USING DAE, DAE-DAE AS WELL AS THE CLASSIFICATION RESULT OF CNN, DAE-CNN, DAE-DAE-CNN AND THE PROPOSED MODEL

Noise Level	Input Image	Reconstructed Image		Classification through Individual Structure			Classification of Proposed RSDAECNN Combining Individual Structures
		DAE	DAE-DAE	CNN	DAE-CNN	DAE-DAE-CNN	
0%				3	8	8	3
10%				3	3	8	3
20%				3	3	8	3
50%				5	3	3	3

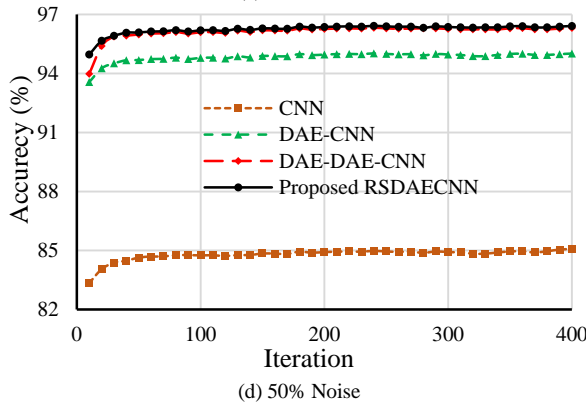
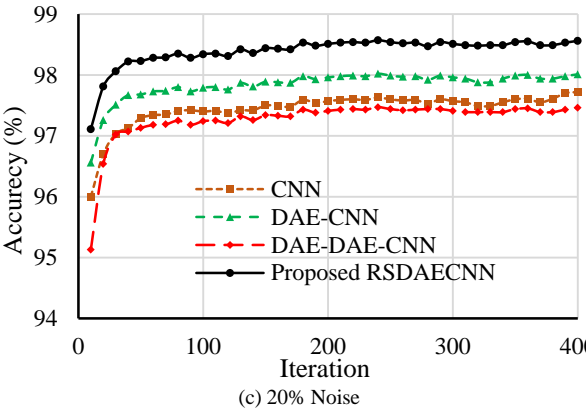
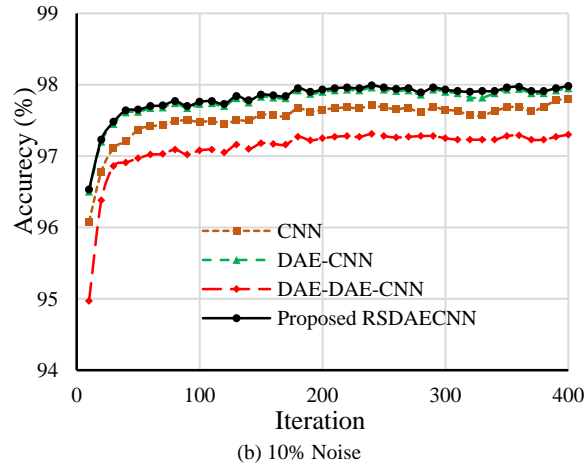
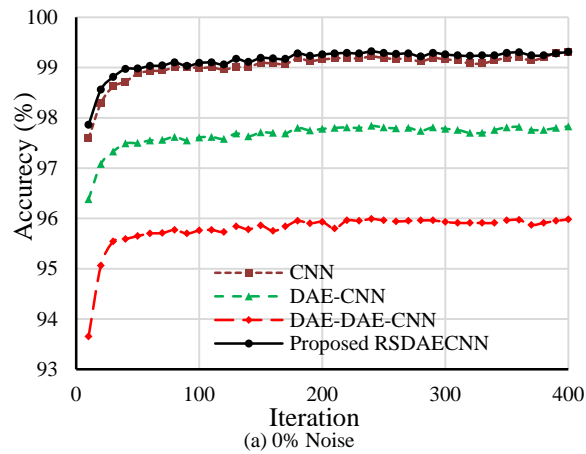
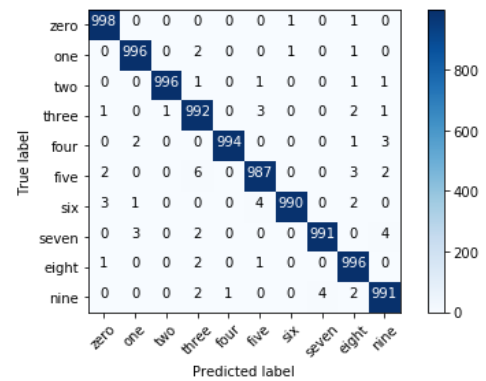
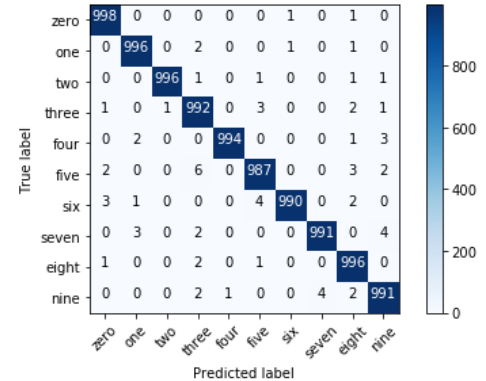


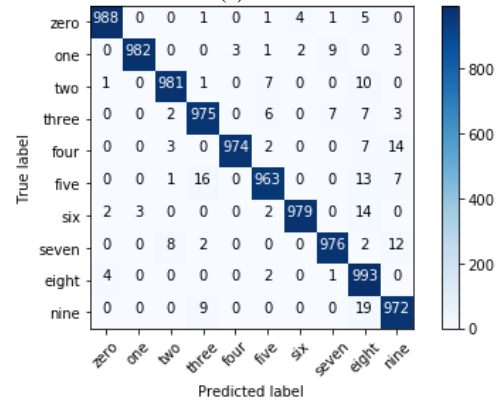
Fig. 4. Test set recognition accuracy with different noise levels.



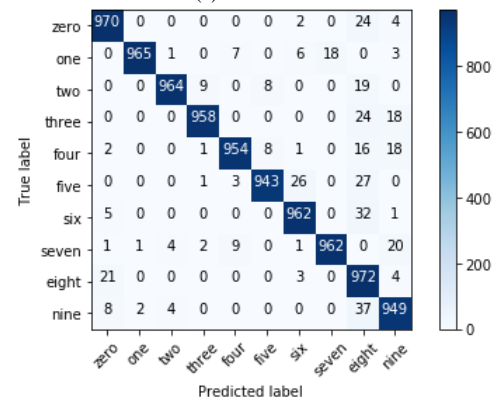
(a) Proposed RSDAECNN



(b) CNN



(c) DAE-CNN

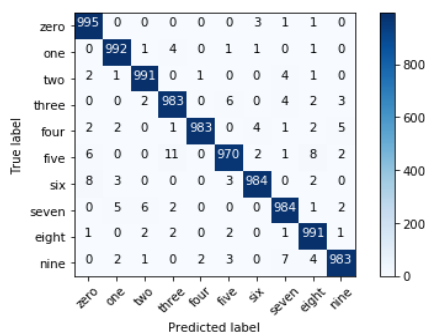


(d) DAE-DAE-CNN

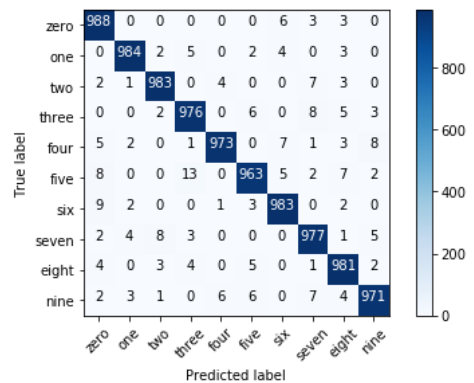
Fig. 5. Confusion matrices of test set with 0% noise for proposed RSDAECNN and individual structures (CNN, DAE-CNN and DAE-DAE-CNN).

Fig. 5 shows the confusion matrixes of test set with 0% noise for the proposed RSDAECNN system along with individual structures after 400 epochs. It clearly observed from the figure for the noiseless case that single CNN (Fig. 5(b)) and proposed model (Fig. 5(a)) performs better than DAE-CNN and DAE-DAE-CNN achieving a fair accuracy. It is also visible from the figure that all the individual structures (i.e., CNN, DAE-CNN and DAE-DAE-CNN) and the proposed system performed worst for the numeral “5”. Among them DAE-DAE-CNN performs worst misclassifying this numeral 57 times out of 1000 samples. Single CNN, DAE-CNN and the proposed model classifies it correctly for 987, 963 and 987 times accordingly. In case of DAE-DAE-CNN it is noticeable that most of the digits are misclassified as numeral “8”; 179 samples out of 1000 samples are misclassified as numeral “8”. This incident is more frequent for numeral “9”, “6”, “5”, “3” and “2”. The reason behind this incident is that the two layered cascaded DAE is fed with the noiseless images this time, where both the DAEs are trained with 20% noisy data. So, this denoiser deforms the shape of the images even if the images contain no noise which in the end misled the CNN classifier. The scenario is almost same for the DAE-CNN also. Still, proposed model performs well because of using winner-takes-all in the end for final selection process. As this method chooses the one, based on maximum node value, misclassification by two models doesn't affect the overall performance of the proposed model. The best classification accuracy is found for numeral “0”. The proposed model, CNN, DAE-CNN, DAE-DAE-CNN classify it 998, 998, 988, 970 times accordingly.

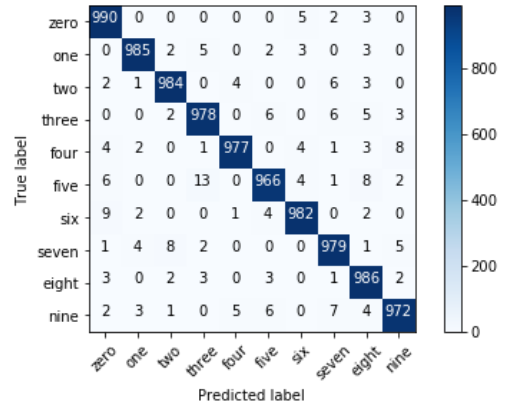
Fig. 6 shows the confusion matrixes of test set with 20% noise for the proposed RSDAECNN system along with individual structures (i.e., CNN, DAE-CNN, DAE-DAE-CNN) after 400 epochs. In such noisy case, all four models performed best for digit ‘0’ and worst for digit ‘5’. The proposed model, CNN, DAE-CNN, and DAE-DAE-CNN classify numeral ‘0’ correctly in 995, 988, 990 and 986 cases out of 1000 cases, respectively. On the other hand, for ‘5’ the true classifications by the methods are 970, 963, 966 and 959 for proposed model, CNN, DAE-CNN, and DAE-DAE-CNN, respectively. On the basis of overall performance, DAE-CNN is the best and DAE-DAE-CNN is the worst among individual structures. In this case CNN performs better than the DAE-DAE-CNN but performance degraded with respect 0% noise case (Fig. 5) architecture which is logical as explained earlier. On the other hand, proposed RSDAECNN is better than best individual structure (i.e., DAE-CNN).



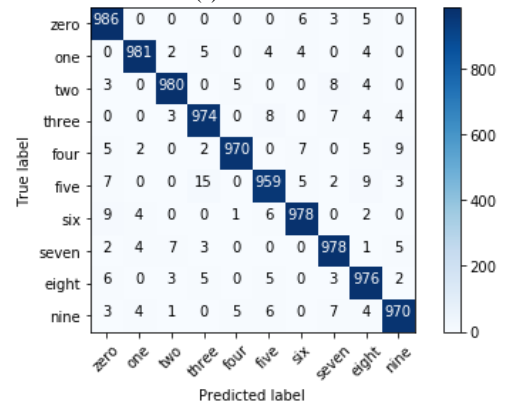
(a) Proposed RSDAECNN



(b) CNN



(c) DAE-CNN

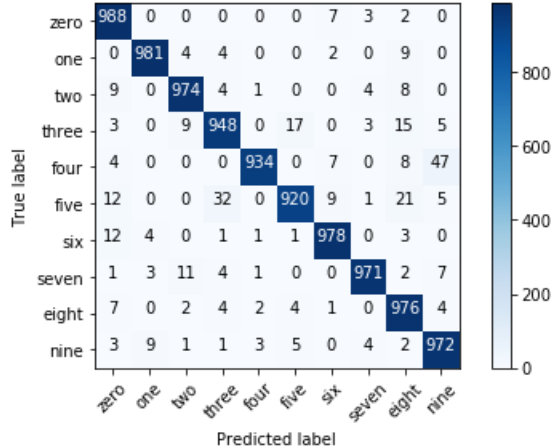


(d) DAE-DAE-CNN

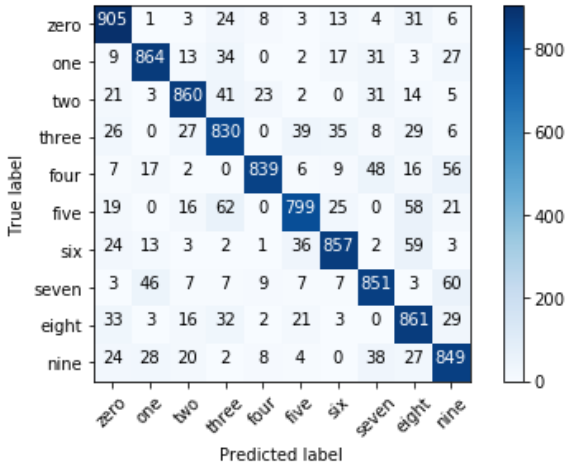
Fig. 6. Confusion matrixes of test set with 20% noise for proposed RSDAECNN and individual structures (CNN, DAE-CNN and DAE-DAE-CNN).

Fig. 7 shows the confusion matrixes of test set with 50% noise for the proposed RSDAECNN system along with individual structures (i.e., CNN, DAE-CNN, and DAE-DAE-CNN) after 400 epochs. This confusion matrix gives the evidence of the fact that proposed model is best suited even if the image is distorted by massive proportion of noise. This time CNN performs the worst; it classifies numerals “1” to “9” correctly only on 864, 860, 830, 839, 799, 857, 851, 861, 849 cases out of 1000 samples for each of them. Apart from classifying numeral “0”, in each and every time its classification accuracy is below 90% and for the numeral “5” its accuracy is even below 80%. In such huge noise, DAE-DAE-CNN is the best among individual structures. On the

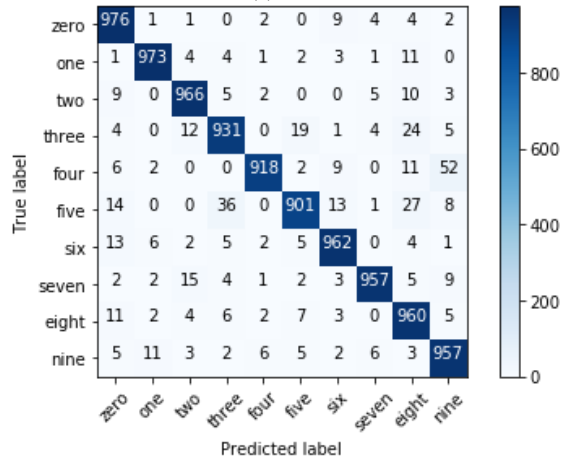
other hand proposed model classifies the numerals “0” to “9” on 988, 981, 974, 948, 934, 920, 978, 971, 976 and 972 cases accordingly. The performance of the proposed model is comparably better than the DAE-DAE-CNN structure in case of classifying such massive noisy image data. The confusion matrixes presented in Fig. 5 to 7 clearly revealed the effectiveness of the proposed system to work well to classify images with noise free to heavy noise scenario.



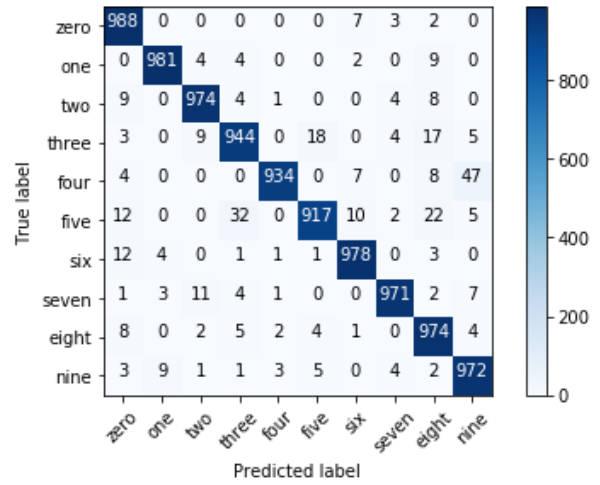
(a) Proposed RSDAECNN



(b) CNN



(c) DAE-CNN



(d) DAE-DAE-CNN

Fig. 7. Confusion matrixes of test set with 50% noise for proposed RSDAECNN and individual structures (CNN, DAE-CNN and DAE-DAE-CNN).

TABLE II. A COMPARATIVE DESCRIPTION OF PROPOSED RSDAECNN NOISY IMAGE CLASSIFIER WITH SOME CONTEMPORARY METHODS

The Work Reference	Classification	Noise Level	Recognition Accuracy		
Bengio et al. [19]	DBN	0%	98.50%		
	Deep net	0%	98.40%		
	Shallow net	0%	95.00%		
Glorot [45]	Sparse rectifier neural network	25%	98.43%		
Vincent et al. [35]	DAE	10%	97.20%		
Vincent et al. [36]	SVM	25%	98.37%		
	SDAE-3	25%	98.50%		
Self-Implemented	CNN	CNN	0%	99.31%	
		CNN	10%	97.88%	
		CNN	20%	97.76%	
		CNN	50%	85.15%	
	DAE-CNN [42]	CNN	0%	97.83%	
		CNN	10%	97.95%	
		CNN	20%	98.01%	
	DAE-DAE-CNN [42]	CNN	50%	95.01%	
		CNN	0%	95.99%	
		CNN	10%	97.31%	
	Proposed RSDAECNN	CNN	CNN	20%	97.47%
			CNN	50%	96.32%
CNN			0%	<b>99.31%</b>	
CNN			10%	<b>97.98%</b>	
CNN	20%	<b>98.56%</b>			
CNN	50%	<b>96.41%</b>			

Table II shows the comparative result analysis of the proposed RSDAECNN model with some other prominent noisy image classifiers. In extent, it describes the particular feature(s) of particular models while classifying noisy images. It is a highly mentionable issue that most of the existing models employ additional feature extraction techniques, whereas, proposed model overcomes the necessity of applying additional feature extraction techniques. The results presented in the table for CNN, DAE-CNN, DAE-DAE-CNN and proposed RSDAECNN are the tabular forms which have already been explained in the previous section. Results of other existing methods are collected from corresponding papers. It is notable that existing methods are tested for different individual noise levels. However, the proposed RSDAECNN has outperformed other models for any noise level. For noise-free case, as an example, Bengio et al. showed accuracy 98.50% and proposed method showed 99.31% accuracy. For 10% noise case, Vincent et al. [35] showed 97.20% accuracy and proposed method showed 97.98%. On the other hand, no existing method presented accuracy for heavy noise (i.e., 50%) and their outcome might be dramatically worse. However, for 50% noise, the performance of proposed method degraded little but outperformed other individual structures CNN, DAE-CNN, and DAE-DAE-CNN. The achieved accuracy for such heavy noisy case is 96.41%. Finally, the results presented in the table clearly revealed the effectiveness of the proposed system for classifying noisy images adulterated with variable level of noise.

#### IV. CONCLUSIONS

Considering real life scenario, it is usual for an image data to be noisy. Pre-processed noiseless images can be classified at ease with the help of existing classification methods. However, for a supervised classifier, it is difficult to deal with the noisy data directly fed to it and failure to classify is quite certain. In this paper, autoencoders are implemented to restore the image from its noisy version and then the reconstructed image is forwarded to a classifier. Another important consideration is that having prior knowledge about the proportion of noise carried by image data is not possible. Keeping all these facts in mind, an innovative model is investigated which includes CNN, DAE-CNN, and DAE-DAE-CNN. This model excludes the necessity to train it for different levels of noise. Being noise independent, the proposed model showed better performance on MNIST dataset compared to other models in terms of classifying images with noises ranging from zero to massive which also ensures its capability of learning hierarchical representations.

Several future research directions are opened from this study. The three-layered architecture investigated in this study is found efficient. Future researches can be conducted by stacking layers with some optimization algorithms to get better performance. Various AEs rather than DAE can also be employed to check whether the image reconstruction process improves or not. Furthermore, proposed model is noise level independent but not noise type independent. The method is tested with images corrupted by only random noise and might perform well for only one type of noise by which it is trained with. To make the system more robust and more applicable in real life scenarios it should be further upgraded so that it would

be both noise level independent as well as noise type independent.

#### REFERENCES

- [1] T. M. Lillesand and R. W. Kiefer, "Remote Sensing and Image Interpretation," *Geological Magazine*, vol. 132, issue 2, pp. 248-249, 1995.
- [2] M. A. H. Akhand, M. Ahmed, M. H. Rahman and M. M. Islam, "Convolutional Neural Network Training incorporating Rotation based Generated Patterns and Handwritten Numeral Recognition of Major Indian Scripts," *IETE Journal of Research (TIJR)*, Taylor & Francis, vol. 63, pp. 1-19, 2017.
- [3] F. J. Huang, and Y. LeCun, "Large-scale learning with svm and convolutional nets for generic object recognition" *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1-8, 2006.
- [4] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "High-performance neural networks for visual object classification", *arXiv Preprint arXiv:1102.0183*, 2011.
- [5] D. C. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN)*, pp. 1918-1921, 2011.
- [6] J. C. Bezdek, L. O. Hall, and L. Clarke, "Review of MR image segmentation techniques using pattern recognition," *Medical Physics*, vol. 20, issue. 4, pp. 1033-1048, 1993.
- [7] Y. Bar, I. Diamant, L. Wolf and H. Greenspan, "Deep learning with non-medical training used for chest pathology identification," *Proceedings of Society for Optics and Photonics*, pp. 94140V-94140V. doi: 10.1117/12.2083124, 2015.
- [8] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda, "Subject independent facial expression recognition with robust face detection using a convolutional neural network," *Neural Networks*, vol. 16, issue 5, pp. 555-559, 2003.
- [9] H. Bourlard, and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological Cybernetics*, vol. 59, issue 4, pp. 291-294, 1988.
- [10] Y. Bengio, "Learning deep architectures for AI. *Foundations and trends® in Machine Learning*," vol. 2, issue 1, pp. 1-127, 2009.
- [11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Parallel distributed processing: explorations in the microstructure of cognition*, vol. 1, pp. 318-362, 1986.
- [12] H. C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, issue 8, pp. 1930-1943, 2013.
- [13] M. Norouzi, M. Ranjbar, and G. Mori, "Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2735-2742, 2009.
- [14] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 609-616, 2009.
- [15] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, issue 7, pp. 1527-1554, 2006.
- [16] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2528-2535, 2010.
- [17] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high-level feature learning," *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*, pp. 2018-2025, 2011.

- [18] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37 issue 23, pp. 3311-3325, 1997.
- [19] Y. Bengio, P. Lamblin, D. Popovici and H. Larochelle, "Greedy layer-wise training of deep networks", In Proc of Advances in 19 th neural information processing systems, pp. 153-160, Dec 2007.
- [20] S. Behnke, "Hierarchical Neural Networks for Image Interpretation", volume 2766 of *Lecture Notes in Computer Science*. Springer, 2003.
- [21] D. Scherer, A. Müller, S. Behnke, "Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition," 20th International Conference on Artificial Neural Networks (ICANN), Thessaloniki, Greece, Springer. pp. 92-101, 2010.
- [22] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks", in Proc. Neural Information Processing Systems, pp. 1097-1105, 2012.
- [23] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 806-813, 2014.
- [24] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv Preprint arXiv:1312.6199, 2013.
- [25] D. Lu, and Q. Weng, "A survey of image classification methods and techniques for improving classification performance." *International Journal of Remote Sensing*, vol. 28, issue 5, 823-870, 2007.
- [26] M.C. Motwani, M.C. Gadiya, R.C. Motwani, F.C. Harris, "Survey of Image Denoising Techniques", Proc. of GSP 2004, Santa Clara, CA, pp. 27-30, 2004.
- [27] R. R. Coifman, and D. L. Donoho, "Translation-invariant de-noising," *Wavelets and Statistics*, vol. 103, pp. 125-150, 1995.
- [28] P. Perona, and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, issue7, pp. 629-639, 1990.
- [29] L. I. Rudin, and S. Osher, "Total variation based image restoration with free local constraints," Proceedings of the IEEE International Conference on Image Processing, vol. 1, pp. 31-35, 1994.
- [30] O. Subakan, B. Jian, B. C., Vemuri and C. E. Vallejos, "Feature preserving image smoothing using a continuous mixture of tensors," Proceedings of the 11th International Conference on Computer Vision (ICCV), pp. 1-6, 2007.
- [31] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, issue 12, 3736-3745, 2006.
- [32] J. Mairal, F., Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding", Proceedings of the 26th Annual International Conference on Machine Learning, pp. 689-696, 2009.
- [33] A. K. Singh, V. P. Shukla, S. R. Biradar., and S. Tiwari, "Multiclass Noisy Image Classification Based on Optimal Threshold and Neighboring Window Denoising." *International Journal of Computer Engineering Science (IJCES)*, vol. 4, issue 3, pp. 1-11, 2014.
- [34] Cheema, T.A., I. Qureshi and M. Naveed A., "Blur and Image Restoration of Nonlinearly Degraded Images Using Neural Networks Based on Nonlinear ARMA Model", Proc. INMIC, pp. 102-107.
- [35] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," Proceedings of the 25th International Conference on Machine Learning, 1096-1103, 2008.
- [36] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, issue 3371-3408, 2010.
- [37] F. Agostinelli, M. R. Anderson, and H. Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," Proceedings of the Advances in Neural Information Processing Systems, pp. 1493-1501, 2013.
- [38] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction." , In Proc. International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, pp. 52-59, 2011.
- [39] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," Proceedings of the 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 241-246, 2016.
- [40] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," Proceedings of the Advances in Neural Information Processing Systems, pp. 1790-1798, 2014.
- [41] B. Du, W. Xiong, J. Wu, L. Zhang, L. Zhang, and D. Tao, "Stacked convolutional denoising auto-encoders for feature representation," *IEEE Transactions on Cybernetics*, vol. 47, issue 4, pp. 1017-1027, 2017.
- [42] S. S. Roy, M. Ahmed and M. A. H. Akhand, "Classification of massive noisy image using auto-encoders and convolutional neural network," 2017 8th International Conference on Information Technology (ICIT), pp. 971-979, 2017.
- [43] M. A. H. Akhand, and K. Murase, "Ensembles of Neural Networks based on the Alteration of Input Feature Values," *International Journal of Neural Systems*, vol. 22, issue 1, pp. 77-87, 2012.
- [44] M. A. H. Akhand, Md. Monirul Islam, and K. Murase, "A Comparative Study of Data Sampling Techniques for Constructing Neural Network Ensembles," *International Journal of Neural Systems*, vol. 19, issue 2, pp. 67-89, 2009.
- [45] X. Glorot, A. Bordes, and Y. Bengio. "Deep Sparse Rectifier Neural Networks." , In Proc of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11). Vol. 15. pp. 315-323, 2011.

# A New Healthcare Context Information: The Social Context

Isra'a Ahmed Zriqat  
Department of Computer Science  
Applied Science Private University  
Amman, Jordan

Ahmad Mousa Altamimi  
Department of Computer Science  
Applied Science Private University  
Amman, Jordan

**Abstract**—During the treatment process, medical institutes collect context information about their patients and store it in their healthcare systems. The collected information describes the measurable, risk, or medication information and used to improve the performance of the institutes healthcare systems by allowing diverse knowledge about patients. Being said that some other information is needed as they influence patients' life style such as education and income as the high level of education or income reflected positively to the patient's life, and probably resulting in reducing likelihood disease or incidence of infectious diseases. In this paper, a new type of healthcare context information (Social Context) is proposed to address this need. It can be divided into four main categories: related people, behavior, income and education of the patient. We believe that the new proposed context information should be considered in the designing process of the context-aware medical informatics systems beside to the well-known context information.

**Keywords**—Context information; social context; healthcare; medical information

## I. INTRODUCTION

The concept of Context is treated differently in the literature, a review discloses large number of definitions. However, authors of [1] proposed the most refer definition that is “any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object)”. The “situation” here refers to a description of the states of relevant entities. According to this definition, different types of information are needed to be gathered and analyzed to characterize an entity's context such as: Identity, Location, Time, Status, Physical context, Emotional context, etc. [1]-[3]. In fact, all the context information cannot be listed here due to their diversity types and numerous number. Therefore, in this work, we will describe only the context information relevant to the health system.

Here, the well-known context information (e.g., Personal Information, location and time) [4]-[13] are considered along with our new context aspect (e.g., Social Context). Together, they describe the patient's social context and stores data about his/her surrounding peoples. For example, tracking the location of patients assists caregivers, especially in emergencies. This will off course improve the performance of healthcare monitoring system.

In fact, there are many cases where tracking the patients place or time is needed. For example, elderly patients

vulnerable to fall more than others due to fall risks. A hard-physical exercise will influent the medical decision as the blood pressure will be high. In this regard, technology such as Real Time Locating System (RTLS) could be used to accurately detecting falls, and can measure many parameters (e.g., Body temperature, respiration rate, heart rate, blood pressure, and ECG) of patient continuously, and then transmitted wirelessly [10], [11]. This results in improving patient care and reducing associated health care costs [12], [13].

In this paper, a new context aspect (e.g., Social Context) is proposed that includes information influence the patient's health such as: the patient's social information and related people. The proposed factor can be categorized into four main categories: related people of the patients, behavior of the patients, income and the education of the patients. Such data should be considered in the designing of medical context-aware systems as they influence the patient's life style. For example, the high level of education or income reflected positively to the patient's life, and probably resulting in reducing likelihood disease or incidence of infectious diseases. Furthermore, some patients may have communicable diseases that may spread and infect all who have contact them.

The structure of this paper is as follow. Section II presents the related work. The research methodology of our work is then given in Section III. The proposed Social Context Information is then discussed in Section IV. Finally, the conclusion and the future work are offered in Section V.

## II. RELATED WORK

Context information has been extensively considered in the literature. Patient's personal information, location, and time context for instance have reviewed in [7]-[11]. Because our work is related to healthcare, and because of the lack of simplicity, we will consider the Context Information that is used for describing Heart Diseases as illustrated in Fig. 1. In fact, context information is primarily divided into two main categories: Personal and location and time context.

### A. Personal Context Information

There are different types of patient's personal information to be considered such as patient's vital signs, the associated risk factors with disease, medical symptoms, and current prescribed medications. Next, we give brief description for each type.

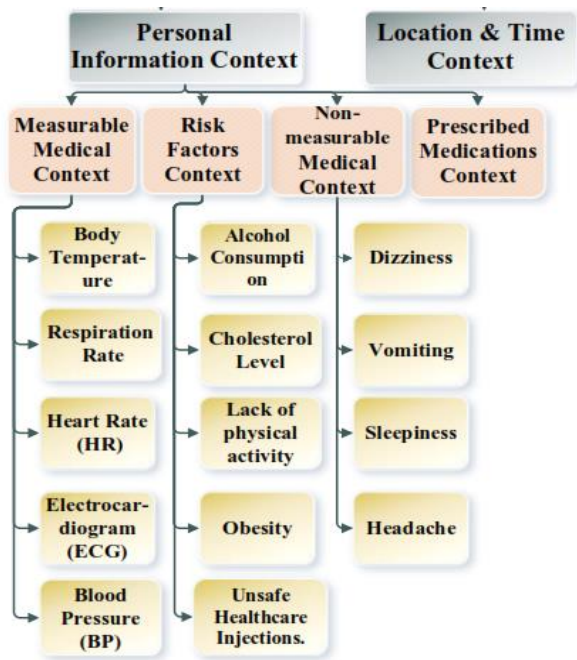


Fig. 1. Types of heart diseases' context information.

1) *Measurable Medical Context Information*: includes vital signs that can be measured and continually monitored such as: body temperature, heart rate, blood pressure, respiration rate, etc. [4], [5], [8].

2) *Risk Factors Context*: covers the health factors that changed infrequently [4], [7]. According to the World Health Organization, these factors are countless, and a set of these risk factors can be combined leading cause of a specific disease. For instance, there are seven risk factors associated with cancer disease (e.g., tobacco, alcohol, air pollution, low fruit and vegetable intake, physical inactivity, unsafe sex, and unsafe healthcare injections).

3) *Non-measurable medical context information*: which are symptoms that are difficult to be measured such as: headache, vomiting, sleepiness, and dizziness. Despite their measuring difficulties, they should be considered along with the measurable context in the healthcare context [9], [14].

4) *Medications context information*: provide the current medications given for a patient as it affects the readings of patient's vital signs [9], [14], [15]. Thus, healthcare givers evaluate the patient's response to take the appropriate medical decisions [9].

**B. Location and Time Context Information**

It is considering the location of patients to improve the performance of healthcare monitoring system. There are many cases where tracking the patients place is needed. For example, elderly patients vulnerable to fall more than others due to fall risks. In this regard, technology such as Real Time Locating System, an embedded microcontroller is connected to a set of medical sensors and a wireless channel to measure and transfer the patient's sign continuously [10], [11], can be employed to accurately detecting falls, which improves healthcare services and reduces the associated costs [12], [13].

**III. METHODOLOGY**

There are many types of context information that have been considered in the literature [4]-[13]. These types can be considered together for describing the patient's context, and designing the medical informatics systems. Such systems should contain all the relevant patient's information to deliver proper service. Beside the well-known contexts information, other context information such as the social aspect context should be considered as it is describing the patient's social context and stores data about his/her surrounding peoples. The proposed aspect consists of data about behavior, income, education of the patient, and data about his/her related people that influences the patients' health.

**IV. SOCIAL CONTEXT INFORMATION**

Social context relates to the social and community that surrounded patients. Those patients may have communicable diseases that may spread to infect all who have relationship with them. In addition, aspects related to social functioning are considered as highly relevant by patients and their family members. Regardless of these aspects, pharmacotherapy and any doubts about their illness, caregivers, home nursing, behavior, and other social aspects present private characteristics that need to be secured from misusing or unauthorized actions [16].

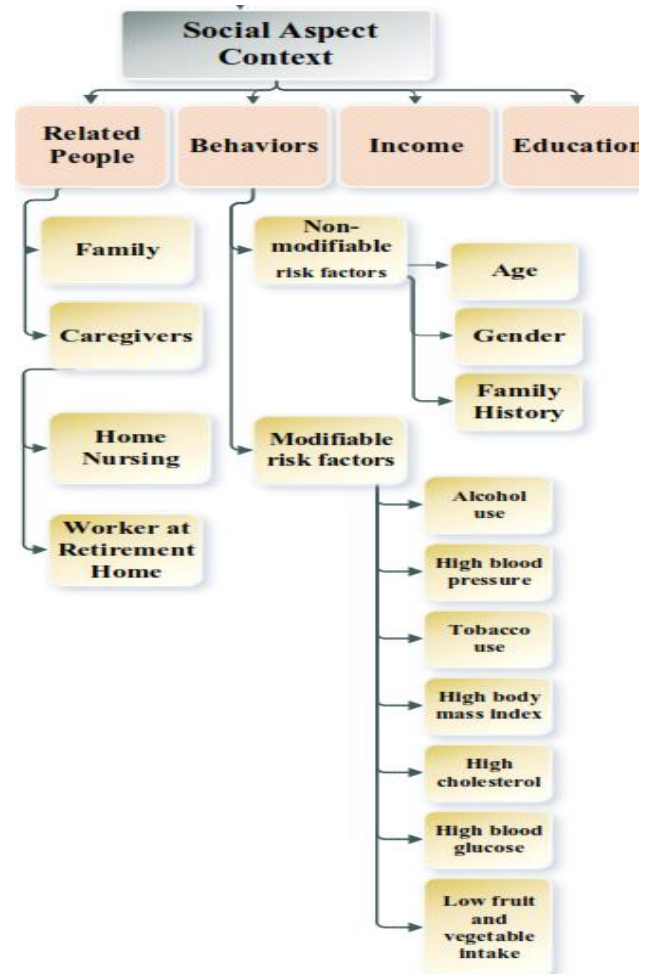


Fig. 2. Proposed social information context.



The proposed social aspect context could include: related people of the patients, behavior of the patients, income and education as illustrated in Fig. 2. Next, a brief description for each category is given.

#### A. Related People

Related peoples refer to those who have a relationship with the patient, such as family members, nurses, and other caregivers. This category could be:

- Family members: refer to individuals who have direct connect with their patients. Since those people live with them at the same place, their information should be kept secret if it is stored in the eHealth system.
- Caregivers: refer to anyone who take care of patients or their work require to deal with patients. Caregivers could be: home nursing who live at home and provide round-the-clock care or long-term medical treatment [17]. Or, workers at retirement home who work at place for the individuals who are still able to care for themselves independently, but have difficulty managing an entire house [17]. We iterate that the information of such related people should be kept secure (if stored in the eHealth systems) as patient may suffer from several types of infectious diseases and could affect the caregiver or those who live with them at the same house.

#### B. Behavior

According to World health federation, it can be divided into Modifiable and Non-modifiable risk factors. The Modifiable health risk factors that associated with heart disease could contain alcohol use, high blood glucose, low fruit and vegetable intake, tobacco use, high body mass index, high cholesterol, high blood pressure, and physical inactivity. As long these risk behaviors have a negative effect on individual health and account for 61% of heart disease deaths [17]. All these risk behaviors considered as private information that should not be exposed to unauthorized people. In the other hand, the modifiable risk factors refer to the factors that cannot be changed such as age, gender and family history [18]. Whereas, the heart disease becomes increasingly common with advancing age, since the heart undergoes subtle physiological changes, as a person gets older. Also, your gender is significant, as a female is at less risk of heart disease than a male. As well as, the risk will increase if a first-degree relative suffers heart diseases, as the family's history could reveal your possibilities. Family's history of heart disease could reveal your possibility [18].

#### C. Income and Education

Many comparative studies found that income primarily influences health [19], [20]. Those studies showed that high income allowing the individuals for buying better quality material goods such as food and shelter, and allowing them to access health services and leisure activities, which make their bodies healthy that in turn reduces the likelihood of disease [19], [20]. On the other hand, there are specific interpretations explain that the education has a large influence on health [19], [21]. Patients with high level education could improve mutual comprehension between them and their physicians, resulting in

improving quality of care [20]. As well as, several studies found that correlation between high level education and the personal management of several diseases like spreading of sexually transmitted diseases.

#### V. CONCLUSION AND FUTURE WORK

In this work, we propose new context that includes the social context patients' information. Such information influences patients' health and life style. For example, the high income provides decent life and reflected positively on the patient's health, resulting in reduces the likelihood of disease. In addition, patients may have infectious diseases, which will be spread and infect all who have relationship with patients. All these data should be considered in the designing of healthcare systems and should be kept secured and protected from unauthorized accessing. As a future work, we are planning to develop a context-aware access control model for eHealth systems that will be built specifically for healthcare systems.

#### ACKNOWLEDGMENT

The authors are grateful to the Applied Science Private University, Amman-Jordan, for the full financial support granted to cover the publication fee of this research article.

#### REFERENCES

- [1] Dey, A.K., Abowd, G.D., and Salber, D.: 'A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications', *Human-computer interaction*, 2001, 16, (2), pp. 97-166.
- [2] A. Dey: "Context-aware computing: The CyberDesk project", *Proceedings of the AAAI 1998 Spring Symposium on Intelligent Environments*, AAAI Press, Menlo Park, CA, 1998, pp. 51-54.
- [3] Ch. Bornträger: "Contextual Influence on the Usability of Different Media Types", *Diploma Thesis*, Technische Universität Ilmenau, Ilmenau, Jan. 2003.
- [4] Roy, N., Gu, T., and Das, S.K.: 'Supporting pervasive computing applications with active context fusion and semantic context delivery', *Pervasive and Mobile Computing*, 2010, 6, (1), pp. 21-42.
- [5] Copetti, A., Loques, O., Leite, J.C., Barbosa, T.P., and da Nobrega, A.C.: 'Intelligent context-aware monitoring of hypertensive patients', in *Editor: 'Book Intelligent context-aware monitoring of hypertensive patients' (IEEE, 2009, edn.)*, pp. 1-6.
- [6] Smith, S.F., and Duell, D.: 'Clinical nursing skills: Nursing process model, basic to advanced skills' (McGraw-Hill/Appleton & Lange, 1992).
- [7] Organization, W.H.: 'Global health risks: mortality and burden of disease attributable to selected major risks' (World Health Organization, 2009. 2009).
- [8] Gardner, R.M., and Shabot, M.M.: 'Patient-monitoring systems': 'Biomedical Informatics' (Springer, 2006), pp. 585-625.
- [9] Al-Bashayreh, M.G., Hashim, N.L., and Khorma, O.T.: 'Context-Aware Mobile Patient Monitoring Frameworks: A Systematic Review and Research Agenda', *JSW*, 2013, 8, (7), pp. 1604-1612.
- [10] Shaikh, R.A.: 'Real time health monitoring system of remote patient using ARM7', *Control and Automation (IJICA) ISSN*, 2012, pp. 2231-1890.
- [11] Al-Aubidy, K.M., Derbas, A.M., and Al-Mutairi, A.W.: 'Real-time patient health monitoring and alarming using wireless-sensor-network', in *Editor: 'Book Real-time patient health monitoring and alarming using wireless-sensor-network' (IEEE, 2016.)*, pp. 416-423.
- [12] Bowen, M.E., Craighead, J., Wingrave, C.A., and Kearns, W.D.: 'Real-Time Locating Systems (RTLS) to improve fall detection', *Gerontechnology*, 2010, 9, (4), pp. 464-471.

- [13] Boulos, M.N.K., and Berry, G.: 'Real-time locating systems (RTLS) in healthcare: a condensed primer', *International journal of health geographics*, 2012, 11, (1), pp. 25.
- [14] Koutkias, V.G., Chouvarda, I., Triantafyllidis, A., Malousi, A., Giaglis, G.D., and Maglaveras, N.: 'A personalized framework for medication treatment management in chronic care', *IEEE transactions on information technology in biomedicine*, 2010, 14, (2), pp. 464-472.
- [15] Mohamed, I., Misra, A., Ebling, M., and Jerome, W.: 'Harmoni: Context-aware filtering of sensor data for continuous remote health monitoring', in Editor: 'Book Harmoni: Context-aware filtering of sensor data for continuous remote health monitoring' (IEEE, 2008, edn.), pp. 248-251
- [16] Organization, W.H.: 'The economics of social determinants of health and health inequalities: a resource book' (World Health Organization, 2013. 2013).
- [17] Alzheimer's Association®. Available at: <http://www.alz.org/care/alzheimers-dementia-residential-facilities.asp#ixzz4L0xPe3kH>
- [18] World health federation. <http://www.world-heart-federation.org/press/fact-sheets/cardiovascular-disease-risk-factors/>. Accessed on 7 June 2017.
- [19] Organization, W.H.: 'A conceptual framework for action on the social determinants of health', 2010.
- [20] Lepage, B., Schieber, A.-C., and Lamy, S.: 'Social determinants of cardiovascular diseases', *Public Health Reviews*, 2011, 33, (2).
- [21] Woolf, S.H., and Braveman, P.: 'Where health disparities begin: the role of social and economic determinants—and why current policies may make matters worse', *Health affairs*, 2011, 30, (10), pp. 1852-1859.

# Brainwaves for User Verification using Two Separate Sets of Features based on DCT and Wavelet

Loay E. George, Hend A. Hadi

College of Science, Computer Science Department, Baghdad University,  
Baghdad, Iraq

**Abstract**—This paper discusses the effectiveness of brain waves for user verification using electroencephalogram (EEG) recordings of one channel belong to single task. The feature sets were previously introduced as features for EEG-based identification system are tested as suitable features for verification system in this paper. The first considered feature set is based on the energy distribution of DCT's or DFT's power spectra, while the second set is based on the statistical moments of wavelet transform, three types of wavelet transforms is proposed. Each set of features is tested using normalized Euclidean distance measure for the matching purpose. The performance of the verification system is evaluated using FAR, FRR, and HTER measures. Two publicly available EEG datasets are used; first is the Colorado State University (CSU) dataset which was collected from seven healthy subjects and the second is the Motor Movement /Imagery (MMI) dataset which is a relatively large dataset was collected from 109 healthy subjects. The attained verification results are encouraging when compared with the results of other recent published works, the best achieved HTER is (0.26) when the system was tested on CSU dataset, while the best achieved HTER is (0.16) when the system was tested on MMI dataset for the features which based on the energy of DFT spectra.

**Keywords**—Electroencephalogram (EEG); wavelet transforms; DCT; DFT; energy features; statistical moments; Euclidean measure

## I. INTRODUCTION

New biometric traits based on physiological signals, such as EEG and ECG signals were recently explored instead of traditional biological traits. The perfect biometric trait should have the following characteristics: very low intra-class variability, very high inter-class variability, stability over time and universality [1]. Typical biometric traits such as fingerprint, voice, and retina, are subject to physical damage such as dry skin, loss or changes of voice, severe injuries such as missing hands or figures, aniridia (i.e. loss of the iris), or burned fingers, etc. [2]. Recent studies have shown that the EEG signals have biometric possibility because the brain signals are distinctive and impossible to replicate and/or steal. Person identification and verification are two different types of biometric applications, the goal of person identification is to identify unknown individual from a group of persons (i.e. matching the input pattern of one person against all the records in a templates database), while the goal of person verification is to confirm or deny the claimed identity [3]. The previous work [4] focused on the person identification, while this paper is particularly interested in person verification.

Palaniappan [5] proposed two stage authentication approach using AR coefficients, channel spectral powers, differences of

inter-hemispheric channel spectral power, inter-hemispheric channel linear and non-linear complexity as features, after filtering the signals with Finite Impulse Response (FIR) filter, and then he used Principal Component Analysis (PCA) to reduce feature vector size. Finally he tested five subjects from CSU dataset using Manhattan distance, he achieved best result with FAR and FRR equal to zero.

Altahat et al. [6] explored the reduction of EEG channels to reduce the complexity and cost of EEG-based authentication system. In this work the signal Power Spectral Density (PSD) was considered as features. They proved that the reduced channels set enhanced the system performance and achieved total HETR (14.69%) when it was tested on (106) subjects from MMI dataset.

Fraschini, et al. [7] introduced an approach based on phase synchronization, to explore individual distinctive brain network organization. Their proposed method is based on four main steps. The first step is band-pass filtering in which "eegfilt" function was used to filter the raw EEG signals. The second step is "functional connectivity estimation" which was performed using PLI for estimating pair-wise statistical interdependence between EEG time series. The third step is "brain network reconstruction" in which the functional network is represented as a weighted graph, where each node in the graph represented EEG channel, and each edge represented functional connection, where the PLI value was used as the strength of the connection. The fourth step "characterization" is to characterize the functional brain organization, in order to estimate the significance of each node in the network they focused on a centrality measure. The best EER was achieved in gamma band; it is (0.044%) for (109) subject.

Bajwa and Dantu [8] proposed the use of EEG signals for both authentication and cryptographic key generation. They used Fast Fourier Transform (FFT) and then Daubechies wavelet (db8) to extract features by calculating statistical information on the wavelet sub bands, in this paper DFT is proposed as a separate extraction method by calculating the energy averages of DFT's power spectra as well as wavelet Daubechies (db4) is proposed as a separate method by calculating the statistical moments to all sub bands. Two types of classifiers were tested: Support Vector Machine (SVM) and Bayesian network, they achieved best accuracy rate (100%) when the system was tested on 7 subjects.

Despite the encouraging achieved results on EEG-based authentication system, the related works have faced complications in feature extraction stage and the fusion of

features from multiple channels or tasks, also the using of many techniques starting from noise removing until classification or matching step.

The main addressed problems in this paper: 1) number of required electrodes and mental tasks; where the feature sets are extracted under the adopted condition (i.e., single channel and single task) in [9], [4] and tested in the verification mode in this paper; 2) the complexity of feature extraction and noise removal; all the proposed methods make the system fast, and simple using fast code for DFT and DCT without need for preprocessing step; 3) the normalized Euclidean distance measures used instead of the complex classification algorithms.

This paper is organized as follows: Section 2 presents the description of used datasets and the proposed methods, Section 3 discusses the experiments result, Section 4 discusses previous works related to this paper, and Section 5 presents conclusions.

## II. MATERIALS AND METHODS

The proposed EEG-based verification system is based on the following main stages for verification purpose just like the proposed identification system:

- Mapping stage.
- Feature extraction stage.
- Feature analysis and selection stage.
- Matching stage.

Different transform algorithms is proposed to perform the mapping in the literature; in this paper the input EEG signal is mapped to frequency domain using DCT, DFT, and three different wavelets algorithms in order to extract the main discrimination features. Feature extraction stage is aimed to extract the most discriminate features from the transformed EEG signal. The task of feature analysis and selection stage is to select the best combination of discriminative features.

In matching stage, the normalized Euclidean distance measures are used to verify the claimed identity of input pattern.

### A. Dataset

Two public datasets are used in the conducted tests. The first one is *Colorado State University dataset* which is a public dataset collected by Keirn and Aunon [10]. It is a small dataset consists of the EEG recordings of seven healthy subjects. Each subject was performed some mental tasks. These tasks are: Baseline task, Letter composing task, mathematics task, rotation task, counting task. Signals were recorded from the positions C<sub>3</sub>, C<sub>4</sub>, P<sub>3</sub>, P<sub>4</sub>, O<sub>1</sub> and O<sub>2</sub>; see Fig. 1. The taken EEG signals duration is 10 sec. with sampling rate of (250 sample/sec) [11]. This dataset holds an error that occurred in one of subjects (i.e., 4<sup>th</sup>) in letter composing trails [12], [10].

Second EEG dataset is *Motor Movement /Imagery dataset* which is a relatively large dataset consists of EEG recordings for 109 healthy volunteers; it was described in [13]. In this dataset the participants performed 14 trails of the following tasks: two Baseline tasks with eyes open and eyes closed, Task1 (open and close left/ right fist), Task2 (imagine the opening and

closing of left/right fist), Task3 (open and close both fists and both feet), Task4 (imagine opening and closing both fists and both feet). The dataset contains the recordings of 64 channel based on 10-20 international system of electrodes placement as shown in Fig. 1. The recording duration is ranging from 1 minute to 2 minutes except for subject (106) who performed task3 for (36 sec. and 294 msec.) in attempt 5; the EGG recording was sampled at (160 Hz) [13], [10].

Table I shows the number of samples for each subject class in CSU and MMI datasets (Note: subject 4 has 9 samples for the letter-composing task because of the error that above mentioned), and the number of samples for each subject class in MMI dataset (Motor Movement/imagery dataset).

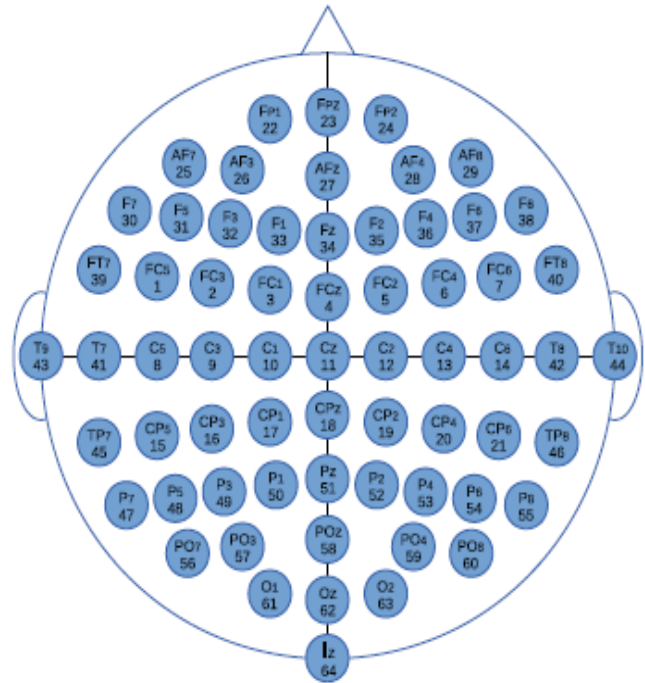


Fig. 1. The (10-20) international system of electrodes placement [14].

TABLE I. THE NUMBER OF SAMPLES FOR EACH CLASS IN CSU AND MMI DATASET

CSU dataset			
Class No.	No. of Samples	Class No.	No. of Samples
1	10	5	15
2	5	6	10
3	10	7	5
4	10 (only 9 for compose task)		
MMI dataset			
109	3 samples for each class		

**B. Features Sets**

In this stage, two separate sets of features were used to generate the feature vectors and tested for the verification purpose, they are energy based features and/or the statistical moments.

1) *Energy of DFT's and DCT's spectra*: Discrete Fourier and Discrete Cosine transforms are considered to map the input EEG signal from a time-domain to frequency domain. DFT's power spectra consist of the sine and cosine components, while DCT use only the cosine functions, it is a Fourier related but just using real numbers [15], [11]. The DCT's general mapping equation is given by (1), while DFT's general mapping equation is given by (2):

$$C(u) = \alpha(u) \sum_{i=0}^{N-1} s(i) \cos\left(\frac{u\pi(2i+1)}{2N}\right) \quad (1)$$

where,  $\alpha(u) = \begin{cases} \sqrt{1/N} & \text{if } u = 0 \\ \sqrt{2/N} & \text{if } u \neq 0 \end{cases}$

$$F(u) = \frac{1}{N} \sum_{i=0}^{N-1} s(i) \left[ \cos\left(\frac{2\pi i u}{N}\right) - j \sin\left(\frac{2\pi i u}{N}\right) \right] \quad (2)$$

Where  $C(u)$  and  $F(u)$  is the  $u^{th}$  coefficient of the DCT and DFT, respectively, and  $s()$  is the input EEG signal.

After the mapping step, the obtained AC coefficients (i.e., coefficients with  $u > 0$ ) are divided into a number of blocks (or bands) and the energy of each block is calculated using (3) [16]:

$$en(j) = \frac{1}{L} \sum_{i=jL+1}^{jL+L} |T(i)|^2, \quad (3)$$

Where,  $T(i)$  represents the transform,  $F(u)$  or  $C(u)$ , coefficients array;  $en(j)$  is the energy of  $j^{th}$  block;  $L$  is number of coefficients belong to each block;  $j=0 \dots P-1$ ;  $P=(N-1)/L$  is the total number of blocks. The array  $en()$  is considered the feature vector.

2) *Statistical moments of discrete wavelet transforms*: The second set of features is the statistical moments of Discrete Wavelet Transforms sub bands. The wavelet transform computes the inner products of a signal with a family of wavelets to decompose the EEG signal (to scale-shift domain) with keeping location in time information; unlike DFT and DCT which maps the input signal to frequencies that making it up regardless of time information. DWT uses two filters (i.e., high pass filter and low pass filter) [17], [11]. Three types of wavelet transform were proposed in the previous work [4]; the first one is Haar Wavelet transform which is the simplest wavelet type, it computes the sums and differences of input signal, the low and high filters of HWT is given by (4) and (5) [17]:

$$L(i) = s(2i) + s(2i + 1) \quad (4)$$

$$H(i + N/2) = s(2i) - s(2i + 1) \quad (5)$$

Where,  $i=0 \dots N/2$ ;  $N$  is the length of input signal.  $L(i)$  is the  $i^{th}$  approximation coefficient,  $h(i)$  is the  $i^{th}$  detailed coefficient.

The second type of proposed wavelet transform is Daubechies (db4) transform, it has four wavelet and scaling coefficients [18], [19]:

$$\alpha_1 = (1 + \sqrt{3}) / (4\sqrt{2}), \quad \alpha_2 = (3 + \sqrt{3}) / (4\sqrt{2})$$

$$\alpha_3 = (3 - \sqrt{3}) / (4\sqrt{2}), \quad \alpha_4 = (1 - \sqrt{3}) / (4\sqrt{2})$$

$$\beta_1 = \alpha_4, \quad \beta_2 = -\alpha_3$$

$$\beta_3 = \alpha_2, \quad \beta_4 = -\alpha_1$$

The low coefficients of first level are given by (6), while the high coefficients of first level can be given by (7):

$$L(i) = \sum_{k=0}^{N/2} \alpha(k) s(j + k) \quad (6)$$

$$H(i + N/2) = \sum_{k=0}^{N/2} \beta(k) s(k + j) \quad (7)$$

Where  $i \in \{0, \dots, (N/2)-1\}$ ,  $j \in \{0, \dots, N-3\}$ , and  $k \in \{0, \dots, 3\}$ .

The third type of wavelet transform is bi-orthogonal (Tap9/7), it transforms the input EEG signal by applying three consecutive phases: (i) split phase (ii) lifting phase and (iii) scaling phase [20].

The four lifting steps and two scaling steps are described by the following equations:

**Lifting phase:**

$$Y(2n+1) = s(2n+1) + a[s(2n) + s(2n+1)] \quad (8)$$

$$Y(2n) = s(2n) + b[s(2n-1) + s(2n+1)] \quad (9)$$

$$Y(2n+1) = Y(2n+1) + c[Y(2n) + Y(2n+2)] \quad (10)$$

$$Y(2n) = Y(2n) + d[Y(2n-1) + Y(2n+1)] \quad (11)$$

**Scaling phase:**

$$Y(2n) = Y(2n) / k \quad (12)$$

$$Y(2n+1) = -k \times Y(2n) \quad (13)$$

Table II shows the coefficients {a, b, c, d, and k} values.

After transforming the input signal using wavelet transform, one of the following two set of statistical moments is adopted to be applied on the obtained sub bands. They are described by the following equations:

The 1<sup>st</sup> Statistical Moments Set:

$$Mom(n) = \frac{1}{k} \sum_{i=0}^{p-1} [S(i) - \bar{S}]^n \quad (14)$$

TABLE II. TAP 9/7 LIFTING COEFFICIENTS

Coefficient	Value
A	- 1.586134342
B	- 0.052980118
C	0.8829110762
D	0.4435068522
K	1.230174105

Where,  $S(i)$  is the  $i^{th}$  sample,  $k$  is the signal length, and  $\bar{S}$  is the mean which is determined as:

$$\bar{S} = \frac{1}{k} \sum_{i=0}^{p-1} S(i) \quad (15)$$

The 2<sup>nd</sup> Statistical Moments Set:

$$Mom(n) = \frac{1}{k} \sum_{i=0}^{p-2} [\Delta S(i) - \bar{\Delta S}]^n \quad (16)$$

Where,  $\Delta S(i)=S(i)-S(i+1)$  for  $(i=0, \dots, p-2)$ , and  $\bar{\Delta S}$  is similar that given in (15) but instead of  $S(i)$  it is  $\Delta S(i)$ . The power  $n$  is taken (0.5, 0.75, 1, 2, and 3).

### C. Features Analysis and Selection Stage

This step is applied to reduce the feature pool size and to select most related and discriminative features with lowest within distance and highest between discrimination, then combining the best set of features that led to best verification accuracy [21], [22].

### D. Matching Stage

The input pattern is matched with the template(s) of the class subject that the user claims to be in order to verify his identity; normalized Euclidian distance measure given by (17) is used to calculate the distance between the input pattern and the class template(s) [23], and similarity distance threshold is checked to accept or deny the claimed identity:

$$nMSD(S_i, T_j) = \sum_{k=0}^{p-1} \left( \frac{s_i(k) - t_j(k)}{\sigma_j(k)} \right)^2 \quad (17)$$

Where,  $S_i = \{s_i(0), s_i(0), \dots, s_i(p-1)\}$  is the feature vector of a sample belong to  $i$ th class,  $T_j = \{t_i(0), t_i(0), \dots, t_i(p-1)\}$  is the template feature vector of  $j$ th class and  $\sigma_j = \{\sigma_i(0), \sigma_i(0), \dots, \sigma_i(p-1)\}$  is the standard deviation vector of  $j$ th template.

## III. RESULT AND DISCUSSION

The accuracy of verification system with all proposed feature extraction methods was tested on the two adopted public datasets. Each set of features is extracted from EEG signal belong to single task and single channel. The best attained system HTER was 0.26 for CSU dataset, while the best achieved HTER for MMI data set is 0.16. The results of the tests are described in details in the following sections:

### A. Verification Results

The Receiver Operating Characteristic (ROC) Curve illustrates the performance of verification system by plotting the False Rejected rate (FRR) which is given by (19) and measures the proportion of incorrectly rejected genuine patterns, against the false Accepted rate (FAR) which is given by (18) and measures the proportion of incorrectly accepted imposter patterns, at various threshold settings to check the intersection point between FRR and FAR in which the Half Total Error rate (HTER) is calculated using (20) to evaluate the performance of the system [8], [24]:

$$FAR = \frac{\text{No.of accepted imposters}}{\text{Total No.of imposters}} \quad (18)$$

$$FRR = \frac{\text{No.of rejected genuines}}{\text{Total No.of genuines}} \quad (19)$$

$$HTER = \frac{1}{2} (FAR + FRR) \quad (20)$$

While the accuracy of the verification system can be determined using the following equation:

$$Accuracy = \frac{TP+TN}{P+N} \quad (21)$$

Where  $P$  is the number of genuine patterns, and  $N$  is the number of imposter patterns [16].

1) *Energy of sliced DFT and DCT spectra's results:* Table III shows the results of the verification system which was proposed in [9] using Energy of Sliced DFT Spectra when tested on CSU dataset when some enhancements were made to the system, while Table IV shows the verification results of the system when tested on MMI dataset. The best achieved HTER is 0.26 at threshold 16.6 for channel P4 belong to Rotation task, while the best achieved HTER is 0.16 at threshold 26.1 for channel C2 belong to Task1. Fig. 2 and 3 show the ROC curve of the P4\_Rot and C2\_Task1 feature sets.

TABLE III. FRR, FAR, ACCURACY, AND HTER OF THE ENERGY OF SLICED DFT SPECTRA FEATURES, CSU DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
P4-Rot	16.6	0	0.52	99.56%	0.26
P3-Rot	15.4	2.38	2.99	97.14%	2.68
P3-Math	15.4	2.86	4.26	96.04%	3.56
P3-Base	10.8	3.81	3.69	96.26%	3.75
C3-Base	23.6	3.81	3.90	96.04%	3.85

TABLE IV. FRR, FAR, ACCURACY, AND HTER OF THE ENERGY OF SLICED DFT SPECTRA FEATURES, MMI DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
C2-Task1	26.1	0	0.31	99.69%	0.16
Fz-Task4	26.4	0.31	0.37	99.63%	0.34
Cpz-Task1	23.9	0.61	0.51	99.49%	0.56
O2-Task4	20.1	0.61	0.51	99.49%	0.56
Cp3-Task1	25.3	0.31	0.56	99.44%	0.43
Oz-Task4	26.5	0.31	0.56	99.44%	0.43
P3-Task1	24.3	0.61	0.58	99.42%	0.59
Fc1-Task4	16.2	0.61	0.58	99.42%	0.6
Cp1-Task4	20.6	0.61	0.6	99.40%	0.61
Po7-Task1	20.9	0.61	0.61	99.39%	0.61

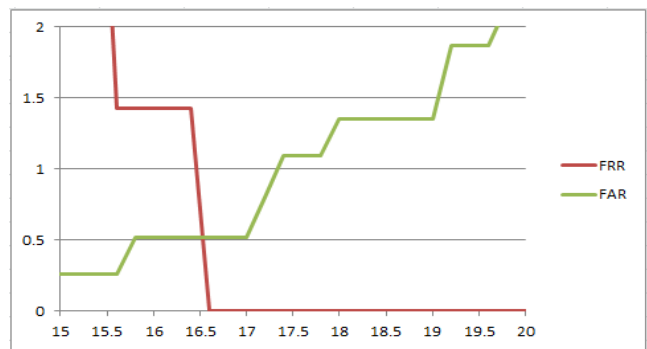


Fig. 2. ROC curves show the interception of FRR and FAR at optimal threshold for the feature set (P4-Rot).

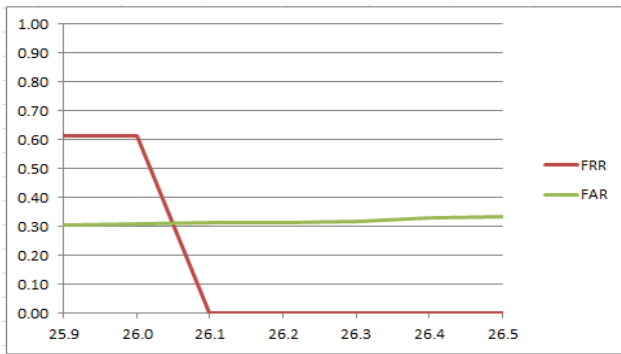


Fig. 3. ROC curves show the interception of FRR and FAR at optimal threshold for the feature set (C2-Task1).

TABLE V. FRR, FAR, ACCURACY, AND HTER OF THE ENERGY OF SLICED DCT SPECTRA FEATURES, CSU DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
P4-Rot	20.2	0	0.81	99.34%	0.4
P3-Math	10.6	2.86	3.71	96.48%	3.29
P3-Rotat	11.9	2.86	2.1	97.80%	2.48
C3-Baseline	13.2	3.81	3.14	96.70%	3.48
P3-Baseline	11.7	4.29	4.76	95.38%	4.52

TABLE VI. FRR, FAR, ACCURACY, AND HTER OF THE ENERGY OF SLICED DCT SPECTRA FEATURES, MMI DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
Cz-Task1	25.2	0.31	0.39	99.62%	0.35
Oz-Task4	24.4	0.31	0.39	99.62%	0.35
C1-Task1	24.4	0.31	0.43	99.57%	0.37
Cpz-Task1	25.4	0.31	0.44	99.56%	0.37
C2-Task1	24.9	0.31	0.48	99.53%	0.39
P4-Task1	26.9	0.31	0.49	99.51%	0.40
Pz-Task4	19.1	0.61	0.49	99.51%	0.55
Pz-Task1	26.7	0.61	0.54	99.46%	0.58
O2-Task1	23.1	0.61	0.55	99.45%	0.58
O1-Task1	23.4	0.61	0.57	99.43%	0.59

Tables V and VI show the attained verification results of the system based on the energy of sliced DCT spectra. The best achieved HTER is 0.4 at threshold (20.2) for the feature set extracted from channel P<sub>4</sub> and Rotate task from CSU dataset, while best achieved HTER is 0.35 at threshold (25.2) for the channel Cz belong to Task<sub>1</sub> from MMI dataset.

2) *Statistical moments of wavelet sub-bands features results:* In the following sections the results of HWT, db4, and Tap9/7 features which based on the statistical moments of the sub-bands are showed. The conducted tests show that the Haar and db4 wavelets show performance less than the features based on the energy of DFT and DCT, and Tap9/7.

Tables VII and VIII show some conducted tests of Haar wavelet transform using 2<sup>nd</sup> set of statistical moments on CSU and MMI datasets, respectively.

TABLE VII. FRR, FAR, ACCURACY AND HTER OF THE STATISTICAL MOMENTS FOR 2<sup>ND</sup> SET OF HWT FEATURES, CSU DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
P3-Rot	11.8	0	1.90	98.46%	0.95
P4-Rot	11.6	2.86	3.14	96.92%	3.00
P4-Math	10.4	3.81	4.21	95.82%	4.01
P3-Math	9.8	5.71	5.51	94.73%	5.61
C3-Baseline	9.80	5.24	5.14	95.16%	5.19

TABLE VIII. FRR, FAR, ACCURACY, AND HTER OF THE STATISTICAL MOMENTS FOR 2<sup>ND</sup> SET OF HWT FEATURES, MMI DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
Iz-Task1	16.5	0.61	0.63	99.37%	0.62
Iz-Task4	11.9	0.61	0.73	99.27%	0.67
O1-Task1	14.8	0.61	0.83	99.18%	0.72
Oz-Task4	21.5	0.92	0.75	99.25%	0.84
Cp4-Task4	15.7	0.92	0.87	99.13%	0.89
Cz-Task4	17.3	0.92	0.98	99.02%	0.95
Po4-Task4	16.3	1.53	0.71	99.28%	1.12
C4-Task1	19.6	1.22	1.07	98.93%	1.15
P4-Task4	18.6	1.22	1.23	98.77%	1.23
P6-Task4	17.9	1.53	1.03	98.97%	1.28

Table IX shows results of some conducted tests of db4 using 2<sup>nd</sup> set of statistical moments on CSU dataset, while

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
C3-Base	13.4	2.38	2.42	97.58%	2.40
P4-Rot	29.8	2.86	2.65	97.36%	2.75
P3-Rot	20.8	2.38	2.39	97.58%	2.39
P3-Base	8.4	3.33	4.24	95.82%	3.79
C4-Rot	10.0	5.24	5.01	95.16%	5.13

shows the result of MMI dataset when the Statistical Moments 1<sup>st</sup> set was applied because it achieved better results on MMI dataset than 2<sup>nd</sup> set.

The best results of the verification system based on Statistical Moments of Tap9/7 Sub-bands are showed for CSU dataset using Statistical Moments 2<sup>nd</sup> set in Table XI, and for MMI dataset using Statistical Moments 1<sup>st</sup> set in Table XII.

TABLE IX. FRR, FAR, ACCURACY, AND HTER OF THE STATISTICAL MOMENTS 2<sup>ND</sup> SET OF DB4 FEATURES, CSU DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
C3-Base	13.4	2.38	2.42	97.58%	2.40
P4-Rot	29.8	2.86	2.65	97.36%	2.75
P3-Rot	20.8	2.38	2.39	97.58%	2.39
P3-Base	8.4	3.33	4.24	95.82%	3.79
C4-Rot	10.0	5.24	5.01	95.16%	5.13

TABLE X. FRR, FAR, ACCURACY, AND HTER OF THE STATISTICAL MOMENTS 1<sup>ST</sup> SET OF DB4 FEATURES, MMI DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
Po4-Task4	14	0.61	0.67	99.33	0.64
O1-Task4	17	0.61	0.69	99.31	0.65
Iz-Task1	14	0.61	0.73	99.27	0.67
Iz-Task4	11.9	0.61	0.73	99.27	0.67
Oz-Task4	21.5	0.92	0.75	99.25	0.84
O1-Task1	14.7	0.92	0.81	99.19	0.86
Cp4-Task4	15.7	0.92	0.87	99.13	0.89
Fc2-Task4	15.5	0.92	0.89	99.11	0.9
Pz-Task4	13.8	0.92	0.89	99.11	0.9

TABLE XI. FRR, FAR, ACCURACY AND HTER OF THE STATISTICAL MOMENTS 2<sup>ND</sup> SET OF TAP9/7 FEATURES FOR CSU DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
P4-Rot	12.8	0.00	0.78	99.34%	0.39
C3-Baseline	20.6	2.38	3.43	96.70%	2.90
P4-Math	20.0	3.81	3.97	96.04%	3.89
P3-Rot	17.8	4.76	3.14	96.70%	3.95
P3-Math	10.0	4.29	3.97	96.04%	4.13

TABLE XII. FRR, FAR, ACCURACY, AND HTER OF THE STATISTICAL MOMENTS 1<sup>ST</sup> SET OF TAP9/7 FEATURES FOR MMI DATASET

Feat. Set	Thr.	FRR	FAR	Accuracy	HTER
O2-Task4	13.7	0.61	0.60	99.40%	0.61
Pz-Task4	16.3	0.61	0.62	99.38%	0.61
Poz-Task4	16.9	0.61	0.67	99.34%	0.64
Pz-Task1	12.3	0.61	0.67	99.33%	0.64
Cz-Task4	14.4	0.61	0.68	99.32%	0.65
Po3-Task4	20.8	0.61	0.74	99.26%	0.68

### B. Processing Time Parameter

In this section; the elapsed processing time on the introduced recognition system is presented. Table XIII shows the average processing time, (in terms of milliseconds) of the proposed methods; when they applied on CSU data set. Table XIV is the average processing time when the methods are applied on MMI datasets. Taking into account the recording time for CSU dataset is (10 sec) with sampling rate (250 Hz), and the taken recoding time for MMI CSU datasets is (1 minute) and sampling rate is (160 Hz); the determined matching time is for one-to-many comparisons. The Computer specification that used in the tests is Intel® Core™ i5-2450M CPU with (4GB) RAM, the operating system is windows7 (64bit), and the development programming language is Microsoft visual C#.

TABLE XIII. THE AVERAGE PROCESSING TIME RESULTS (IN MSEC) FOR CSU DATASET

Proposed Method	Feature Extraction (msec)	Matching (in msec)	Total (in msec)
DFT	13.359	0.002	13.361
DCT	22.0991	0.002	22.1011
HWT	1.7495	0.002	1.7515
Daub4	0.95797	0.002	0.95997
Tap9/7	1.007	0.002	1.009

TABLE XIV. THE AVERAGE PROCESSING TIME RESULTS (IN MSEC) FOR MMI DATASET

Proposed Method	Feature Extraction (msec)	Matching (in msec)	Total (in msec)
DFT	217.364	0.001	217.365
DCT	323.864	0.001	323.865
HWT	3.559	0.001	3.56
Daub4	3.389	0.001	3.39
Tap9/7	3.719	0.001	3.72

### IV. COMPARISON WITH RECENTLY RELATED WORKS

Some of the related published works on EEG-based verification system have achieved good results, some of them reached 100% on CSU dataset but many of them used more than one channel or task for verification tasks. Table XV shows that the attained results in this paper is competitive when compared with the results of other published works on CSU dataset and Motor Movement/Imagery dataset; taking into account that all proposed methods in this article has low computational complexity, they require very small execution time because the system uses single channel and single task, and fast algorithms.

TABLE XV. COMPARISONS WITH OTHER PUBLISHED WORKS ON CSU DATASET AND MMI DATASET BASED ON NUMBER OF SUBJECTS, NUMBER OF USED CHANNELS AND TASKS.

Author	No. of Subjects	# of Ch.	# of Tasks	Accuracy (%)
CSU dataset				
[5]	5	6	1	FAR=0 FRR=0
[8]	7	6	1	Acc=100%
Proposed work	7	1	1	Acc=99.56% HTER=0.26 FAR=0.52 FRR=0
MMI dataset				
[7]	109	64	2	EER=0.044
[6]	106	8	1	HTER=14.64
Proposed work	7	1	1	Acc=99.69% HTER=0.16 FAR=0.31 FRR=0

All published works haven't mentioned the elapsed processing time clearly, so we can't compare with them.

### V. CONCLUSION AND FUTURE WORK

In this paper the proposed feature extraction methods for verification purpose were tested, and make a comparison among them. For each proposed method the system was fast, simple and achieved encouraged results. The conducted tests showed that the best achieved HETR is 0.26 for DFT feature set when was applied on CSU database, and 0.16 when was applied on MMI dataset. DFT, DCT, and Tap9/7 showed performance better than Haar and Daubechies (db4) wavelet transforms methods, but WT methods showed complexity and processing time less than of that DFT and DCT.



In order to enhance the performance of wavelet based methods, another approach based on fusion of features from two channels belong to same task (i.e. two channels but single task) can be explored to increase the degree of discrimination among subjects with keeping the complexity as low as possible.

#### REFERENCES

- [1] N. V. Boulgouris, K. N. Plataniotis, and E. Micheli-Tzanakou, *Biometrics: theory, methods, and applications*.: John Wiley & Sons, 2009, vol. 9.
- [2] P. Campisi and D. La Rocca, "Brain waves for automatic biometric-based user recognition," *IEEE transactions on information forensics and security*, vol. 9, no. 5, pp. 782-800, 2014.
- [3] H. A. Shedeed, "A new method for person identification in a biometric security system based on brain EEG signal processing," in *Information and Communication Technologies (WICT), 2011 World Congress on*, 2011, pp. 1205-1210.
- [4] H. A. Hadi and L. E. George, "EEG Based User Identification Methods Using Two Separate Sets of Features Based on DCT and Wavelet," sent for publication to *Journal of Theoretical and Applied Information Technology*, vol. xx, no. x, pp. xx-xx, 2017.
- [5] R. Palaniappan, "Two-stage biometric authentication method using thought activity brain waves," *International Journal of Neural Systems*, vol. 18, no. 1, pp. 59-66, 2008.
- [6] S. Altahat, M. Wagner, and E. M. Marroquin, "Robust electroencephalogram channel set for person authentication," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 997-1001.
- [7] M. Frascini, A. Hillebrand, M. Demuru, L. Didaci, and G. L. Marcialis, "An EEG-based biometric system using eigenvector centrality in resting state brain networks," *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 666-670, 2015.
- [8] G. Bajwa and R. Dantu, "Neurokey: Towards a new paradigm of cancelable biometrics-based key generation using electroencephalograms," *Computers & Security*, vol. 62, pp. 95-113, 2016.
- [9] H. A. Hadi and L. E. George, "EEG Based User Identification and Verification Using the Energy of Sliced DFT Spectra," *International Journal of Science and Research (IJSR)*, vol. 6, no. 9, pp. 46-51, 2017.
- [10] Z. A. Keim and J. I. Aunon, "A new mode of communication between man and his surroundings," *IEEE transactions on biomedical engineering*, vol. 37, no. 12, pp. 1209--1214, 1990.
- [11] M. Abo-Zahhad, S. M. Ahmed, and S. N. Abbas, "State-of-the-art methods and future perspectives for personal recognition based on electroencephalogram signals," *IET Biometrics*, vol. 4, no. 3, pp. 179-190, 2015.
- [12] P. Kumari and A. Vaish, "Feature-level fusion of mental task's brain signal for an efficient identification system," *Neural Computing and Applications*, vol. 27, no. 3, pp. 659-669, 2015.
- [13] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: a general-purpose brain-computer interface (BCI) system," *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034--1043, 2004.
- [14] D. Rodrigues, G. F.A. Silva, J. P. Papa, and A. N. Marana, "EEG-based person identification through binary flower pollination algorithm," *Expert Systems with Applications*, vol. 62, pp. 81-90, 2016.
- [15] N. AHMED, T. Natarajan, and K. R. RAO, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90-93, 1974.
- [16] A. M.J. Abbas and L. E. George, "Palm Vein Identification and Verification System Based on Spatial Energy Distribution of Wavelet Sub-Bands," *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, no. 5, pp. 727-734, may 2014.
- [17] R. S. Stanković and B. J. Falkowski, "The Haar wavelet transform: its status and achievements," *Computers and Electrical Engineering*, vol. 29, no. 1, pp. 25-44, 2003.
- [18] J. S. Walker, *A primer on wavelets and their scientific applications*: CRC press, 2008.
- [19] J. Shen and G. Strang, "Asymptotics of daubechies filters, scaling functions, and wavelets.," *Applied and Computational Harmonic Analysis*, vol. 5, no. 3, pp. 312-331, 1998.
- [20] M. Beladgham, A. Bessaid, A. M. Lakhdar, and A. T. Ahmed, "Improving quality of medical image compression using biorthogonal CDF wavelet based on lifting scheme and SPIHT coding," *Serbian Journal of Electrical Engineering*, vol. 8, no. 2, pp. 163-179, 2011.
- [21] A. P. James and S. Dimitrijević, "Ranked selection of nearest discriminating features," *Human-Centric Computing and Information Sciences*, vol. 2, no. 1, pp. 1-14, 2012.
- [22] S. N. Mohammed and L. E. George, "Subject Independent Facial Emotion Classification Using Geometric Based Features," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 11, no. 9, pp. 1030-1035, 2015.
- [23] W. K. Pratt, *Digital Image Processing: A Wiley-Inter Science Publication*, 2001.
- [24] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861-874, 2006.

# FARM: Fuzzy Action Rule Mining

Zahra Entekhabi

Department of Computer Engineering  
Fars Science and Research Branch, Islamic Azad University  
Marvdasht, Iran

Pirooz Shamsinejadbabki

Department of Computer Engineering and Information  
Technology  
Shiraz University of Technology  
Shiraz, Iran

**Abstract**—Action Mining is a sub-field of Data Mining that concerns about finding ready-to-apply action rules. The majority of the patterns discovered by traditional data mining methods require analysis and further work by domain experts to be applicable in target domain while Action Mining methods try to find final cost-effective actions that can be applied immediately in target domain. Current state-of-the-art methods in AM domain only consider discrete attributes for action rule mining. Therefore, one should discretize continuous attributes using traditional discretization methods before using them for action rule mining. In this paper, the concept of Fuzzy Action Rule has been introduced. In this type of action rule, continuous attributes can be presented in fuzzy form. So that they can suggest fuzzy changes for continuous attributes instead of discretizing them. Because the space of all fuzzy action rules can be so huge a Genetic Algorithm-based Fuzzy Action Rule Mining (GA-FARM) method has been devised for finding the most cost-effective fuzzy action rules with tractable complexity. The proposed method has been implemented and tested on different real datasets. Results confirm that the proposed method is successful in finding cost-effective fuzzy action rules in acceptable time.

**Keywords**—Action mining; fuzzy action rule mining; genetic algorithm

## I. INTRODUCTION

Nowadays, the increase in computer usage has led to the sharp growth of databases and aggregation of data in the majority of organizations. In recent years, tendency to discovering actionable knowledge from data with the aim of decision making has seen a dramatic increase. Due to the large size of the data, human experts cannot investigate and analyze all data; on the other hand, the capability of fast and effective response to customer needs requires fast enough decisions which cannot be made without accurate evaluation and analysis of information and data. Therefore, discovering knowledge from database in almost real-time is of paramount importance. Data mining is a procedure which uses data analysis tools to find patterns and relationships between data. Data mining methods focus on discovering and extracting descriptive patterns and do not take into account the actionability of these patterns in target domain. For instance, consider the customer loyalty detection system. The conventional data mining systems predict which customer will leave the company in the close future; however, the companies need to know how they can prevent this and the subsequent losses. In fact, companies need profitable actions instead of just frequent patterns.

Action mining is among the proposed feasible solutions which have attracted attention recently. Actions determine for the user what should be done in order to reach high profits. Action is a type of knowledge which aims to maximize profits in desired areas by suggesting a number of alterations to a sample's condition. The methods proposed for action mining can be categorized into transductive and inductive [1]. In inductive methods, such as what was done by Ras et al. [2]-[9], the obtained action rules are general rules which can be applied to a group of things with similar attributes. Ras method, which has been introduced as DEAR [2], initially extracts the conventional classification rules in data mining and later, by combining compatible classification rules, creates action rules (rules that suggest alterations in attributes).

Tsay et al. [10] improved the DEAR algorithm by DEAR2 system which initially classifies the rule table based on the number of existing decision attribute values and consequently creating a sub-tree based on stable attributes. At the end, leaves are compared with each other to extract action rules.

Tsay et al. [11] also proposed DEAR3 algorithm for incomplete databases which can reduce the errors of undetermined, unreliable and outlier data to obtain more reliable action rules.

Another method which was suggested by Ras to extract action rules is called action rule discovery (ARD) [3]. This method extracts action rules without combining classification rules. The advantage of this method is its lower time complexity that makes it proper for large databases.

Meat-actions are actions that should be triggered to activate another action rule. In [12] Meta-actions have been used for mining business action rules and in [13] surgical meta-actions have been mined for medical diagnosis.

Inductive methods in other hand try to find most profitable actions for each case. The most prominent method of this type is presented by Yang et al. [14]-[15] which uses a traditional decision tree to extract actions. In summary, at first, a decision tree is inferred from the data. Next, for each new sample, it searches a leaf node and computes the net profit made from the movement of the new sample to the other leaves. Finally, the leaf that has the highest profit is selected as the destination leaf and the necessary alterations are performed for transferring the sample from the current node to the destination node. In order to extract action rules, Shamsinejadbabki et al. [1] proposed a new action mining method by using causal networks. This

method can obtain practical actions in the real world by considering causal relationships between the attributes.

Traditional decision making trees (used by Yang) in [15] require data with discrete values and the continuous values need to be discretized before being put in the algorithm. Despite conventional in data mining, the continuous to discretized domain transformation is little compatible with the real world and it is therefore difficult to define a precise border for continuous attributes' values. For instance, consider the question of how the border between middle age and old can be determined. If it is 60 years, a person who is 59 years old is considered middle-age and a 60-year-old person is considered old. Obviously this classification is not compatible with the real world actualities. Using fuzzy logic in traditional fuzzy logic capabilities makes it possible to work with continuous data. Kalanat et al. [16] use fuzzy set theory to improve Yang method when we face continuous attributes in data. They have shown that although the final rules are not fuzzy but their actionability is higher than Yang rules.

In this study, the concept of Fuzzy Action Rule (FAR) has been introduced. This type of action suggests fuzzy alterations for continuous attributes and therefore, it is expected that the resulting actions will be more applicable in the real world circumstances. Besides, a method has been devised for derivation of fuzzy action sets with maximum net profit from data using Genetic Algorithm.

The rest of paper is as follows: In Section 2 preliminaries are explained. Fuzzy Action Rule will be introduced in Section 3 and GA\_FARM, a GA-based FAR mining method will be presented in Section 4. Experimental results are argued in Section 5 and paper will be concluded in Section 6.

## II. PRELIMINARIES

Set and element are the two primary concepts of set theory. In classic set theory (crisp),  $x$  either can or cannot be a member of set  $S$ . As a result, the membership function of element  $x$  to set  $S$  ( $\mu_S$ ) can be shown as:

$$\mu_S(x) = \begin{cases} 1 & \text{if } x \in S, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

However, a large number of the real world sets are inherently fuzzy sets. Fuzzy set theory is a mathematical structure which allows an element to belong to more than one set. The membership degree of element  $x$  in fuzzy set  $S$  as  $\mu_S$  can be shown as  $x \rightarrow \{0,1\}$  where  $0 \leq \mu_S \leq 1$  [17].

Suppose three fuzzy sets (young, middle age, old) which are attributed to age variable. Now if in a classic set, 60 is considered as the border between middle-aged and old people, this is tantamount to considering a 59 years-old person as a middle age person and a 60 years-old person as old. However, fuzzy theory puts narrow borders between the values so that an element can be classified with various degrees. For instance, a 55 years-old person can belong to the middle-age and old sets at the same time with different membership degrees.

### A. Fuzzy Decision Tree

A fuzzy decision tree is a combination of fuzzy theory with traditional decision making trees. The generality in learning

from samples, clarity of the knowledge obtained from the decision tree and also the ability of working with fuzzy imprecise and incomplete data are among the benefits of using this combination. It can also increase robustness and applicability in inaccurate and unclear areas. The knowledge obtained from FDT more closely resembles human decision making and therefore, is more practical in real world circumstances. For example, in [18] FDT has been used for software cost estimation to handle imprecise data in describing software.

Fig. 1 depicts an example of FDT from [17] where attribute  $x_1$  is the root and attribute  $x_2$  is the internal node. The  $v^{\text{th}}$  membership function can be determined from the  $j^{\text{th}}$  attribute by  $\mu_{jv}$ . The  $K^{\text{th}}$  probable class in the  $L^{\text{th}}$  leaf is determined with  $\beta_{kL}$ . The number of shown classes in tree is Equal to 2. As a result, each leaf includes two fuzzy variables:  $\beta_{1L}$  and  $\beta_{2L}$

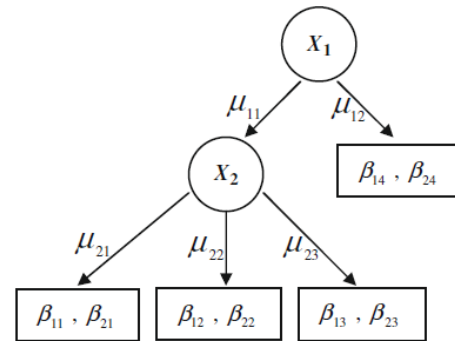


Fig. 1. A sample fuzzy decision tree [17].

### B. Action, Action Set and Action Rule

An action has a  $(A, a \rightarrow a')$  structure in which  $A$  is an attribute except the goal ( $A$  is called the action attribute) and  $a, a' \in \text{Dom}(A)$ . This action indicates change of  $A$  from value  $a$  to  $a'$  by a force outside of the system [1]. Action set is a set of actions. For example, (Service, low  $\rightarrow$  high; Bill, notPaid  $\rightarrow$  paid) is an action set. Action Rule is type of rule that its antecedent is an action set and its consequent is a change in goal attribute. It is shown an action rule below:

(Service, low  $\rightarrow$  high; Bill, notPaid  $\rightarrow$  paid)  $\Rightarrow$  (loyalty, no  $\rightarrow$  yes)

## III. FUZZY ACTION RULE

In this work the fuzzy action concept is proposed. In fuzzy action, unlike the conventional actions, fuzzy alterations are suggested. For instance, fuzzy action suggests the reduction in smoking rather than cessation for pulmonary disease patients. The same is true with fuzzy action and soft drinks for patients. The motivation beyond this is that if fuzzy changes is considered it is possible to find more cost-profitable actions. Suppose attribute  $A$  has  $a_i$  value in a sample. Then, fuzzy action on  $A$  is defined as:

$$\varphi_A = (A, a_i \rightarrow a_i + k\Delta t) \quad (2)$$

Where  $\Delta t$  is the change granularity of attribute  $A$  and can be computed using the following formula:

$$\Delta t = \frac{A_{max} - A_{min}}{n} \quad (3)$$

$A_{max}$ : maximum value of attribute A in dataset.

$A_{min}$ : minimum value of attribute A in dataset.

$n$ : determines to what extent  $\Delta t$  will be fine-grained or coarse-grained. Larger  $n$  results in smaller  $\Delta t$  and therefore more fine-grained fuzzy action sets.

$k$ : is change magnitude and can be positive, negative or zero.

If  $k$  is positive, the value of A increases and if it is negative, it decreases. It is obvious that if  $k$  is considered zero, there is no change in attribute value.

The following example can clarify the concept of fuzzy action:

Consider three attributes  $X_1, X_2$  and  $X_3$  with the values (12, 3, 5). Suppose that  $\Delta t$  and  $K$  are 0.6 and -2 for  $X_1$ , respectively. By the fuzzy action,  $X_3, 5 \rightarrow 5 + (-2 \times 0.6)$ ,  $X_3$  will change from 5 to 3.8. A set of fuzzy actions suggest alterations to multiple attributes and we name it Fuzzy Action Set. For example:

$$\left[ \begin{array}{l} (X_1, 12 \rightarrow 12 + (0 \times 0.8)) \\ \wedge (X_2, 3 \rightarrow 3 + (5 \times 0.4)) \\ \wedge (X_3, 5 \rightarrow 5 + (-2 \times 0.6)) \end{array} \right]$$

Since  $\Delta t$  is constant for each attribute, an action set can be written as:

$$[(X_1, 12, 0) \wedge (X_2, 3, 5) \wedge (X_3, 5, -2)]$$

The above action set suggests that value of attribute  $X_1$  does not change, value of attribute  $X_3$  with an initial value of 5 decreases by  $2\Delta t_{x_3}$  and value of attribute  $X_2$  with an initial value of 3 increases by  $5\Delta t_{x_2}$ .

Fuzzy Action Rule is type of action rule that its antecedent is a fuzzy action set. E.g. a Fuzzy Action Rule can be like this:  
 $[(X_1, 12, 0) \wedge (X_2, 3, 5) \wedge (X_3, 5, -2)] \Rightarrow (G, n \rightarrow y)$

#### IV. GA-FARM

The major problem in finding fuzzy action sets is the search space. Even in a dataset with a few attributes and small domain ranges of attributes the space of all possible fuzzy action sets is huge. So that exhaustive searching of this space is intractable. As a meta heuristic solution genetic algorithm has been used that is quite strong in solving such NP-Hard problems. We are looking for such rules that net profit, i.e. profit minus cost, will be maximum by applying them. At the following, different parts of proposed GA algorithm will be described.

##### A. Chromosome Structure

In GA, each solution is shown as a chromosome and each chromosome contains multiple genes. In our work each chromosome represents a Fuzzy Action Set (FAS) in which each gene is a Fuzzy Action. So that number of genes indicates

number of attributes in FAS and because number of actions in FAS is various chromosomes' length is also variable.

Each gene has three components: attribute name A, current attribute value V and change magnitude  $k$ . So that a gene with structure (A,V,k) represents fuzzy action(A,V $\rightarrow$ V+k $\Delta t$ ).

##### B. Fitness Functions

The most important part of a GA is its fitness function. It is defined as:

$$\text{Fitness}(\text{ch}) = \sum_{i=1}^m \text{profit}(i) \times (\text{md}_{\text{after}}(\text{ch}, i) - \text{md}_{\text{before}}(\text{ch}, i)) - \text{cost}(\text{ch}) \quad (4)$$

Where  $m$  is the number of values of goal attribute,  $\text{profit}(i)$  is the profit of being in class  $i$  that will be provided by domain experts.  $\text{md}_{\text{before}}(\text{ch}, i)$  and  $\text{md}_{\text{after}}(\text{ch}, i)$  indicate the membership degree of the instance to class  $i$  before and after applying the rule, respectively. For example, consider the following chromosome:

$$\text{ch}: [(A, 1, 3), (B, 2, 0), (C, 3, -1)]$$

Also assume  $\Delta t_A = 2, \Delta t_B = 3, \Delta t_C = 1$  and goal attribute G has two values  $G_1$  and  $G_2$ .  $\text{md}_{\text{before}}(\text{ch}, G_1)$  is the membership degree of instance (A = 1, B = 2, C = 3) to class  $G_1$  and  $\text{md}_{\text{after}}(\text{ch}, G_1)$  is the membership degree of instance (A = 1 + 3 \* 2, B = 2 + 0 \* 3, C = 3 + (-1) \* 1) to class  $G_1$ . Here fuzzy decision tree will be used for computing these values.

Cost will be computed using the following formula:

$$\text{cost} = \sum_{i=1}^n (k \times \text{cost}_i(\Delta t)) \quad (5)$$

Where  $\text{cost}_i(\Delta t)$  is the cost of changing attribute  $i$  with value  $\Delta t$ . Cost values will be fed into model by domain experts.

##### C. Crossover

Crossover is responsible for exploiting search space in GA. It tries to combine two good chromosomes with the hope of generating more fitted offspring. Here a one-point crossover operation has been designed for combining the most fitted chromosomes probabilistically.

For each pair of chromosomes, which are selected using roulette wheel selection strategy, two offspring will be generated by exchanging their genes around crossover point which is determined probabilistically. Designed crossover has been depicted in Fig. 2 for two sample chromosomes.

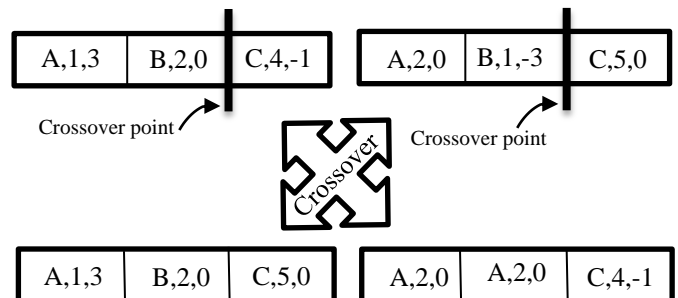


Fig. 2. One-point crossover operation used in GA-FARM.

D. Mutation

Mutation is the core of exploration mechanism of GA. It helps GA exit from local optimums. Mutation operator tries to mutate some genes of a chromosome with the hope of generating a new chromosome from other parts of the search space.

In GA-FARM a predefined percentage of chromosomes are muted by changing value *k* of some of their genes by a random value. In Fig. 3, the result of mutation on a sample chromosome has been depicted.

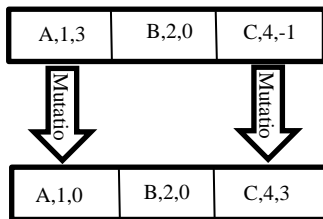


Fig. 3. Mutation operation used in GA-FARM.

V. EXPERIMENTAL RESULTS

For experimental results GA-FARM has been tested on three different datasets: Bank Loan dataset which has gathered from one of the most prominent banks in IRAN (dataset description has been shown in Table II), Pima Indians Diabetes [19] and Turkey Student Evaluation [20] both from UCI Machine Learning Repository. The summary of these datasets can be seen in Table I.

One of the major problems in action mining domain is evaluating the mined actions because it is necessary to apply them in real world. To investigate the effect of proposed method we have designed two research questions: How logical action rules will be mined using FARM and whether GA can converge in acceptable time or not. The former question is subjective while the latter is objective.

Fig. 4 shows the results of applying the GA-FARM on bank data sets. It shows the convergence of GA for this dataset. As it can be seen the GA is successful in improving the overall fitness of population. Elapsed time for this dataset is 281.31 seconds which confirms that GA-FARM has acceptable time complexity.

TABLE I. DATASET DETAILS

Dataset	No. of Instances	No. of Attributes	No. of Continuous Attributes	No. of class attributes
Bank Loan	199	10	7	3
Pima Indians Diabetes	768	8	7	2
Turkey Student Evaluation	5820	33	0	5

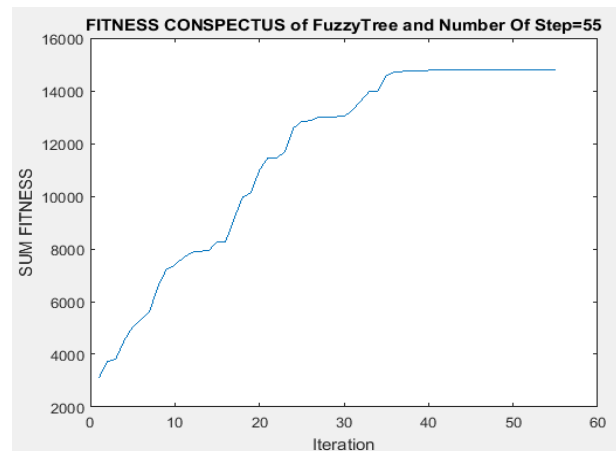


Fig. 4. Results of applying the GA-FARM on bank loan dataset.

One of the mined rules using GA-FARM from bank loan dataset has been depicted below:

$$\left[ \begin{array}{l} (X_1,0,0) \wedge (X_2,36,0) \wedge (X_3,0,-2) \\ \wedge (X_4,1,1) \wedge (X_5,2,-1) \wedge (X_6,0,3) \\ \wedge (X_7,0,3) \wedge (X_8,770000,0) \wedge \\ (X_9,770000,0) \wedge (X_{10},32,-2) \end{array} \right] \Rightarrow [\text{Loyal, No} \rightarrow \text{Yes}]$$

This rule shows alterations of the values of two attributes: marital status and the number of deferred payments. There is no change in the attributes where *k* is zero. For instance, for attribute #10 which is the number of deferred installments, -2 is suggested for *k* and 0.5 is computed for Δ*t*. Based on this rule, the customer must reduce his or her deferred paid installments by -2 × 0.5, i.e. from 32 to 31. By applying this rule, the bank may reach a net profit of 190. This action rule shows that it is possible to achieve good profit but with less changes in attributes. This is the motivation behind fuzzy action rule.

TABLE II. BANK LOAN DATASET ATTRIBUTES

X	Name of Attribute
X1	sex
X2	Age
X3	Marital Status
X4	Degrees of Education
X5	Job
X6	Number of bank transactions during last 3 months
X7	Number of Electronic Services used during last 30 transactions
X8	6-month average balance
X9	12-month average balance
X10	Number of deferred instalments

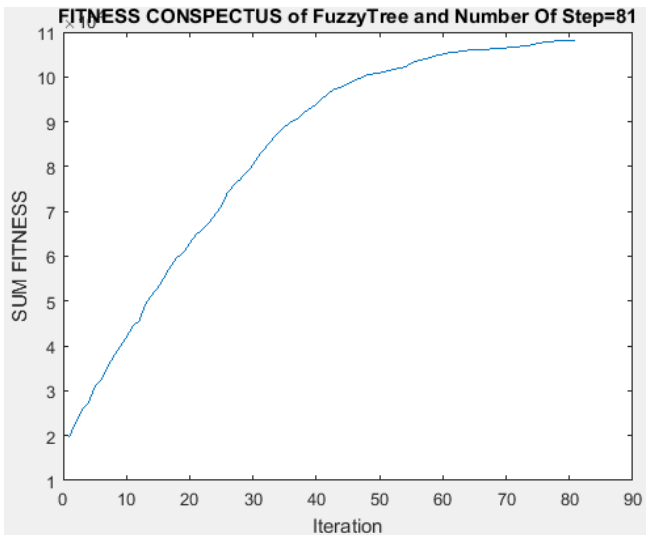


Fig. 5. Results of applying the GA-FARM on Pima Indians Diabetes dataset.

By applying proposed method on Pima Indian Diabetes dataset it is succeeded to find the most profitable fuzzy action Rules. Fig. 5 shows the convergence of GA-FARM on this dataset. One of the mined action rules is:

$$\left[ \begin{array}{l} (X_1,4,0)\wedge(X_2,129,0)\wedge(X_3,60,-3) \\ \wedge(X_4,12,-9)\wedge(X_5,231,0) \\ \wedge(X_6,27.5000,-7)\wedge(X_7,0.5270,-1) \\ \wedge(X_8,31,0) \end{array} \right] \Rightarrow [\text{Diabetes, Yes} \rightarrow \text{No}]$$

This rule suggests that the patient can reduce the probability of getting diabetes by losing 7Δt(4.697) kg weight.

Finally, GA-FARM is applied on Turkey Student Evaluation dataset. The GA convergence graph has been depicted in Fig. 6 which shows it succeeded in finding local optimum fuzzy action sets from this dataset. One of the mined action rules has been depicted below.

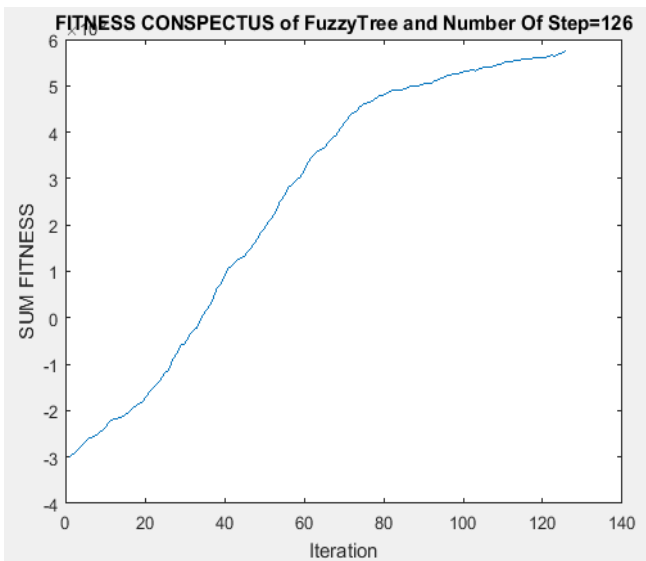


Fig. 6. Results of applying the GA-FARM on Turkey Student Evaluationdataset.

$$\left[ \begin{array}{l} (X_1,1,0)\wedge(X_2,1,0)\wedge(X_3,1,0) \\ \wedge(X_4,3,0)\wedge(X_5,3,0)\wedge(X_6,4,0) \\ \wedge(X_7,4,0)\wedge(X_8,4,0)\wedge(X_9,4,0) \\ \wedge(X_{10},4,0)\wedge(X_{11},4,1)\wedge(X_{12},4,0) \\ \wedge(X_{13},4,2)\wedge(X_{14},4,0)\wedge(X_{15},4,2) \\ \wedge(X_{16},4,0)\wedge(X_{17},4,1)\wedge(X_{18},4,0) \\ \wedge(X_{19},4,2)\wedge(X_{20},4,0)\wedge(X_{21},4,0) \\ \wedge(X_{22},4,0)\wedge(X_{23},4,0)\wedge(X_{24},4,3) \\ \wedge(X_{25},4,0)\wedge(X_{26},4,0)\wedge(X_{27},4,0) \\ \wedge(X_{28},4,0)\wedge(X_{29},4,3)\wedge(X_{30},4,0) \\ \wedge(X_{31},4,0)\wedge(X_{32},4,0)\wedge(X_{33},4,0) \end{array} \right] \Rightarrow [\text{raise points, low} \rightarrow \text{high}]$$

Fuzzy action rules mined from datasets show that GA-FARM can extract meaningful rules in corresponding domains. The elapsed time for GA and number of epochs before convergence show that GA operations succeeded.

## VI. CONCLUSION AND FUTURE WORKS

Action Rule is a set of change suggestions with the aim of converting an instance from less profitable class to more profitable one. Current state of the art methods in this domain only suggest crisp changes in attribute values. In this paper the concept of Fuzzy Action Rule has been introduced which suggests fuzzy changes. Applying fuzzy changes help us to gain profit with less cost comparing to crisp changes.

A big burden against finding fuzzy action rules is the huge search space. So that, GA-FARM has been devised for extracting fuzzy action rules using GA. The proposed method then has been implemented and tested on different datasets. Results confirmed that GA-FARM can find optimum fuzzy action sets in acceptable time.

The most important limitation in the field of action mining is about evaluating the results because checking the effect of an action needs to apply it in real world. It is necessary to design some frameworks that can evaluate the effect of actions more realistically. For future works we will also try to devise new FARM methods that can find more profitable fuzzy rules. Incorporating fuzzy type-2 into action rules can be another promising research area.

## REFERENCES

- [1] P. Shamsinejadbabaki, M. Saraee, and H. Blockeel. "Causality-based Cost-effective Action Mining", International Journal of Intelligent Data Analysis , Volume 17 Issue 6, Pages 1075-1091, 2013.
- [2] Z. W. Ra's and L. S. Tsay, "Discovering Extended Action-Rules (System DEAR)", Intelligent Information Processing and Web Mining. Advances in Soft Computing, Springer Berlin Heidelberg, vol. 22, pp. 293-300,2003.
- [3] Z. W. Ra's and A. Dardzińska, "Action Rules Discovery, a New Simplified Strategy", Foundations of Intelligent Systems. ISMIS 2006. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, Vol. 4203, pp. 445-453,2006.
- [4] Z. W. Ra's, A. Dardzińska, L. S. Tsay and H. Wasyluk, "Association Action Rules". IEEE International Conference on Data Mining Workshops, vol. 166, pp. 283-290, 2008.
- [5] Z. W. Ra's, L. S. Tsay and A. Dardzińska, "Mining E-Action Rules", Journal of Mining Complex Data, Springer-Verlag Berlin Heidelberg , 2008.
- [6] Z. W. Ra's, and A. Wiczorkowska, "Action-Rules: How to increase profit of a company", European Conference on Principles of Data

- Mining and Knowledge Discovery, Springer Berlin Heidelberg ,Vol. 1910, pp. 587–592, 2000.
- [7] Z. W. Ra´s, and A. Dardzinska, “From Data to Classification Rules and Actions”, International Journal of Intelligent Systems, Vol. 26(6), pp. 572–590, 2011.
- [8] Z. W. Ra´s, A. Tzacheva, L. S. Tsay and O. G¸urda, “Mining for interesting action rules”, international conference on intelligent agent technology, IEEE Computer Society, 187–193, 2005.
- [9] Z. W. Ra´s, A. Dardzi ´nska, L.S. Tsay, L.S and H Wasyluk, “Association action rules”. In IEEE International conference on data mining workshops, 283–290, 2008.
- [10] L.S. Tsay and Z. W. Ra´s, “Action rules discovery: system DEAR2, method and experiments”. Journal of Experimental & Theoretical Artificial Intelligence, Vol. 17, No. 1–2, pp. 119–128, 2005.
- [11] L.S. Tsay and Z. W. Ra´s, “Action Rules Discovery System DEAR3”, International Symposium on Methodologies for Intelligent Systems, Springer, pp. 483–492, 2006.
- [12] J. Kuang and Z. W. Ra´s, “In Search for Best Meta-Actions to Boost Businesses Revenue”, Flexible Query Answering Systems, Springer, 2016.
- [13] H. Touati, Z. W. Ra´s, J. Studnicki, and A. A. Wiczorkowska, “Mining Surgical Meta-actions Effects with Variable Diagnoses’ Number”, International Symposium on Methodologies for Intelligent Systems Springer, pp. 254–263 , 2014.
- [14] Q. Yang and C. Hong, “Mining Case Bases for Action Recommendation”, IEEE International Conference on Data Mining, pp. 522 – 529, 2002.
- [15] Q. Yang, Jie Yin, C.X. Ling and T. Chen, “Postprocessing decision trees to extract actionable knowledge”. Third IEEE International Conference on Data Mining, pp. 685–688, 2003.
- [16] N. Kalanat, P. Shamsinejad, and M. Saraee, "A Fuzzy Method for Discovering Cost-effective Actions from Data", Journal of Intelligent and Fuzzy Systems, pp.757-765, 2014.
- [17] H. Sattar and Y. Yang, “Flexible Decision Trees for Mining High Speed Data Streams at Presence of Noise and Uncertainty”, Data Mining and Knowledge Discovery, vol. 19, pp. 95–131, 2009.
- [18] A. Idri and S. Elyassami, “A Fuzzy Decision Tree to Estimate Development Effort for Web Applications” (IJACSA) International Journal of Advanced Computer Science and Applications, 2011.
- [19] G. Gunduz and E. Fokoue, “UCI Machine Learning Repository”, Irvine, CA: University of California, School of Information and Computer Science, 2013.
- [20] V. Sigillit, UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2013.

# A Secured Interoperable Data Exchange Model

A. Bahaa, A. Sayed and L. Elfangary

Department of Information Systems, Faculty of Computers and Information,  
Helwan University, Egypt

**Abstract**—Interoperability enables peer systems to communicate with each other and use the functionality of peer systems effectively. It improves ability for different systems to exchange information between cooperative systems. It plays a vital role in educational information system institutions. Practically, there are two main technical reasons that restrain the interoperability of the system. First, these systems may be developed under various operating systems, programming languages and different database management systems. Second, the obsessions of security greatly impact the execution of interoperability among various educational institutions. This paper proposes a new RESTful secured interoperable model for data exchange among different information system. This will help educational information system to exchange data among them with a pre-defined standard format of messages. Additionally, this paper designed Cross Platform Web Application Interoperability Protocol (CPWAIP) to facilitate the interaction among components of the proposed model.

**Keywords**—Data sharing; security; integrity; and protection

## I. INTRODUCTION

The IEEE defines interoperability as the “Ability of a system or a product to work with other systems or products without special effort on the part of the customer” [1], [2]. Interoperability means that different systems are related together with removed incompatibilities between them. It improved ability for different systems to exchange information between cooperative systems. Semantic interoperability, structural interoperability and syntactic interoperability are three different levels of interoperability [3].

Hence, Interoperability plays a vital role for educational information system to make the content accessible in different systems as well as by collaborative and cross-organizational learning and teaching [4]. Interconnection between various educational systems in order to exchange information about students who move from campuses to another are needed nowadays [5]. Practically, there are two main technical reasons that restrain the interoperability of the system. First, these systems may be developed under various operating systems, programming languages (i.e., Java, and Dot.Net), and different database management systems (DBMS) (i.e., SQL Server, MySQL, and Oracle) and standards that make it difficult to achieve data sharing and interoperability among them [6]. Second, the obsessions of security greatly impact the execution of interoperability among various information system institutions.

Many models or systems have been proposed to solve data exchange among various information systems. However, most of the proposed models are based on peer-to-peer

communication among information system [7], [8]. A peer-to-peer communication imposes new security challenges to interoperability among information systems. Thus, there is a need for new models that handle these security issues.

In this paper, the researchers propose a new RESTful secured interoperable model among different educational information system. This will help educational information system to exchange data among them with standard format of messages. Limitation of proposed model is suitable only for educational information system. Additionally, the researchers were designed Cross Platform Web Application Interoperability Protocol (CPWAIP) to facilitate the interaction between internal component and external components of the proposed model.

The rest of the paper is organized as follows: Section 2 provides a background overview for cloud computing and web service. Section 3 describes the proposed model components. Section 4 Applying proposed model between different educational information systems. The last section concludes the paper with final remarks.

## II. BACKBOARD OVERVIEW

This section consists of two parts. The first part presents introduction for cloud computing. The second part of introduction is for web service. The third part presents the related work focusing on interoperability between different educational information systems.

### A. Cloud Computing

Recently, cloud computing is a hot topic all over the world [9], [10]. In 2006, Google’s CEO, Eric Schmidt, proposed the word “cloud” to describe the business model of providing services across the internet. The term cloud was used as marketing concept [11]. Cloud computing means to provide remote service to users and customers to store and process data without the need to having hardware equipment. It, also, provides the ubiquitous network access anywhere, anytime and from any platforms [12]. It is considered as a sharing architecture of the IT trends, in which a third party provides highly scalable, reliable on demand software, hardware, and infrastructure services with agile management capabilities [13], [14]. Cloud computing is divided into three major types of services; public, private and hybrid [15]-[17].

Public cloud provides an open environment that enables any user to access the service over the Internet. Private cloud concerns data security and provides smooth control that is not available in public cloud. Hybrid cloud is the combination of public and private cloud.



Services offered by cloud providers are of three types; Infrastructure as a Service (IaaS), Platform as aService (PaaS) and Software as aService (SaaS) [18]-[20].

### B. Web Service

Web service is defined as an interface that helps desperate, heterogeneous environments to communicate among each other effectively, in the form of XML messages (Extensible Markup Language) or JSON [21]. Web services have become a popular way of offering online services by businesses [22], [23]. Simple Object Access Protocol (SOAP) over HTTP is traditionally web service which provides a decentralized, distributed XML-based messaging framework between peers. SOAP is an xml based Remote Procedure Call (RPC) solution while HTTP is a much more lightweight solution where resources are managed by HTTP interactions [24].

REST (Restful Sate transfer) is another inherently resource oriented service [25], [26]. REST is an architectural style when used in applications that utilize HTTP features (URI, response code, and query-methods GET, POST, PUT, and DELETE) to work on the API users [27].

### C. Related Work

There are many studies conducted to solve the problem of interoperability among different information systems. For example:

- The Ministry of Education of China [28] proposed Education Management Information System Interoperability Framework (EMIF) to address the challenges of sharing data and integrating different colleges and departments. EMIF used SOA (Service Oriented Architecture) in integrating various EMIS in tertiary education.
- Z. Xiao-guang, et al. [29] uses SOA to apply interoperability between medical information systems (MISs). So, Each MISs build services interface without modifying the existing systems in each hospital. The proposed model was used to exchange information between different MISs.
- D.Zhou [30] proposed SOA-based education information system interoperability model to improve the interoperability of educational information systems. The proposed model is based on WCF and SOA. This proposed model is composed of the Education Information Interoperability Center (ZIS) and Agents. ZIS is used to exchange information between registered agents.
- SIF Association [31] proposed School Interoperability Framework (SIF) to enable interoperability and data sharing among different educational information systems. SIF specification consists of two key parts: SOA and XML. SOA specification aimed for sharing information between institutions. XML specification aimed for modeling educational data according to the educational locale.
- A. A.Chandio, et al. [32] proposed a system integration of interconnectivity of information system (i3) for the

University of Sindh (UoS) Pakistan. This system is designed to share and exchange the information associated with students of different departments in the institution. The system i3 is based on SOA and XML.

TABLE I. SUMMARY OF THE PREVIOUS WORK

Author	Year	Focus	Shortcomings
The Ministry of Education of China [28]	2009	Proposed interoperability solution among different educational information systems	Each educational information system established special connection with other educational information system database.
Z.Xiaoguang [29]	2009	Apply interoperability between MISs in each hospital	Propose a system is only suitable for healthcare institutions and did not secure message transmitted.
D.Zhou [30]	2011	Proposed SOA-based education information system interoperability model	Proposed model did not pre-defined standard format for sending and receiving messages between agents and did not secure message transmitted.
SIF Association [31]	2014	Proposed interoperability solution among different educational information systems	Lack of efficiency in transmitting large data messages among different educational information systems.
A.A.Chandio [32]	2014	Propose a system integration of interconnectivity of information system (i3) for the University of Sindh (UoS) Pakistan	Propose a system is only suitable for institutions facing similar problems of University of Sindh (UoS).
R.Jessadapatharakul [7]	2015	proposed data exchange protocol for healthcare service in Thailand	Each healthcare system established direct access to the systems of the other healthcare institutions.
A. Sayed [8]	2016	Proposed data exchange protocol for educational information system.	Each educational system established peer-to-peer communication among all systems of educational information system

- R. Jessadapatharakul, et al. [7] proposed data exchange protocol for healthcare service in Thailand. This model is used to data exchange system by using cloud-based service platform. This platform is based on PaaS (Platform as a service) to provide a service for health institutes. Healthcare data is exchange between medical institutes under pre-defined standard.
- A. Sayed, et al. [8] proposed interoperable architecture for educational software systems. This paper introduced data exchange platform for educational information system based on RESTful web service. The proposed data exchange platform for educational system is based on a cloud-based service platform.

All proposed architectures are a good starting point for addressing the problem of data exchange among different kind of information system institutions. Unlike these researches, our work tries to:

- Proposes a new RESTful secured interoperable model among different educational information system.
- Design Cross Platform Web Application Interoperability Protocol (CPWAIP) to facilitate the interaction between internal component and external components of model.

Table I presents a summary of all the works that were reviewed. This table includes the authors, the publication year, the focus of the study and its main shortcomings.

### III. PROPOSED MODEL

The main objective of proposed model is to exchange data among multiple educational system institutions. In the proposed model, a cloud-based service platform is using for data exchange platform for educational which is based on PaaS (Platform as a service) to provide a service for exchange and conversion of data into a pre-defined standard format. Additionally, Cross Platform Educational Application protocol (CPEAP) designed to facilitate the interaction between internal component and external components.

#### A. Component of the Proposed Model Architecture

The component of the proposed model can be classified into two categories: internal component and external component. The internal component includes services which provide on this platform. The external component is used to communicate with internal component. As shows in Fig. 1, the proposed model consists of the following components:

- 1) Message Queue
- 2) Directory services
- 3) Web service endpoint
- 4) Web Application Interoperability System
- 5) Web based –API/Services
- 6) Cross Platform Web Interoperability Application System
- 7) Information Conversion Services
- 8) Data Base

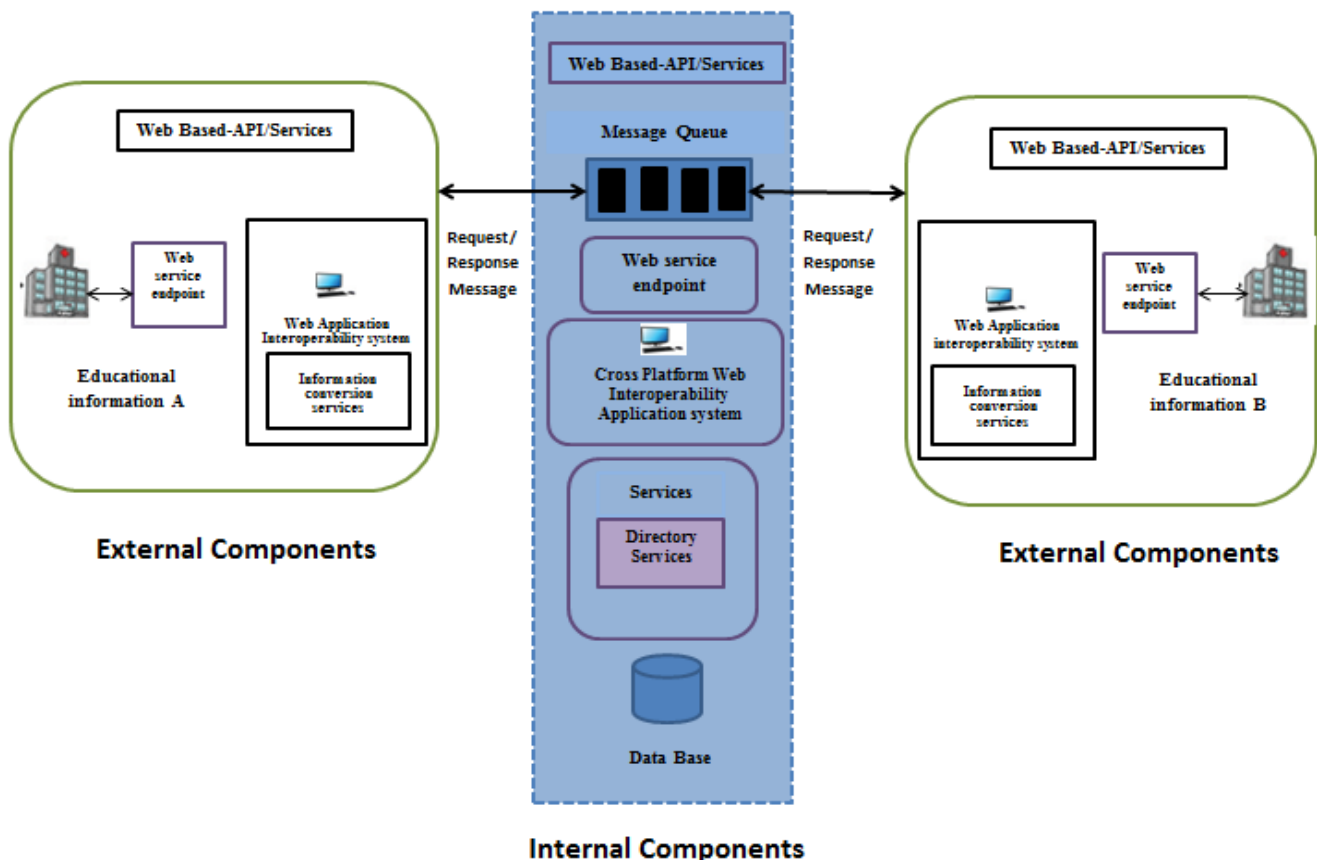


Fig. 1. General interoperable model.

TABLE II. PROPOSED WEB SERVICE ENDPOINT AT INTERNAL COMPONENT

NO	HTTP Method	URI	Operation
1	GET	/CPWAIS/Search	Invoking search operation of different WAI system

TABLE III. PROPOSED WEB SERVICE ENDPOINT AT EXTERNAL COMPONENT

NO	HTTP Method	URI	Operation
1	GET	/WAI/Search	Search for data

The main components of the proposed model are discussed briefly in the following subsections:

*a) Message queue*

Message Queue provides an asynchronous messaging service that facilitates huge amount of concurrent messages among various external components.

*b) Directory services*

It is used for a central directory that keeps the educational information data. It contains educational UUID, educational name, URL methods invoke and public key infrastructure (PKI).

*c) Web service endpoint*

Web service endpoint is a web address (URL) which will return response messages with a pre-defined standard to client according to request message. Both internal component and external component build its own web service endpoint to exchange messages with each other. Table II shows proposed web service endpoint at internal component. Table III shows proposed web service endpoint at external component.

*d) Web application interoperability system*

Web application Interoperability System (WAI) is web application interface used for communicating with internal components to retrieve response messages.

*e) Web Based-API/Services*

Web Based-API/Services is an application programming interface (API) used when internal and external components application needs to access web service endpoint of internal and external components.

*f) Cross platform web interoperability application system*

Cross Platform Web interoperability Application System (CPWAIS) is web application interface that enables a system to communicate with the other authorized systems in order to exchange data.

*g) Information conversion services*

It is used for compose and decompose educational data into a pre-defined standard. This data is based on JSON that is proposed in this model.

*h) DataBase*

Database is used to store registration data of educational information system like educational name, region, country, URL methods invoke and PKI. It also is used to cache request message of educational information system and response messages from other educational information system. Cached data will be released after 6 months. In registration form, each educational information system determines if the data will be cached on database.

*B. Cross Platform Web Interoperability Application Protocol*

CPWAIP uses HTTP as a protocol to communicate by defining a request and response message between internal and external components. This section presents application protocol standard format, security and privacy and mechanism and usage scenario.

*1) Application Protocol Standard Format:* In proposed CPWAIP, a request and response message contains HTTP Header and HTTP Body. HTTP Header contains some meta-data of sender educational information system. HTTP Body contains request and response query from educational information system. Table IV shows proposed HTTP headers which starts with a prefix “WAI-Service”.

TABLE IV. PROPOSED HTTP HEADERS FOR CPWAIP

Header keys	Description
WAI-Service- UUID	Unique identifier for a certificate to identify WAI system.
WAI -Service- Public key	Public key that is used to encrypt body of message.
WAI -Service-WAI Receiver	Value of WAI system that we want to get data from

TABLE V. DATA STANDARD FORMAT FOR SEND QUERY

Attribute	Type	Null option	Description
UUID(PK)	Integer	not null	ID Query
Student ID	Integer	not null	ID student
Prenome	String	Null	Title
Name	String	Null	Name, middle name
Surname	String	Null	Last name
Level	Integer	not null	Student level, level: 1,2,3,4
Educational system	String	not null	ID of Educational system that we want to query From.
Student type	String	not null	Type of student , student Type: credit hours, student Type: general
Search Type	String	not null	status, transcript, year grades

TABLE VI. DATA STANDARD FORMAT FOR RESPONSE QUERY

Attribute	Type	Null option	Description
UUID(PK)	Integer	not null	ID Query
Student ID	Integer	not null	ID student
Prename	String	Null	Title
Name	String	Null	Name, middle name
Surname	String	Null	Last name
Level	Integer	not null	Student level ,level :1 ,2,3,4
GPA	Integer	Null	Student GPA
Status	String	not null	pass, fail
Year grades	JSON	Null	{ "level1": "70", "level 2": "80", "level 3": "90", "level 4": "90" }
Hours	Integer	Null	Student Hours
Subject	JSON	Null	{ "subject 1": "70", "subject 2": "80", "subject 3": "90", "subject 4": "90", ... }
Total Grades	Integer	Null	Student total Grades

According to use different format for storing data, each educational information system have to pre-defined standard format for sending and receiving messages between them. Table V shows proposed data standard format for sending a query to other educational information system. The first column is the attribute. The second is the type of each attribute. The third column is the Null option, which specifies whether the field can be empty in some case. Finally, it is a description of attributes.

Table VI shows proposed data standard format for response from educational information system. The first column is the attribute. The second is the type of each attribute. The third column is the Null option, which specifies whether the field can be empty in some case. Finally, it is a description of attributes.

2) *Security and Privacy*: The security and privacy of educational information system data is the important issue for educational information system, so this model need to secure data transmitted between external components and internal components as following (Fig. 2):

- Each educational information system needs to register with CPWAIS by creating an account to log into the system. The CPWAIS generates unique identifier (UUID) for a certificate.
- Using PKI keys (Public Key Infrastructure) encrypt all data before transmission. Each registered educational information system generates two PKI key sets.

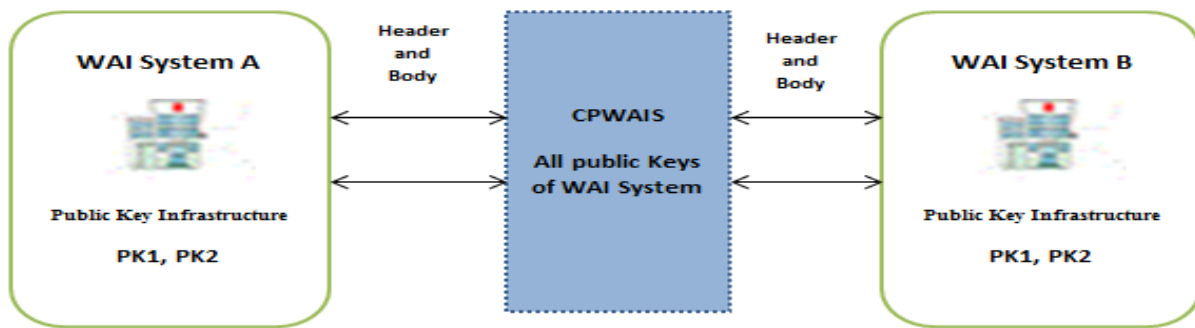


Fig. 2. Security and privacy in model.

Fig. 3 shows secure information flow when educational system A request data from educational system B through CPWAIS. (M= original message, Enc = encryption function, Dec = decryption function, ES= educational system, CPWAIS= Cross Platform Web interoperability Application System)

$$\text{Enc}_{ES_A}(\text{Privatekey1}_{ES_A}, M) \Rightarrow M' \Rightarrow \text{Dec}_{CPWAIS}(\text{public Key1}_{ES_A}, M') \Rightarrow M \Rightarrow \text{Enc}_{CPWAIS}(\text{public key1}_{ES_B}, M) \Rightarrow M'' \Rightarrow \text{Dec}_{ES_B}(\text{private Key 1}_{ES_B}, M'') \Rightarrow M$$

Fig. 3. Secure information flow for requested data.

Fig. 4 shows secure information flow response data from educational system B after educational system, A retrieves requested data. (M = original message, Enc = encryption function, Dec = decryption function, ES= educational system, CPWAIS= Cross Platform Web interoperability Application System).

$$\text{Enc}_{ES_B}(\text{Publickey2}_{ES_A}, M) \Rightarrow M' \Rightarrow \text{Dec}_{ES_A}(\text{PrivateKey2}_{ES_A}, M') \Rightarrow M$$

Fig. 4. Secure information flow for response data.

$$\text{Enc}_{\text{ES}_B}(\text{Privatekey1}_{\text{ES}_B}, M) \Rightarrow M' \Rightarrow \text{Dec}_{\text{CPWAIS}}(\text{PublicKey1}_{\text{ES}_B}, M') \Rightarrow M \Rightarrow \text{Enc}_{\text{CPWAIS}}(\text{Publickey2}_{\text{ES}_A}, M) \Rightarrow M' \Rightarrow \text{Dec}_{\text{ES}_A}(\text{PrivateKey2}_{\text{ES}_A}, M') \Rightarrow M$$

Fig. 5. Secure information flow for cached response data.

Fig. 5 shows secure information flow response data from educational system B after educational system A retrieves requested data and response data cache in CPWAIS. (M = original message, Enc = encryption function, Dec = decryption function, ES= educational system, CPWAIS= Cross Platform Web interoperability Application System).

3) Mechanism and Usage Scenario: Fig. 6 shows sequence diagram for full registration of information system institutes. Information system institutes fills application form and submits data to CPWAIS. CPWAIS auto validate enter data if data is true CPWAIS generates Universally Unique Identifier (UUID) and sends to register Information system institutes. Educational system receives and stores UUID in its own transaction database then it generates two keys set (PK1, PK2) then sends public key of first key set to CPWAIS which shall be stored it in its own transaction database.

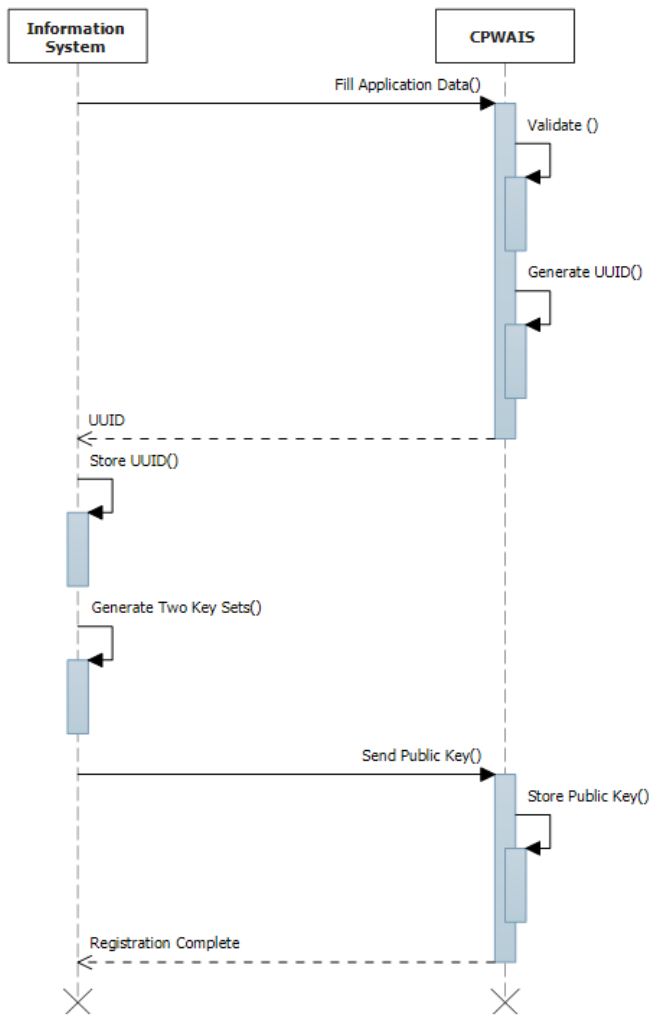


Fig. 6. Sequence diagram for full registration with CPWAIS.

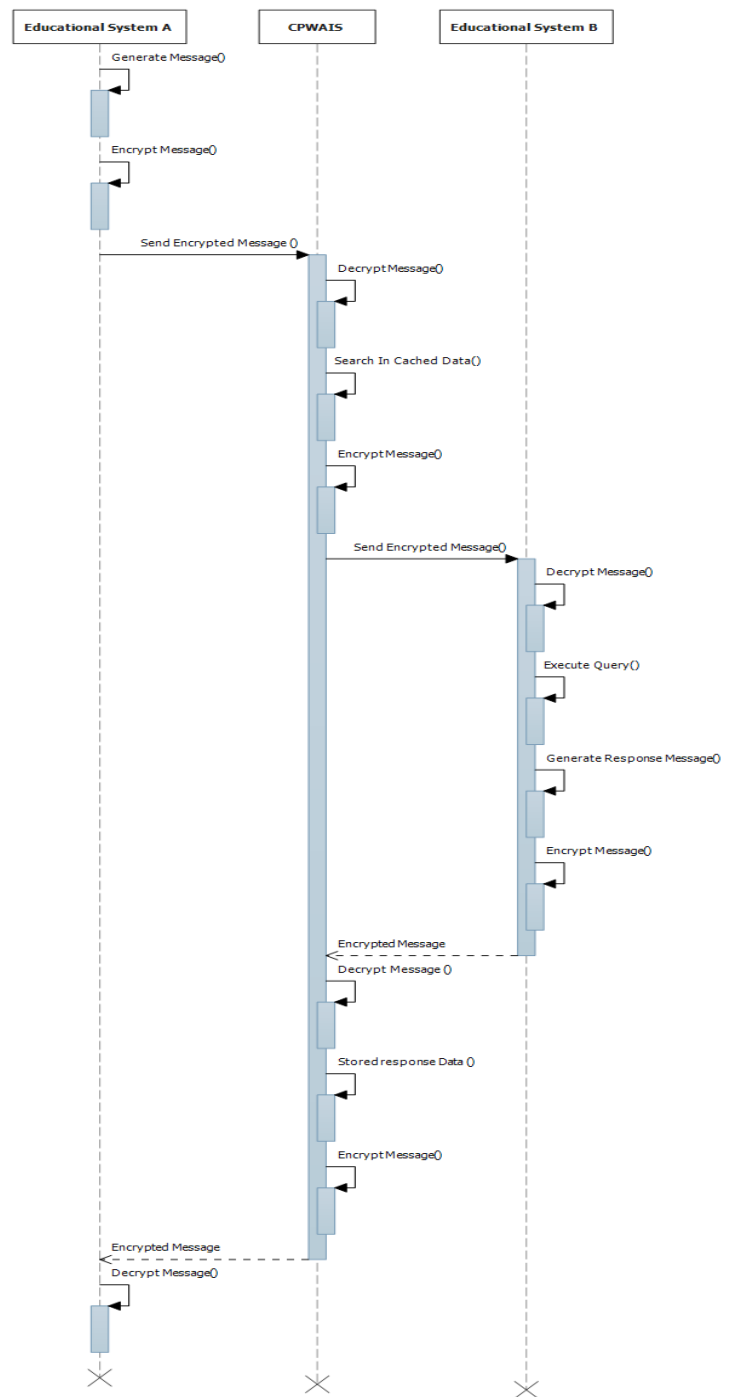


Fig. 7. Sequence diagram for requesting not cached data in CPWAIS.

Fig. 7 shows sequence diagram when registered educational system A requests from other registered educational system B and requested data is not cached in CPWAIS and educational system B allows caching data at CPWAIS. Educational system A identifies requested data using proposed a pre-defined standard format for sending message. Educational system A encrypts message and HTTP header and sends it to CPWAIS. CPWAIS decrypts message and header. CPWAIS don't find requested data in cached data. CPWAIS encrypts message and header using public key of educational system B. Educational

system B receives encrypted message and encrypts it using his own private key (PK1). Educational system B executes query for requested data and format response with propose a pre-defined standard format for response message. Educational system B encrypts response message with own private key (PK1) and sends it to CPWAIS. CPWAIS encrypts message using stored public key (PK1) of WAI system B. CPWAIS caches data in its own data base and encrypts message using public key (PK2) of educational system A. educational system A receives encrypted message and decrypts message using its own private key (PK2).

Fig. 8 shows sequence diagram when registered educational system A requests same data from other registered educational system B, which is cached in CPWAIS. Educational system A identifies request data using propose a pre-defined standard format for sending message. Educational system A encrypts message and HTTP header and sends it to CPWAIS, then it decrypts message and header. CPWAIS search for the request in cached data on CPWAIS data base and finds requested data on cached database. CPWAIS encrypts requested data using public key (PK2) of educational system A and will send encrypted message to it. Educational system A receives encrypted message and will decrypts message using own private key (PK2).

#### IV. APPLING PROPOSED MODEL AMONG DIFFERENT EDUCATIONAL SYSTEM

The proposed model was applied between two universities in Egypt. Cross Platform web Interoperability Application System CPWAIS was hosted on a cloud environment. Amazon is the service provider that was selected for hosting the system. Fig. 9 shows the registration web form used for the system users to enter their educational name, region, country, URL method, etc. Then, CPWAIS generated UUID for the registered universities as shown in Fig. 10. Fig. 11 shows the public keys generated for securing data exchange. The web form presented in Fig. 12 was used for retrieving the data form the different registered systems.

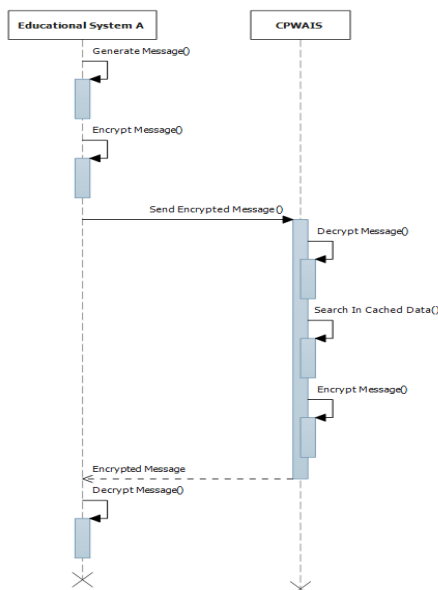


Fig. 8. Sequence diagram for requesting cached data in CPWAIS.

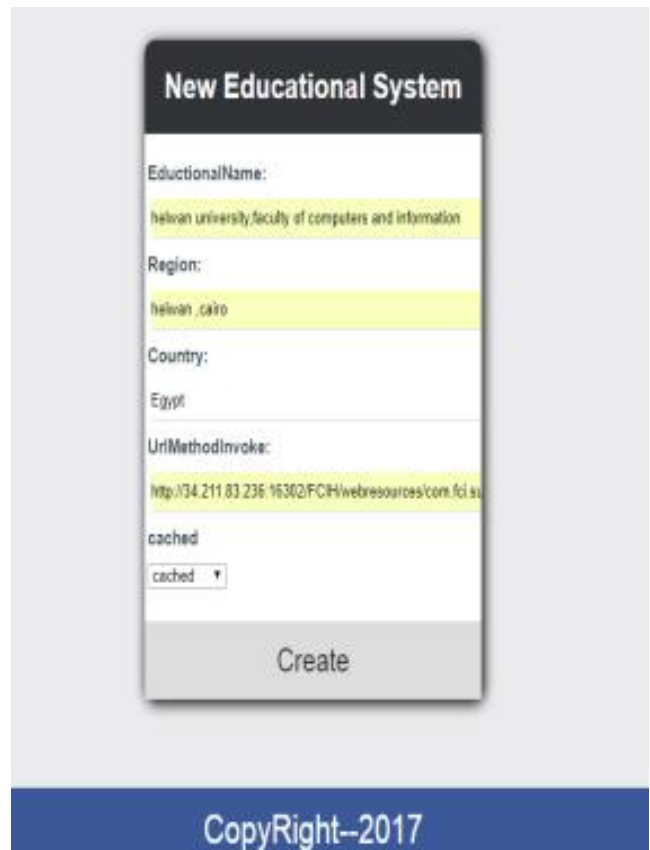


Fig. 9. Registration form in CPWAIS.

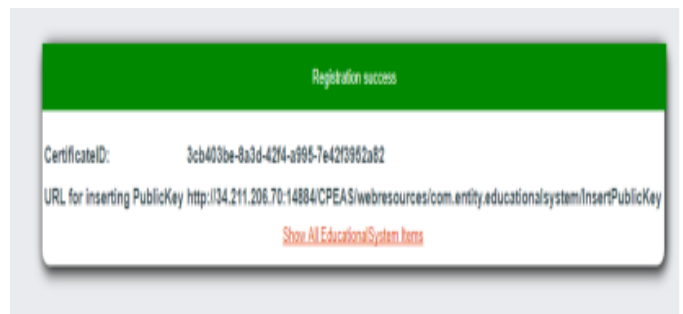


Fig. 10. UUID for educational information system.

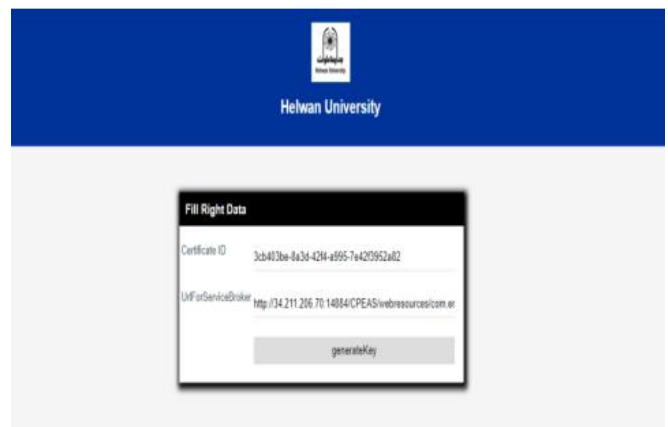


Fig. 11. Two PKI sets for educational information system.



Fig. 12. Retrieve data from educational information system.

## V. CONCLUSION AND FUTURE WORK

Interoperability plays a vital role in information system institutions. However, most of the proposed models are based on peer-to-peer communication among information system. This paper proposed a novel interoperable model to secure exchanging data among different system. Additionally, Cross Platform Educational Application protocol (CPEAP) designed to facilitate the interaction between CPWAIS and different information systems. The proposed model was applied on the educational information systems in Egypt. This model enhanced the security aspects for data exchange among different information systems. It is recommended as a future work to apply the proposed model in other environments, such as healthcare, e-government, etc.

### REFERENCE

- [1] L. Cardoso, F. Marins, C. Quintas, F. Portela, M. Santos, A. Abelha and J. Machado, "Interoperability in Healthcare", Cloud Computing Applications for Quality Health Care Delivery, 2014.
- [2] "The Role of Standards in Engineering and Technology", IEEE Standards Position Paper. June 2007.
- [3] S. Yunmei, L. Xuhong, X. Yabin and J. Zhenyan. "Semantic-Based Data Integration Model Applied to Heterogeneous Medical Information System", (ICCAE) The 2nd International Conference on Computer and Automation Engineering. Singapore, vol. 2, pp. 624- 628, February 2010.
- [4] A. W. P. Fok and H. H. S. Ip, "Educational ontologies construction for personalized learning on the web", In Evolution of teaching and learning paradigms in intelligent environment, Springer Berlin Heidelberg, 2007.
- [5] M. Yarime, G. Trencher, T. Mino, R. W. Scholz, L. Olsson, B. Ness and J. Rotmans, "Establishing sustainability science in higher education institutions: towards an integration of academic development", 2012.
- [6] Bo Zhou and Wen Liang Liu, "Integrated Query of Multi-Resources Information in the Way of SaaS Application", 2012.
- [7] Ruedeemart Jessadapatharakul, Santitham Prom-on, Chularat Tanprasert and Tiranee Achalakul, "Data exchange protocol for healthcare service in Thailand", 2015.
- [8] A.Sayed, A. Bahaa and L. Elfangary, "A Proposed Interoperable Architecture For Educational Software System", 2016.
- [9] Majed AlOtaibi, Lo'ai A. Tawalbeh and Yaser Jararweh, "Integrated sensors system based on IoT and mobile cloud computing", 2016.
- [10] Qassim Bani Hani and Julius P. Ditcher, "Mobile-Based Location Tracking without Internet Connectivity Using Cloud Computing Environment", 2017.
- [11] Qi Zhang, Lu Cheng and Raouf Boutaba, "Cloud Computing: State-of-the-art and research challenges", 2010.
- [12] Nancy Jain and Sakshi Choudhary, "Overview of virtualization in cloud computing", 2016.
- [13] Ayush Agarwal, Siddhant Siddharth and Pratosh Bansal, "Evolution of cloud computing and related security concerns", 2016.
- [14] Jayachander Surbiryala, Chunlei Li and Chunming Rong, "A framework for improving security in cloud computing", 2017.
- [15] Kaiping Xue and Peilin Hong, "A Dynamic Secure Group Sharing Framework in Public Cloud Computing", 2014.
- [16] Justin Riley, John Noss and Wes Dillingham, "A High-Availability Cloud for Research Computing", 2017.
- [17] Song Li, Yangfan Zhou and Lei Jiao, "Towards Operational Cost Minimization in Hybrid Clouds for Dynamic Resource Provisioning with Delay-Aware Optimization", 2015.
- [18] Zhang Kun, Li Qing-Zhong and Shi Yu-Liang, "Research on Data Combination Privacy Preservation Mechanism for SaaS[J]", 2010.
- [19] R. Dua, A. Raja and D. Kakadia, "Virtualization vs containerization to support PaaS", 2014.
- [20] H. Liu and B. He, "Reciprocal resource fairness: Towards cooperative multiple-resource fair sharing in IaaS clouds", 2014.
- [21] W3.org, "Web Services @ W3C", 2011. [Online]. Available at: <http://www.w3.org/2002/ws/>. [Accessed: 19 Jan 2016].
- [22] K. Fysarakis, D. Mylonakis, C. Maniavas and I. Papaefstathiou, "Node-DPWS: Efficient Web Services for the Internet of Things", 2016.
- [23] Arne Wall, Vlado Altmann and Johannes Müller, "Decentralized configuration of embedded web services for smart home applications", 2017.
- [24] E. MacLennan and J. P. V. Belle, "Factors affecting the organizational adoption of service-oriented architecture (SOA)", 2014.
- [25] R. Th. Fielding, "Architectural Styles and the Design of Networkbased Software Architectures", 2000. [Online] Available: <http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm> [Accessed: January.15, 2016].
- [26] Urjita Thakar, Amit Tiwari and Sudarshan Varma "On Composition of SOAP Based and RESTful Services", 2016.
- [27] Agon Memeti, Besnik Selimi, Adrian Besimi and Betim "A Framework for Flexible REST Services: Decoupling Authorization for Reduced Service Dependency", 2015.
- [28] W. Yanfeng, B. Xinlong and H. Zhengbing, "The structure of EMIF and its model for data exchange and message processing", 2009.
- [29] Z. Xiao-guang, L. Jing-song, Z. Tian-shu, Y. Yi-bing, C. Yun-qi, X. Wang-guo and Z. Jun-ping, "Design and Implementation of Interoperable Medical Information System Based on SOA", 2009.
- [30] Dongdai Zhou, Ling Qin, Pan Xie, Zhuo Zhang and Hongyan Tao "SOA-based Education Information System Interoperability Model", 2011.
- [31] School Interoperability Framework, Available at: <https://www.a4l.org/Pages/default.aspx>. (Accessed 02 March 2017), 2014.
- [32] Aftab Ahmed Chandio, Dingju Zhu, Ali Hassan Sodhro, and Muhammad Umer Syed, "An implementation of web services for interconnectivity of information systems", 2014.

# Iterative Removing Salt and Pepper Noise based on Neighbourhood Information

Liu Chun

College of Computer Science and Information Technology  
Daqing Normal University  
Daqing, China

Sun Bishen

Twenty-seventh Research Institute  
China Electronic Technology Group Corporation  
Zhengzhou, China

Liu Shaohui\*

School of Computer Science and Technology  
Harbin Institute of Technology  
Harbin, China

Tan Kun, Ma Yingrui

College of Computer Science and Information Technology  
Daqing Normal University  
Daqing, China

**Abstract**—Denoising images is a classical problem in low-level computer vision. In this paper, we propose an algorithm which can remove iteratively salt and pepper noise based on neighbourhood while preserving details. First, we compute the probability of different window without free noise pixel by noise ratio, and then determine the size of window. After that the corrupted pixel is replaced by the weighted eight neighbourhood pixels. If the neighbourhood information does not satisfy the denoising condition, the corrupted pixels will recover in the subsequent iterations.

**Keywords**—Salt and pepper noise; noise detection; neighbourhood similarity; detail preserving denoising

## I. INTRODUCTION

Salt and pepper noise (SPN) usually comes from the image sensor, transmission channel and decoding processing. The corrupted pixels take either maximum or minimum grey value, contributing to black and white dots on image. Most of tasks about computer vision, for example, image segmentation, feature extraction, image recognition, are strongly influenced by impulse noise. Therefore, it is necessary for removing the salt and pepper noise on image.

In the field of SPN reduction, most of the proposed methods are based on two phases method, namely, noise detection and filter. Currently, there are many noise detection and judgment methods, for example, Laplacian convolution [7], maximum gradient difference in window [5], minimum gradient difference in window [3]. For the filter of the second phase removing SPN, many approaches are proposed in recently years. For example, the median filter (MF), the switch median filter [1], the adaptive median filter (AMF) [2], the noise adaptive fuzzy switching median filter (NAFSMF) [3], [6], [8], the adaptive weighted mean filter (AWMF) [4], the using long-range correlation filter (LRC) [9]. Most of them are based on the improvement of traditional median filter or mean value in window. Thus, they will cause inevitably fuzzy edge

in image restoration. Moreover, the size of filtering window has remarkable impact on the effect of SPN image filtering. There are many ways to choose size of filtering window, for example,  $3 \times 3$ ,  $5 \times 5$ ,  $7 \times 7$ , or even if the current filtering window does not have noise-free pixel, the filtering window will be expanded until  $9 \times 9$ . The way of keeping the same size of filtering window is clearly unreasonable. It is obvious that there are no enough clean pixels to restore the corrupted pixels in very high noise ratio window if choosing the window  $3 \times 3$ . However, if choosing a big filtering window ( $9 \times 9$ ) or much larger one, that means amount of calculation in the low noise image recovering. Thus, how to choosing a most appropriate size of the window is very important for removing SPN in images.

In this paper, we proposed a new algorithm for SPN called the iterative restoration based on neighbourhood information. As we know, if the noise pixel has enough neighbourhood information, we can determine the corrupted pixel which belongs to edge area or flat area approximately, then based on this information, a perfect recovery can be achieved. When eight-neighbourhood information of the destroyed pixel are not complete or even are not available, the recovery needs through good blocks from global of image to assist the process. It will be an iteratively process until the eight-neighbourhood can remove noise pixels. The simple example is shown in Fig. 1. It is worth noting that the aforementioned algorithms may be bad in high or low SPN. But our present pattern has better recovery both in case of high or low SPN. The superiority of our algorithm is very obvious.

The rest of this paper is organized as follows. The Section II will give the details of proposed algorithm. The detection of noise level is first analyzed, and then the removing noise algorithm is proposed based on the neighbourhood pixels. And experiment results are conducted in Section III. The Section IV gives the conclusion.

This work is partially supported by the Science Research Foundation of Daqing Normal University under Grant No.14ZR02.



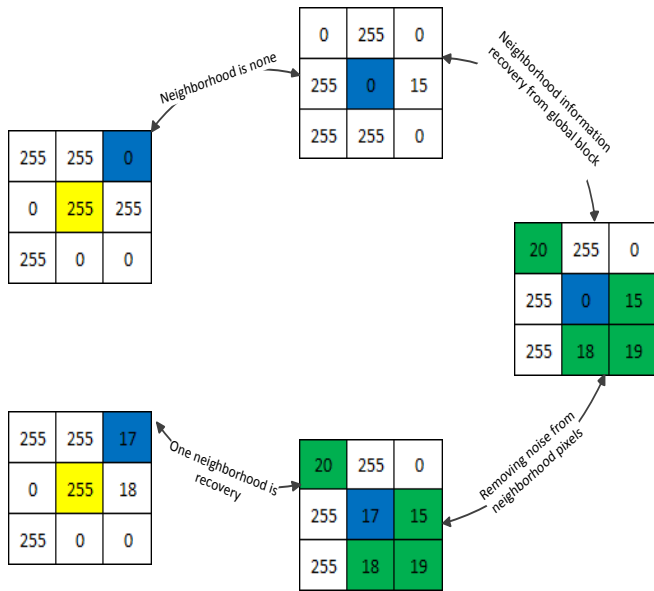


Fig. 1. The noise pixel and neighbourhood information recovery process.

## II. PROPOSED ALGORITHM

Following, we will illustrate the algorithm from the choosing window, noise detection, iteratively removing noise with neighbourhood pixels.

### A. Choosing a Window

Assuming that SPN is random distribution, and the ratio of noise is  $r$ ,  $r = \frac{\text{\#noise pixel}}{\text{\#total pixels}}$ ,  $r$  can be acquired from priori or according to the noise density of image. Select an  $m \times m$  window, the probability of pixel in window is  $P(m) = r^{m \times m}$ ,  $0 \leq r < 1$ . It means the probability of all pixels are noise in window  $m \times m$ . For example, if  $m = 3$ , the probability of no undamaged in window  $3 \times 3$  is  $r^{3 \times 3}$ . Then we can estimate the relationship among  $P(m)$ ,  $r$  and  $m$ . The Table I shows the result.

TABLE I. RELATIONSHIP AMONG  $P(m)$ ,  $r$  and  $m$

$m$ $r(\%)$	3	5	7	9	11
30	$2.0 \times 10^{-5}$	$8.5 \times 10^{-14}$	$2.4 \times 10^{-26}$	$4.4 \times 10^{-43}$	$5.4 \times 10^{-64}$
40	$2.6 \times 10^{-4}$	$1.1 \times 10^{-10}$	$3.1 \times 10^{-20}$	$5.8 \times 10^{-33}$	$7.1 \times 10^{-49}$
50	0.0020	$3.0 \times 10^{-8}$	$1.8 \times 10^{-15}$	$4.1 \times 10^{-25}$	$3.8 \times 10^{-37}$
60	0.010	$2.8 \times 10^{-6}$	$1.3 \times 10^{-11}$	$1.1 \times 10^{-18}$	$1.4 \times 10^{-27}$
70	0.040	$1.3 \times 10^{-4}$	$2.6 \times 10^{-8}$	$2.8 \times 10^{-13}$	$1.8 \times 10^{-19}$
80	0.13	0.0038	$1.8 \times 10^{-5}$	$1.4 \times 10^{-8}$	$1.9 \times 10^{-12}$

From the Table II, if selecting  $3 \times 3$  window, the window contains free noise pixels at least 99.9% when  $r$  below 0.3. If selecting  $9 \times 9$  window, the window contains free noise pixels at least 99.9% when  $r$  below 0.7. The formula of chosen window defined as

$$m = \begin{cases} 3 & \text{if } r < 0.3 \\ 5 & \text{if } 0.3 \leq r < 0.5 \\ 7 & \text{if } 0.5 \leq r < 0.6 \\ 9 & \text{if } 0.6 \leq r < 0.7 \\ 11 & \text{if } r \geq 0.7 \end{cases} \quad (1)$$

If it is difficult to get value of  $r$ , the default value of  $m$  set 7.

### B. Noise Detection

$X(i, j)$  denotes the pixel value of coordinate  $(i, j)$  on image.  $N(i, j)$  records  $(i, j)$  whether noise-free pixel or not (0 means undamaged pixel, 1 means damaged pixel). For the SPN, how to further determine whether it is real noise becomes intractability. It has two cases shown in Fig. 2.

In the case of (1), there is no undamaged pixel to remove SPN. But in the previous step, we know the probability of containing uncorrupted pixels at least 99.9%. Therefore, it can be sure that the pixels in window are white block or black block. Updating the pixel at location  $(i, j)$  using

$$X(i, j) = \operatorname{argmax}_{k \in \Omega_{i,j}^m} (c(k)) \quad (2)$$

where  $c(k)$  is the number of pixel value for  $k$ ,  $\Omega_{i,j}^m$  is the pixels in  $m \times m$  window which centre of location  $(i, j)$ .

In the case of (2), we use  $G_\Omega$  indicating the noise free pixels in window, defined as

$$G_\Omega = \{X(k, l) | X(k, l) \neq 0 \text{ and } 255, |k - i| \leq w, |l - j| \leq w\} \quad (3)$$

$w = (m - 1)/2$ , calculating the mean of not maximum and minimum pixel values and standard deviation (SD) in window.

$$SD = \sqrt{\frac{\sum_{X(k,l) \in G_\Omega} (X(k,l) - \text{mean})^2}{c_{G_\Omega}}} \quad (4)$$

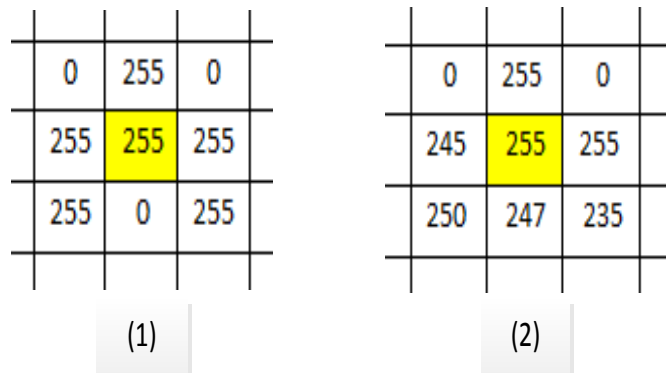


Fig. 2. Two cases of extreme value in window.

where  $c_{G_\Omega}$  is the number of clean pixels in window. This step only adapt for flat area in image. So the  $SD$  must less than a threshold  $\delta$ .  $\delta$  which sets to smaller is reasonable, or according to algorithm of OTSU may be more accuracy. In this paper, we simply set  $\delta = 10$ . The detailed algorithm is as follows:

### Noise Detection Algorithm

- 1) **for** each  $\text{pix}(i, j)$  in noise image X
- 2) **if**  $X(i, j) == 0$  or  $X(i, j) == 255$
- 3) **if**  $\forall X(r, t) \in \Omega_{i,j}^m, X(r, t) == 0$  or  $X(r, t) == 255$
- 4) **then**  $X(i, j) = \text{argmax}_{k \in \Omega_{i,j}^m} (c(k)), N(i, j) = 0$
- 5) **else if**  $SD < \delta$  &  $|X(i, j) - \text{mean}| < SD$
- 6) **then**  $N(i, j) = 0$
- 7) **else**
- 8)  $N(i, j) = 1$
- 9) **endif**
- 10) **else**
- 11)  $N(i, j) = 0$
- 12) **endif**
- 13) **end for**

### C. Noise Removing

Assuming that  $X(i, j)$  is the noise pixel at location  $(i, j)$ . Its eight-neighbourhood is defined as

$$\Omega_{EN} = \{(k, l) | (k, l) \neq (i, j), |k - i| \leq 1, |l - j| \leq 1\} \quad (5)$$

Setting up eight  $m \times m$  windows centered as  $\Omega_{EN}$ . They are respectively compared the  $m \times m$  window centered as  $(i, j)$ , using the similarity between windows and weighted filtering. The similarity comparison is given as

$$KL(\alpha || \beta) = \alpha \log \frac{\alpha}{\beta} + (1 - \alpha) \log \frac{1 - \alpha}{1 - \beta} \quad (6)$$

In this paper,  $\alpha$  is the pixels of window centered at  $(i, j)$ ,  $\beta$  is the pixels of window centered as  $\Omega_{EN}$ . For avoiding overflow and dividing by zero, the image pixels scaling to  $(0, 1)$  before using formula (6), given as

$$X(i, j) = \frac{X(i, j) + \mu}{256} \quad (7)$$

$\mu$  sets a smaller value 0.001, avoiding  $X(i, j)$  equal to 0 or 1.

Taking partial derivatives of formula (6)

$$\frac{\partial KL}{\partial \alpha} = \log \left( \frac{\alpha}{\beta} \times \frac{1 - \beta}{1 - \alpha} \right) \quad \frac{\partial KL}{\partial \beta} = \frac{1 - \alpha}{1 - \beta} - \frac{\alpha}{\beta} \quad (8)$$

By letting formula (8) equals to zero. Then it can be obtained easily that the minimum value of formula (6) is 0 at  $\alpha = \beta$ . So the more similar between  $\alpha$  and  $\beta$ , the smaller value of  $KL(\alpha || \beta)$ . If  $\beta = 0.3$ , the graph shows as Fig. 3.

The similarity between window centered at  $(i, j)$  and window centered at  $(k, l)$  defined as

$$s(k, l) = \sum_{p=-w}^w \sum_{q=-w}^w (1 - N(i + p, j + q)) (1 - N(k + p, l + q)) \times KL(X(i + p, j + q) || X(k + p, l + q)) \quad (9)$$

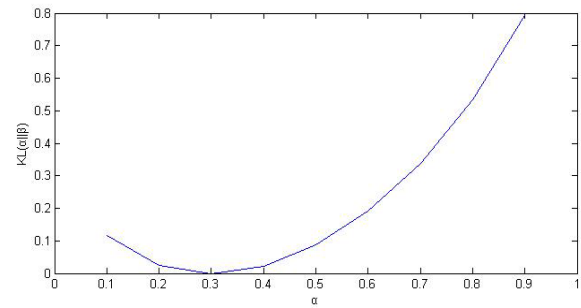


Fig. 3. The relationship between  $\alpha$  and  $\beta$ .

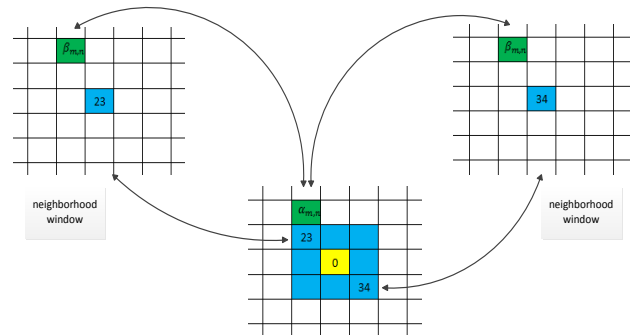


Fig. 4. The similarity comparison and computing.

From this equation, we can observe that the smaller  $s(k, l)$ , the stronger the correlation and the filtering will apt to use larger weight value. As shown in Fig. 4, illustrates the calculation of similarity between  $\Omega_{i,j}^m$  and  $\Omega_{k,l}^m$  with same location, where  $(k, l) \in \Omega_{EN}$ . It is noted that the calculation of similarity only uses clean pixels in window.

Let  $Y(i, j)$  denote filtering pixel value for the image  $X(i, j)$ , it can be defined as

$$Y(i, j) = (1 - N(i, j)) \times X(i, j) + N(i, j) \times Z(i, j) \quad (10)$$

where

$$Z(i, j) = \frac{\sum_{k,l \in \Omega_{EN} s(k,l)} \frac{1}{s(k,l)} \times X(k,l)}{\sum_{k,l \in \Omega_{EN} s(k,l)} \frac{1}{s(k,l)}} \quad (11)$$

In the formula (9), we could not consider the situation of  $s(k, l) = 0$ , to illustrate this problem, given formula as

$$\text{comp}_{k,l} = \sum_{p=-w}^w \sum_{q=-w}^w (1 - N(i + p, j + q)) (1 - N(k + p, l + q)) \quad (12)$$

It means the total number of undamaged pixels in same location between window centered at  $(i, j)$  and window centered at  $(k, l)$ . If  $\text{comp}_{k,l} = 0$ , we do not compute  $s(k, l)$ . If  $\text{comp}_{k,l} \neq 0$ , but  $s(k, l) = 0$ ,  $s(k, l)$  is set to the value of  $\sigma$ ,  $\sigma$  is a very small value of 0.0001. If  $\forall (k, l), (k, l) \in \Omega_{EN}$ ,  $\text{comp}_{k,l} \equiv 0$ , i.e., the neighbourhood information is not enough for removing noise. The corrupted pixels at location  $(i, j)$  will be recovered in the subsequent iterations.

The convergence condition is that any  $Y(i, j)$  has not been updated in one iteration or any  $Y(i, j)$  does not equal the initial value. If some  $Y(i, j)$  equals initial value end of iteration, then the iteration is terminated. And  $Y(i, j)$  is set to the mean of clean pixels ( $G_\Omega$ ) in  $m \times m$  window centered at  $(i, j)$ .

### Iterative Recovery Algorithm

```

1) initialize iter = true, flag = true  $Y(i, j) = 1$ 
2) while iter and flag
3)   iter = false
4)   flag = false
5)   for each pix( $i, j$ ) in noise image X
6)     if  $N(i, j) == 1$ 
7)       then for ( $k, l$ ) in  $\Omega_{EN}$ 
8)         compute  $comp_{k,l}, s(k, l)$ 
9)         if  $\forall (k, l), (k, l) \in \Omega_{EN}, comp_{k,l} \equiv 0$ 
10)          then iter=true
11)        else
12)           $Y(i, j) = Z(i, j) N(i, j) = 2$ 
13)          flag = true
14)        endif
15)      else
16)         $Y(i, j) = X(i, j)$ 
17)        flag = true
18)      end for
19)    update  $N(i, j)$ 
20)    for each ( $i, j$ ) in N
21)      if  $N(i, j) == 2$ 
22)        then  $N(i, j) = 0, X(i, j) = Y(i, j)$ 
23)      endif
24)    end for
25)  end while
26) if flag == false
27)   for each pix( $i, j$ ) in noise image Y
28)     if  $Y(i, j) == 1$ 
29)        $Y(i, j) = mean(G_\Omega)$ 
30)     endif
31)   end for
32) endif

```

### III. EXPERIMENTAL RESULTS

In our experiment tests, five different methods include median filtering (MF), NAFSMF [3], LRC [9], AMF [2], AWMF [4] are used to evaluate the performance of our proposed algorithm. The extensive experiments are conducted to verify the performance. In this paper, only the four tested images “Lena”, “boat”, “bridge” and “Elaine” are reported. All the image size is  $512 \times 512$ . We use peak signal to noise ratio (PSNR) to evaluate the performance of different methods. The PSNR is defined as

$$PSNR = 10 \log_{10} \frac{255^2}{MSE} \quad (13)$$

$$MSE = \frac{1}{M \times N} \sum_{i=0}^{i=M} \sum_{j=0}^{j=N} (Y(i, j) - M(i, j))^2 \quad (14)$$

where  $M(i, j)$  is the pixel value of original image. The Table II gives the results compared with other five different

methods. From this table, proposed method achieves the best results in most of cases. Actually, there are only two cases which the proposed algorithm ranks the second position for 4 images with 9 noise levels. And the AWMF algorithm achieves the best results for the Bridge image with 80% and 90% noise levels. The black font in the table means the best results. In most of case, the proposed algorithm achieves the PSNR over 20 db even the noise level is larger than 90%.

Assuming that SPN is random distribution, the ratio between salt and pepper is 1, namely, the number of salt pixels is almost equal to that of pepper pixels. In Fig. 5, the original image with 90% SPN, the whole process of iteration is shown. The each iterated middle-result image with removing noise pixel by fully using neighbourhood information is shown, the progressive quality is rather obvious from the (a) to (f). The experimental results demonstrate that the method possesses good vision effect.

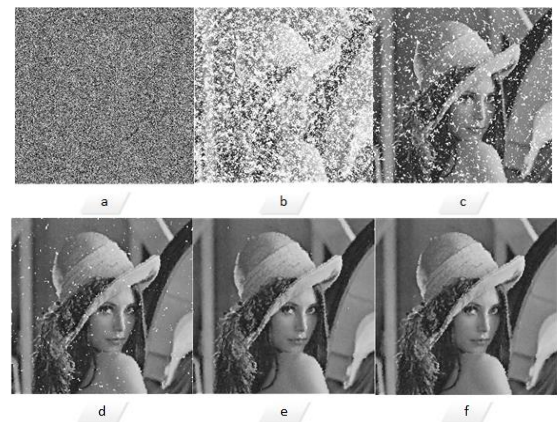


Fig. 5. The iterative restoration for “Lena” with 90% SPN.(a) Original image (b) First iteration (c) Second iteration (d) Third iteration (e) Fourth iteration (f) Convergent image.

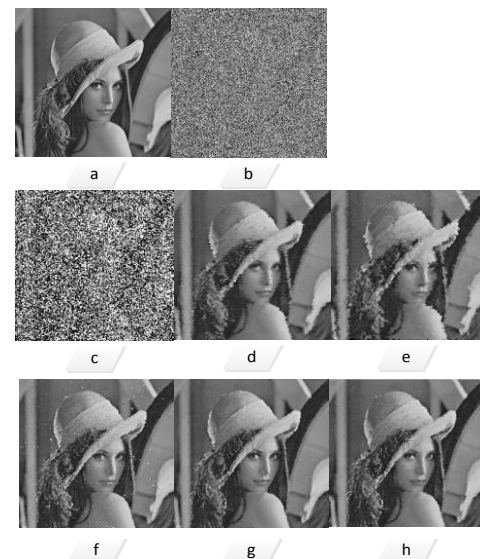


Fig. 6. The restoration for “Lena” with different methods (a) Original image (b) Image with 90% SPN (c) MF (d) LRC (e) AMF (f) NAFSMF (g) AWMF (h) Proposed.

TABLE II. RESULT OF PSNR (DB) FOR DIFFERENT ALGORITHM WITH VARIOUS NOISE LEVEL

Image	Algorithms	10%	20%	30%	40%	50%	60%	70%	80%	90%
Lena	MF	32.50	29.49	23.79	19.14	15.32	12.41	10.02	8.15	6.60
	LRC	40.96	36.69	32.99	29.38	26.17	23.81	22.21	21.34	21.18
	NAFSM	41.17	37.87	35.27	33.36	31.72	30.03	28.22	26.18	22.08
	AMF	36.58	34.75	32.69	31.12	29.35	27.91	26.35	24.35	21.73
	AWMF	38.16	36.22	34.99	33.80	32.42	31.03	29.59	27.84	25.37
	<b>Proposed</b>	<b>43.49</b>	<b>39.73</b>	<b>37.94</b>	<b>35.87</b>	<b>34.16</b>	<b>32.34</b>	<b>30.36</b>	<b>28.30</b>	<b>25.58</b>
Boat	MF	29.55	27.14	22.96	18.88	15.17	12.34	10.04	8.14	6.64
	LRC	38.59	34.29	30.08	26.64	23.61	21.42	19.97	19.24	19.00
	NAFSM	38.14	34.62	32.33	30.31	28.73	27.30	25.72	24.02	20.93
	AMF	33.36	31.59	29.69	28.30	26.71	25.33	23.94	22.24	19.95
	AWMF	35.01	33.34	32.05	30.69	29.58	28.17	26.89	25.24	23.01
	<b>Proposed</b>	<b>40.98</b>	<b>36.98</b>	<b>35.03</b>	<b>32.95</b>	<b>31.17</b>	<b>29.33</b>	<b>27.53</b>	<b>25.55</b>	<b>23.33</b>
Bridge	MF	26.05	24.53	21.51	18.03	14.68	11.95	9.68	7.88	6.38
	LRC	33.24	29.34	26.33	23.43	21.09	19.34	18.16	17.48	17.17
	NAFSM	34.57	31.29	28.95	27.25	25.74	24.22	22.86	21.43	18.89
	AMF	30.03	28.74	27.20	25.76	24.43	23.15	21.85	20.36	18.35
	AWMF	32.58	30.82	29.37	28.10	26.88	25.67	24.39	22.91	21.06
	<b>Proposed</b>	<b>35.68</b>	<b>32.17</b>	<b>30.48</b>	<b>28.78</b>	<b>27.30</b>	<b>25.87</b>	<b>24.40</b>	<b>22.81</b>	<b>20.78</b>
Elaine	MF	31.58	28.93	23.75	19.19	15.36	12.45	10.06	8.21	6.68
	LRC	38.64	35.19	32.51	22.19	26.91	24.66	23.00	22.33	22.13
	NAFSM	40.52	37.20	35.16	33.51	31.93	30.57	28.81	26.99	23.02
	AMF	35.06	34.13	32.90	31.56	30.35	28.90	27.46	25.76	23.08
	AWMF	38.80	36.94	35.44	34.06	32.73	31.50	30.18	28.82	26.82
	<b>Proposed</b>	<b>41.20</b>	<b>37.73</b>	<b>36.09</b>	<b>34.43</b>	<b>33.25</b>	<b>31.87</b>	<b>30.38</b>	<b>28.83</b>	<b>26.84</b>

In Fig. 6, we show the PSNR of different methods for “Lena” with 90 percent SPN. Obviously, MF, LRC, AMF and NAFSMF are not efficient for high noise ratio in image. But AWMF is not ideal for low noise ratio in image. The details of filtering results are in Table II.

In Fig. 7, we show the average time of different methods for image filtering. MF, NAFSMF, AWMF keep a low filtering time no matter noise ratio. Although more time than other when noise is high in our algorithm, the PSNR is highest from our filtering. From the trend of curve, our approach is reasonable. Because it is an iterative process, the process time will be longer with high SPN.

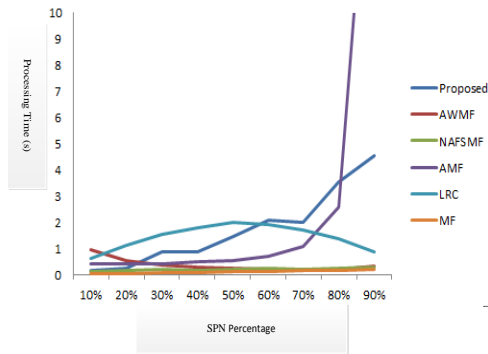


Fig. 7. The average time of different methods for tested images.

#### IV. CONCLUSIONS

In this paper, a highly efficient denoising method is presented no matter high or low SPN. It is an iterative process by effectively using neighbourhood information and global information. Our present pattern has a better recovery both in case of high or low SPN by experiment. Importantly, our ideal

is novel in the field of SPN and obtained good result. In addition, the method can also be used in 3D models by simple extension. The future work includes how to apply this algorithm into the practical application, especially the depth image denoising and 3D points denoising. In addition, how to accelerate the speed is also another research point.

#### REFERENCE

- [1] Z. Wang and D. Zhang, “Progressive switching median filter for the removal of impulse noise from highly corrupted images,” *IEEE Trans. Circuits Syst. II: Analog Digital Signal Process.*, vol. 46, no. 1, pp. 78–80, 1999.
- [2] H. Hwang and R. A. Haddad, “Adaptive median filters: New algorithms and results,” *IEEE Trans. Image Process.*, vol. 4, no. 4, pp. 499–502, 1995.
- [3] K. K. V. Toh and N. A.M. Isa, “Noise adaptive fuzzy switching median filter for salt-and-pepper noise reduction,” *IEEE Signal Process. Lett.*, vol. 17, no. 3, pp. 281–284, 2010.
- [4] P. Zhang and F. Li, “A New Adaptive Weighted Mean Filter for Removing Salt-and-Pepper Noise,” *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1280–1283, 2014.
- [5] C. Lu, T. Chou, “Denoising of salt-and-pepper noise corrupted image using modified directional-weighted-median filter,” *Pattern Recognition Letters*, vol.33, pp. 1287–1295, 2012.
- [6] H. Xu, G. Zhu, H. Peng, and D. Wang, “Adaptive fuzzy switching filter for images corrupted by impulse noise,” *Pattern Recognition Letters*, vol.25, pp. 1657-1663, 2004.
- [7] S. Tai, S. Yang, “A Fast Method For Image Noise Estimation Using Laplacian Operator and Adaptive Edge Detection,” *International Symposium on Communications, Control and Signal Processing*, Malta, pp. 1077-1081, 2008.
- [8] K. K. V. Toh, H. Ibrahim, and M. N. Mahyuddin, “Salt-and-pepper noise detection and reduction using fuzzy switching median filter,” *IEEE Trans. Consumer Electron.*, vol. 54, no. 4, pp. 1956–1961, Nov.2008.
- [9] Z. Wang and D. Zhang, “Restoration of Impulse Noise Corrupted Images Using Long-Range Correlation” *IEEE Signal Process. Lett.*, vol. 5, no. 1, pp. 4–7, 1998.

# Attendance and Information System using RFID and Web-Based Application for Academic Sector

Hasanein D. Rjeib  
Faculty of Engineering  
University of Kufa  
Al-Najaf, Iraq

Nabeel Salih Ali, Ali Al Farawn, Basheer Al-Sadawi,  
Haider Alsharqi  
IT-RDC Center  
University of Kufa  
Al-Najaf, Iraq

**Abstract**—Recently, students attendance have been considered as one of the crucial elements or issues that reflects the academic achievements and the performance contributed to any university compared to the traditional methods that impose time-consuming and inefficiency. Diverse automatic identification technologies have been more in vogue such as Radio Frequency Identification (RFID). An extensive research and several applications are produced to take maximum advantage of this technology and bring about some concerns. RFID is a wireless technology which uses to a purpose of identifying and tracking an object via radio waves to transfer data from an electronic tag, called RFID tag or label to send data to RFID reader. The current study focuses on proposing an RFID based Attendance Management System (AMS) and also information service system for an academic domain by using RFID technology in addition to the programmable Logic Circuit (such as Arduino), and web-based application. The proposed system aims to manage student's attendance recording and provides the capabilities of tracking student absentee as well, supporting information services include students grading marks, daily timetable, lectures time and classroom numbers, and other student-related instructions provided by faculty department staff. Based on the results, the proposed attendance and information system is time-effective and it reduces the documentation efforts as well as, it does not have any power consumption. Besides, students attendance RFID based systems that have been proposed are also analyzed and criticized respect to systems functionalities and main findings. Future directions for further researchers are focused and identified.

**Keywords**—Student attendance; Attendance Management System (AMS); information service; RFID; IoT; radio-frequency identification; Arduino

## I. INTRODUCTION

Information Technology (IT) has played a significant role in developing several aspects in academic sectors and domains such as student monitoring and management systems [1], [2]. Therefore, it is a critical subject to tracking and manages student's attendance in school, college, and university environment. Since it can be helped to urge students to attend on time, amend the efficiency of the learning, increase learning grade, and finally boosting and improving the education level [3], [4]. Calling student's name or taking student's signature are two traditional methods for tracking the attendance of the students in the classroom and they were more time-consuming [5], [6]. Nevertheless, the academic

performance influenced by student's presentation. So, there is a need to manage the student attendance records automatically by using information technology management system in a faculty to assist the maintaining attendance [1], [7]. Hence, the attendance systems can be useful to reduce administrative complexity and cost rather than increase the efficiency of the education [8], [9]. In the digital era, technologies have been developed and emerged recently, and that could change the future of sciences to affect people everyday life such as Wireless Sensor Networks (WSNs) [10]. Biometrics techniques are used to verify identification through their characteristics like face recognition, signatures, fingerprint, voice recognition, irises, barcode, Bluetooth, Near-Field Communication (NFC), RFID and so on [11], [8]. Identification, tracking, and counting are different applications for these technologies based attendance systems. RFID is an automation technology used to identifying and positioning an object [12]. Healthcare industry, financial institutions, cars, books, mobile phones, computer equipment, are several applications that they used RFID technology to positioning and managing people, assets, and inventory [11].

Diverse studies have been conducted to propose students attendance system to manage, record, and track the presenting of the students in an academic sector. These systems used several technologies that are ranging from Quick Response (QR) code, Ethernet and Wi-Fi interfaces to RFID with Liquid Crystal Display (LCD), or General Packet Radio Service (GPRS). Related works proposed and developed student system attendance such as, in 2012, Patel et al. proposed student attendance system based on RFID technology to compact lightweight and inexpensive used to record students' attendance and displayed on the screen and integrated good system [13]. Likewise, Yuru et al., 2013 is presented an integrated student attendance system which based on RFID technology and the hardware node of the system, and the development processes of related application have been presented in details [1]. In addition to, student's attendance system with RFID designed by Kurniali et al., 2014 that collected web-based with RFID readings and the main findings of the proposed system was to reduce or eliminate the manual labour requirements. As well as, the system provided faster processes, less inventory, fewer efforts, and better quality via providing direct cost savings while it caused some technical issues and slow system deployment [7]. Furthermore, development of a student attendance

management system using RFID and face recognition proposed by Patel and Priya in 2014. The developed system log contained an RFID tag ID and captured the image by a camera [11]. On the other hands, QR Code technology proposed by Miran in 2014 to develop and check the student's attendance system at the University of Sulaimaniyah. The advantage of the conducted system is to determine students absentee rate regularly, but it required each student has a smartphone that is capable of image capturing which considered as slow method because the teacher read the names over the phone and then sent to the database [14]. While NFC technology with the embedded camera on a mobile device that proposed by Dae in 2014 to develop attendance system. The conducted system recorded students' attendance by using Bluetooth, but the limitation of the proposed system is that the phones must have Bluetooth technology within the operating system [15].

Several traditional methods for student's attendance management imposed time consumption, increasing workforce requirements, and duplication of the efforts respectively. On the other hands, these mechanisms were boosted and improved the education level and amended the efficiency of the learning in the academic sector such as college or university. In this article, a student attendance management and information service system is proposing. The system prevents to manage student's records and provides the capabilities of tracking student attendance, supporting an information service about student grading marks, daily timetable, lectures time and classroom numbers, and other student-related instructions that provided by faculty department staff by using RFID technology with web-based application hybrid scheme. Students attendance RFID based systems that have been proposed are also analyzed and criticized respect to systems functionalities and main findings to identify and focus on the critical and vital systems or technology that need further attempts by future researchers through which the advantages of high efficiency and effectiveness can be obtained. The system functionality include data management, tracking students, sending reports, monitoring records, maintenance records, and finally providing information services. The remainder of the article structures as follows. Section 2 discusses the methodology and the materials used in the proposed system architecture. Section 3 provides all steps of the implementation methods and measures, and Section 4 presents the results. Reviews and analyzes the attendance systems that have been conducted before based on several metrics in Section 5. The conclusions, remaining challenges, and future directions for this system are presented in Section 6.

## II. ATTENDANCE AND INFORMATION SYSTEM ARCHITECTURE

In this section, the system will be presented and described, also, the equipment used for developing and designing the electronic circuit that includes software and hardware requirements will be displayed in Fig. 1, as well, methods and implementation steps to conduct and monitor the student attendance and information system. The proposed circuit aims to investigate student's services that provide presence and

information services based on the internet of things applications and technologies by literature review that gives an overview of what has been done. To implement the presented system, hardware and software components are required to establish the implementation process that has been chosen based on three criteria and metrics such as, cost, availability, and easy programming. The RFID reader connected to Arduino Uno microcontroller device which is open circuit system by pins and Ethernet shield device that connected with Arduino board. The Arduino circuit sends the signal to a server-based via using Ethernet cable as well using Wamp server, PHP and MySQL for the server to archive the student information attendance records and present student records via a using web-based application like a computer at the front end of the attendance records and information management end to present students attendance records and to students registration via the staff in a faculty. Besides, the proposed system provides information service for students by displaying their information such as grading marks, daily timetable, lecture time, classroom number, and other related instructions via LCD screen. These information services demonstrated shown in the block diagram for the proposed system in Fig. 1.

## III. SYSTEM IMPLEMENTATION

### A. Phase One (Student Attendance and Information Management Phase)

This section gives a clear description of all processes of the system. In this stage, all steps and procedures for conducting the student attendance management part of the current system are described and presented in Fig. 2. The student scans (RFID Tag) into (RFID Reader) where (RFID Reader) reads the (ID) for the student in particularly via student ID (Reading Process) and then transfer information via Arduino board (Microcontroller Process) and Ethernet shield (Transmission Process) to send data to the Wamp server (MySQL and PHP) by wired (Server Process) to record, manage, and display student attendance records by a web-based application.

### B. Phase Two (Student Information Service Phase)

In this phase, the RFID reader reads the student's ID (Reading Process), Arduino UNO (Microcontroller process) is used to transfer student's information to the Wamp server through the cable via Ethernet shield card (Transmission Process). Server (MySQL and PHP) is used to identify student ID and to send student's information to the screen (see Fig. 3). The student scans the (RFID Tag) to the (RFID Reader) where (RFID Reader) reads (ID) for the student and then send it through Arduino board and to the server side (MySQL and PHP) where it searches for the ID of the particular student and fetches his data from database then the information can be presented on the screen or LCD (see Fig. 3). These data contain information regarding the student such as student name, stage, and group as well as daily timetable that encompasses classroom number, lecture time, subject name, and lecturer name. As well, the system presents all instructions and roles which are sent by the administrator to a particular student (see Fig. 4).

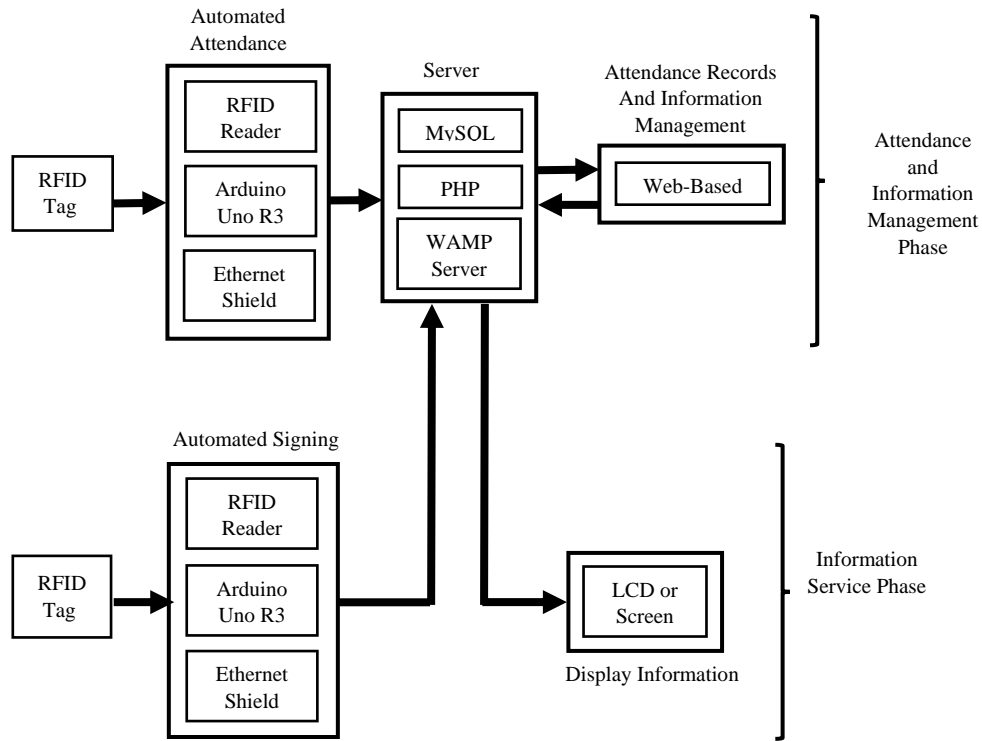


Fig. 1. Block diagram for the proposed system architecture.

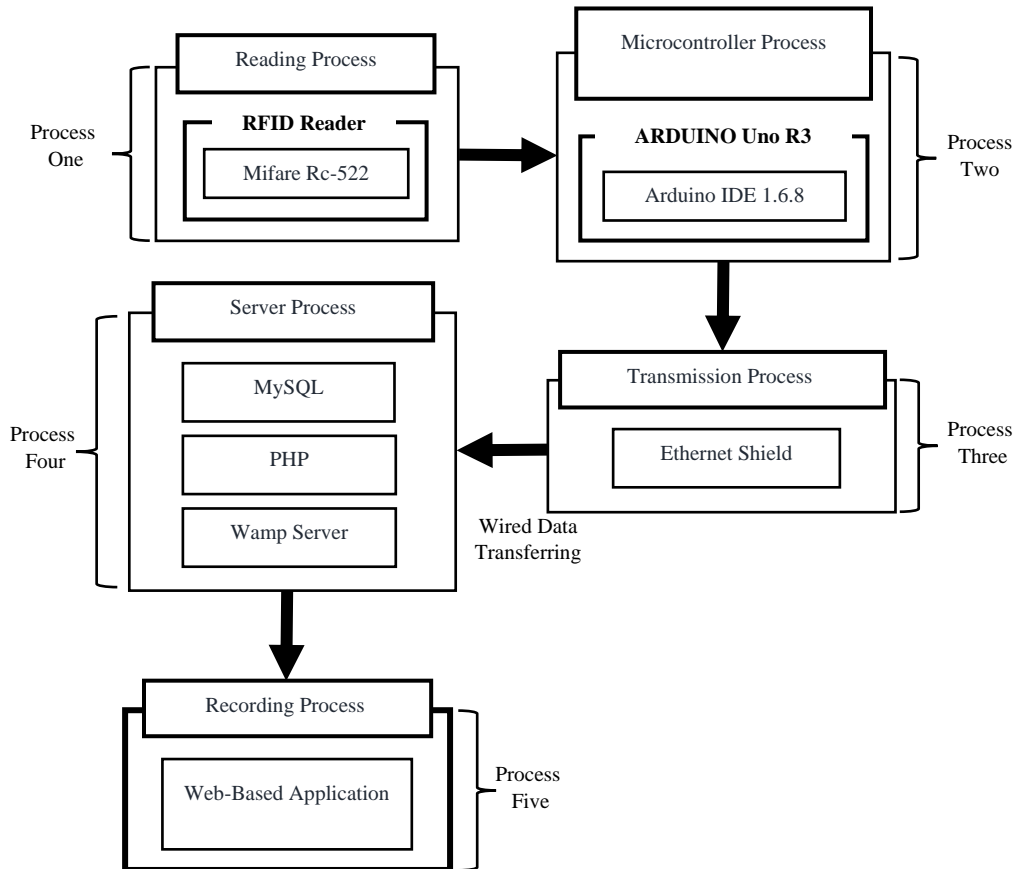


Fig. 2. Procedures steps for student attendance and information management phase.

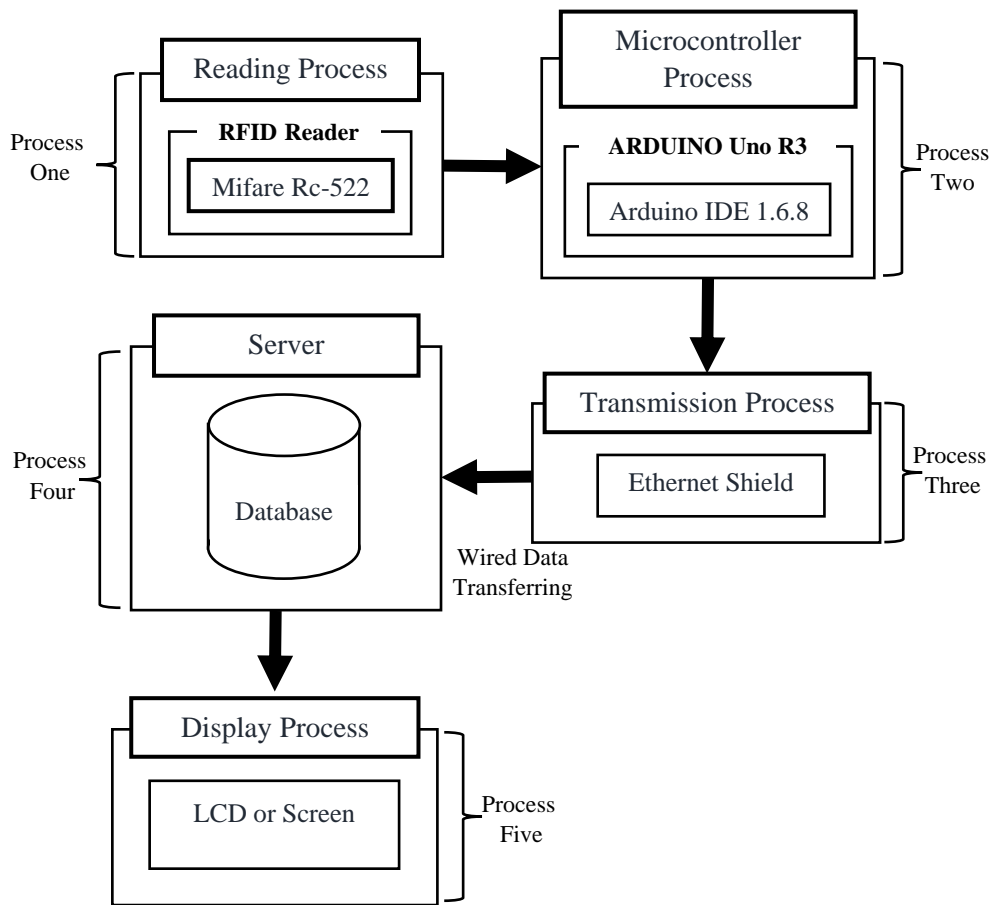


Fig. 3. Procedure steps for the student information service phase.

Student Information Service  
ECE Department, Faculty of Engineering, University of Kufa

**Mujtaba Basheer**

Student ID: 504689  
Student Stage: Fourth  
Student Group: 5

**Administrator Notes**

- ▶ Network Test Result: 7 over 10
- ▶ Next Monday: Architecture Test

**Student Notes**

- ▶ You should activate your Email
- ▶ Your order no. 124 is completed

**Student's Timetable | Forth Stage | 21 - 05 - 2017**

Subject Name	Teacher Name	Classroom No.	Time Start	Time End
Microwave	Dr. Maithem N.	5	8:30 AM	10:30 AM
Networking	Haider A. Hassan	2	11:00 AM	1:00 PM
Communication	Ahmad Nidhal	1	1:30 PM	3:30 PM

Fig. 4. Student information service displayed screen.



The parts of RFID reader is connected to the Arduino device's pins (the first pin to 3.3v, the second (RESET) is connected to D9, the third pin to the ground, the four (NC) is not used, the fifth pin (MISO) is connected to (D12), the sixth (MOSI) is connected to (D11), the seventh (SCK) connects to (D13) and the last pin (SAD) connects to (D10)). Ethernet device is setup with Arduino device. The signal that input to the Arduino is processed inside it. Then, the signal is sent to the server via Ethernet cable. MySQL is used to archive the student records and information in which it will be shown by using PHP and Arduino IDE with the aid of the graphical user interface for student information via web-based application as we demonstrated in Fig. 4.

IV. RESULTS AND DISCUSSION

The proposed system is achieving two aims, the first objective is to register, record, and manage a student attendance using RFID tag, and the second aim is to provide student information service such as timetable, lecture time and classroom number, and other student-related data that displayed in screen or LCD. The traditional method for taking student absence report is usually done by using paper-work and handwriting on the advertisement wall. Hence, paper-work method consumes workforce requirements, duplication of the efforts, and imposes time-consuming and inefficiency.

Table I presents and lists a comparison between traditional attendance system (paper-work) and the proposed system based on different parameters [8], [16]-[20].

On the other side, several types of automatic attendance systems such as a barcode, magnetic stripe, biometrics, and RFID attendance system are suited for different needs and requirements. To differentiate between the most standard types of automatic attendance systems, Table II discusses and describes the current generation of the common automated attendance registration systems with concerning different parameters [11], [17], [19], [21]-[25].

Traditional technology such as QR code, Barcode, and Magnetic stripe imposed a long time for registration and error-prone, low data accuracy and resources, artificial identification, traditional manual management and individual personnel statistics for attendance management records, and it is not eco-friendly due to paper attendance cards and documentation. While, the proposed system based on RFID technology can achieve several advantages such as user-friendliness, affordability, security, flexibility, high resources and data accuracy, automatic and tag identification without human interference, indicating work status and generating the attendance report automatically, and it does not need to spend extra time and efforts.

TABLE I. COMPARISON BETWEEN TRADITIONAL SYSTEM AND PROPOSED SYSTEM

Parameters	Human Interference	Time-Consuming	Efforts Spend	Speed	System Security	Resources (Documents)	Data Accuracy	Registration Time	User Friendly
<b>Traditional System</b>	Yes	More than 5 minutes	Yes	Slow (human)	More vulnerable	More paper work	Low	More than 8 minutes for each student	No
<b>Proposed System</b>	No	Less than 2 minutes	No	High (computer)	Authenticated persons only	Only one electronic record	High Accuracy	1-2 minute	Yes

TABLE II. THE CURRENT GENERATION OF THE COMMON AUTOMATIC ATTENDANCE REGISTRATION SYSTEMS

Parameters	Barcode	Magnetic Stripe	Biometric	RFID
Resources and data Accuracy	High	High	High	Very High
Data Density	Low	Low	High	Very High
Purchasing Cost	Low	Low	High	Low
Speed and Security	High	High	High	High
Influence Covering the Data carrier	Total failure of system	Total failure of system	complete failure as system works on contact Not	No control
Functionality	Wide	Wide	Wide	Wide
Operating Cost	Low	Low	High	Low
Power Consumption	High	High	Moderate	Moderate
Influence Direction of reader and Data Carrier	Failure - if no line-of-sight communication	Failure - if no slot- of communication	Not applicable as direct contact is needed	No influence as data are transferred via radio waves
Distance of Reader and Data Carrier ( in Centimeters)	0-50 cm Direct	Direct contact	Direct contact	0-6 m depending on the frequencies used

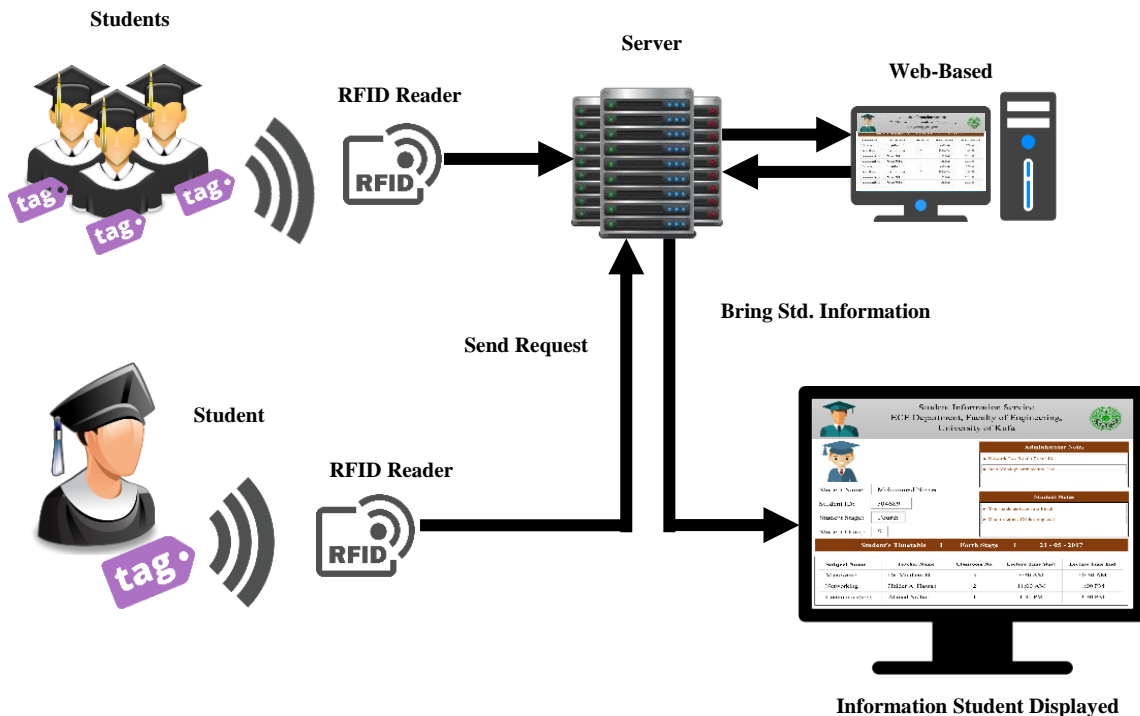


Fig. 5. Flowchart of the proposed attendance and information service system.

The proposed system provides facilities for both students and staff by reducing time to take absence, as well as, providing a database system that holds all the student's information (i.e. there is no need for archiving shelf and paper works). The system introduces facilities for registering new students, sending reports and warnings to them; displaying results for tests and homework, and other notifications such as staff appointment schedule, lecture cancelling, and so on. From the student perspective, the student will be informed of all the required information via the screen, as well as the absence report warning if any, also the system provide student attendance mechanism. From the staff or faculty management perspective, it would be much easier to track the student report by just clicking the required student name to add or edit specific information or to add warning and notification, and there is no need to check and review dozens of papers to collect information about the students. As well as, the system provides a weekly statistical report regarding student absences which is got by the computer-based application. Fig. 5 shows the two phase's flowchart of the proposed attendance and information system.

#### V. COMPARISON OF THE PREVIOUS RFID BASED ATTENDANCE SYSTEMS BASED ON CRITICAL REVIEW

This section presents a critical review of the works and attempts by previous authors that implemented Attendance Management System (AMS) in an academic sector in details and highlights their systems functionality, schemes and main findings. The feature of the proposed system includes several characteristics such as data management, tracking students, sending reports, monitoring records, maintenance records, and finally providing information services. The section

conceptually provides insights into the standard (previous and existing) works. These works, which have been conducted from 2013 until 2017, were aimed to determine and eliminate the lacks for the traditional methods and techniques regarding RFID based systems. Table III summarizes the functionality metrics of the previous systems in practice. We use two different types of markings to indicate system functionality. The symbol '✓' denotes that the system has one of the metrics. Conversely, the symbol '✗' signifies that a system does not have any metrics of functionality. We assess each system concerning its functionality based on the functionality of the proposed systems like data management, tracking students, sending reports, monitoring records, maintenance records, and finally providing information services. As we have shown in Table III, several results are found through comparing attendance management systems. We compared the functionality for each system based on the following criteria or metrics:

- 1) Does the system manage the student's attendance data or records?
- 2) Does the system have any technology to track the student's position and location?
- 3) Does the system have to send notifications or reports services?
- 4) Does the system monitor the performance of the students via monitoring the attendance student records?
- 5) Does the system maintain student attendance records and data when an error occurred?
- 6) Does the system provide any information service for students which displayed via the screen?

TABLE III. COMPARISON OF THE PREVIOUS STUDIES FOR RFID BASED ATTENDANCE SYSTEMS

Author (s)	System Functionality						Main Findings
	Data Management	Tracking Students	Sending Reports	Monitoring Records	Maintenance Records	Information Services	
Yuru et al., 2013[1]	✓	✗	✗	✗	✗	✗	Designed an attendance checking system of class based on embedded of ARM and RFID technology.
Yadav& Nainan, 2014 [26]	✓	✓	✓	✓	✗	✗	An automatic attendance management system presented for students and teachers as well using GSM to sending notifications to parents.
Arbain et al., 2014 [27]	✓	✗	✗	✓	✗	✓	Proposed an attendance system to record and manage student attendance automatically in the lab by using RFID-ARDUINO approach in web-based laboratories settings.
Tiwari et al., 2014 [28]	✓	✓	✗	✓	✓	✓	Conducted GPRS based student attendance system which it can be easily accessed by the lecturers via the web to check and monitor student attendance recording.
Kurniali, 2014 [7]	✓	✗	✗	✗	✗	✗	Developed a student attendance management system by using RFID with a web-based approach to managing student's attendance in an Indonesian higher education institution.
Pranali et al., 2015 [29]	✓	✓	✓	✓	✗	✗	Adopted and developed an attendance monitoring system based on Mifare technology and server-based to tracking and positioning students in campus settings.
Farpat et al., 2015 [30]	✓	✗	✓	✗	✓	✗	Managed student's attendance via proposing an automatic computing system in classrooms by using RFID technology.
Shengli et al., 2015 [16]	✓	✗	✗	✗	✗	✗	Conducted automatic attendance system based on RFID card connected to Arduino microcontroller via the real-time database environment to manage the employee's attendance in an enterprise.
Kuriakose & Vermaak, 2015 [21]	✓	✗	✗	✗	✗	✗	Proposed an automate attendance registration system by using Java-based RFID technology to monitor, management students attendance at the Central University of Technology, South Africa.
Praveen Kumar and Mani Kumar, 2015 [31]	✓	✗	✓	✗	✗	✗	Presented attendance management system based on RFID and Internet of Things (IoT) applications which can be accessed from anywhere and stored attendance records in the cloud and sending SMS to several smart-phones.
Srinidhi & Roy , 2015 [32]	✓	✓	✓	✓	✓	✓	Developed and adopted an automation attendance monitoring and management system based on web-based applications using RFID and biometrics technologies for an academic college and university environment with safe and secure system advantages.
Ya'acob et al., 2016 [33]	✓	✗	✗	✗	✓	✓	RFID Efficient student attendance management system presented with a web portal to reduce the time of taking the attendance compared with the traditional method.
Proposed System	✓	✓	✓	✗	✓	✓	The system provides attendance managing records to evaluating their performance as well support an information services for students and faculty staff.

The answers to these questions are summarized in Table III. Based on Table III, there are several results regarding previous systems functionality such as all of the proposed attendance systems have data management of their records. While, [26], [28], [29], [32], and the proposed system is providing tracking students in their school, college or university to improve their student performance. Diverse systems have generating information reports for students or sending notifications to the parents or lecturers and so on like [26], [29]-[32], and the proposed system. A few of the proposed attendance systems have monitored and maintained records as we see in studies [28] and [32]. However, [26], [27], and [29] have monitored their attendance records whereas [30], [33], and the proposed system have maintained their attendance records. Four of the previous AMS have information services functionality such as [27], [28], [32], [33], and the proposed system.

## VI. CONCLUSION AND FUTURE WORKS

A student attendance and information system are designed and implemented to manage student's data and provide capabilities for tracking student attendance, grading student marks, giving information about timetable, lecture time, room number, and other student-related information. Also, the proposed system provides easiness for the staff where there is no need for extra paper works and additional lockers for saving data. Results achieved the innovation of developing the system proved reliable to support the attendance management system for an academic sector in the usage of the RFID technology and microcontroller board. It can be considered as a successful implementation. We found two general trends in the results of the comparison study in Section 5. Two of the proposed AMS have most of the system functionality criteria which are Tiwari et al., 2014 [28] and the proposed system. While, the AMS which presented by Srinidhi and Roy, 2015 [33] have all of the system functions. Two primary goals for future directions, the first goal is to extend the proposed system to include staff information as well. The second one is to extend the system to encompass more than one faculty with the insertion of face detection mechanism in the attendance monitoring system to control card replacements among different students.

## REFERENCES

- [1] Yuru, Z., Delong, C., & Liping, T. (2013). The Research and Application of College Student Attendance System based on RFID Technology. *International Journal of Control and Automation*, 6(2), 273-282.
- [2] Sunehra, D., & Goud, V. S. (2016, October). Attendance recording and consolidation system using Arduino and Raspberry Pi. In *Signal Processing, Communication, Power and Embedded System (SCOPES)*, 2016 International Conference on (pp. 1240-1245). IEEE.
- [3] Sayanekar, P., Rajiwate, A., Qazi, L., & Kulkarni, A. (2016). Customized NFC enabled ID card for Attendance and Transaction using Face Recognition. *International Research Journal of Engineering and Technology*, 3(9), pp. 1366- 1368.
- [4] Noor, S. A. M., Zaini, N., Latip, M. F. A., & Hamzah, N. (2015, December). Android-based attendance management system. In *Systems, Process and Control (ICSPC)*, 2015 IEEE Conference on (pp. 118-122). IEEE.
- [5] Kohalli, S. C., Kulkarni, R., Salimath, M., Hegde, M., & Hongal, R. (2016). Smart Wireless Attendance System. *International Journal of Computer Sciences and Engineering*, 4(10), pp. 131-137.
- [6] Jacob, J., Jha, K., Kotak, P., & Puthran, S. (2015, October). Mobile attendance using Near Field Communication and One-Time Password. In *Green Computing and Internet of Things (ICGCIoT)*, 2015 International Conference on (pp. 1298-1303). IEEE.
- [7] Kurniali, S. "The Development of a Web-Based Attendance System with RFID for Higher Education Institution in Binus University." *EPI Web of Conferences*. Vol. 68. EDP Sciences, 2014.
- [8] Walia, H., & Jain, N. (2016). Fingerprint Based Attendance Systems-A Review. *International Research Journal of Engineering and Technology*, 3(5), pp. 1166- 1171.
- [9] Prince, N., Sengupta, A., & Unni, M. K (2016). Implementation of IoT Based Attendance System on a Dedicated Web-Server. *International Journal of Scientific & Engineering Research*. 7(6), pp. 351- 355.
- [10] Ali, N. S., & Alyasseri, Z. A. A. (2017). Wireless Sensor Network and Web Application Hybrid Scheme for Healthcare Monitoring. *Journal of Soft Computing and Decision Support Systems*, 4(5), 1-7.
- [11] Patel, U. A., & Swaminarayan Priya, R. (2014). Development of a student attendance management system using RFID and face recognition: a review. *International Journal of Advance Research in Computer Science and Management Studies*, 2(8), 109-19.
- [12] Kumar, Jay, and Amit Kumar. (2016). Automatic Attendance Monitoring and Tracking System Using Bluetooth and Face Identification. *International Journal of Advanced Research in Electronics and Communication Engineering*, 5(4), pp. 1166-1170.
- [13] Patel, R., Patel, N., & Gajjar, M. (2012). Online students' attendance monitoring system in classroom using radio frequency identification technology: a proposed system framework. *International Journal of Emerging Technology and Advanced Engineering*, 2(2), 61-66.
- [14] Baban, M. H. M. (2014). Attendance checking system using quick response code for students at the University of Sulaimaniyah. *Journal of Mathematics and Computer Science (JMCS)*.
- [15] Subpratatsavee, P., Promjun, T., Siriprom, W., & Sriboon, W. (2014, May). Notice of Violation of IEEE Publication Principles Attendance System Using NFC Technology and Embedded Camera Device on Mobile Phone. In *Information Science and Applications (ICISA)*, 2014 International Conference on (pp. 1-4). IEEE.
- [16] Shengli, K., Jun, Z., Guang, S., Chunhong, W., Wenpei, Z., & Tao, L. (2015). The Design and Implementation of the Attendance Management System based on Radio Frequency Identification Technology.
- [17] Arulmozhi, P., Rayappan, J. B. B., & Raj, P. (2016). The design and analysis of a hybrid attendance system leveraging a twofactor (2f) authentication (fingerprint-radio frequency identification). *Biomedical Research*.
- [18] Azasoo, J. Q., Engmann, F., & Hillah, K. A. (2014, October). Design of RF based multithreaded RFID student attendance management information system. In *Adaptive Science & Technology (ICAST)*, 2014 IEEE 6th International Conference on (pp. 1-5). IEEE.
- [19] Abas, M. A., Tuck, T. B., & Dahlui, M. (2014, October). Attendance Management System (AMS) with fast track analysis. In *Computer, Control, Informatics and Its Applications (IC3INA)*, 2014 International Conference on (pp. 35-40). IEEE.
- [20] Younis, M. I., Al-Tameemi, Z. F. A., Ismail, W., & Zamli, K. Z. (2013). Design and Implementation of a Scalable RFID-Based Attendance System with an Intelligent Scheduling Technique. *Wireless personal communications*, 1-19.
- [21] Kuriakose, R. B., & Vermaak, H. J. (2015, November). Developing a Java based RFID application to automate student attendance monitoring. In *Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2015 (pp. 48-53). IEEE.
- [22] Chiagozie, O. G., & Nwaji, O. G. (2012). Radio frequency identification (RFID) based attendance system with automatic door unit. *Academic Research International*, 2(2), 168.
- [23] Benyo, B., Sodor, B., Doktor, T., & Fördös, G. (2012, April). Student attendance monitoring at the university using NFC. In *Wireless Telecommunications Symposium (WTS)*, 2012 (pp. 1-5). IEEE.
- [24] Soewito, B., Gaol, F. L., Simanjuntak, E., & Gunawan, F. E. (2015, August). Attendance system on Android smartphone. In *Control, Electronics, Renewable Energy and Communications (ICCEREC)*, 2015

- International Conference on (pp. 208-211). IEEE.
- [25] Liang, X. Q., Li, W. Y., & Lu, T. C. (2017, August). A Study of the Multi-Organization Integrated Electronic Attendance System. In International Conference on Intelligent Information Hiding and Multimedia Signal Processing (pp. 53-61). Springer, Cham.
- [26] Yadav, R., & Nainan, S. (2014). Design of RFID based student attendance system with notification to parents using GSM. International Journal of Engineering, 3(2).
- [27] Arbain, N., Nordin, N. F., Isa, N. M., & Saaidin, S. (2014, December). LAS: Web-based laboratory attendance system by integrating RFID-ARDUINO technology. In Electrical, Electronics and System Engineering (ICEESE), 2014 International Conference on (pp. 89-94). IEEE.
- [28] Tiwari, A. S., Tiwari, A. S., Ade, N. M., Sheikh, S., Patel, N. R., & Khan, A. R. (2014, January). Optimized design of student attendance system using rfid. In International Conference on Machine Learning, Electrical and Mechanical Engineering (pp. 8-9).
- [29] Pranali, S., Mayuri, S., Manisha, T., & Varsha, J. (2015). An Intruder Tracking and Attendance Monitoring System Using Mifare Technology.
- [30] Farpat, S., VYAS, D., & Chavan, S. (2015). Monitoring Of Attendance Using RFID and GSM Technology, 3(2), pp. 1-8
- [31] M.Praveen Kumar and B.Mani Kumar. (2015). RFID based Attendance monitoring system Using IOT with TI CC3200 Launchpad. International Journal & Magazine of Engineering, Technology, Management and Research, 2(7), pp. 1465-1467.
- [32] Srinidhi, M. B., & Roy, R. (2015, January). A web enabled secured system for attendance monitoring and real time location tracking using Biometric and Radio Frequency Identification (RFID) technology. In Computer Communication and Informatics (ICCCI), 2015 International Conference on (pp. 1-5). IEEE.
- [33] Ya'acob, N., Adnan, S. F. S., Yusof, A. L., Azhar, A. E., Naim, N. F., Mustafa, N., & Mahmon, N. A. (2016). RFID lab management system using Arduino microcontroller approach associate with webpage.

# Social Network Link Prediction using Semantics Deep Learning

Maria Ijaz, Javed Ferzund, Muhammad Asif Suryani, Anam Sardar  
Department of Computer Science  
COMSATS Institute of Information Technology  
Sahiwal, Pakistan

**Abstract**—Currently, social networks have brought about an enormous number of users connecting to such systems over a couple of years, whereas the link mining is a key research track in this area. It has pulled the consideration of several analysts as a powerful system to be utilized as a part of social networks study to understand the relations between nodes in social circles. Numerous data sets of today's interest are most appropriately called as a collection of interrelated linked objects. The main challenge faced by analysts is to tackle the problem of structured data sets among the objects. For this purpose, we design a new comprehensive model that involves link mining techniques with semantics to perform link mining on structured data sets. The past work, to our knowledge, has investigated on these structured datasets using this technique. For this purpose, we extracted real-time data of posts using different tools from one of the famous SN platforms and check the society's behavior against it. We have verified our model utilizing diverse classifiers and the derived outcomes inspiring.

**Keywords**—Link prediction system; post analysis; semantic similarity; data analysis; social network analysis; dictionary; co-similar links

## I. INTRODUCTION

The social network is an online platform where peoples use to create informal communities or social relations with other peoples who share alike or different interests, views, genuine links, and experiences. It allows discussions and relations with other people online. Examples of SN platforms are Facebook, Instagram, Snapchats, etc. It is the modern technique to model the relations between the people in a group or community. Social network analysis (SNA), a rising branch started after sociology [1]. To predict the link between the networks is its main determination. For example, to perceive who take the most "important" role in a circle we can design the social network link between individuals and the link between two individuals shows connection like working on the same task. As soon as the network is built it can be used for information gathering of individuals the most active user, the common interest, followers, likes, etc.

Study this valuable information in the social networks has open the door for research where researcher used the different models of analysis such as sentiment analysis, link prediction, semantic analysis and many tools are existing to get the clear picture of analysis results.

In the beginning, the majority of the research in the social network has been done by social scientists and psychologists

and lately Computer Scientists contributed a lot. SNA is now use for different research purposes as the hidden conceptual model [2].

In order to show relationship in the network between the links nodes and edges are used where nodes represent persons and edge shows a relation between nodes. Edge data can be lost due to many reasons i.e. partial information gathering methods or ambiguity of links or source restrictions [3]. The variations in short period cause many problems and generate many challenging questions like:

- Two heads will be linked together for how much time?
- Does the link between two heads are formed by others?
- Peoples that are not linked, is it expected that they will get linked at some point later?

The examples that we address in this study is to predict the future relationship between two heads, realizing that there is no relationship between the people in the existing state. Hence, to predict such deviations with high correctness is important for the future of social networks.

Data can be extracted from Social Networks using different techniques which can typically emit only few information about nodes due to privacy. All information about nodes can't be gathered without permission.

In this paper, we propose a method to predict the relationship between people that may appear or fade with time, based on their behavior towards the Social Network. To explore such biased behavior of nodes we extracted special data from one of the famous social platform using various tools. Section II covers the related work, Section 3 covers methodology of and purposed framework followed by experimental setup and results.

## II. LITERATURE REVIEW

The SN has received extensive consideration in the analysis work. It has been adequately improved to study the changed applications such as malicious networks [4], online societies and professional groups between other networks. Numerous workshops were enfolded in the mid-1990s to unite the artificial intelligence (AI) and investigation of links groups. During the conference in 1998 on AI were reduced and Link Analysis started working for the first time with a direct focus on covering AI techniques to related data [5].

Basically, structural link analysis from profiles and groups approach considered the problems of foreseeing, categorizing marking friends' relations in SN by application feature constructing approach [6].

In [7], [8], the authors cover the advances in probabilistic models, manifold, and deep learning. This encourages longer-term unanswered inquiries regarding the proper objectives for learning of good representations, computing representations (i.e., inference), and similarly the geometrical influences between representation learning. It also tells about the learning of density estimation and manifold by using the links to predict classes or qualities of entities [9], [10].

Whereas other work has been done by using the location based on high-dimensional space. In [11] authors have identified features set that are the solution of superior performance under the supervised learning set up and explain the effectiveness of the features of their class density distribution.

The work defined in [12] presented the link prediction method which is created on comparison of the nodes. It is the significant utilization of nodes that consider the correspondence of nodes such as age, gender, etc.

In [13] practice the SN which gives the best approach to seek and get customized, reliable health advice from peers at wherever and at any time, by tracing dental health information searched for got on Twitter. In [14], the authors proposed a link prediction model that can forecast links that may exist or vanish later. The model has been effectively practiced in two distinct spaces (health care and a stock market).

Whereas in [15] the authors utilize the unique way for semantic analysis which is called Wikipedia Link Vector Model or WLVM that practices only the hyperlink composition of Wikipedia instead of full written material.

Nowadays there is a new trend of finding the online friends as well as offline friends as done in [16] which help users to off line contacts, known and find new groups online by utilizing the machine learning classifier. The classifier distinguishes missing associations even when practical checked on the tough problem of categorizing associations between individuals who have at least one common friend.

The approach used in [17] based on recommended that the combination of topological structures and node traits improve association forecast. For this purpose, they used Covariance Matrix Adaptation Evolution Strategy (CMA-ES) to optimize weights of nodes.

In [18], max-margin learning technique for nonparametric latent component social models is used. The author creates a bond among max-margin learning and Bayesian nonparametric to determine different latent structures for link prediction.

In [19], the author examined the innovative issue of negative connection with just positive connections. The authors proposed a principled system NeLP by observing the negative links, which can misuse positive connections and component driven interaction to foresee negative connections.

In [20], author revisit on weighting technique and proposed the two new weighting methods for link prediction min-flow and multiplicative. Also, find that these two methods give different prediction results on data sets.

In [21], author worked on Twitter and location-based SN to check the manifold link that may exist across the network between binary users and find that binary users connected on both platforms having more neighborhood then users which are connected only on one platform.

In [22], author proposed that Bi-directional links are more useful than uni-directional links for testing the real data set. For this purpose, the author proposed a new directing randomized technique to study the part of direction for predicting links in a network.

In [23], researcher has proposed a CF-based web service which provides predictions for various substances by using the ratings and opinions of people providing on Facebook and other social media sites.

In [24], author used the different algorithms such as support vector machine and decision tree and different topology structure to check the either the links exist between two nodes and also checked the load on the link and link type.

In [25], author calculated the similarity by calculating its different features between two homepages and computed the possibility when these pages are more related to each other.

The author Popescul and Ungar [26] proposed the statistical learning model of reference prediction where model learn the link prediction from queries of database which also involves joins, aggregation and selections.

### III. PROPOSED STUDY AND DESIGN

In the previous section it has been described that there are different studies that have been performed for Link Prediction by considering various factors. This section gives comprehensive visions about the proposed study and design for Link Prediction Framework. For our research, we targeted the highly active social platform and performed the semantic analysis on it.

The purposed approach consists of two frameworks Post analysis and Post kind analysis to predict the links. The theme of the purposed work is to explore page follower's behavior against pages posts in the form of likes and comments, checking the kind of the posts and checking the page trends to predict the links between page followers.

#### A. Conceptual Framework for Data Selection

In Fig. 1, a conceptual framework is given for data extraction. Initially, data collection is one of the major tasks as there is neither any database publicly available nor any tool is available to extract data directly. So, for our research we selected the publicly available pages of Facebook and extracted the data related to posts from these pages.

The Facebook is picked as data source because of its diversity in nature of data, as it provides rich functionalities alongside humongous audience. There are certain other social networks which also provide different functionalities

according to user interest i.e. LinkedIn the users are closely connected while on Instagram only the images and video data is present.

Principally, for data collection, we used different methods and techniques to make the properly consolidated dataset. After the extraction of data, we applied different scripts and procedures to make a proper authenticated dataset as extracted data contains huge crude information. Preprocessing step contains the several stages such as removing of stop words, white spaces, emojis, URLs, etc. This process required ample effort and time.

**B. Analysis**

Our first approach for Link Prediction was by using the semantic framework on Posts data and making the relationships between page followers based on it. Fig. 2, describes semantic similarity kernel framework on the dataset. After preprocessing we applied the Semantic kernel on the targeted dataset.

Primarily, applied the TF-IDF method to get term document matrix as it defines the frequency of input words in the dataset and term by term co-relation. It helps to identify which post is more identical to another post on the same page. We applied the KNN approach by Semantic Kernel to determine the relationships in the dataset.

The Semantic kernel also comprises the GVSM (Generalized Vector Space Model) [27] by accepting the vectors which are independent linearly and gives the results in form term by term relationship. Let suppose if X is matrix which consists of n documents and m terms than using GVSM gives the semantic kernel.

$$K = X^T G X \tag{1}$$

In this expression K shows the gram matrix of rows, G is the gram matrix of Columns. G must be semi positive and must show the internal product of vector terms.

$$G_{ASSC} = Q Q^T \tag{2}$$

$$G_{ASSC\_N} = L_Q^{1/2} Q Q^T L_Q^{1/2} \tag{3}$$

Where  $L_Q$  is a  $m \times m$  diagonal matrix whose terms are the diagonal elements of  $Q Q^T$ . The semantic kernel that parallels to different estimates of K are:

$$K_{ASSC} = X^T Q Q^T X \tag{4}$$

$$K_{ASSC\_N} = X^T L_Q^{1/2} Q Q^T L_Q^{1/2} X \tag{5}$$

We can calculate the cosine similarity by

$$Sim = L^{1/2} K L^{1/2} \tag{6}$$

It expands the clustering approach and gives which terms are more co-related. If the XYZ is a Facebook page by applying the semantic kernel on posts, post A terms matches with the post D terms. Let A contains 20 comments by page followers and D have the 40 comments by page followers than we can predict that it might be possible that post A followers create the relation in future with post D followers as both posts having the same content.

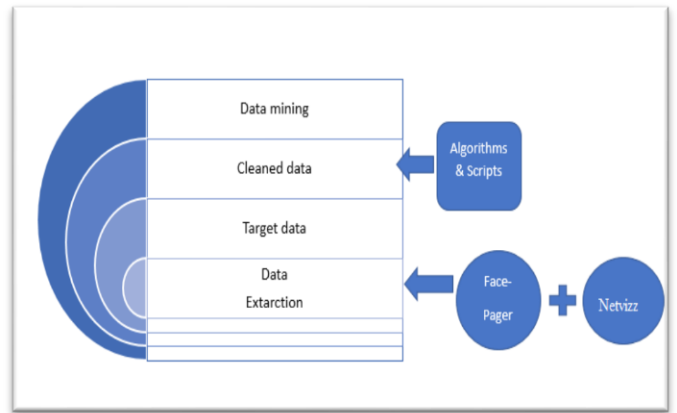


Fig. 1. A conceptual framework for data gathering.

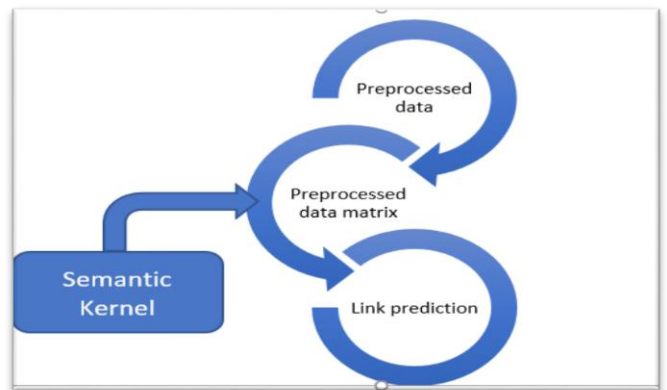


Fig. 2. Framework for semantic kernel.

**C. Link Prediction based on Post Kind Analysis**

Our second proposed method for link prediction is the Post kind analysis. Facebook page followers can comment on the posts and posts consist of URLs, emojis, digits, hashtags, etc. These comments and likes depend upon the post nature. Users behavior varies from post to post. It is not necessary if a user liked the one post of the page may or may not like all the page posts. For our framework, we first changed the preprocessed posts dataset into the matrix and compared that matrix with the dictionary [28], [29]. As the results, we get three categories of posts text as shown in Fig. 3.

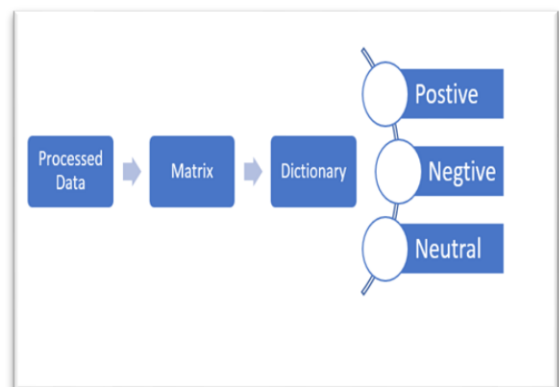


Fig. 3. A conceptual framework for post kind analysis.



- Positive
- Negative
- Neutral

We made the user’s classes against these categories to predict the relations between them. It is also beneficial to identify the page trend. As more optimistic posts on the page the more positive page, it is.

If the XYZ is a dataset of Facebook posts by comparing the dataset with the dictionary we categorized the posts. Let’s suppose there are 2 posts A and B and both posts belong to be optimistic category. Or 20-page followers showed the response against the post-A or 30-page followers showed the response against the post B on the bases of its responses we predicted the links that might be developed with time. It also helps to identify and predict the link in such a way that if 2 posts having the same nature then the post follower of A must show response to the post B.

#### IV. EXPERIMENTAL DESIGN AND SETUP

In this section, complete details are given related to the collection of data set and also complete results are shown in this chapter.

##### A. Data Collection and Analysis

This study covers the phases that require meeting the problem of Link Prediction based on semantic analysis and deep learning.

Consequently, for data collection, we used the different methods and techniques to make the properly consolidated dataset. Gathered data contains multiple columns such as.

- Page followers like against posts.
- Query Time and its duration.
- Page follower’s comments against posts.
- Followers Ids.
- Post creating time, date, etc.

Where we pick out the only those data columns which meet the condition. Data preprocessing was applied to targeted data to remove the raw and unstructured dataset. The preprocessing of data was very time-consuming.

##### B. Targeted Pages

There are multiple pages on Facebook related to the movies, games, education, celebrities, news, books, poetry, online shopping and many more. On the bases of its content, we can categories these pages. For experimentation, we targeted the four categories as shown in the Table I.

Initially, against each category, we extracted the two different pages data where data extraction sections contain two parts.

IDs are gathered using Netvizz [30] which is one of the Facebook applications. It supports data extraction, data collection from different parts of Facebook. Against the selected page IDs, we extracted the data from Facebook using

Facepager [31] and R tool. Facepager can fetch the publicly available data from Twitter, Facebook and web pages. Facepager takes the pages and groups IDs and allows users to access the data. It permits users to retrieves the Facebook posts, albums, pictures data also its metadata in the form of likes, shares, comments, and tags etc. For this work, we choose only Pages posts as mentioned above. We created new databases for targeted Pages and retrieve the data of Posts and its metadata and stored in the form of CSV as shown in Fig. 4.

TABLE I. TARGETED PAGES

Category	Page 1	Page 2
Games	Minecraft	Scrolls
Scholarship	Scholarship Networks	Scholarship Networks of Pakistan
Online Shopping	Darazpk	MyGerrys
Social	Humans of Pakistan	Humans of Kinnaird

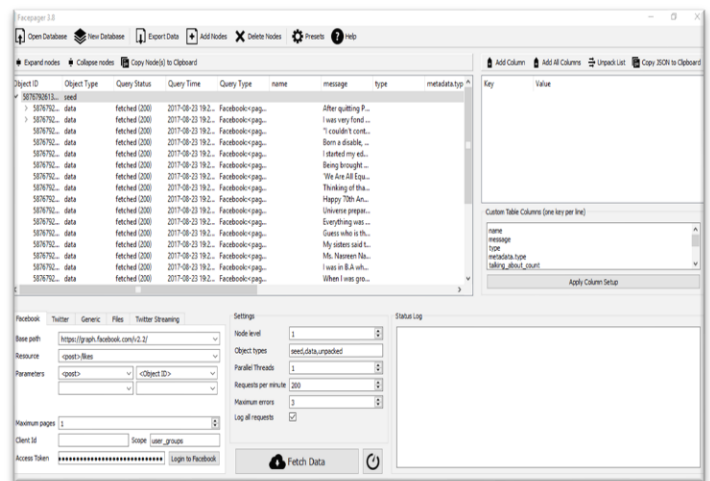


Fig. 4. Facepager.

##### C. Preprocessing of Data

The Extracted dataset was not in purified form as the posts were written in different format e.g. good was written as gud. Intelligent replacements of such words were made to change the data in the purified form by using scripts and algorithms which gives the text free from irrelevant content and reduce the size of the dataset for better processing. It consists of three steps where each step its own importance.

Step 1: Removed the stop words (most frequently used words) as it’s not directly related to the content, Remove the extra spaces in the text, digits etc.

Step 2: Making term-document matrix.

Step 3: Removing URLs, digits, emojis, etc.

Step4: Intelligent replacements of words.

We used the C++, Java scripts and R tool for this purpose the transformation of the dataset was very time-consuming

and another difficult task whereas preprocessing results are shown in Fig. 5, which shows that extracted Scrolls dataset contains 18% raw data and only 32% was useful. After preprocessing, was applied the semantic kernel on the selected dataset which sequentially compares page posts with each other. As page follower’s behavior varies from post to post even on the same page. If a follower acts positively against one post on the page it’s not necessary he/she act similarly on other posts. As the semantic kernel tells similarity which helps to identify which posts have the same content. We assigned the temporary number to the Posts IDs for easy processing of dataset and in graph representation phase we used again the original IDs. This framework 1 gives the content based post similarity. The following form of results is produced on the dataset of Scrolls as shown in figures. We interpreted the all possible posts co-relate with other posts where few pots create the direct relations and few generated the co-similar relation. We picked the one node(post) form the posts dataset and checked its co-similarity with another node(post). Let’s suppose we select the node(post) 8 as shown in Fig. 6. Now by checking its similarity with other posts it leads to the post 4 and 15 post as shown in Fig. 7 and 10, respectively. By continuing this example following relations are formed.

Post 4 has the most co-similarity with the post 34 as shown in Fig. 8 where 34 have co-similarity with post 27 vice versa as presented in Fig. 9. So we end the process here as both post are bi-relational.

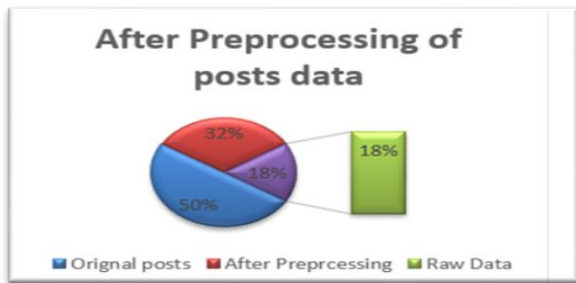


Fig. 5. After reprocessing the useful dataset.

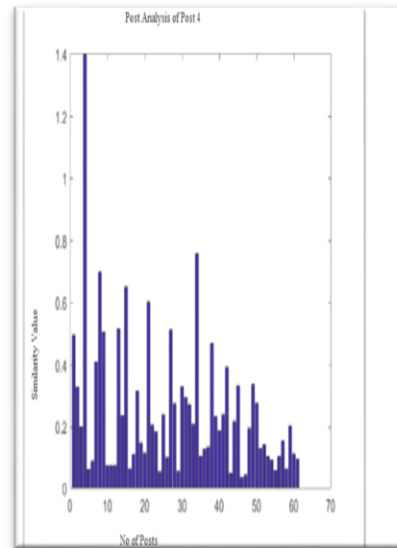


Fig. 7. Post 4 of Scrolls dataset.

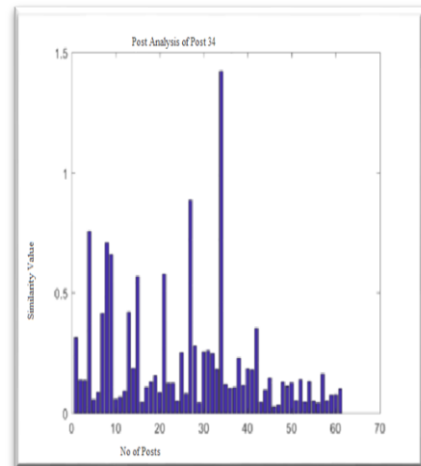


Fig. 8. Post 34 of Scrolls dataset.

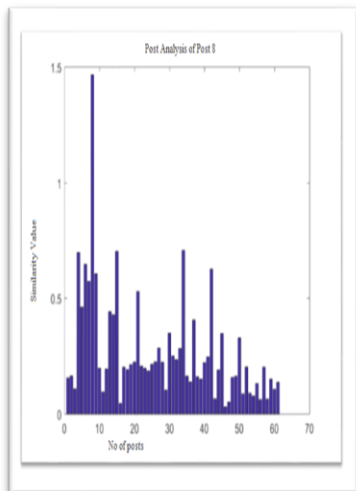


Fig. 6. Post 8 of Scrolls dataset.

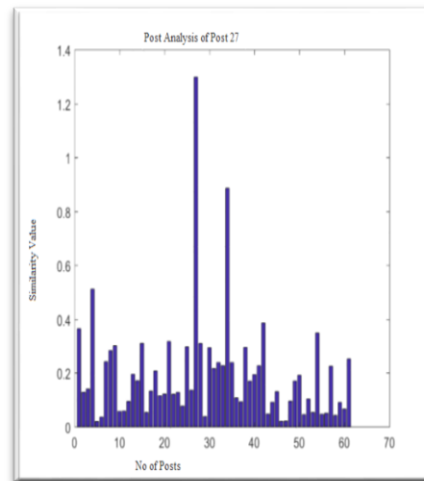


Fig. 9. Post 27 of Scrolls dataset.

First relation is shown:

8 ->4 ->34 ->27

As 8 also created the co-similar relation with 15 posts as shown in Fig. 6. Now continuing this relation process by post 15 as shown below.

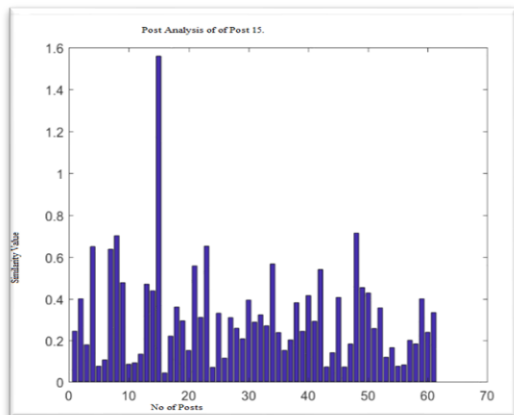


Fig. 10. Post 15 of Scrolls dataset.

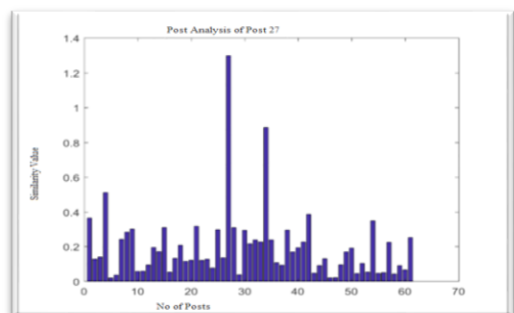


Fig. 11. Post 27 of Scrolls dataset.

Post 15 has the co-similarity post 48 vice versa.so we end the process here as both post are bi-relational as shown in Fig. 11.

Here the nodes are 8,4,34,27,15,48. The direct relation created between nodes are as follows:

- 8 ->4,
- 4 ->34
- 34 ->27.
- 8 ->15
- 15 ->48.

By using these directed relations of nodes, we find the co-similar links in the nodes as shown below:

- 8 ->34
- 4->27
- 8-> 48.

For representation of the results of this proposed framework we used the Graph technique as shown in Fig. 12.

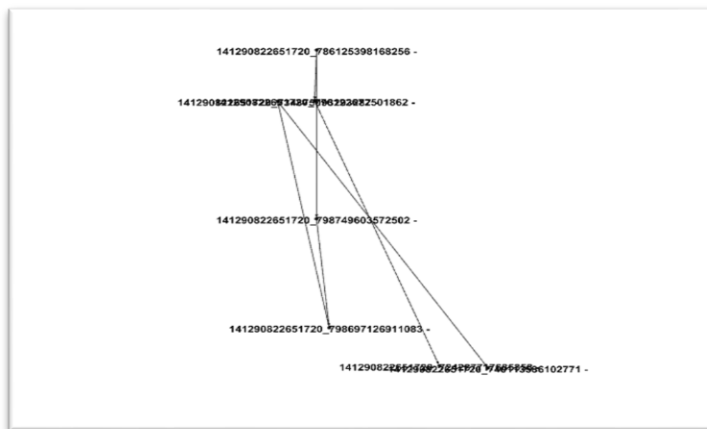


Fig. 12. Results of semantic analysis framework of scrolls above example.

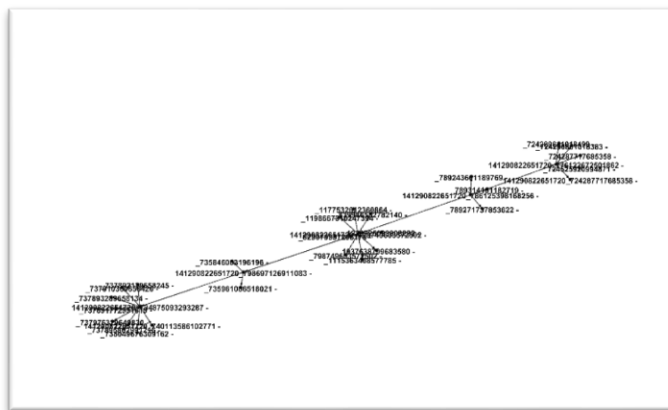


Fig. 13. Followers against these posts where post followers are attached to Posts IDs.

The first half or the IDs shows the Page ID where other half is that person ID who posted on the page. IDs i.e. 141290822651720\_781519135295549

Fig. 13 shows the Post followers against these posts who commented on these posts.

A similar experiment has been done on all above-mentioned Pages and results are incredible. It is most commonly used technique for the network analysis and provides the actual representation of a network in the form of graphs. As the graph consists of Nodes (V) and Edges (E):

$$G = (V, E)$$

We picked out the one Node we called it starting Node and against that Node we checked which post is more co-related with it. We did the same with the second Node and continued this process until 2 posts created the bidirectional graph. Consequently, by using these graphs we find the Co-Similar links and giving the links prediction on the bases of these Co-Similar links. Fig. 14 and 15 signifies the Co-Similar Links of Humans of Pakistan.



Fig. 14. Co-similar link of human of Pakistan.

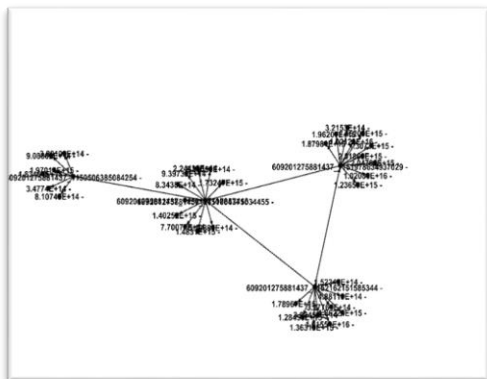


Fig. 15. Co-similar link of humans of Pakistan where post followers are attached to Posts IDs.

Where, the first node is 9 and following relation is formed 9->20,20->16,16->7 are bi-directional. node 9 also have Co-Similar content with post 13 where 13 have with post 9 and making a bi-directional relation. Co-Similar Links are the 9-> 16, 20->7,13->20.

Fig. 16 and 17 shows the Co-Similar links results of Human of Kinnaird.

Where, starting node is 9 ->8, 8->5 and with 8 ->10 where 10 is having again Co-Similarity with post 5. Post 8,10,5 having triad relation. Here Co-Similar links are the post 9-> 5 and 10->9.

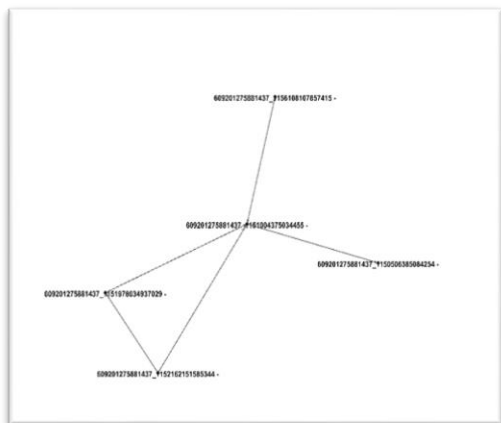


Fig. 16. Co-Similar Link of Human of Kinnaird.

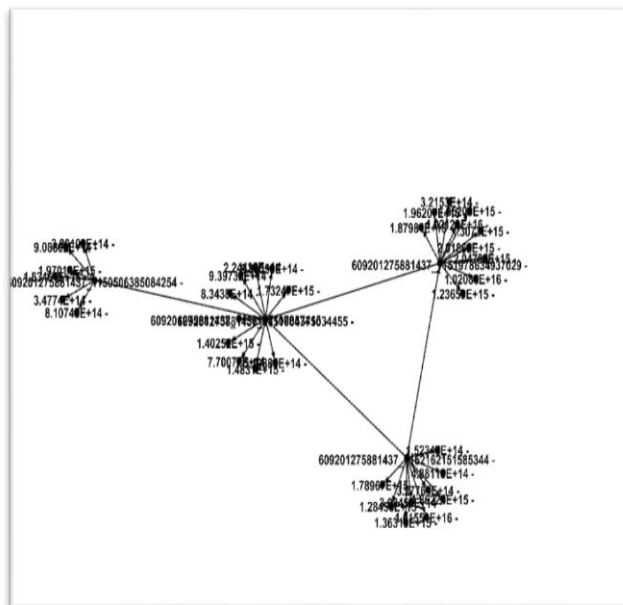


Fig. 17. Co-similar link of human of Kinnaird where post followers are attached to Posts IDs.

TABLE II. POSITIVE AND NEGATIVE USERS

Sr. No	Positive Users	Negative Users
1	141290822651720_687206098060187	141290822651720_776122849168511
2	141290822651720_697046457076151	141290822651720_776122272501902
3	141290822651720_695465677234229	141290822651720_729318370515626
4	141290822651720_729318370515626	141290822651720_728448620602601
5	141290822651720_716699601777503	141290822651720_729331113847685
6	141290822651720_687207418060055	141290822651720_729331113847685
7	141290822651720_687206098060187	141290822651720_697290067051790
8	141290822651720_710416342405829	141290822651720_687206098060187
9	141290822651720_729318370515626	141290822651720_776122672501862

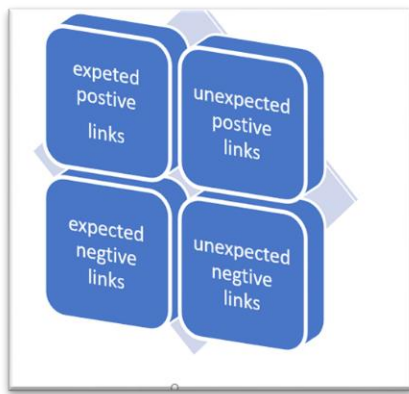


Fig. 18. Possibility of links creates between page followers based upon the dictionary results.

#### D. 2<sup>nd</sup> Method

After comparing the Posts of Scrolls with dictionary, following results are produced. On the bases of these Positive and Negative posts, we predicted that 1st post follower may create the relation with post follower of the same category. The same experiment is done on all other categories of Facebook pages. For the experiment, we take the 62 posts after analysis it divided the posts 9 in the Positive category and 9 in Negative and remaining all are the Neutral as shown in Table II. Fig. 18 expresses the expected type of links.

#### V. CONCLUSION

The theme of this work is to exploit Social Networks for prediction of nature of relationships among users that are not directly connected. For this purpose, alike pages from famous Social Networks was selected and data was gathered according to nature of work by using our proposed framework, results have been achieved. Moreover, our proposed framework indicates the involvement of semantic approach.

The proposed framework was including the involvement of dictionary in order to find the nature of post which is also playing the vital role in categorization posts as well as the links among users. For future work, a comprehensive tool should be developed that has the capability to exploit the public available data from SN. Thus, results are creating links among users belong to different networks. Moreover, the data can be used to monitor the activities of users on certain page or group. In addition, this activity will also enable us to find the Groups or Page with public sentiments. Finally, the timing in our approach certainly enables us to find spam pages or groups as well as users across any social network.

#### REFERENCES

- [1] Scott, J. (2017). Social network analysis. Sage.
- [2] Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology*, 58(7), 1019-1031.
- [3] Volkova, S. LINK PREDICTION IN SOCIAL NETWORKS.
- [4] Ressler, S. (2006). Social Network Analysis as an Approach to Combat Terrorism: Past, Present, and Future Research. *The Journal of Naval Postgraduate School Center for Homeland Defense and Security*, 2(2).
- [5] Colgrove, C., Neidert, J., & Chakoumakos, R. : Using network structure to learn category classification in Wikipedia. (2014-01-09),2011.

- [6] Hsu W. H., Lancaster J., Paradesi M. S. R., & Weninger T. :Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach, In Proceedings of the International Conference on Weblogs and Social Media ,March 26-28, 2007.
- [7] Bengio, Y., Courville, A. C., & Vincent, P.: Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 1. 2012.
- [8] L. Getoor & C. P. Diehl. Link mining: a survey. *SIGKDD Explorations*, Special Issue on Link Mining, 2012.
- [9] Link Prediction in Relational Data, B. Taskar, M. F. Wong, P. Abbeel and D. Koller. *Neural Information Processing Systems Conference (NIPS03)*, Vancouver, Canada, December 2003.
- [10] Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. :Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*,2006.
- [11] Miller, K., Jordan, M. I., & Griffiths, T. L. : Nonparametric latent feature models for link prediction. In *Advances in neural information processing systems* (pp. 1276-1284),2009.
- [12] Liu, W., & Lu, L:Link prediction using random walk Available at <http://arxiv.org/abs/1001.2467>,2010.
- [13] Almansoori, W., Gao, S., Jarada, T. N., Elsheikh, A. M., Murshed, A. N., Jida, J., ... & Rokne, J: Link prediction and classification in social networks and its application in healthcare and systems biology. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 1(1-2), 27-36,2012.
- [14] Milne, D. :Computing semantic relatedness using wikipedia link structure. In *Proceedings of the new zealand computer science research student conference* (pp. 1-8),2007.
- [15] Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., & Elovici, Y. :Link prediction in social networks using computationally efficient topological features. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, 2011 IEEE Third International Conference on (pp. 73-80). IEEE,2011.
- [16] Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. *Proc. of the international conference on Information and knowledge management*, 556-559.
- [17] Bliss, C. A., Frank, M. R., Danforth, C. M., & Dodds, P. S. (2014). An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science*, 5(5), 750-764.
- [18] Zhu, J., Song, J., & Chen, B. (2016). Max-margin nonparametric latent feature models for link prediction. *arXiv preprint arXiv:1602.07428*.
- [19] Tang, J., Chang, S., Aggarwal, C., & Liu, H. (2015, February). Negative link prediction in social media. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 87-96). ACM.
- [20] Sett, N., Singh, S. R., & Nandi, S. (2016). Influence of edge weight on node proximity based link prediction methods: An empirical analysis. *Neurocomputing*, 172, 71-83.
- [21] Hristova, D., Noulas, A., Brown, C., Musolesi, M., & Mascolo, C. (2016). A multilayer approach to multiplexity and link prediction in online geo-social networks. *EPJ Data Science*, 5(1), 24.
- [22] Shang, K. K., Small, M., & Yan, W. S. (2017). Link direction for link prediction. *Physica A: Statistical Mechanics and its Applications*, 469, 767-776.
- [23] Esparza, S.G., M.P. O'Mahony, and B. Smyth, Mining the real-time web: a novel approach to product recommendation. *Knowledge-Based Systems*, 2012. **29**: p. 3-11.
- [24] Popescul, A & Ungar, LH 2003, Structural Logistic Regression for Link Analysis, Proceedings of KDD Workshop on Multi-Relational Data Mining, 2003, viewed 12 June 2006.
- [25] Popescul, A and Ungar, LH 2004, Cluster-based Concept Invention for Statistical Relational Learning, Proceedings of Conference Knowledge Discovery and Data Mining (KDD-2004), 22-25 August 2004, viewed 12 June 2006.
- [26] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in Proceedings of the 19th

- International Conference on World Wide Web, pp. 641–650, New York, NY, USA, April 2010.
- [27] Farahat, A.K. and M.S. Kamel. Document clustering using semantic kernels based on term-term correlations. in Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on. 2009.
- [28] Opinion Lexicon: Positive (<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>)
- [29] Opinion Lexicon: Negative(<http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>)
- [30] Netvizz: <https://tools.digitalmethods.net/netvizz/facebook/netvizz/>
- [31] Facepager: <https://github.com/strohne/Facepager/tree/master/build>

# Matrix Clustering based Migration of System Application to Microservices Architecture

Shahbaz Ahmed Khan Ghayyur

Department of Computer Science & Software Engineering  
Faculty of Basic and Applied Sciences  
International Islamic University  
Islamabad, Pakistan

Abdul Razzaq

Department of Computer Science & Software Engineering  
Faculty of Basic and Applied Sciences  
International Islamic University  
Islamabad, Pakistan

Saeed Ullah

Department of Computer Science,  
Federal Urdu University of Arts Science and Technology  
Islamabad, Pakistan

Salman Ahmed

Department of Computer Science & Software Engineering  
Faculty of Basic and Applied Sciences  
International Islamic University  
Islamabad, Pakistan

**Abstract**—A microservice architecture (MSA) style is an emerging approach which is gaining strength with the passage of time. Micro services are recommended by a number of researchers to overcome the limitations and issues encountered by usage of aging method of monolithic architecture styles. Previously the monolithic applications cannot be decomposed into smaller and different services. Monolithic styles application was the one build application. The issue resolution has the focus on lightweight independent application services in the form of sizable services, self-contained units with primary focus on maintenance, performance, scalability, and online services eliminating dependency. All quality factors have been thoroughly discussed in literature, system application migration is becoming an emerging issue with different challenges. This study is addressing the tight coupling to reducing this issue. Moreover, this literature review indicates some complex problems about the migration or conversion of system application into microservice. In architecture, dependency is a big challenge and issue in recent technology. Microservices are recommended by a number of researchers to overcome the limitations issue about how to migrate the existing system application to microservice. The need for a systematic mapping is essential in order to recap the improvement and identify the gaps and requirements for future studies. This study shows open issues first, new findings of quality attributes of microservices and then this study helps to understand the difference between previous traditional systems and microservices based systems. This research study creates awareness about system migration to microservices.

**Keywords**—*Monolithic architecture; microservices architecture; systematic mapping; system migration; application transformation; traditional application development; emerging challenges; API*

## I. INTRODUCTION

### A. Microservices

Microservice possesses important characteristics like it claims the responsibility of a single task, it meets all the requirements of a single business, it can be individually

deployed, it is loosely coupled and it is independently responsible as it is self-contained [9], [10]. An enterprise application which is designed for a particular organization consists of different microservices which are responsible to communicate with one another with the help of a light-weight protocol and the API contract [2]. MSA design is generally preferred to the conventional Monolithic Architecture due to the fact that it can be continuously deployed and its scalability has no parallel while the conventional Monolithic Architecture lacks all these important features. Because of this undeniable charm of MSA design, most of the enterprises tend to prefer this design [6].

In the beginning, the developers introduced the concept of service orientation with the help of SOA. Later, the evolutionary process took place and service orientation became capable of supporting the easy and swift operability of the applications designed as per requirement [12]. Now, with more research being carried out in this field, researchers have started building independent, multiple and self-contained services to meet the challenges of the market [23]. Because of these features, there is no denying the fact that Software Architecture plays very important role in software lifecycle to support the quality and vital attributes of the software [13]. This approach helps the developers to make sure that quality attributes are up to the complete satisfaction and there exists no defect in the design of the software systems [23]. If maintenance and development of information system needs to be improved, Component Based Development is useful and viable solution for these requirements [13]. Need of SOA approach coupled with its products was felt because of the reasons that a Component Based Distributed Architecture was required in the market [30]. Moreover, a solution was required to make the business agile and meet the challenges that arise when a particular business need is to be met. Moreover, a compatible and flexible solution was required which may become capable of keeping pace with the evolution that takes place with every single day that passes [8].

In service oriented software companies, the micro services have become architecture style that is inspired by service oriented computing. [3], [6]. Microservices architecture helps to develop the complex application along with the distribution of the application in chunks or units by composing it [1]. Nowadays, in any system, the scalability, service discovery and communication among services that are being supported by microservices architecture in development phase are two important sections [45]. Simultaneously, microservices architecture also handles a heavy concurrency during input load [2]. In fact, the purpose for using microservices is that it works on latest platform and is independently deployable [1]. Microservices API can be written in any language. Then, Microservices architecture would automatically make all the languages compatible to display the desired output [4].

There is no denying the fact that the ever-changing evolutionary process of styles of communication and integration has proved to be cyclic. At times, some of its concepts seem to fall apart and looks as if these concepts would become obsolete with the passage of time but these concepts resurface in different and refined forms as the time elapses. Out of these two styles; Service Oriented Architecture is relatively an older concept because of obvious reasons [8].

#### B. Migration

System application migration is becoming an emerging issue with different challenges. Transform that migration is the procedure of moving from the usage of one functional environment to another operating environment with alike functionalities [32]. The migration procedure contains, and making sure the new environment's features are exploited, old settings do not require changing and that present applications continue to work.

#### C. Migration to Microservices

The focus of migration is that it indicates some complex problems about the migration or conversion of system application into microservice. Migration of the system to microservice optimizes decentralization, replace-ability and autonomy of software architectures [32]. Although, researchers are not convinced on any specific definition of microservice, its modelling techniques, and its properties [7], it is aware about system migration to microservices.

The components which are used in these vital software applications are made up of basic blocks which can be combined together depending upon the requirement [17]. Microservice architecture is preferred owing to the reasons that it has the capability to address all the concerns starting from requirement of the enterprise to the operations to be performed by the software of a particular business for which it is designed. Moreover, it can also claim the responsibility for individual teams [25], [38]. In this type of approach to find out the solution, the architecture, open source development, organizational structure and responsibility is vertically decomposed [14], [43].

#### D. Clustering

In this technique, reverse engineering also produces desired results. The technique used for the purpose of reverse engineering is clustering which is the considered the simplest

and fundamental technique used in engineering and science [17]. Main and most important objective of implementing this technique is to make the observations clearer to develop a better understanding. This better understanding makes it easy to develop complex knowledge structure from given features. Clustering technique or method is generally preferred to identify all the related components of System Software Application along with their responsibilities. As the input used in this technique highlights the interconnectivity of all these components, this clustering technique is quite useful to minimize the interconnection among different components to produce optimum results [30]. Clustering is a technique in which large systems are decomposed into chunks and smaller and manageable systems in a distinct way that the entities which bear similarity with one another belong to the same subsystem while the entities with difference among one another are classified into different subsystems [17]. Clustering technique is generally used in identifying the software components which generally adopts one metric so that the similarity of components may be measured. The main advantages of this technique are that low coupling and high cohesion of components are achieved. These advantages play very important role to solve the problems which require software evolution [21, [27].

There exist a lot of clustering techniques out of which Hierarchical Agglomerative Clustering (HAC) and K-means clustering stand out. Hierarchical Agglomerative Clustering (HAC) plays very important role to find out the number of clusters or segments which do not work well or cause inconvenience because of malfunctioning in practice [16]. Moreover, K-means is also used to locate the numbers of clusters or segments which do not work well but the only problem that occurs is the fact that it cannot be applied in HAC algorithms.

#### E. Need of Systematic Mapping

Many different software companies have recently migrated to microservices or are considering migrating to microservices. These services are known as a style of an architecture that develops an application as a set of small services independently [7]. Now, microservices are becoming very popular with cloud platform which is an emerging style in the context of application development due to its independency, scalability, flexibility, performance, and manageability [3], [5]. There is a lot of research in this area that needs to be address. In the previous a few years, the software product companies and software consultancy firms have found the microservices approach useful because it allows the team and software organizations to increase the productivity [6].

Ever-changing needs of customers due to ever-changing situational contexts and business needs inspire the enterprises to introduce evolutionary concepts in software products to compete the market. Due to these developments, most of the Software Development Organizations and the businesses which include Software Production are facing bursting pressure to improve their Software Intensive Systems on daily basis. They can achieve this goal if they develop and release valuable and compatible software in a very short span of time to meet the challenges of the market [11].



## II. BACKGROUND KNOWLEDGE

Literature review sheds light upon the importance and architecture of microservice. Moreover, this literature review indicates some complex problems about the migration or conversion of system application into microservice. This discussion in literature answers the very first question of this research paper. It highlights all the issues which involve migration to microservice [32]. The main features which are creating and promoting the demand of such a technology are scalability, security, reliability, fast progress, and speed of the network [20]. So, the researchers are trying hard to introduce new software architecture styles and software development methods to meet all the demands of enterprises [6]. Migration of the system to microservice optimizes decentralization, replace-ability, traceability and autonomy of software architectures. Although, researchers are not convinced on any specific definition of microservice, but it is modelling techniques, and its properties [7].

Microservice plays very important role to capture software maintenance, architecture and evolution [15]. If software architecture recovery is taken into account, it becomes clear that the prevailing techniques in this field are quite limited because all these techniques are based upon reverse engineering [5].

Software designer or developer generally encounters two types of problems in practical. First issue is embedded in the fact that it is quite tough to determine specific cluster which is used for highly coupled components [15]. Second problem in line is to determine the cluster mapping which is applied on software modules [17]. Upon investigation, the technique of decomposition of software has made sure that the source code of software is in accordance with all the requirements gathered.

Main drawback of the traditional monolithic services is its lack of scalability when a certain task is to be executed within the service [9]. Long software release cycle because of the complexity of system is also a hurdle in traditional monolithic services. Because of these limitations monolithic approach has shifted towards the development of modern cloud application [35], [36].

ICT is making a name and becoming a reliable partner in demand driven and dynamic market environment because of its customer experienced, customer centered and ever-changing demand driven competitive market [38]. Because of this competitive race of different enterprises, most of the companies are transforming themselves into virtually organized bodies with pure digital styles. These virtually organized bodies are supported and enabled by the applications based on microservice [18]. Genuinely, microservice in any application is responsible to execute a single task, i.e. it works on only one business requirement at a time. Moreover, it is self-contained that it can complete its responsibility without depending upon any other software [6]. For example, it contains business and data layers, and presentation all together. Additionally, it is loosely coupled, light-weighted and autonomous [1].

If the applications are to be run in cloud with efficiency, it requires much more skill than what is necessary to deploy any type of software in virtual machines. It is always recommended to manage cloud applications continuously in order to utilize their resources according to the incoming load and to face the failures in order to replicate and restate all the components to provide resilience in case of unreliable infrastructure [42]. Once a program or software is designed keeping in view all the requirements, it becomes extremely tough for the designer to introduce radical changes which are later on demanded by business models or user frequently because it becomes more complicated for the developer to make changes when the code starts expanding because of the involvement of different people or specialist who make changes in the software [14], [26]. As more and more effort is required to coordinate for updating in tightly coupled model of monolithic approach, this whole process ultimately makes the release cycle of the application slow [37]. It also makes the model fragile and unreliable. Scalability is also a vital feature which is required in the operation and development of enterprise applications [9], [14], [22].

## III. RESEARCH METHOD

### A. Planing of Mapping

This mapping study, researcher is combining the knowledge for all issues which are related to system migration. There are different types of systems migration techniques but researcher is migrating the system migration that is based on components. This knowledge will help us to migrate the system application and mitigate these issues during migration and why researcher needs migration of system application. This study will aware about microservices understanding and its characteristics.

### B. Search Strategy

The term of 'microservice' keyword and microservice architecture that found in the published articles journals, and conferences but rest were excluded. Our selected research papers 48 which are published between 2010 and 2017. Selected research papers' electronic digital libraries are included which are Four (IEEEExplore, ACM DL, DirectScience, ResearchGate, GoogleScholar) (Table I). In this systematic mapping study the selected papers are maximum from the IEEEExplore.

TABLE I. SELECTED ELECTRONIC DATABASES

Electronic Database	URL
IEEE	<a href="http://ieeexplore.ieee.org/Xplore/">http://ieeexplore.ieee.org/Xplore/</a>
ACM	<a href="http://dl.acm.org/">http://dl.acm.org/</a>
ScienceDirect	<a href="http://www.sciencedirect.com/">http://www.sciencedirect.com/</a>
GoogleScholar	<a href="https://scholar.google.com.pk/">https://scholar.google.com.pk/</a>

TABLE II. ELECTRONIC DATABASE

Digital Library	Publications	Selected
IEEE explorer	208	32
ACM Digital Library	796	4
Science Direct	140	1
ResearchGate	3	1
Other		10
Total	1147	48

### C. Keywords

These keywords which are used for finding all the studies are:

((({Microservice} OR {Monolithic} OR {Traditional}) AND {Architect\*}) AND ({System Migrat\*} OR {Transform\*} OR {Component} OR {API} OR {Cloud}) AND year >= 2010 AND year <= 2017

### D. Selection of Primary Study

This section suggests that many studies were deeply checked before the selection of this study. Moreover, relevance to the research question was also given due consideration. At first, the papers were included after carefully reading title and abstract. In case of any ambiguity about the paper in title and abstract section, the researcher reviewed the complete paper by applying inclusion and exclusion criteria.

### E. Search Engine

The term of 'microservice' keyword and microservice architecture that found in the published articles journals, and conferences but rest were excluded. Selected research papers' electronic digital libraries are included which are four (IEEEExplore, ACM DL, DirectScience, GoogleScholar). In this systematic mapping study the selected papers are maximum from the IEEEExplore.

### F. Inclusion Criteria

- Studies had been published in journals, conferences, and workshops.
- Studies must be written in English.
- Studies must be accessible electronically.
- Collected studies must be published after 2010.
- Research papers will be included which are based on the expert opinion
- Research papers related to the topic, will be included as weak evidence which do not provide evidence

### G. Exclusion Criteria

- Non peer reviewed studies (tutorials, slides, editorials, posters, keynotes) are also excluded.
- Peer reviewed but not published in journals, or conferences (e.g. Book, and blogs articles).
- Publications not in English
- Electronically non-accessible.

### H. Conducting Mapping Study

Research papers which are published in different conferences or journals that would be a complete version, on the basis of studies, discussed in this article, will be included. Selected primary studies are 48 (Table II). But the further evaluation for these studies researcher has included the studies that are most appropriate to the topic.

1) *Challenges of Microservices (RQ1)*: These challenges are shown in the challenges keyword graph in Fig. 2. Selected papers have discussed about challenges in depth. Researchers have shown a list of open issues in table of current challenges that are open issues of microservices architecture. These open challenges are not discussed in detail in literature. Table IV shows the challenges of microservices.

2) *Quality Attributes of Microservices (RQ2)*: These are quality attributes Scalability, Independency, Maintainability, Deployment, Performance, Reusability, Security, and Load Balancing have been discussed in this mapping study [9], [42]. But researcher have identified few more attributes of microservices which are Reliability, Portability, Availability, these are also important attributes. Other previous quality attributes have been discussed in this mapping study [6], but researcher find out different few more.

3) *Motivation for Microservice Architecture (RQ3)*: Microservices is a new emerging style which is becoming very familiar adopting by industries. It helps the developer to develop the large and complex application to distribute the application in chunks or unit by composing this application [1]. It can be written in language by using APIs for microservices [3]-[5]. Mostly papers are discussing about its independent services that can be upgrade or new addition or services any time. Table III shows the motivations of Microservices.

TABLE III. RESEARCH QUESTIONS

No.	Research Question	Motivation
1	What challenges has been reported in literature about microservices architecture?	This question MSA will elaborate the current challenges. It will discuss in detail about research challenges of microservices.
2	What are the new quality attributes of microservice architecture?	This question aim to identify the new quality attributes of microservice architecture.
3	What are the main motivations for using MSA?	In this question will discuss the benefits of microservices and the aim is to get insight in what are the main reasons for organizations to architect in a microservices style.
4	What are the existing techniques to migrate the application to microservices?	The main to explore this question to highlight the techniques and methods which are helping to migrating the system application from traditional to microservices.

TABLE IV. COMPARISON BETWEEN TRADITIONAL & MICROSERVICES ARCHITECTURE

#	Traditional	SOA	Microservices
1	Single large application	Several applications sharing services	<i>Small autonomous services</i>
2	Single deployment unit	Multiple units depending on each other	<i>Independently deployable units</i>
3	Limited clustering possibilities	Distributed deployment	<i>Distributed deployment</i>
4	Homogeneous technologies	Heterogeneous technologies	<i>Heterogeneous technologies</i>
5	Shared data storage	Shared data storage	<i>Independent data storage</i>
6	Single point of failure	Single point of failure ESB	<i>Resilient to failures</i>
7	In-memory function calls	Remote calls through ESB	<i>Lightweight remote calls</i>
8	Single large team	Multiple teams with shared knowledge	<i>Independent teams owning full lifecycle</i>

4) *Migrating to Microservices (RQ4)*: There is not proper technique that helps to migrate the complete application to microservices. One paper introduces its method to migrate the application to microservices by independent components but not dependent components [24], [36]. But this method does not help to migrate the complete system to microservices.

## Challenges Keywords of Architecture

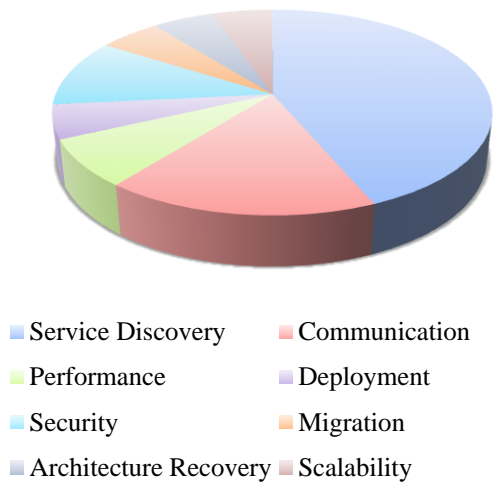


Fig. 1. Related to challenges keywords in architecture.

## Microservices Factors No. of papers

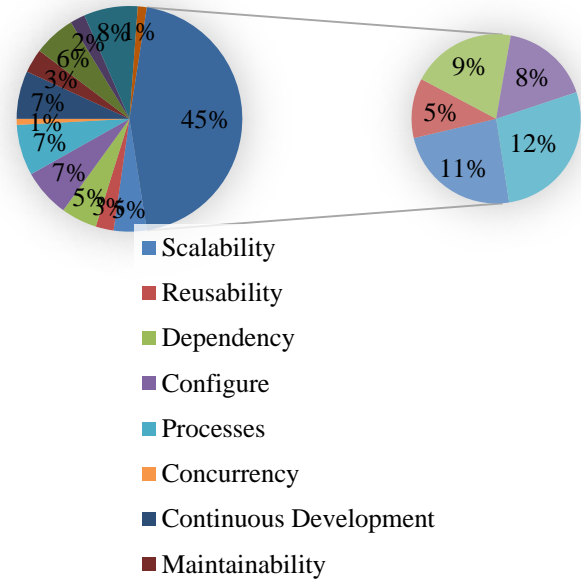


Fig. 2. No. of factors in papers of microservices.

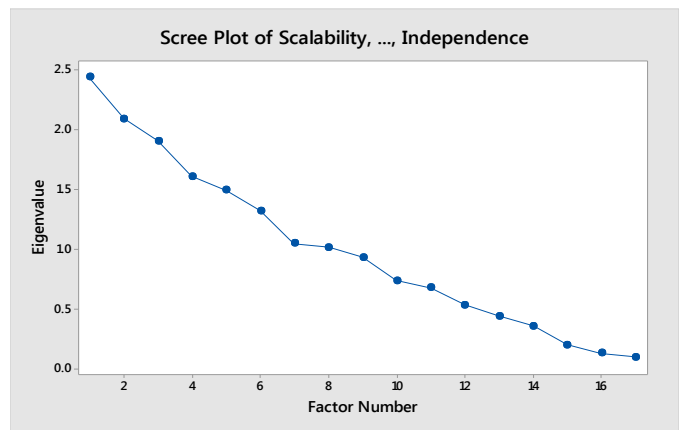


Fig. 3. Microservices factor list analysis variation graph.

### 5) Discussions of Figures

Fig. 1 shows all the challenges keywords of architecture, this figure shows the growth of keywords Table VI. All these keywords have been reported in literature.

Fig. 2 shows all factors of microservices architecture that can be easily measured. Graph shows the importance of factors by percentages.

Fig. 3 shows the variation of all factors. Researchers use the Minitab stats tool to find the variations of all factors.

Fig. 4 shows the analysis result of all factors that how many papers have discussed each factor.

TABLE V. TOP FIVE EMERGING CHALLENGES

# Microservices factors analysis

- Scalability
- Reusability
- Dependency
- Configure
- Processes
- Concurrency
- Continuous Development
- Maintainability
- Load balancing
- Portability
- Security
- Modularity
- Performance
- Reliability
- Cost
- Availability
- Independence

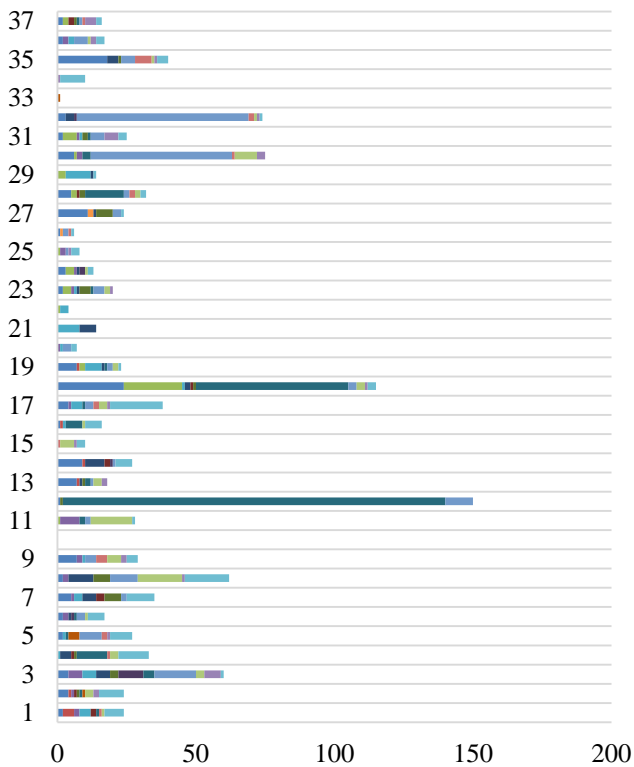


Fig. 4. Factors list of microservices and analysis result graph.

#	Challenge	Description	Ref.
1	<b>Challenges of reimagining and re-architecting a software product.</b>	It is the big challenge for software architect considering microservices is the need to reimagine and also re-think how the application will work.	[1], [5], [7], [12], [18], [31], [34]
2	<b>Testing can become challenging.</b>	Integration testing, it is necessary for the quality assurance engineer to clearly understand each of the different services in order to write the test cases effectively. Debugging meanwhile can mean the QA engineer having to analyze logs across different microservice environments.	[6], [31], [33], [34]
3	<b>System migration to microservices</b>	Old system application needs to be migrated to microservices.	[3], [9], [24], [31], [33], [34]
4	<b>Databases need to be completely decoupled from each other.</b>	It's easy to be decoupled but also a big challenge because previous database schemas worked with different table by its relations now in microservices it need to be changed this. When transitioning to cloud microservices, you need database models 100% decoupled from each other.	[6], [9], [31]
5	<b>Performance monitoring under continuous software change</b>	It's part of microservice architecture that need to be continually changes. Performance does matter very much when following this microservice architecture.	[2], [3], [4], [6], [10], [22], [31], [33], [34]

Fig. 5 found the frequency of all factors that how many papers discussed the each factor.

TABLE VI. FACTORS RELATED TO CHALLENGES

No.	Challenges	Factors
1	Challenges of reimagining and re-architecting a software product.	Dependency, Deployment, Configure, Process, Continues, Development, Maintainability.
2	Testing can become challenging.	Security, Dependency.
3	System migration to microservices	Performance, Independence, Cost, Scalability, Reusability, Continues, Development.
4	Databases need to be completely decoupled from each other.	Independency, Load Balancing, Process.
5	Performance monitoring under continuous software change	Performance, Continues Development, Availability, Load Balancing, Deployment.

**MICROSERVICES FACTORS FREQUENCY**

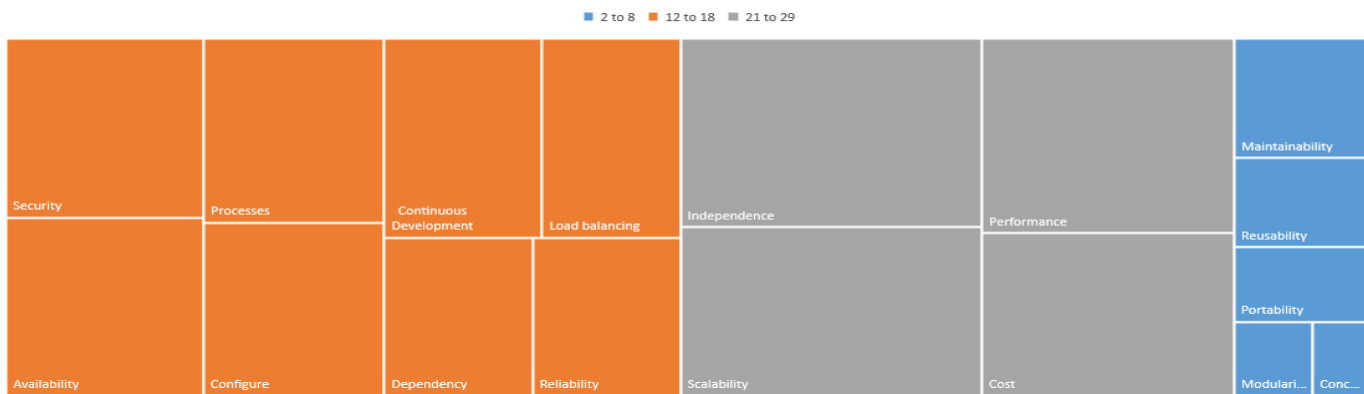


Fig. 5. Frequency of microservices' factors.

TABLE VII. IDENTIFIED FACTORS LIST OF MICROSERVICES

No.	Factors	Paper Reference
1	Scalability	2, 3, 4, 5, 8, 9, 10, 12, 14, 18, 19, 20, 26, 31, 37, 38, 39, 41, 42, 44, 15, 45, 47, 48
2	Reusability	18, 26, 37, 38, 48
3	Dependency	1, 3, 4, 5, 9, 11, 12, 19, 20, 46
4	Configure	1, 4, 8, 10, 12, 19, 20, 29, 37, 38, 39, 42, 44, 15, 45, 46
5	Processes	10, 11, 12, 8, 9, 18, 19, 26, 29, 37, 39, 40, 41, 45
6	Concurrency	2
7	Continuous Development	2, 7, 9, 11, 14, 19, 20, 26, 31, 39, 40, 42, 44, 15, 45, 48
8	Maintainability	3, 5, 9, 37, 38, 40, 44
9	Load balancing	2, 3, 5, 9, 12, 14, 15, 19, 38, 39, 40, 44, 47, 48
10	Portability	20, 31, 39, 42
11	Security	3, 4, 5, 12, 8, 9, 18, 19, 26, 37, 38, 39, 40, 41, 42, 46, 47, 48
12	Modularity	38, 41
13	Performance	1, 2, 3, 4, 5, 8, 9, 10, 12, 14, 15, 19, 26, 29, 31, 39, 41, 42, 44, 45, 46, 47, 48
14	Reliability	3, 4, 5, 7, 8, 14, 31, 37, 40, 41, 45
15	Cost	3, 4, 7, 8, 9, 10, 14, 15, 18, 19, 20, 26, 31, 37, 38, 39, 42, 45, 46, 48
16	Availability	1, 4, 5, 7, 8, 9, 10, 12, 14, 15, 19, 31, 38, 39, 41, 45, 48
17	Independence	1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 14, 15, 18, 20, 24, 26, 29, 31, 37, 38, 39, 40, 41, 42, 44, 45, 46

6) *Factors analysis method*: In this section, researcher used a Minitab tool of stats to analyze the factors of microservices. This tool creates the four different graphs based on analysis result. Graph of scree plot tell us about variation among each factor and show it by dotted line that where the variation is occurring and how much it is. Researcher analyze the result of factors by values, these values mean that how many time each factor is used in literature that is counted as a value of factor for each paper. It means it will help to find the importance of factor and literature focus on factor. Researcher attached the results of all factors in Appendix A section, factor analysis result table (Tables VIII and IX). This research brings forth the number of factors in row and name of factors in column. And in Appendix B section, Fig. 6, 7 and 8 show the result in different view.

#### IV. CONCLUSION

This critical evaluation and mapping study has reviewed carefully the given studies on microservices architecture and the relevant architectural challenges reported in literature. Researchers have discussed in details about microservices. Write the planning of mapping study to produce the results that how it will be shown in this study and the major keywords that support to find the literature related to microservices. Research flow diagram is showing the flow of this study the selection of research papers. Research Questions is a major part of this study these are impact on result of this study. The first research question addresses the different challenges in microservices that is shown in Fig. 1 challenges keywords. These challenges keywords are discussed in depth in literature, but open issues are not deeply discussed. The second question discusses about quality attributes of microservices, most of them quality attributes are discussed in previous literature, but the researcher has identified few more quality attributes which are also important in microservices. The third question discusses motivations of microservices that can be seen in literature and comparison Table IV. The last fourth question is very important of this study is migration of system application to microservices, it shows the importance of migration to microservices in the comparison Table IV. Researcher found the list of emerging challenges in Table V. And highlight the factors of microservices in a list form of Table VII then use the Minitab static tool to analyze the factors of microservices and produce the result in the form of quantitative values and different graph. Scree plot is the major graph of this analysis and is discussed above in Fig. 5. Other graph and results are shown in Appendix sections which are Fig. 6, 7, and 8.

#### V. FUTURE WORK

Analyzed material is based on the state of the art research mined for migration, clustering, and services of Microservices. There is need for performing a detailed empirical analysis of system migration based on software industry input to establish a gap between theory and practice.

Further plans include for proposal of a migration technique for practitioners of micro-services, by which guidelines for software firms shall be proposed to increase their scalability

with productivity. Future research in the area shall consider comparison of proposed methods to similar methods in literature, using a suitable framework.

#### REFERENCES

- [1] Luca Florio, Elisabetta Di Nitto. Gru: an Approach to Introduce Decentralized Autonomic Behavior in Microservices Architectures. 2016 IEEE International Conference on Autonomic Computing. IEEE, 2016.
- [2] Nam H. Do, Tien Van Do, Xuan Thi Tran, Lorant Farkas, Csaba Rotter. A Scalable Routing Mechanism for Stateful Microservices. IEEE, 2017.
- [3] Christian Esposito, Aniello Castiglione, Kim-Kwang Raymond Choo. Challenges in Delivering Software in the Cloud as Microservices. IEEE Cloud Computing published by the IEEE computer society. IEEE, 2016.
- [4] Hamzeh Khazaei, Cornel Barna, Nasim Beigi-Mohammadi, Marin Litoiu. Efficiency Analysis of Provisioning Microservices. 2016 IEEE 8th International Conference on Cloud Computing Technology and Science. IEEE, 2016.
- [5] G. Granchelli, M. Cardarelli, Towards Recovering the Software Architecture of Microservice based system, IEEE, 2017
- [6] Nuha Alshuqayran, Nour Ali and Roger Evans. A Systematic Mapping Study in Microservice Architecture. 2016 IEEE 9th International Conference on Service-Oriented Computing and Applications. IEEE, 2016.
- [7] Sara Hassan, Andreas Oberweis, Rami Bahsoon. Microservices and Their Design Trade-offs: A Self-Adaptive Roadmap, 2016 IEEE International Conference on Services Computing, 2016.
- [8] Zhongxiang Xiao, Andreas Oberweis, and Thomas SchXinjian Qiang. Reflections on SOA and Microservices, 2016 4th International Conference on Enterprise Systems. IEEE, 2016.
- [9] Mohsen Ahmadvand and Amjad Ibrahim. Requirements Reconciliation for Scalable and Secure Microservice (De)composition. 2016 IEEE 24th International Requirements Engineering Conference Workshops. IEEE, 2016.
- [10] Stefan Haselböck, Rainer Weinreich, Decision Guidance Models for Microservice Monitoring. 2017 IEEE International Conference on Software Architecture Workshop. IEEE, 2017.
- [11] Rory V. O'Connor, Peter Elger, Paul M. Clarke, Exploring the impact of situational context – A case study of a software development process for a microservices architecture. IEEE, 2016.
- [12] Csaba Rotter, Gergely. Telecom Strategies for Service Discovery in Microservice Environments. IEEE, 2017.
- [13] Gholam Reza Shahmohammadi, Saeed Jalili. Identification of System Software Components Using Clustering Approach. 2010.
- [14] Wilhelm Hasselbring, Guido Steinacker. Microservice Architectures for Scalability, Agility and Reliability in E-Commerce. 2017 IEEE International Conference on Software Architecture Workshops. IEEE, 2017.
- [15] Mario Villamizar, Oscar Garces, Harold Castro. Evaluating the Monolithic and the Microservice Architecture Pattern to Deploy Web Applications in the Cloud. IEEE. 2015
- [16] Abdulaziz Alkhalid, Chung-Horng Lung, Duo Liu, Samuel Ajila. Software Architecture Decomposition Using Clustering Techniques. 2013 IEEE 37th Annual Computer Software and Applications Conference. IEEE, 2013
- [17] Duo Liu, Chung-Horng Lung, Samuel A. Ajila. Adaptive Clustering Techniques for Software Components and Architecture. 2015 IEEE 39th Annual International Computers, Software & Applications Conference.
- [18] Yale yu, Haydn Silveira, Max Sundaram. A Microservice Based Reference Architecture Model in the Context of Enterprise Architecture. IEEE, 2016.
- [19] Giovanni Toffetti, Sandro Brunner, Martin Bl ochlinger. An architecture for self-managing microservices. ACM, 2015.
- [20] David Jaramillo, Duy V Nguyen. Leveraging microservices architecture by using Docker technology. IEEE, 2016.

- [21] Ibrar Hussain, Aasia Khanum, Abdul Qudus Abbasi, Muhammad Younus Javed. A Novel Approach for Software Architecture Recovery using Particle Swarm Optimization. 2014.
- [22] Nam H. Do, Tien Van Do. A Scalable Routing Mechanism for Stateful Microservices. IEEE, 2017.
- [23] Ben Horowitz. Website: "Adapting the Twelve-Factor App for Microservices". July 28, 2016. Copied date: July 13, 2017.
- [24] Alessandra Levcovitz, Ricardo Terra, Marco Tulio Valente. Towards a Technique for Extracting Microservices from Monolithic Enterprise Systems. Google Scholar. Website 1, 2. Copied date: July 14, 2017
- [25] Bc. Tomáš Livora. Thesis: "Fault Tolerance in Microservices". Masaryk University, 2016.
- [26] Sascha Alpers, Christoph Becker, Andreas Oberweis. Microservice based tool support for business process modelling. IEEE, 2015.
- [27] Jagdeep kaur, Pradeep tomar Validation of Software Component selection algorithms based on Clustering. Indian Journal of Science and Technology, 2016.
- [28] Kamran Sartipi. Software Architecture Recovery based on Pattern Matching. IEEE ICSM.
- [29] Matthias Vianden, Horst Lichter, Andreas Steffens. Experience on a Microservice-based Reference Architecture for Measurement Systems, 2014 21st Asia-Pacific Software Engineering Conference. IEEE, 2014.
- [30] Suresh Marru, Marlon Pierce. Apache Airavata as a Laboratory: Architecture and Case Study for Component-Based Gateway Middleware. ACM, 2015.
- [31] Robert Heinrich, André van Hoorn, Holger Knoche, Fei Li, Lucy Ellen Lwakatare, Claus Pahl, Stefan Schulte. Performance Engineering for Microservices: Research Challenges and Directions. ACM, 2017.
- [32] Armin Balalaie, Abbas Heydarnoori, Pooyan Jamshidi. Microservices Migration Patterns.
- [33] Nicola Dragoni, Saverio Giallorenzo, Alberto Lluch Lafuente, Manuel Mazzara Fabrizio Montesi, Ruslan Mustafin, Larisa Safina. Microservices: yesterday, today, and tomorrow. 2017.
- [34] Armin Balalaie, Abbas Heydarnoori, and Pooyan Jamshidi. Migrating to Cloud-Native Architectures Using Microservices: An Experience Report. ResearchGate, 2017
- [35] Jyhjong Lin, Lendy Chaoyu Lin, S. Huang. Migrating Web Application To Cloud with Microservice Architecture.
- [36] Holger Knoche. Sustaining Runtime Performance while Incrementally Modernizing Transactional Monolithic Software towards Microservices. ACM, 2016
- [37] Mazedur Rahman, Jerry Gao. A Reusable Automated Acceptance Testing Architecture for Microservices in Behavior-Driven Development. IEEE. 2015
- [38] Alexandr Krylovskiy\*, Marco Jahn\*, Edoardo Patti. Designing a Smart City Internet of Things Platform with Microservice Architecture. IEEE. 2015
- [39] Hui Kang, Michael Le, Shu Tao. Container and Microservice Driven Design for Cloud Infrastructure DevOps. IEEE. 2016
- [40] Gabor Kecskemeti, Attila Csaba Marosi and Attila Kertesz. Microservices validation: Mjólnir platform case study. IEEE. 2015
- [41] Joao Rufino, Muhammad Alam, Joaquim Ferreira, Abdur Rehman. Orchestration of Containerized Microservices for IIoT using Docker. IEEE. 2017
- [42] Dong Guo, Wei Wang\*, Guosun Zeng, Zerong Wei. Microservices Architecture based Cloudware Deployment Platform for Service Computing. IEEE. 2016
- [43] Gabor Kecskemeti, Attila Csaba Marosi and Attila Kertesz. The ENTICE Approach to Decompose Monolithic Services into Microservices. IEEE. 2016
- [44] Björn Butzin, Frank Golatowski, Dirk Timmermann. Microservices Approach for the Internet of Things. IEEE. 2016
- [45] Gustavo Sousa, Walter Rudametkin, Laurence Duchien. Automated Setup of Multi-Cloud Environments for Microservices Applications. IEEE. 2016
- [46] Srikanta Patanjali, Benjamin Truninger, Piyush Harshand Thomas Michael Bohnert. A Micro Service based approach for dynamic Rating, Charging & Billing for cloud.
- [47] Yuqiong Sun, Susanta Nanda, Trent Jaeger. Security-as-a-Service for Microservices-Based Cloud Applications. IEEE. 2015
- [48] Tomislav Vresk\* and Igor Čavrak. Architecture of an Interoperable IoT Platform Based on Microservices. 2016

## FACTOR ANALYSIS RESULT REPORT

### APPENDIX A

**Factor Analysis:** Scalability, Reusability, Dependency, Independence

#### A. Principal Component Factor Analysis of the Correlation Matrix

TABLE VIII. UN-ROTATED FACTOR LOADINGS AND COMMUNALITIES

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Scalability	0.109	-0.687	0.507	0.066	-0.263	-0.097	-0.170
Reusability	-0.151	-0.165	-0.495	-0.214	-0.340	0.041	-0.558
Dependency	-0.023	-0.468	0.527	-0.324	-0.273	0.299	0.069
Configure	0.697	0.350	-0.107	-0.075	-0.031	0.338	-0.277
Processes	0.042	0.034	-0.328	-0.558	-0.141	-0.258	0.300
Concurrency	-0.110	-0.218	0.196	0.095	0.729	-0.275	-0.359
Continuous Development	0.561	-0.469	-0.238	0.009	0.058	-0.170	0.392
Maintainability	0.017	-0.626	-0.394	-0.146	-0.241	-0.041	-0.224
Load balancing	0.477	-0.531	-0.041	0.059	0.527	-0.103	0.024
Portability	0.647	0.201	0.087	-0.406	-0.061	-0.373	0.012
Security	-0.165	-0.201	0.418	-0.206	-0.051	0.382	0.263
Modularity	-0.199	0.149	-0.071	0.358	-0.204	-0.248	0.227
Performance	0.345	0.217	0.393	0.241	-0.159	-0.155	-0.000
Reliability	-0.012	-0.077	0.187	0.614	-0.423	-0.340	-0.105

Cost	0.625	0.058	-0.059	0.353	0.027	0.586	-0.037
Availability	0.574	0.201	0.281	-0.130	-0.258	-0.289	-0.157
Independence	0.198	-0.328	-0.542	0.436	-0.156	0.089	0.212
Variance	2.4409	2.0895	1.9048	1.6085	1.4945	1.3187	1.0477
% Var	0.144	0.123	0.112	0.095	0.088	0.078	0.062
<b>Variable</b>	<b>Factor8</b>	<b>Factor9</b>	<b>Factor10</b>	<b>Factor11</b>	<b>Factor12</b>	<b>Factor13</b>	<b>Factor14</b>
Scalability	-0.104	0.243	0.123	-0.036	-0.089	-0.080	0.020
Reusability	-0.060	-0.147	0.343	0.035	0.085	-0.154	0.221
Dependency	0.140	0.318	-0.031	0.246	0.020	-0.067	-0.088
Configure	0.073	0.162	0.143	-0.135	-0.108	0.158	-0.207
Processes	-0.307	0.318	0.247	0.053	0.295	0.217	-0.066
Concurrency	0.006	0.128	0.251	-0.020	0.107	-0.072	-0.144
Continuous Development	-0.264	-0.116	0.060	0.049	-0.260	-0.090	0.151
Maintainability	0.184	-0.310	-0.183	0.084	-0.127	0.265	-0.215
Load balancing	0.194	-0.072	0.003	0.017	0.157	0.212	0.138
Portability	0.134	-0.052	0.116	-0.250	-0.199	-0.213	-0.104
Security	0.127	-0.442	0.292	-0.408	0.157	0.063	0.014
Modularity	0.620	0.121	0.431	0.172	-0.146	0.108	0.042
Performance	-0.259	-0.485	0.186	0.425	0.145	-0.029	-0.173
Reliability	-0.292	0.110	-0.007	-0.343	0.029	0.206	-0.006
Cost	-0.117	0.152	0.141	0.097	0.020	0.065	0.130
Availability	0.350	-0.009	-0.280	0.012	0.301	0.008	0.195
Independence	0.178	0.074	-0.050	-0.116	0.311	-0.312	-0.211
Variance	1.0140	0.9291	0.7360	0.6743	0.5294	0.4405	0.3555
% Var	0.060	0.055	0.043	0.040	0.031	0.026	0.021
<b>Variable</b>	<b>Factor15</b>	<b>Factor16</b>	<b>Factor17</b>	<b>Communality</b>			
Scalability	0.011	0.165	-0.139	1.000			
Reusability	0.059	-0.021	0.054	1.000			
Dependency	0.050	-0.121	0.130	1.000			
Configure	0.021	0.148	0.101	1.000			
Processes	-0.035	0.002	-0.043	1.000			
Concurrency	-0.187	-0.052	0.039	1.000			
Continuous Development	-0.145	0.041	0.117	1.000			
Maintainability	-0.110	-0.057	-0.053	1.000			
Load balancing	0.255	-0.002	0.009	1.000			
Portability	0.091	-0.142	-0.070	1.000			
Security	-0.072	0.009	0.009	1.000			
Modularity	-0.024	0.006	0.000	1.000			
Performance	0.054	0.017	0.001	1.000			
Reliability	0.037	-0.101	0.080	1.000			
Cost	-0.102	-0.145	-0.114	1.000			
Availability	-0.167	0.034	0.020	1.000			
Independence	0.017	0.024	0.001	1.000			
Variance	0.1976	0.1253	0.0938	17.0000			
% Var	0.012	0.007	0.006	1.000			



TABLE IX. FACTOR SCORE COEFFICIENTS

Variable	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
Scalability	0.044	-0.329	0.266	0.041	-0.176	-0.074	-0.163
Reusability	-0.062	-0.079	-0.260	-0.133	-0.228	0.031	-0.532
Dependency	-0.009	-0.224	0.277	-0.201	-0.183	0.227	0.066
Configure	0.286	0.168	-0.056	-0.047	-0.021	0.257	-0.264
Processes	0.017	0.016	-0.172	-0.347	-0.094	-0.196	0.287
Concurrency	-0.045	-0.105	0.103	0.059	0.488	-0.208	-0.343
Continuous Development	0.230	-0.224	-0.125	0.006	0.039	-0.129	0.374
Maintainability	0.007	-0.299	-0.207	-0.091	-0.161	-0.031	-0.214
Load balancing	0.195	-0.254	-0.022	0.037	0.352	-0.078	0.023
Portability	0.265	0.096	0.046	-0.252	-0.041	-0.283	0.011
Security	-0.067	-0.096	0.220	-0.128	-0.034	0.289	0.251
Modularity	-0.081	0.071	-0.038	0.223	-0.136	-0.188	0.217
Performance	0.141	0.104	0.206	0.150	-0.106	-0.118	-0.000
Reliability	-0.005	-0.037	0.098	0.382	-0.283	-0.258	-0.100
Cost	0.256	0.028	-0.031	0.220	0.018	0.444	-0.035
Availability	0.235	0.096	0.147	-0.081	-0.173	-0.219	-0.150
Independence	0.081	-0.157	-0.285	0.271	-0.104	0.067	0.202
Variable	Factor8	Factor9	Factor10	Factor11	Factor12	Factor13	Factor14
Scalability	-0.102	0.262	0.167	-0.054	-0.168	-0.183	0.056
Reusability	-0.059	-0.158	0.466	0.052	0.161	-0.349	0.622
Dependency	0.138	0.342	-0.042	0.365	0.038	-0.152	-0.247
Configure	0.072	0.175	0.194	-0.200	-0.203	0.358	-0.582
Processes	-0.303	0.342	0.336	0.078	0.558	0.493	-0.185
Concurrency	0.006	0.138	0.341	-0.030	0.202	-0.163	-0.406
Continuous Development	-0.260	-0.125	0.081	0.073	-0.491	-0.204	0.425
Maintainability	0.182	-0.334	-0.249	0.125	-0.240	0.601	-0.605
Load balancing	0.191	-0.077	0.004	0.026	0.296	0.482	0.388
Portability	0.132	-0.056	0.157	-0.371	-0.376	-0.483	-0.294
Security	0.125	-0.476	0.397	-0.605	0.297	0.142	0.041
Modularity	0.611	0.130	0.586	0.255	-0.276	0.246	0.119
Performance	-0.255	-0.522	0.253	0.631	0.273	-0.067	-0.485
Reliability	-0.288	0.118	-0.010	-0.508	0.054	0.467	-0.016
Cost	-0.115	0.163	0.192	0.144	0.038	0.147	0.366
Availability	0.346	-0.010	-0.381	0.018	0.569	0.019	0.550
Independence	0.176	0.080	-0.067	-0.171	0.587	-0.707	-0.594

Variable	Factor15	Factor16	Factor17
Scalability	0.056	1.316	-1.480
Reusability	0.300	-0.164	0.572
Dependency	0.254	-0.968	1.383
Configure	0.104	1.179	1.074
Processes	-0.175	0.020	-0.456
Concurrency	-0.945	-0.412	0.418
Continuous Development	-0.732	0.326	1.245
Maintainability	-0.555	-0.454	-0.562
Load balancing	1.289	-0.013	0.099
Portability	0.459	-1.131	-0.749
Security	-0.364	0.075	0.093
Modularity	-0.124	0.050	0.001
Performance	0.275	0.132	0.015
Reliability	0.189	-0.810	0.850
Cost	-0.519	-1.158	-1.215
Availability	-0.847	0.275	0.216
Independence	0.087	0.191	0.007

APPENDIX B

B. Analysis Graph section

Fig. 6 shows us relationship among factors in form of groups. It tells the different relationships in order to positive and negative. This graph allows us to rapidly locate similar observations.

Fig. 7 tells us the co-relationships in two ways among factors horizontally and vertically. It tells us the relationship just between two components.

Fig. 8 shows us relationships among factors in the form of pairs.

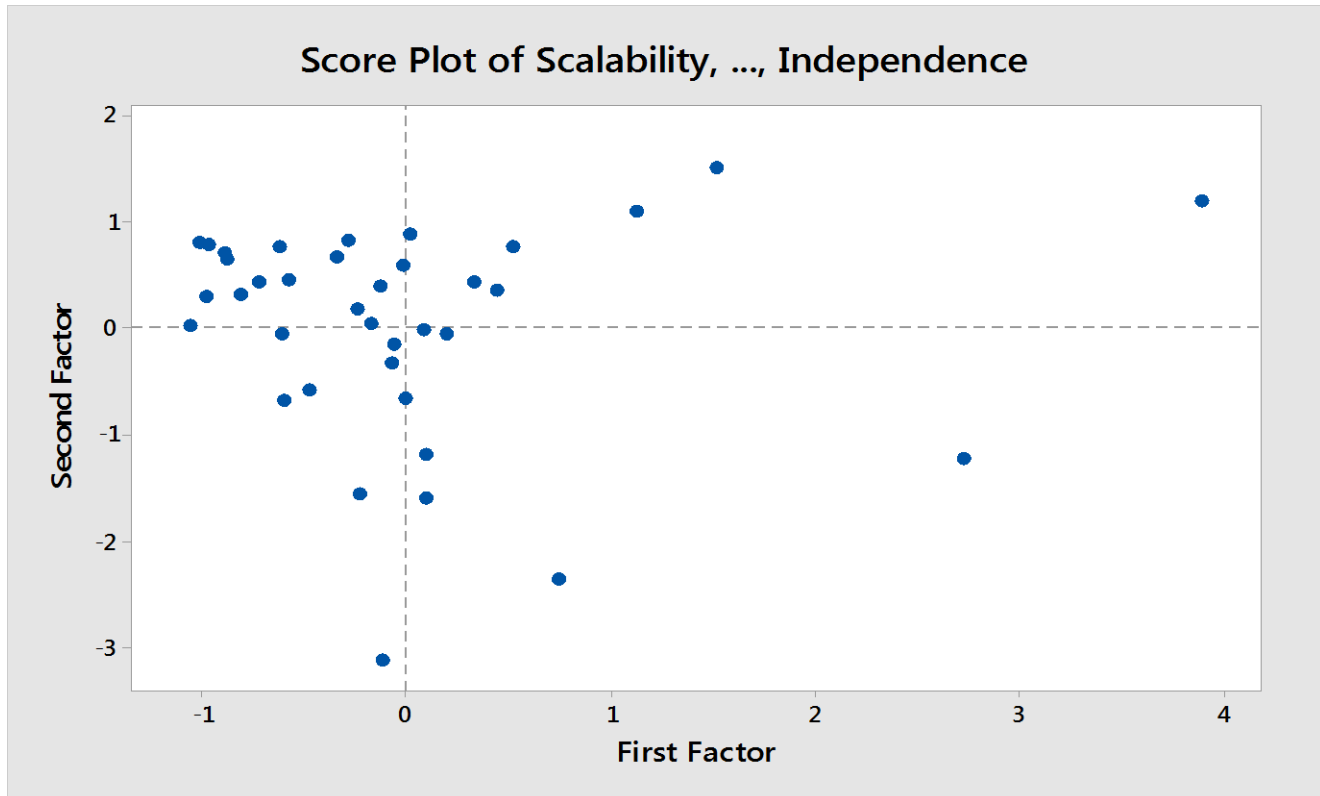


Fig. 6. Score plot.

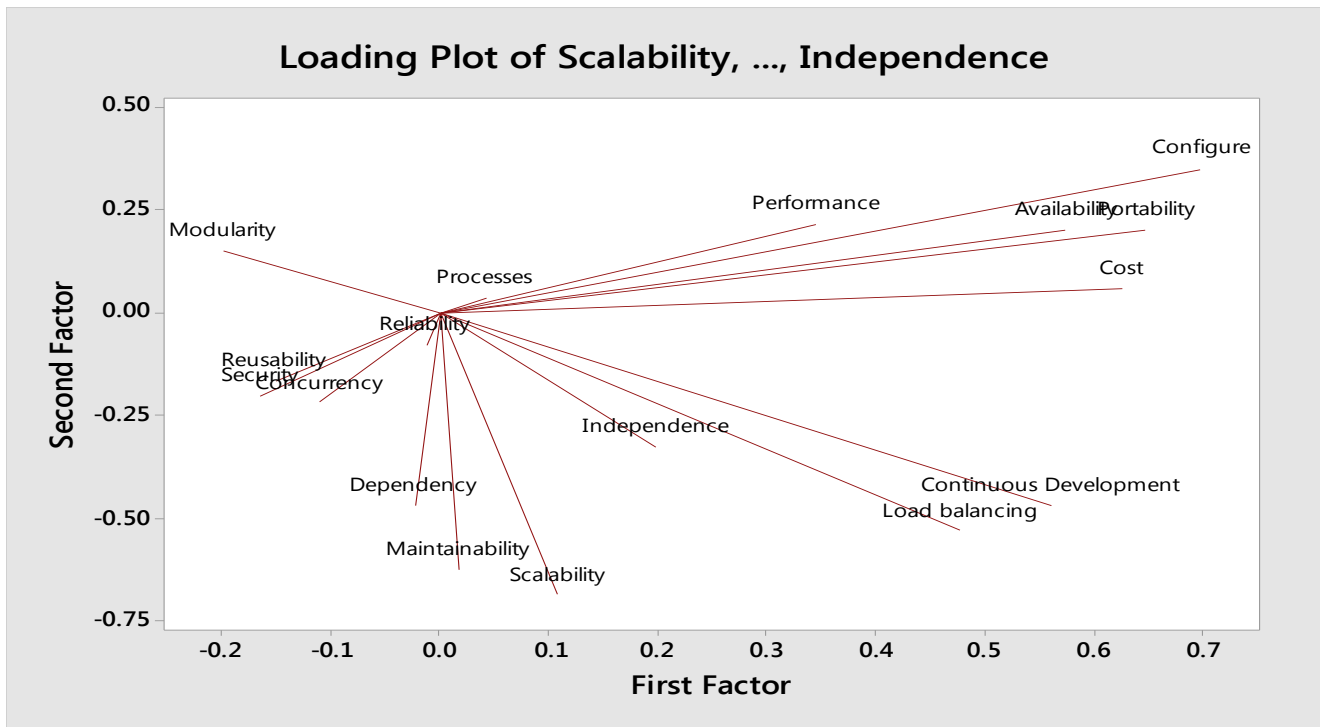


Fig. 7. Loading plot.

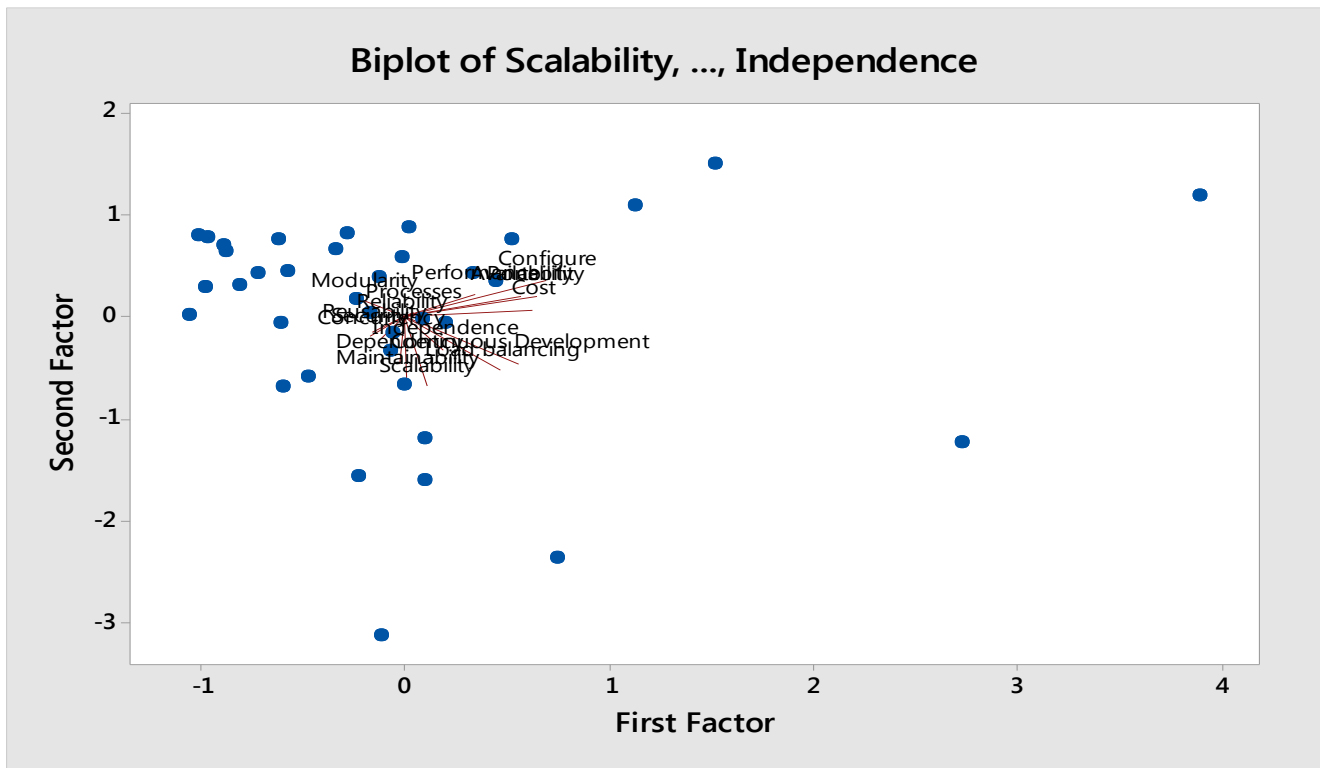


Fig. 8. Biplot.

# DoS/DDoS Detection for E-Healthcare in Internet of Things

Iftikhar ul Sami

Graduate School of Science and Engineering  
PAF-Karachi Institute of Economics and Technology  
Karachi

Maaz Bin Ahmad

Graduate School of Science and Engineering  
PAF-Karachi Institute of Economics and Technology  
Karachi

Muhammad Asif

Lahore LEADS University  
Lahore

Rafi Ullah

Graduate School of Science and Engineering  
PAF-Karachi Institute of Economics and Technology  
Karachi

**Abstract**—Internet of Things (IoT) has emerged as a new horizon in communication age. IoT has provided platform to various emerging technologies and applications for growth. E-Health services have also been integrated and greatly benefitted from IoT. Due to the increased use of computer technology, computer networks have faced serious security challenges and IoT is also facing the same security threats. As IoT has provided platform to other fields, like E-Health, these services are also prone to such threats. Denial of Service (DoS) and Distributed Denial of Service (DDoS) attacks on E-Health servers in IoT would endanger real-time monitoring of patients and also overall reliability of the E-Health services. In this paper, existing solutions to DoS/DDoS attacks in IoT have been reviewed and a reliable solution is presented for securing the servers against these attacks.

**Keywords**—E-Healthcare; DDoS attack; Internet of Things

## I. INTRODUCTION

Internet of Things (IoT) integrates various fields of life ranging from Environmental Monitoring, Infrastructure Monitoring, Energy Management, Traffic Management and Healthcare Management. Today, new advancements are being observed in these fields. Basic motives behind these advancements are to create simplicity in infrastructure and extend reliable solutions to the consumers. Fig. 1 shows the overview of IoT. Due to heterogeneous nature of these technologies, critical applications like traffic system monitoring and healthcare monitoring require special care for transferring their data in timely and secure fashion. Various Metropolitan cities in the world are now benefitting from IoT for real-time traffic monitoring and also integrating various hospitals into IoT for extending patient's health monitoring. Patients are also benefitting from these technologies as they are being continuously monitored without regularly visiting hospitals.

Presently, many hospitals are offering E-Healthcare facilities to their patients and their doctors are continuously engaged in monitoring of these patients. These hospitals are in formal agreement with these patients for extending required services. As these hospitals are geographically located at

specific locations of a country, their patients also reside closer or at moderate distance from these hospitals. If these patients need physical checkups then they can easily visit these hospitals.



Fig. 1. Internet of Things.

E-Healthcare is a modernization of medical services, primarily designed to benefit E-Health consumers and health professionals. Fig. 2 shows E-healthcare scenario. Patients who are using sensors on their bodies for monitoring of health conditions are the consumers of E-Health system and doctors, nurses and allied staff who are responsible for extending medical services are the health professionals. Confidentiality of patient's data is very critical and must be secured to prove the reliability of the system. Servers in E-Health systems are very critical as live monitoring of patients is carried out with the help of these servers. If these servers become unavailable for a moment or longer, health monitoring of patients could be jeopardized. Expansion of networks has witnessed increased network vulnerabilities as well as launching of sophisticated attacks by professional hackers on critical resources. E-Health is critical and most challenging field in which availability of network is of prime importance and if the network becomes

under massive attack like DDoS, lifesaving operations of patients may be very difficult. So, there must be a mechanism to ensure reliability and to prevent/detect such kind of attacks.

The contents of this paper are arranged as: Section II is about literature review. In Section III, proposed solution is presented. Section IV is about conclusion and future work.

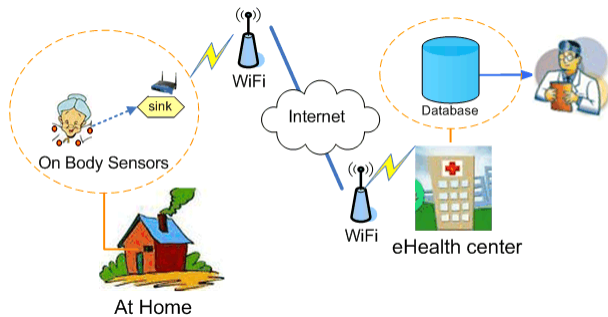


Fig. 2. E-Healthcare System.

## II. LITERATURE REVIEW

Authors of [1] have classified DDoS attacks in two categories as flooding and logical (software) attacks. In flooding attacks, they have highlighted SYN flooding, ICMP attack, UDP flooding and in logical attack they have identified ping of death, teardrop attack, and land attack. They have suggested preventing DDoS attacks at edge routers by installing up-to-date patches and by applying filtering at these edge routers. Firewalls can be efficiently utilized by denying protocols, IP addresses and ports to counter the DDoS attacks. They have also highlighted DDoS detection techniques such that signature based detection and anomaly based detection. In this paper, they have suggested countermeasures like using load balancer, Fault tolerance and Quality of Service techniques to counter the threat. Authors of [2] have presented a survey in which they have highlighted various types of DoS/DDoS attacks on network in which they identified UDP flood, ICMP/PING flood/, SYN flood, ping of death. They have also identified the scenarios in which DoS attack can be launched such as Jamming, kill command attack and de-synchronization attacks. In this paper they didn't present any method to prevent Dos/DDoS attack on the network. Authors of [3] have discussed various attacks like Jamming and it can be avoided by using cryptographic techniques (Attribute based and fuzzy attribute based). They have suggested spread spectrum, priority message and cycle duty to counter the jamming and eavesdropping attacks. Collision attack on network can be avoided by applying error correcting codes. Hello flooding attack is used to create confusion in the network. Authors of [4] have identified various defense methods against DDoS like filtering the attack packets, single/multi-source attacks and application of IDS systems. They have suggested adaptive defense mechanism that can adaptively adjust itself according to the attack severity. In adaptive approach they have focused on the value of traffic rate at specific time which obviously is very high at the time of

attack. Authors of [5] have analyzed attack pattern on application layer based on entropy of HTTP GET request per source IP address by applying support vector machine classifier. Authors of [6] have proposed a detection scheme based on Information theory in which they have used user browsing behavior. On the basis of entropy, suspicious requests are identified. They have suggested rate limiter to downgrade services to malicious users. Authors of [7] have explored the scope of DDoS flooding attack in various situations and explored various countermeasures according to the situation in which one of the countermeasure is packet dropping, based on the level of congestion.

Authors of [8] have analyzed security measures in collaborative environment and identified various tools and surveys the existing traceback mechanisms to identify the real attacker. Authors of [9] have suggested network architecture and algorithms for countering DDoS attack on IoT server. They introduced a router throttle technique by proposing-leaky bucket rate at server that is under stress. In their proposed solution, they used level-k max-min fairness technique for allocating server capacity among routers.

In above referenced papers, researchers have tried different techniques to counter the attacks as there is no specific solution for diverse network situations. With the passage of time, nature and severity of attack is also changing which also needs state of the art techniques to cope with this problem. Some researchers have also tried machine learning approaches to learn the different models for detection of attack which reflects that the use of conventional network analysis techniques alone are not sufficient and it requires various other integrated efforts for securing network resources against these attacks. The basic scenario of DDoS attack is presented in the Fig. 3.

## III. PROPOSED METHODOOGY

In our proposed solution we have taken few assumptions such that Server has some normal buffer utilization i.e. 70% or 75% under normal operations. But when its utilization increases from normal range to maximum, it is suspected that server is under DoS attack. Under such situations, an adaptive measures like selective packet dropping methodology to be taken by server to escape from this attack.

In the proposed solution, we have used packet buffer utilization rate of a server and the TTL value of arriving packet and matched both values with pre-determined values for analyzing attack pattern.

As we know that when packet is in the path, its mutable fields are changing at various routers and its TTL value is also decremented at each router. So it is heuristics that packet having less Time To Live value is coming from far distance in the network which indicates that such packets are coming from geographically far distance. These packets might be generated by bots or by some subnets using various DoS attack techniques. Following algorithm have been suggested to check the DoS attack

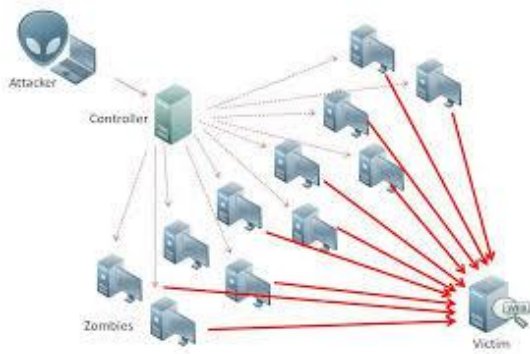


Fig. 3. DDoS Attack.

Algorithm:

1. Read server buffer utilization
2. If buffer utilization > Threshold Value  
// the server is suspected to under DoS attack
3. If (packet TTLvalue < min Value)  
Discard those packets
4. else  
go to step after some specific time

We can also consider geographical locations of nodes. As we know that coordinators (Controller Body Area Network) must be predefined at some geographical locations. By using this information, we can calculate the number of routers (estimated) which may come in the path of a packet to detect DoS attack. As legitimate nodes have some predefined time to send data to a server, and attackers nodes might take longer than normal calculated time.

$$MTU = 1500 \text{ bytes (1460 bytes payload + 40 bytes header)}$$

$$\text{Data rate of link} = d\_rate$$

At coordinator side "A is coordinator"

$$\text{Time to get data on link by node "A"} = TG_A$$

$$\text{Packet size in bits} = Psize$$

$$\text{Data rate in bits} = Dr$$

$$TG_A = Psize / Dr \quad (1)$$

$$\text{Speed or Rate at which data is sent to server} = V$$

$$\text{Time taken by packet to reach server} = T\_travel$$

$$T\_travel = total\_dist / V$$

$$\text{Time to get packet from link by server (Receiver)} = TG_s$$

Note that  $TG_s = TG_A$

$$\text{Total time taken} = Total\_time$$

$$Total\_time = TGA + T\_travel + TG_s \quad (2)$$

Using this Total\_time we can detect DoS attack. By using heuristics, that our actual nodes must have some deterministic time to reach server. We will discard all other packet whose Total\_time is greater than some threshold value.

Algorithm:

1. Calculate Total\_time for incoming packets
2. if Total\_time >= Threshold Value  
// the server is suspected under attack  
discard those packets
3. else  
accept packets

Our proposed solution is based on the assumption that all communicating nodes are present in specific geographical locations (except few, who are temporarily out from the actual zone due to some reasons) and their most likely routes are calculated on pre-commissioning trials. Total packet travel time for a packet at different hours of a day is also taken and threshold rate is calculated based on these packet travel times. Buffer utilization of a server is also calculated based on pre-commissioning trials of the system.

#### IV. CONCLUSION

In our proposed solution, we have focused on enhancing server abilities to observe attack pattern and take adaptive measures to handle DoS/DDoS attack. We have also tried not to over burden the edge/path routers for combing these attacks. In our proposed solution, legitimate packets can also be identified by their pre-calculated TTL values and attack packets can easily be identified and dropped before attack reaches its peak point. In this way normal traffic of legitimate users also possible and attack packets filtering in real time is also carried out.

#### V. FUTURE WORK

In our future work, we will apply machine learning approach to enhance the abilities of our system to cope with DDoS attacks. Daily log of our system will be utilized for training of our model and based on training examples, attack behavior could be timely observed and adaptive defense mechanism can be effectively utilized.

There is a problem in our proposed solution. If attacker is from same distance as our legitimate users then our algorithm will detect attack here. For such scenario we proposed multi-layer check. We set some range in which our proposed solution will use some existing methods.

#### REFERENCES

- [1] A. Srivastava, B.B. Gupta, A. Tyagi, Anupama Sharma and Anupama Mishra, "A Recent Survey on DDoS Attacks and Defense Mechanisms," *Advances in Parallel Distributed Computing*, pp. 570-580, 2001.
- [2] Krushang Sonar and Hardik Upadhyay, "A Survey: DDOS Attack on Internet of Things," *International Journal of Engineering Research and Development*, pp.58-63. 2014.
- [3] Masdari, Tahmineh Haddadi Bonab and Mohammad, "Security attacks in wireless body area networks: challenges and issues," *Academie Royale Des Sciences D Outre-Mer Bulletin Des Seances*, pp. 100-107, 2015.
- [4] Muhai Li and Ming Li, "An Adaptive Approach for Defending against DDoS Attacks," *Mathematical Problems in Engineering*, 2010.
- [5] Tongguang Ni, Xiaoqing Gu, Hongyuan Wang and Yu Li, "Real-Time Detection of Application-Layer DDoS Attack Using Time Series Analysis," *Journal of Control Science and Engineering*, 2013.
- [6] S. Renuka Devi and P. Yogesh, "Detection of Application Layer DDoS Attacks using Information Theory based matrices," *Computer Science and Information Technology (CS & IT)*, pp. 217-223, 2012.

- [7] M. Young, "A Survey of Defense Mechanisms against Distributed Denial of Service (DDoS) Flooding Attacks," *IEEE Communications Surveys & Tutorials*, pp. 2046-2069, 2013
- [8] Arun, Raj Kumar P., and S. Selvakumar. "Distributed denial-of-service (ddos) threat in collaborative environment-a survey on ddos attack tools and traceback mechanisms." In *Advance Computing Conference, 2009. IACC 2009. IEEE International*, pp. 1275-1280, 2009.
- [9] Yau, David KY, John Lui, Feng Liang, and Yeung Yam. "Defending against distributed denial-of-service attacks with max-min fair server-centric router throttles." *IEEE/ACM Transactions on Networking (TON)* 13, no. 1 (2005), pp. 29-42.

# Truncated Patch Antenna on Jute Textile for Wireless Power Transmission at 2.45 GHz

Kais Zeouga, Lotfi Osman, Ali Gharsallah  
Department of Physics, UR "CSEHF" 13ES37  
Faculty of Sciences of Tunis, University of Tunis  
El Manar, Tunis 2092, Tunisia

Bhaskar Gupta  
Department of Electronics and Telecommunication  
Engineering, Jadavpur University  
Kolkata - 700 032, India

**Abstract**—Jute textile is made from natural fibres and is known for its strength and durability. To determine if jute could be used as a substrate for microstrip antennas, its electromagnetic characteristics (permittivity and loss tangent) are measured in the band of 1 GHz to 5 GHz. The obtained data are used to compare the performances of a simple rectangular patch antenna resonating at 2.45 GHz on jute with others using different textiles as a substrate. Comparing the simulation results gives an idea of using jute as a substrate for microstrip antennas. In the second part of this paper, a truncated patch antenna on jute is studied to be used for wireless power transmission at 2.45 GHz. The antenna was simulated and then fabricated. The measured reflection shows a shift in the resonance frequency compared to the simulated one. The frequency shift is explained, and a solution is proposed to correct it; a second antenna was fabricated and measured.

**Keywords**—Jute textile; permittivity measurement; loss tangent measurement; patch antenna, truncated patch antenna; frequency shift; wireless power transmission

## I. INTRODUCTION

Microwave applications are used in different areas and technology advancement gives us the possibility to make new and interesting devices. Many of the applications work in the proximity of the human body to enhance the quality of life [1], [2]. In [3], the authors made a taxonomy of electronic applications on the human body: Body sensor network (BSN), body area network (BAN), wireless body area network (WBAN) [4]. They are used for medical, military and emergency applications, wireless power transmission (WPT) and radio frequency identification (RFID). The devices could be implanted directly into the human body or could be wearable using textile as substrate. In the last years, many antennas have used textile materials as a substrate such as cotton [5], denim [6] and polyester [7]. Every material has its own characteristics which affect the performance of the antennas when they are used as substrate. There are textiles made of natural fibers (cotton, wool) and others made of synthetic materials (polyester, plastic, rubber) [8], [9]. In recent years, using textile as substrate has been used in many microwave applications [10]. In this study, the characteristics of jute are measured in the frequency band [1, 5] GHz. This textile is interesting because it's made from natural fibers, so it's environmentally friendly. It's also known for its strength and durability. After measuring the characteristics of jute, its permittivity and tangential losses were compared to ones of other textiles at the frequency 2.45 GHz, the different values

are used to simulate a rectangular patch antenna resonating at the same frequency. The comparative study is made to define if jute could be used as a substrate for microstrip antennas. Jute is used for the fabrication of clothes and carrying bags that are used for product transport. Therefore, we thought about using jute as substrate to make applications such as WPT or RFID to facilitate the management of storage and transportation of products. In general, textile antennas are related to mobility where a circular polarization is preferred. A truncated patch antenna that easily gives a circular polarization is studied using jute as substrate. The antenna was simulated and optimized to resonate at 2.45 GHz. The measurement of the fabricated antenna shows a shift in the resonance frequency. This frequency shift is explained, and a correction method is then proposed.

## II. MEASUREMENT OF JUTE CHARACTERISTICS

For the measurement of dielectric characteristics, there are two main methods: the non-resonant methods [11] and the resonant methods [12]. The non-resonant methods are based on reflection or reflection/transmission. In the first case (reflection), the permittivity value is given by the information extracted of the reflected electromagnetic wave from free space to the sample. In the second case (reflection/transmission), we use both information of reflection from the substrate and transmission through it to extract the proprieties of the dielectric. The resonance methods are based on the principle that the resonant frequency and the quality factor (Q-factor) of a dielectric resonator with specified dimensions are determined by its permittivity and permeability. Using the resonance perturbation method, the complex permittivity of the measured sample can be extracted. In fact, using a rectangular or circular resonator, its resonant frequency and Q-factor are calculated. Then, the sample is introduced inside the resonator, the resonance frequency and the Q-factor are measured. The frequency shift and the change of the Q-factor give the complex permittivity of the sample using resonance-perturbation theory. The advantage of this method is the narrow band of the cavity resonator, so it is sensitive to perturbations. It also handles high fields which give a substantial variation of the signal even for a small change in permittivity. A new resonance method proposed by S. Sankaralingam is based on a rectangular patch antenna [13], [14]. This method consists of simulating a rectangular patch antenna using an approximate value of permittivity. The antenna is then fabricated, and the resonance frequency is



measured. The value of this frequency gives the effective permittivity of the substrate using the theory of patch antennas [15].

Jute is a textile made of natural fibers [16]. Fig. 1 shows a photo of a sample of jute; it's filled with air, so a low value of permittivity is expected. The thickness of the sample is 0.5 mm.

For the measurement of permittivity and loss tangent, E4991A RF impedance/material analyzer from Agilent Technologies and the dielectric fixture from Novocontrol Technologies are used to measure the permittivity and loss tangent of the sample. Fig. 2 and 3 show the measured permittivity and loss tangent of the jute between 1 and 5 GHz, respectively.



Fig. 1. Sample of a jute textile.

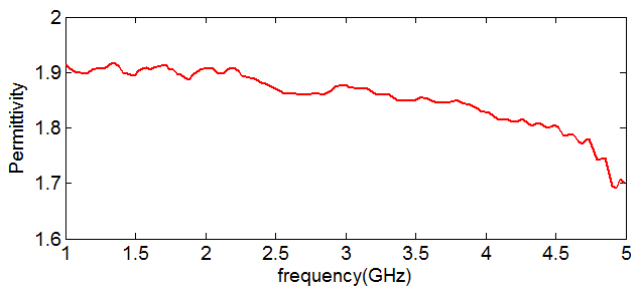


Fig. 2. Measure of the permittivity of the jute versus frequency.

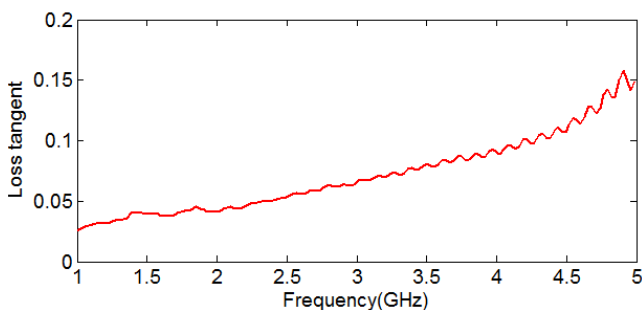


Fig. 3. Measure of the loss tangent of the jute versus frequency.

The permittivity doesn't vary much in the band of measurement, it's between 1.69 and 1.91. This low value of permittivity is expected because jute is filled with air as shown in Fig. 1. However, the loss tangent increases with the frequency, it varies between 0.025 and 0.157. The values of loss tangent are relatively high compared to other substrates or

even other textiles. For this reason, a comparative study based on simulation is made to evaluate the performance of an antenna on jute compared to other textiles.

### III. COMPARATIVE STUDY OF A SIMPLE PATCH ANTENNA USING JUTE AND OTHER TEXTILES AS SUBSTRATE

At the frequency  $f_r = 2.45$  GHz, the jute's permittivity is 1.87. In [17], a width "w" that gives a resonant patch at the frequency  $f_r$  is given by (1).

$$w = \frac{1}{2f_r\sqrt{\mu_0\epsilon_0}} \sqrt{\frac{2}{\epsilon_r+1}} = \frac{c}{2f_r} \sqrt{\frac{2}{\epsilon_r+1}} \quad (1)$$

The finite dimensions of the patch will introduce a phenomenon called fringing effect. It is as if the patch was done on a substrate with a different permittivity called effective permittivity  $\epsilon_{reff}$ . In the case  $w/h > 1$  where  $h$  is the height of the substrate,  $\epsilon_{reff}$  is given by (2) [18].

$$\epsilon_{reff} = \frac{\epsilon_r+1}{2} + \frac{\epsilon_r-1}{2} \left[ 1 + 12 \frac{h}{w} \right] \quad (2)$$

The fringing effect makes the electrical dimensions of the patch different from its physical ones. The electrical length is equal to the physical length adding  $\Delta L$  in each side. Reference [19] gives an approximate value of  $\Delta L$  described by (3).

$$\frac{\Delta L}{h} = 0.412 \frac{(\epsilon_{reff}+0.3)(\frac{w}{h}+0.264)}{(\epsilon_{reff}-2.58)(\frac{w}{h}+0.8)} \quad (3)$$

So,  $L_{eff} = L + 2\Delta L$ , we can deduce the actual length of the patch using (4).

$$L = \frac{1}{2f_r\sqrt{\epsilon_{reff}}\sqrt{\mu_0\epsilon_0}} - 2\Delta L \quad (4)$$

In this section, a simple patch antenna resonating at 2.45 GHz is simulated using different textile substrates to compare jute with natural textiles (cotton, Denim) and with synthetic textiles (Polyester, Cordura). Table I shows the characteristics (permittivity and loss tangent) of each textile at 2.45 GHz [20]. The width (w) and length (L) of the patch depend on the permittivity value of the substrate as are given by (1) and (4).

TABLE I. ELECTROMAGNETIC CHARACTERISTICS OF DIFFERENT TEXTILE MATERIALS AT 2.45 GHz

Textile	Permittivity	Tanδ
Jute	1.87	0.052
Cotton	1.61	0.0138
Denim	1.59	0.031
Polyester	1.44	0.01
Cordura	1.9	0.0098

Different antennas are simulated and optimized using CST Microwave Studio. The same height of 1.5 mm is for all the textiles, so the comparison of the performance is based only on electromagnetic characteristics. Table II exhibits the performance of a patch antenna made on jute compared to others made on different textiles resonating at the same frequency 2.45 GHz. The effect of the permittivity and  $\tan \delta$  values is shown by the given results. When permittivity increases, the patch surface is reduced as explained in equation (1), but the radiation efficiency  $Q_r$  decreases. When the  $\tan \delta$  increases, the gain is reduced, and the bandwidth is enlarged. Table II shows that the performances of antennas made of

synthetic textiles are much better than those made of natural textiles in terms of gain and efficiency. This is due to the low value of their  $\tan \delta$ . Comparing the performance of the patch antenna made of jute with that made of cotton and denim, we notice that the jute patch is the most compact because of its higher permittivity at 2.45 GHz; it has the largest bandwidth and its angular width at -3 dB is slightly larger, but its gain is lower due to the high value of  $\tan \delta$ . This low value of gain

could be a problem for some applications that need high power or deal with long distances. The results given in Table II are correlated to the data given in Table I. At the resonance frequency 2.45 GHz, all the antennas are well matched and the bandwidth (BW) increases with the permittivity and the loss tangent values. This can be explained by the relation  $BW \approx \frac{1}{Q}$  where Q is the quality factor [21].

TABLE II. PERFORMANCE COMPARISON OF ANTENNAS USING DIFFERENT TEXTILE MATERIALS AS SUBSTRATE

Textile	S <sub>11</sub> (dB)	Band Width (MHz)	Angular width (°) (φ=90°)	Angular width (°) (φ=0°)	Directivity (dBi)	Gain (dB)	Efficiency (%)	Patch surface (cm <sup>2</sup> )
Jute	-24.46	117	73.6	77.9	7.9	1.98	25	18.18
Cotton	-28.33	54	66.4	76.1	8.41	6.05	58	20.7
Denim	-32.19	87	68.7	75.3	8.36	4.26	38	21.25
Polyester	-26.8	55	65.1	73.1	8.73	6.9	65	23.66
Cordura	-29.6	52	71	77.8	8	6.12	65	18.18

IV. STUDY OF A TRUNCATED PATCH ANTENNA ON JUTE

For wireless power transmission, a circular polarization is preferred to reduce losses due to polarization mismatch. The truncated patch antenna is a simple structure that easily gives a circular polarization [22], [23]. The dimension and the direction of the truncations are responsible for axial ratio variation and then the polarization. CST MWS software is used for simulation and optimization. The antenna is simulated using one layer of jute and the gain was very low because of the high value of  $\tan \delta$  and the low thickness of the substrate. The low thickness of jute does not allow a constructive superposition of the wave 2.45 GHz. For this reason, the antenna is simulated using different numbers of layers. Table III shows the gain of the simulated antenna for each number of layers. The gain increases with the number of layers. Using six layers (3mm) gives good simulated gain, but the choice is fixed to use only three layers (1.5mm). This choice is made for comfort reasons, as the structure will be used near the human body (on clothes), it is preferred to reduce the thickness of the antenna as possible. The choice is a compromise between the gain and the comfort. The dimensions (in mm) of the optimized antenna using three layers of jute are represented in Fig. 4.

TABLE III. GAIN OF THE ANTENNA FOR DIFFERENT NUMBERS OF USED JUTE LAYERS

Number of layers	Height (mm)	Gain (dB)
1	0.5	-2
2	1	0.4
3	1.5	1.89
4	2	2.8
5	2.5	3.2
6	3	4.23

Fig. 5 to 8 show the simulated results of the reflection coefficient, the radiation pattern in E-plan, the radiation pattern in H-plan and the axial ratio, respectively. At 2.45 GHz S<sub>11</sub> = -30 dB, the antenna is well adapted. From Fig. 6, we can notice

a small deviation in the radiation pattern; the main lobe magnitude is at 5°, the angular width is 75°. For many applications, an antenna is considered circularly polarized when its axial ratio is less than 5 dB. Fig. 8 shows the axial ration of the truncated patch antenna on jute, the antenna is circularly polarized from Theta= -110° to Theta= 62°.

We can say that the antenna is circularly polarized when there is a maximum of radiation. The maximum gain is 1.89 dB. The characteristics of the antenna are suitable for WPT.

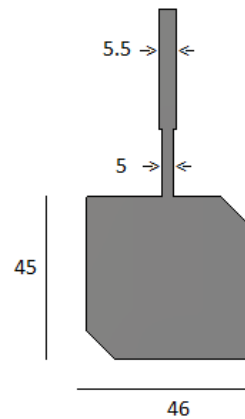


Fig. 4. Dimensions of the truncated patch on jute (in mm).

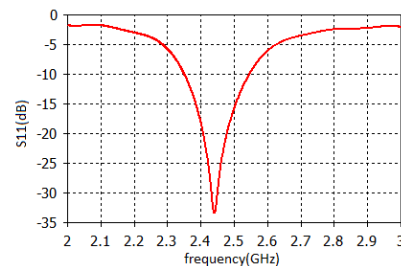


Fig. 5. Reflection coefficient of the simulated patch antenna on jute.

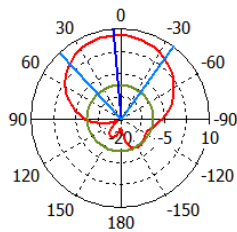


Fig. 6. Radiation pattern E-plan at 2.45 GHz.

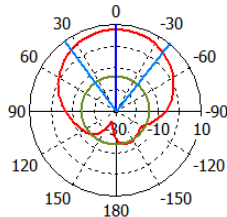


Fig. 7. Radiation pattern in H-plan at 2.45 GHz.

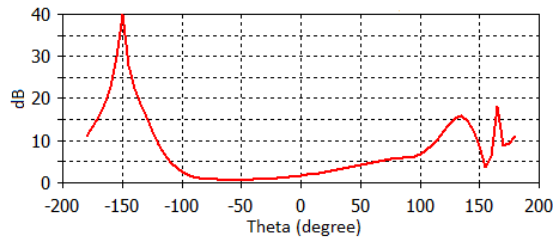


Fig. 8. Axial ratio of the truncated patch on jute at 2.45.

The antenna was fabricated and then measured. When measuring the reflection coefficient, a shift in the resonance frequency is noticed. The measured resonance frequency is 2.8 GHz instead of 2.44 GHz in the simulation. This frequency shift could be explained using three layers. The use of multiple layers induces an air gap between every two layers that reduces the global permittivity of the substrate. In the next section, a method is proposed to find the global permittivity of three layers of jute and correct the frequency shift.

#### V. FREQUENCY SHIFT CORRECTION FOR TRUNCATED PATCH ON THREE LAYERS OF JUTE

A frequency shift is noticed between the simulated and measured results that can be explained by the air gap between the layers. This air gap will reduce the permittivity of the global substrate, the three layers of jute will be considered as a substrate having a lower permittivity than measured for jute. To find the global permittivity, a solution is then proposed. This is to simulate the antenna by decreasing the permittivity value of the substrate until getting a resonating frequency at 2.8 GHz. After optimizations, we got a resonance at 2.81 GHz for a permittivity value  $\epsilon_1=1.31$ . This value will be considered as the permittivity of three layers of jute. Using this value, the antenna is optimized to resonate at 2.45 GHz. Fig. 9 shows the dimensions in mm of the antenna. The antenna was then fabricated and measured. Fig. 10 shows the fabricated antenna under measure. Fig. 11 shows the simulated and measured reflection coefficient. A good agreement is then observed. The proposed method of correction is efficient and the effective permittivity of three layers of jute is 1.31 at 2.45 GHz.

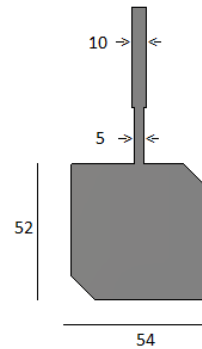


Fig. 9. Dimensions (in mm) of a truncated patch on jute with corrected value of permittivity.

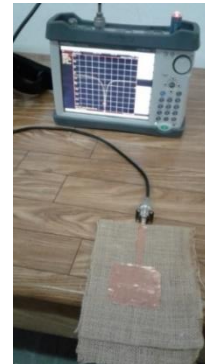


Fig. 10. The fabricated antenna under test.

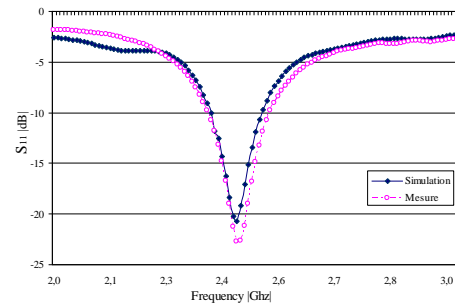


Fig. 11. Reflection coefficient of the truncated patch on jute.

The permittivity of jute is measured from 1 to 5 GHz, but these values could be considered only for one layer. Using more layers will reduce the global permittivity and then will induce errors in calculations and simulations. This variation of permittivity concerns all textiles and must be considered when using more than one layer.

#### VI. CONCLUSION

In this paper, the electromagnetic characteristics of jute are measured to find if it's suitable for WPT and other microwave applications. The measured results were compared to those of other textiles at the frequency 2.45 GHz. The loss tangent of jute is relatively high. Therefore, a simple patch antenna made on natural and synthetic textiles including jute were simulated to compare their performances. Antennas made on synthetic textiles have better gain due to their low  $\tan \delta$ . Comparing performances of jute and other textiles made of natural fibres, jute has a compact size and a large bandwidth, but its gain is

very low. Antennas with circular polarization are very important for WPT and wearable microwave applications. A truncated patch antenna resonating at 2.45 GHz using three layers of jute is designed. The simulated antenna presents the required characteristics for WPT. The measurement of the reflection coefficient shows a shift in the resonance frequency compared to the simulated one due to the use of multiple layers. In fact, using more than one layer, induce more air and then reduce the global permittivity value. To correct this shift, the value of permittivity was changed on the simulated antenna until getting a resonance in the same frequency as the measured one. The obtained permittivity is considered as the global permittivity of three layers of jute. Using this value, we designed another antenna. The simulated and measured results have a good agreement proving the efficiency of the proposed method of correction. Based on the obtained results, we can confirm that jute is a good substrate for WPT and wearable microwave applications.

#### ACKNOWLEDGMENT

This research work is supported by a project under the scheme of Indo-Tunisian Program of Cooperation in Science and Technology sanctioned by DST (India) and MHESR (Tunisia) bearing Sanction Order No. INT/Tunisia/P-02/2012 dated 01/03/2013.

#### REFERENCES

- [1] M. E. Jalil, M. K. A. Rahim, N. J. Ramly, N. A. Samsuri, K. Kamardin, M. A. Abdullah, and H. A. Majid, "Planar Textile Antenna for Body Centric Wireless Communication System," *Progress In Electromagnetics Research C*, Vol. 54, 29–40, 2014.
- [2] H. Nawazi and M. A. B. Abbasi, "On-body Textile Antenna Design and Development for Body-centric Wireless Communication Systems," *12th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, 13-17 Jan. 2015.
- [3] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V. C. M. Leung, "Body Area Networks: A Survey," © Springer Science+Business Media, LLC 2010.
- [4] Fatemeh Rismanian Yazdi, Mehdi Hosseinzadeh and Sam Jabbehdari, "A Review of State-of-the-Art on Wireless Body Area Networks" *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8(11), 2017.  
<http://dx.doi.org/10.14569/IJACSA.2017.081154>
- [5] S. Sankaralingam and B. Gupta, "Development of Textile Antennas for Body Wearable applications and investigations on their performance under bent conditions," *Progress In Electromagnetics Research B*, Vol. 22, 53–71, 2010.
- [6] M. Grilo and F. S. Corra, "Parametric Study of Rectangular Patch Antenna Using Denim Textile Material," *International Microwave & Optoelectronics Conference (IMOC)*, 4-7 Aug. 2013.
- [7] E. G. Lim, Z. Wang, M. Leach, R. Zhou, K. L. Man, and N. Zhang, "Compact Size of Textile Wearable Antenna," *Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol II, IMECS 2014, March 12 - 14, 2014, Hong Kong.*
- [8] A. Priya, A. Kumar, and B. Chauhan, "A Review of Textile and Cloth Fabric Wearable Antennas," *International Journal of Computer Applications (0975 – 8887)*, Vol. 116 – No. 17, April 2015.
- [9] Saadat Hanif Dar, Jameel Ahmed and Muhammad Raees, "Characterizations of Flexible Wearable Antenna based on Rubber Substrate" *International Journal of Advanced Computer Science and Applications (IJACSA)*, 7(11), 2016.  
<http://dx.doi.org/10.14569/IJACSA.2016.071124>
- [10] M. K. Elbasheer, A. Abuelnuor, M. K. A. Rahim, and M. E. Ali, "Conducting Materials Effect on UWB Wearable Textile Antenna," *Proceedings of the World Congress on Engineering 2014 Vol I, WCE 2014, July 2 - 4, 2014, London, U.K.*
- [11] R. Moro, S. Agneessens, H. Rogier, A. Dierck, and M. Bozzi, "Textile Microwave Components in Substrate Integrated Waveguide Technology," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 63, No. 2, February 2015.
- [12] M. T. Jilani, M. Z. Rehman, A. M. Khan, M. T. Khan, and S. M. Ali, "A Brief Review of Measuring Techniques for Characterization of Dielectric Materials," *International Journal of Information Technology and Electrical Engineering*, vol. 1, issue 1, December 2012.
- [13] J. Sheen, "Study of microwave dielectric properties measurements by various resonance techniques," *Elsevier, Measurement 37 (2005) 123–130.*
- [14] S. Sankaralingam and B. Gupta, "Determination of Dielectric Constant of Fabric Materials and Their Use as Substrates for Design and Development of Antennas for Wearable Applications," *IEEE Transactions on Instrumentation and Measurement*, Vol. 59, No. 12, December 2010.
- [15] S. Sankaralingama, S. Dhar, B. Gupta, L. Osman, K. Zeouga, and A. Gharsallah, "Performance of Electro-Textile Wearable Circular Patch Antennas in the Vicinity of Human Body at 2.45 GHz," *Journal of Procedia Engineering by Elsevier*, Vol. 64, pp. 179-184, 2013.
- [16] C. A. Constantine, A. Balanis, "Antenna Theory: Analysis and Design", 2nd edition, John Wiley & Sons, INC.
- [17] L. Ammayappan, L. K. Nayak, D. P. Ray, S. Das, and A. K. Roy, "Functional Finishing of Jute Textiles—An Overview in India," *Journal of Natural Fibers*, 10:390–413, Copyright © Taylor & Francis Group, LLC, 2013.
- [18] I. J. Bahl, P. Bhartia, "Microstrip Antennas," Artech House, Dedham, MA, 1980.
- [19] E. O. Hammerstad, "Equations for Microstrip Circuit Design," *Proc. Fifth European Microwave Conference*, pp. 268–272, 1-4 Sept. 1975.
- [20] R. Salvado, C. Loss, R. Gonçalves, and P. Pinho, "Textile Materials for the Design of Wearable Antennas: A Survey," *Sensors* 2012, 12, pp. 15841–15857.
- [21] M. Capek, L. Jelinek, and P. Hazdra, "On the Functional Relation between Quality Factor and Fractional Bandwidth," *Journal of LATEX class files*, Vol. 6, No. 1, January 2007.
- [22] A. K. Aswad, F. Abdulrazak, T. A. Rahman "Design and development of high gain wideband circularly polarized patch antenna," *IEEE International RF and microwave conference*, 2008.
- [23] Chithra Liz Palson, Ajeena Elza Sunny, D. D. Krishna, "circularly polarized square patch antenna with improved axial ratio bandwidth," *IEEE Annual India Conference (INDICON)*, 2016.

# A Hybrid Approach for Feature Subset Selection using Ant Colony Optimization and Multi-Classifier Ensemble

Anam Naseer

Department of Computer Science  
and Information Technology  
The University of Poonch Rawalakot  
AJK, Pakistan

Waseem Shahzad

Department of Computer Science  
National University of Computer  
and Emerging Sciences,  
Islamabad, Pakistan

Arslan Ellahi

Department of Computer Science  
MY University, Islamabad  
Pakistan

**Abstract**—An active area of research in data mining and machine learning is dimensionality reduction. Feature subset selection is an effective technique for dimensionality reduction and an essential step in successful data mining applications. It reduces the number of features, removes irrelevant, redundant, or noisy features, and enhances the predictive capability of the classifier. It provides fast and cost-effective predictors and leading to better model comprehensibility. In this paper, we proposed a hybrid approach for feature subset selection. It is a filter based method in which a classifier ensemble is coupled with Ant colony optimization algorithm to enhance the predictive accuracy of filters. Extensive experimentation has been carried out on eleven data sets over four different classifiers. All of the data sets are available publically. We have compared our proposed method with numerous filter and wrapper based methods. Experimental results indicate that our method has remarkable ability to generate subsets with reduced number of features. Along with it, our proposed method attained higher classification accuracy.

**Keywords**—Ant colony optimization; predictive; classifier features selection

## I. INTRODUCTION

Data mining [1] is the method of removing hidden predictive information from very large texts, databases, and web etc. Mining huge data can extrapolate data and information that can decrease the chances of fraud, improve audit reactions to potential business changes, and ensure that risks are managed in a proactive fashion [1].

Feature subset selection is mostly applied to high-dimensional data which contains a number of features. Such a large number of features make training and testing of classification methods much more difficult. Some of these features may not be important whereas some of the important features may be redundant. So feature selection technique detects most discriminating features which decreases the of data. FSS also increases the predictive accuracy by removing redundant and irrelevant features and decreases the computational time by reducing data dimensionality [2].

## A. Feature Subset Selection Process

Fig. 1 demonstrates the general feature subset selection process.

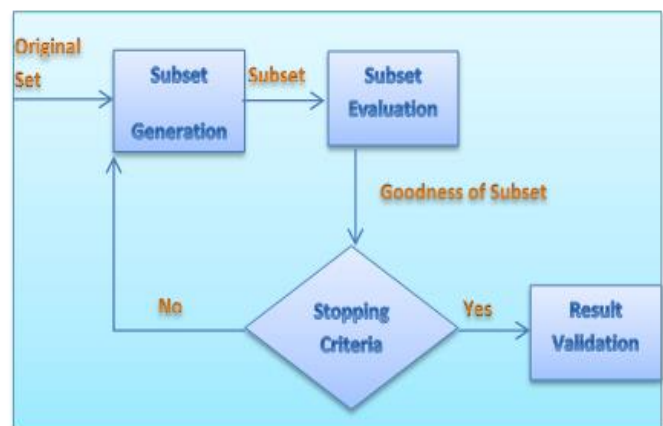


Fig. 1. Feature subset selection process.

## B. Subset Generation

Subset generation is a heuristic search process where search space contains states, each of which specifies a candidate subset for evaluation. Two things must be determined for subset generation, Search starting point and Search strategy [3]. Search starting point can be forward, backward, bi-directional and random. In forwarding selection, thesearch starts with an empty subset and selectively adds those features that are deemed relevant, whereas, in backward elimination, the search starts with full feature subset and selectively discards those features that are useless or irrelevant. In bidirectional selection, search starts with both ends that add and removes features simultaneously and in random search, a feature subset starting point is chosen randomly without any consideration and features are added or deleted as per the requirement. Search may also start with a haphazardly selected feature subset so that it cannot stuck in the local optima [4].

A search strategy must be decided to select the candidate subsets. Different Search strategies have been explored such as exhaustive, heuristic and randomized searching algorithms [5], [6]. The time complexity is exponential for exhaustive search in terms of dimensionality and quadratic for heuristic search. In Random search, complexity can be linear to the number of iterations [5].

### C. Subset Evaluation

An evaluation criterion is used to evaluate each newly generated candidate. Based on the dependency on learning algorithms that will be applied to selected feature set an evaluation criterion is categorized into two groups, one is dependent criteria second one is independent criteria [3].

Wrapper model uses dependent criteria and for feature selection it needs a learning algorithm. It applies that learning algorithm on selected subset and uses its performance to determine best feature subset, whereas filter model use independent criteria. Goodness of feature or its subset is measured with the help of significant features of the training data without linking any learning algorithm. Most common independent criteria are information theoretic measures, dependency measures, distance measures, and consistency measures.

### D. Stopping Criteria

In feature subset selection a stopping criterion governs when feature selection process will stop. Some of the commonly used stopping criteria are as follows:

- Exhaustive search completes.
- A bounds could be used as stopping criteria where a bound can be a specified number. It can be a maximum number of iterations or minimum number of features.
- If a successive addition or removal of any feature does not affect results feature selection process could be stopped.
- If a satisfactorily good subset is selected.

### E. Result Validation

At the end, results are validated by using classification error rate of classifiers as a performance indicator. Experiments are conducted to equate the classification error rate on the full set of the classifier learned on features and that trained on the selected feature subset [7], [8].

Feature subset selection techniques are of two types. Selection based reduction and transformation based reduction. Selection based reduction reduce data using original set of features whereas form new set of features by transforming an original set of features. Proposed approach is based on selection based reduction. Selection based algorithms have two categories, Filters and wrappers. Filter model feature subset evaluation methods are those that perform feature selection using some independent selection criterion, independently of any learning algorithm [9]. The computational cost of filter-based feature subset evaluation methods are less as equated to other methods. Filter based

methods depend on the independent measures that shows the relationships among different features.

Wrapper based feature subset evaluation methods induce learning algorithms during evaluation step to measure the goodness of a selected feature subset based on the algorithm's accuracy so are computationally expensive as compared to filters. In terms of predictive or classification, accuracy wrapper methods are considered superior to filter [10].

The methodology proposed is a hybrid filter based selection method algorithm, where ACO is coupled with the Gain ratio for the first time to cope with biases of other information theoretic measures towards multi-valued attributes. A multi-classifier ensemble is used iteratively for selecting the best subset of different convergence threshold value and also for final subset selection in a novel way. Our proposed approach has used gain ratio as the subset evaluator and an ensemble of classifiers for selecting final best. If the independent measure fails to capture important features, an ensemble of classifiers captures those features. In proposed approach ensemble of classifiers are used iteratively for only selecting a final subset and not used for subset optimization as in wrapper based methods. So proposed algorithm is computationally less expensive as compared to wrapper approaches and yields higher accuracy with many reduced subsets.

The paper is organized as follows. Section II describes some of existing techniques of related to feature subset selection. A detailed description of proposed approach is described in Section III. Section IV presents the results of our experimental studies, methodology, and a comparison with other existing feature selection techniques. Section V provides Conclusion and future work directions.

## II. LITERATURE REVIEW

Literature studied shows that much of the work has already been done on feature subset selection different techniques. The present techniques are grouped into two categories: filters and wrappers on the basis of search strategy and subset evaluation method [9], [10]. Some existing filter and wrapper based approaches are described here:

Feng Tan et al. presented a framework for feature subset selection based on genetic algorithm [11]. The proposed algorithm rank features using entropy and T-statistics as a ranking criterion and select features on the basis of their rank. Top-ranked features are provided as input to GA and then evaluation is done on the basis of fitness function.

Bai Jiang et al. gave hybrid algorithm for feature subset selection [12]. It is composed of two step process. Symmetric Uncertainty of the each individual features is calculated in the first step and features with SU more than the threshold value is selected and features with SU less than the threshold value are discarded. In the second step GA based searching is carried out for the left over features. To assess the quality of the feature subsets Naive Bayes classifier is used by 10 fold cross validation. For subset optimization, Naive Bayes is also used along with symmetric uncertainty [24].

Li-Yeh Chuang et al. proposed hybrid filter-wrapper approach [13] in which an improved binary particle swarm optimization is used as a wrapper feature selection for which information gain is used as filtered model; for the performance evaluation of classification selected gene subsets were used.

Shailendra Kumar Shrivastava proposed a new ensemble technique [14]. The motivation of this approach is to enhance the performance of multiple K-nearest neighbor classifiers. This approach combines multiple K-nearest neighbor classifiers. Each classifier uses a different subset. These subsets are selected through ACO based search procedure. A subset which gets high classification accuracy on a majority of K-Nearest Neighbour classifiers is selected as a final subset.

Md.Monirul Kabir et al. proposed a hybrid technique using ant colony optimization that takes the advantage of both wrapper and filter approaches [15]. Information gain is used as filter approach and neural network as wrapper approach. This research has focused on generating reduced sized subsets. The proposed approach has used a subset size determination scheme that emphasizes not only the selection of a subset of relevant features but also on selecting features of reduced number.

Gang Wang gave a hybrid ensemble method for credit risk assessment problem [16]. In ensemble method, multiple classifiers are used to solve the same problem and also to boost many weak learners. The approach proposed in this paper works through integrating two popular ensemble strategies i.e. bagging and random subspace.

Shunmugapriya Palanisamy et al. gave a hybrid algorithm ABCE [17]. It is the combination of Artificial Bee Colony algorithm with ensemble classifier. This multi ensemble classifier is composed of support vector machine classifier, decision tree classifier, and naïve Bayes classifier. The author used ABC for generating and selecting feature. For evaluating subsets an ensemble made up of Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM) is used.

In 2012 Syed Imran Ali et al. gave a feature subset selection mechanism based on ant colony optimization algorithm and symmetric uncertainty [18]. It is a pure filter based approach which investigated the role of ACO in filter approaches. In this technique, ACO is introduced to generate optimal feature subsets. And symmetric uncertainty is used as an independent statistical measure for subset evaluation. Proposed algorithm selects fewer features and produces comparably higher accuracy.

### III. PROBLEM STATEMENT

Different filter and wrapper techniques and a number of classifier ensemble methodologies for feature selection have been proposed and implemented so far in order to improve the classification accuracy. Filter approaches applied so far mostly used statistical measures to evaluate feature and to measure the goodness of feature subset. Most of the existing techniques have used information gain as a goodness measure. The main limitation of this measure is its biases towards attributes with large number of distinct values. So this drawback should be normalized.

Secondly, most of the existing techniques used learning algorithms or wrapper approach to improve classification accuracy of filter approaches. Some of these approaches have used classifier ensemble to evaluate the fitness of feature subset. These approaches increase classification accuracy but also increase computational complexity. So we will deal with two problems in this thesis: 1) We will compensate drawbacks of information gain. 2) We will try to improve classification accuracy of filter approach using a classifier ensemble without increasing computational complexity.

## IV. PROPOSED SOLUTION

The proposed approach ACO-CE employs Ant colony optimization algorithm as a population based feature subset selection mechanism. Proposed approach is a hybrid filter-based feature selection, where ACO is coupled with the gain ratio for the first time as a filter solution. The gain ratio is used to normalize biases of some of the already used statistical measures towards multi-valued attributes such as information gain and mutual information etc. high split information is penalized using gain ratio an ensemble of the classifier is used iteratively over different convergence threshold values for final subset selection, not for subset optimization.

On each convergence threshold value, some best subsets are selected using gain ratio based fitness function and these subsets are provided to classifier ensemble and on the basis of average mean accuracy one best subset is selected and saved. Then this process is repeated by changing convergence threshold ten times and each time one best subset is selected and saved at the end from ten saved subsets on ten different convergence threshold values one subset with highest average accuracy of classifier ensemble is selected as final subset.

### A. Gain Ratio

The gain ratio is normalized or compensates the biases of information gain towards attributes with a large number of values. It is basically a refinement of information gain. It takes into account split information of every attribute. Large numbers of small partitions in every split are penalized. The gain ratio is defined as:

$$\text{Gain ratio (X)} = \frac{\text{Information gain (X)}}{\text{Split information (X)}} \quad (1)$$

$$\text{Information\_Gain} = H(X) - H(X|Y) \quad (2)$$

Here  $H(X)$  is the entropy of a random variable  $X$  and  $H(X|Y)$  is the conditional entropy of  $X$  given  $Y$ . following are the equations for entropy and conditional entropy of a variable. Where split information is defined as:

$$\text{Split\_information} = \sum_{i=0}^n \frac{n_i}{n} \log \frac{n_i}{n} \quad (3)$$

A feature that will get a high value of information gain and low value of split information will be preferred. Its goal is to maximize information gain and minimize the number of its values.

### B. Ant Colony Optimization

Ant colony optimization (ACO) a population-based probabilistic Meta-Heuristics ACO is based on ants foraging behavior [19]. Foraging behavior of ant is an interesting

phenomenon by which ant colonies find the shortest path between food source and nest through indirect communication called stigmergy. Ants, like many other social insects, communicate with each other by dropping a chemical substance on their path. This chemical substance is called pheromone. It provides a positive feedback mechanism to attract other ants. Those paths which have a higher value of pheromone have a high probability of being selected. Whereas the paths that are not selected their pheromone is decreased by an evaporation process.

In ACO each ant constructs a complete solution using two things (1) node transition probability function which is based on the quantity of pheromone spread by ants and heuristic information about the importance and quality of each individual solutions and (2) already traversed solutions memory. As generations get completed, solutions constructed by each ant are evaluated using some evaluation criteria. After that pheromone evaporation and update mechanism is also used which evaporates intensity of pheromone from the paths with low fitness value and hence discarded gradually. The ACO algorithm requires specifying the following aspects for implementation:

1) Representation of the problem domain in such a way that it lends itself to incrementally building a solution for the problem, usually in the form of a graph.

2) Node transition probability rule based on the amount of pheromone value and of the heuristic function we have employed gain ratio as a heuristic function. Following is the equation for calculating the probability of each node:

$$P_j^i = \frac{[\tau(i,j)]^\alpha [\eta(i,j)]^\beta}{\sum_{k \in S} [\tau(i,k)]^\alpha [\eta(i,k)]^\beta} \quad (4)$$

Where  $P_j^i$  is the probability of the  $i$ th ant to move from node  $i$  to node  $j$  at time  $t$ .  $P_i^j(t) = 0$  means that ants are not allowed to move to any node in the neighbor.

$[\tau(i,j)]^\alpha$  is the amount of pheromone on the edge connecting  $i$  and  $j$ , where  $\alpha$  is a constant which is used to control relative importance of pheromone information. After each iteration, this pheromone information is updated by all the ants and in some versions of ACO only best ant is allowed to update pheromone.

$[\eta(i,j)]^\beta$  is the heuristic function that denotes the heuristic value of edge connecting  $i$  and  $j$ . usually, the heuristic value does not change during execution of the algorithm. In this paper we have used gain ratio to denote heuristic value.  $\beta$  is a constant which is used to control relative importance of heuristic value.

3) A heuristic evaluation function called fitness function dependent on the problem, which provides a goodness measurement for the different solution components. We have used fitness function is based on gain ratio to normalize the biasness of information gain and mutual information towards multi-valued attributes. Following formula being used to compute the value of the selected subset.

$$\text{Fitness}(S) = \frac{(F-S) * (\sum_{i=1}^n (GR))}{F} \quad (5)$$

Where  $S$  is reduced subset selected by ACO,  $GR$  is the gain ratio of feature  $i$  in the subset  $S$  and  $F$  is the total number of features present in the dataset. It will select feature subset with high gain ratio value and with less number of features.

4) Pheromone evaporation and updating rule which takes into account the evaporation and reinforcement of the paths. Once subsets are evaluated using fitness function, pheromone trails are updated. Firstly using an evaporation rate  $\rho$  the pheromone trails on the edges are evaporated or decreased to minimize the effect of a sub-optimal feature to which the ants have previously converged. Secondly amount of pheromone on the edges is updated with amounts proportional to the fitness of the solution. Some approaches for pheromone updating allowed all the ants to update their paths according to the fitness of their solution and in some approaches only best ant is allowed to update pheromone value on its path. In this thesis former approach is used in which all the ants update their path according to the fitness of their solution.

For the pheromone evaporation and updating following equations are used.

$$\tau = (1 - \rho) * \tau \text{ where } \rho \text{ is } 0.15 \quad (6)$$

$$\tau = \tau + (\tau * Q) \quad (7)$$

$$\text{and } Q = \left[ 1 - \left( \frac{1}{1 + \text{Fitness}} \right) \right] \quad (8)$$

Equation (7)-(8): Pheromone evaporation and updating.

5) Where "Fitness" is the value of the selected subset through an independent statistical measure.

6) Stopping/convergence criterion that decides when the algorithm terminates usually depends on maximum number of iterations.

### C. Proposed ACO-CE

This is our proposed approach. In this approach ACO is used for selecting most optimal feature subsets along with Gain Ratio where Gain ratio is used as heuristic function for selecting most relevant features. Fitness function or subset evaluation is also based on gain ratio. It's a pure filter approach, along with it we have also used classifier ensemble to improve predictive performance of filter approaches comparable to the wrapper approaches.

In proposed approach first of all dataset is loaded. Once dataset is loaded, gain ratio of each feature/attribute in data set is computed. Then all the parameters of ant colony optimization algorithm are initialized. Such as number of ants,  $\alpha$  and  $\beta$  values of node transition probability function, path convergence threshold value, pheromone evaporation rate  $\rho$ . and maximum number of generations. A search space is constructed that consists of nodes proportional to the number of features in the dataset. Fixed numbers of ants are generated in each iteration where each ant generates a candidate solution. After each generation, generated solutions are evaluated using a subset evaluator. Subset evaluator is based



on Gain Ratio between selected features and the class. After subset evaluation best solution is gained on the basis of maximum fitness value and is preserved. Then termination criteria of the algorithm are checked which is based on two conditions i.e. on a maximum number of generations and convergence threshold. If termination criteria are not met each ant updates its pheromone value to the quality of solution generated by each ant. Otherwise, if any termination/stopping criterion is met algorithm outputs ten best subsets.

Then these subsets are provided to classifier ensemble and these subsets are provided to classifier ensemble consisting of C4.5 decision tree classifier, Naïve Bayes, and K-Nearest Neighbor classifier.

Then one subset is selected on the basis of the highest average weighted accuracy of classifier ensemble and saved. Then again convergence threshold is checked. If it is less than 500, the whole process is repeated again. Otherwise, the algorithm stops and one best subset with the highest accuracy is selected from all saved subsets and is considered as final subset. Then new ants are produced and this complete process goes iteratively till highest number of epochs is reached or the algorithm convergence to a solution.

Our proposed approach as shown in Fig. 2 is a filter approach which selects features on the basis independent gain ratio measure. So some features that might be less important in terms of independent relevance to class but for a classifier, such features could be important. Therefore, ACO-CE uses classifier ensemble on different convergence threshold values and classification accuracy of subsets is used to provide final feature subset. So our approach improves classification accuracy of filter approaches.

TABLE I. PROPOSED ALGORITHM

1.	Start
2.	Load the dataset
3.	Compute gain ratio (heuristic value) of all the attributes/features in dataset
4.	Initialize ACO parameters
5.	Set convergence threshold(50:50:500)
6.	Do 1: maximum generations
7.	Each ant generate solutions
8.	Evaluate each solution
9.	Keep track of best solutions
10.	Check stopping criteria( yes: go to 14)
11.	Update pheromone for all ants.
12.	Generate new ants
13.	Go to 6
14.	Select 10 best subsets of converged /maximum generations
15.	Run multi classifier ensemble
16.	Select subset with the highest average accuracy
17.	Save it
18.	Check if convergence threshold >500(yes: go to 20)
19.	Go to 5
20.	Select one best subset from all the saved subsets that has got highest average classification accuracy of classifier ensemble
21.	Output best subset as final subset
22.	Stop.

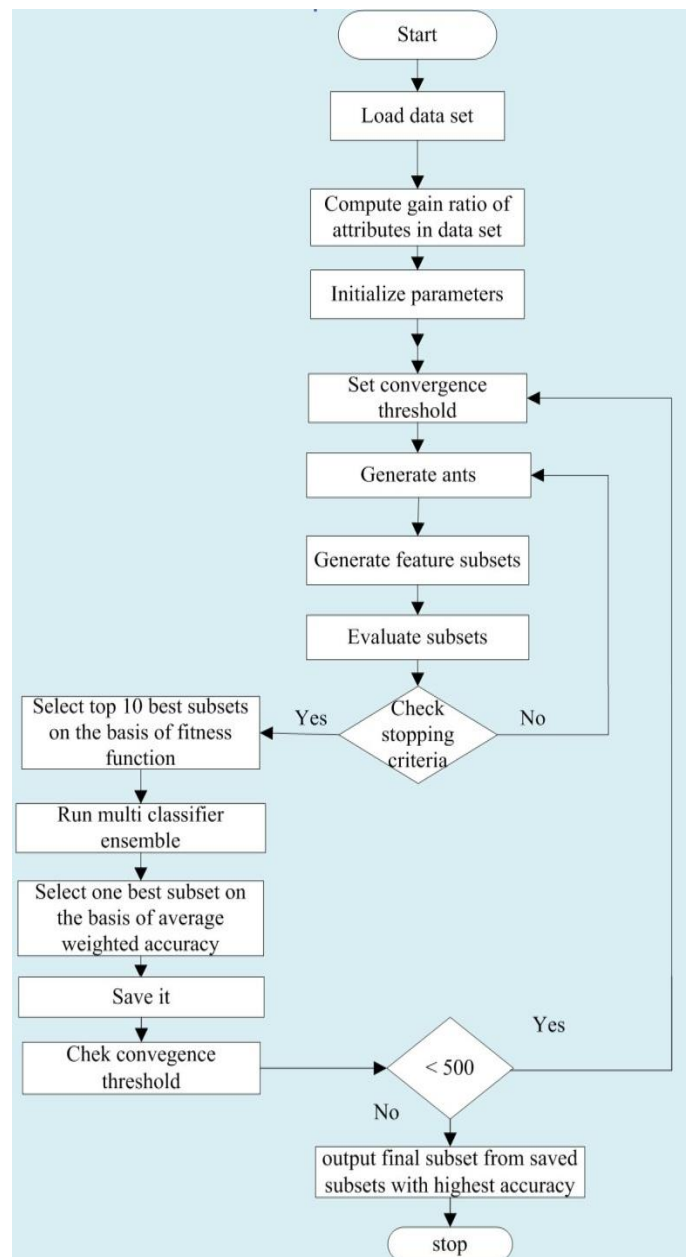


Fig. 2. Flow Chart of ACO-CE.

## V. EXPERIMENTS AND RESULTS

Extensive experiments have been carried out on ACO-CE in order to find out the effectiveness of ACO-CE for feature selection. Feature selection using ACO-CE have been implemented in Matlab 2009. We have used standard parameters of ACO i.e.  $\alpha = \beta = 1$ . A number of ants in proposed ACO-CE are equal to the number of attributes in the dataset. Maximum epochs are 500 and path convergence threshold starts from 50 and stops on 500 with incrementing threshold value 10 times with 50 and Classifier ensemble consists of C4.5 decision trees, K-Nearest Neighbor and Naïve Bayes.

TABLE II. DATASET

Dataset	Total Features	Instances	Class
Iris	4	150	3
Diabetes	8	768	2
Liver disorder	6	345	2
Hepatitis	19	155	2
Colic horse	22	368	2
Ionosphere	34	351	2
Dermatology	34	366	6
Breast cancer	10	699	2
Lymphography	18	148	4
Vote	16	435	2
Labor	16	57	2

We have tested the performance of our technique with a standard implementation of three existing feature selection techniques: Genetic algorithm with consistency measure for subset evaluation [22], PSO using fuzzy rough sets as subset evaluation method [20] and ACO using fuzzy rough sets as subset evaluation method [21]. These algorithms have already been implemented in Weka [24], data mining software. Most of these algorithms are implemented by their respective authors so we have used these with their default values without doing any modification.

#### A. Data Sets

We have used eleven datasets as shown in Table II which are publically available in UCI machine learning repository [23]. Table I shows details about data sets used for experimentation. All of these datasets are discretized using weka 3.7.11 [24].

#### B. Results and Discussion

Table III shows the total features that were selected by our proposed methodology in comparison with the features that are selected by other eleven datasets. It is observed that ACO-CE selected a small number of features for all other datasets having more features.

TABLE III. NUMBER OF FEATURES SELECTED

Dataset	Total	PSO	GA	ACO	ACO-CE
Iris	4	4	4	4	2
Diabetes	8	8	8	8	2
Liver disorder	6	5	5	5	2
Hepatitis	19	15	10	16	7
Colic horse	22	18	9	20	5
Ionosphere	34	17	19	26	7
Dermatology	34	9	14	21	12
Breast Cancer	10	10	7	10	4
Lymphography	18	8	9	8	7
Vote	16	11	10	14	5
Labor	16	10	6	11	6

TABLE IV. CLASSIFICATION ACCURACY ON C4.5

Dataset	PSO	GA	ACO	ACO-CE
Iris	<b>97.33</b>	<b>97.33</b>	<b>97.33</b>	<b>97.33</b>
Diabetes	65.75	65.75	65.75	<b>68.09</b>
Liver disorder	57.39	57.39	57.39	<b>57.68</b>
Hepatitis	83.22	83.22	81.93	<b>83.87</b>
Colic horse	85.32	<b>85.86</b>	85.32	85.59
Ionosphere	86.32	85.75	90.88	<b>91.16</b>
Dermatology	89.89	88.79	<b>93.98</b>	<b>93.98</b>
Breast cancer	94.42	94.42	94.42	<b>94.70</b>
Lymphography	<b>81.75</b>	79.72	77.02	81.08
Vote	<b>96.32</b>	96.32	96.32	95.63
Labor	70.17	75.43	70.17	<b>78.94</b>

Table IV presents the comparison of the classification accuracy of ACO-CE with all algorithms over C4.5 classifier. Classification accuracy is checked in weka by using 10 folds cross-validation process. The Bold value in every column represents the highest value of accuracy. Proposed approach is better in 8 data sets. In iris, all the algorithms have same predictive accuracy but our approach has gained same accuracy with smaller feature set as compared to other approaches.

Table V presents the comparison of the classification accuracy of ACO-CE with all algorithms over K Nearest Neighbor classifier. Classification accuracy is checked by using 10 folds cross-validation process. Proposed approach is better in 9 d.

TABLE V. CLASSIFICATION ACCURACY ON KNN

Dataset	PSO	GA	ACO	ACO-CE
Iris	96.66	<b>97.33</b>	<b>97.33</b>	<b>97.33</b>
Diabetes	65.88	65.88	65.88	<b>68.09</b>
Liver disorder	56.23	56.23	56.23	<b>58.55</b>
Hepatitis	85.80	<b>87.09</b>	83.87	86.45
Colic horse	79.61	84.51	83.47	<b>85.86</b>
Ionosphere	82.33	83.47	85.18	<b>88.31</b>
Dermatology	87.70	87.97	<b>95.08</b>	94.53
Breast cancer	95.27	94.84	95.27	<b>95.99</b>
Lymphography	77.70	72.97	77.70	<b>82.43</b>
Vote	93.56	93.33	92.41	<b>96.09</b>
Labor	77.19	78.94	73.68	<b>82.45</b>

TABLE VI. CLASSIFICATION ACCURACY ON NAIVE BAYES

Dataset	PSO	GA	ACO	ACO-CE
Iris	<b>97.33</b>	<b>97.33</b>	<b>97.33</b>	<b>97.33</b>
Diabetes	<b>69.01</b>	<b>69.01</b>	<b>69.01</b>	68.09
Liver disorder	<b>58.55</b>	<b>58.55</b>	<b>58.55</b>	<b>58.55</b>
Hepatitis	82.58	84.51	83.87	<b>86.45</b>
Colic horse	79.61	81.35	80.16	<b>86.68</b>
Ionosphere	76.92	79.48	88.03	<b>91.16</b>
Dermatology	89.07	91.25	<b>97.26</b>	96.44
Breast cancer	<b>97.28</b>	96.85	<b>97.28</b>	<b>97.28</b>
Lymphography	80.40	77.70	82.43	<b>84.45</b>
Vote	92.87	92.41	91.03	<b>94.94</b>
Labor	84.21	82.45	84.21	<b>89.47</b>

TABLE VII. CLASSIFICATION ACCURACY ON RIPPER

Dataset	PSO	GA	ACO	ACO-CE
Iris	97.33	97.33	97.33	97.33
Diabetes	67.83	67.83	67.83	68.09
Liver disorder	56.23	56.23	56.23	58.55
Hepatitis	83.22	83.22	80.00	83.87
Colic horse	85.05	85.59	85.59	85.05
Ionosphere	86.03	87.74	90.88	91.16
Dermatology	80.05	85.51	93.98	92.07
Breast cancer	94.56	94.56	94.56	94.84
Lymphography	76.35	78.37	77.70	75.67
Vote	95.63	95.17	95.63	95.86
Labor	84.21	80.70	82.45	85.96

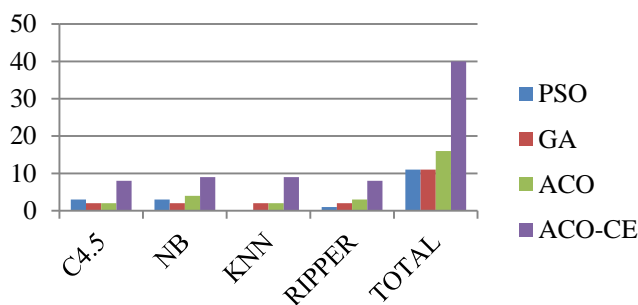


Fig. 3. Summarized comparison.

Classification accuracy is checked by using 10 folds cross-validation process. Proposed approach is better in 8 data sets in both Tables VI and VII. It has been observed that proposed approach has outperformed as compared to other approaches over all classifiers. Fig. 3 is the graphical representation to present the performance rate of ACO-CE and all algorithms. It has shown that our proposed approach has performed much better as compared to all algorithms over all four classifier.

#### ACKNOWLEDGEMENT

The author would like to thanks Dr. Waseem Shahzad Professor NUCES, Islamabad for his assistance and corporation. Also special thanks to HEC digital library for providing research material.

#### VI. CONCLUSIONS

A new hybrid method of ACO-CE has been proposed and implemented in this paper. ACO-CE is proposed by combining the ACO with a Classifier Ensemble (CE) and has been used to optimize the feature subset selection process.

Results showed that proposed approach has outperformed in terms of dimensionality reduction and classification accuracy as compared to other approaches. The Gain Ratio is used as a heuristic measure in ACO-CE which has normalized the biases of another heuristic measure towards multi-valued attributes and selected features that are highly relevant to the class. Secondly, the classifier ensemble has been used in a novel way with ACO. It checks the classification accuracy of

subsets achieved on different convergence threshold. Classifier ensemble helps to opt important features that are not selected by independent measure. We have not used classifier ensemble for optimizing results rather we have used it only for selecting a subset with the highest accuracy so our approach is not computationally costly.

Results showed that our approach has performed superior as compared to other feature selection techniques.

#### REFERENCES

- [1] Y.Ramamohan, K.Vasantharao, C.Kalyana, A.S.K Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process", International Journal of Soft Computing and Engineering vol. 2, Issue 3, pp. 74-79, July 2012.
- [2] ChulminYun, Byonghwa Oh, JihoonYang andJonghio Nang, " Feature Subset Selection Basedon Bio-Inspired Algorithms", Journal of Information Science and Engineering 27, 1667-1686 , 2011.
- [3] Haun Liu, Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering pp. 491-502, Vol. 17, No. 4 April 2005.
- [4] J. Doak, "An Evaluation of Feature Selection Methods and Their Application to Computer Security," technical report, Univ. of California at Davis, Dept. Computer Science, 1992.
- [5] Roberto Ruiz, Jos'e C. Riquelme, and Jes'us S. Aguilar-Ruiz; "Heuristic Search over a Ranking for Feature Selection"; IWANN, LNCS 3512, pp. 742-749; 2005.
- [6] Yao-Hong Chan; "Empirical comparison of forward and backward search strategies in L-GEM based feature selection with RBFNN"; International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 3, pp. 1524 - 1527 ; 2010.
- [7] I.H. Witten and E. Frank, Data Mining-Practical Machine Learning Tools and Techniqueswith JAVA Implementations.Morgan Kaufmann, 2000.
- [8] I.A. Gheyas and L.S. Smith. Feature subset selection in large dimensionality domains, Pattern Recognition, 2010.
- [9] P.A. Est\_vez, M. Tesmer, C.A. Perez, and J.M. Zurada. Normalized mutual information feature selection. Neural Networks, IEEE Transactions on, 20(2):pp189-201, 2009.
- [10] Lei Yu and Huan Liu. "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", In Proceedings of the Twentieth International Conference on Machine Learning, pp. 856-863, 2003.
- [11] Feng Tan, XuezhengFu ,Yanqing Zhang, "A genetic algorithm-based method for feature subset selection", Volume 12, Issue 2, pp. 111-120 Springer-Verlag 2007.
- [12] Bai-Ning Jiang Xiang-Qian Ding Lin-Tao Ma "A Hybrid Feature Selection Algorithm: Combination of Symmetrical Uncertainty and Genetic Algorithms" The Second International Symposium on Optimization and Systems Biology, Lijiang, China, pp. 152-157, 2008.
- [13] Li-Yeh Chuang, Chao-HsuanKe, and Cheng-Hong Yang, "A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification", Proceedings of the International MultiConference of Engineers and Computer Scientists pp. 146-150, Vol I, 2008.
- [14] Shailendra Kumar Shrivastava, "ACO Based Feature Subset Selection for Multiple k-Nearest Neighbor Classifiers", International Journal on Computer Science and Engineering, pp. 1831-1838, Vol. 3 No. 5 May 2011.
- [15] Md.MonirulKabir, Ms. Shahjahan Kazuyuki Murase, "A New Hybrid Ant Colony Optimization Algorithm for Feature SelectionVolume 39, Issue 3, 15 February 2012, Pages 3747-3763.
- [16] Gang Wang, JianMa,"A Hybrid Ensemble Approach For Enterprise Credit Risk Assessment Based on Support Vector Machine", Volume 39, Issue 5, April 2012, pp. 5325-5331, Elsevier journal
- [17] Shunmugapriya Palanisamy1 and Kanmani S2, "Classifier Ensemble Design using Artificial Bee Colony based Feature Selection based Feature Selection" IJCSI International Journal of Computer Science Issues, pp522-529, Vol. 9, Issue 3, No 2, May 2012.

- [18] Syed Imran Ali, and WaseemShahzad “Feature Subset Selection Method Based on Symmetric Uncertainty and Ant Colony Optimization,” in IEEE International Conference on Emerging Technologies (ICET), pp. 1-6, October 2012.
- [19] Hazem Ahmed,” Swarm Intelligence Concepts Models and Applications”, Technical Report, School of Computing Queen’s University Kingston, Ontario Canada, February 2012.
- [20] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, “Feature selection based on rough sets and particle swarm optimization”, presented at Pattern Recognition Letters, pp.459-471, 2006.
- [21] R. Jensen and Q. Shen, “Fuzzy-rough data reduction with ant colony optimization”, presented at *Fuzzy Sets and Systems*, Vol. 149, pp.5-20, 2005.
- [22] David E. Goldberg, *Genetic algorithms in search, optimization and machine learning*, Addison-Wesley, 1989.
- [23] S. Hettich, and S.D. Bay, “*The UCI KDD Archive*”. Irvine, CA: Dept. Inf. Comput. Sci., Univ. California, 1996 [Online]. Available: <http://kdd.ics.uci.edu>.
- [24] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, “The WEKA Data Mining Software: An Update”, *SIGKDD Explorations*, Vol. 11, No. 1, pp. 10-18, 2009.

# Efficient Smart Emergency Response System for Fire Hazards using IoT

Lakshmana Phaneendra Maguluri, Tumma  
Srinivasarao

Department of Computer Science and Engineering  
Koneru Lakshmaiah Education Foundation  
Green Fields, Vaddeswaram, Guntur  
Andhra Pradesh - 522502

Maganti Syamala

Department of Computer Science and Engineering  
Dhanekula Institute of Engineering & Technology  
Ganguru  
Andhra Pradesh - 521139

R. Ragupathy

Department of Computer Science and Engineering  
Annamalai University  
Annamalai Nagar, Chidambaram  
Tamil Nadu – 608002

N.J. Nalini

Department of Computer Science and Engineering  
Annamalai University, Annamalai Nagar,  
Chidambaram  
Tamil Nadu - 608002

**Abstract**—The Internet of Things pertains to connecting currently unconnected things and people. It is the new era in transforming the existed systems to amend the cost effective quality of services for the society. To support Smart city vision, Urban IoT design plans exploit added value services for citizens as well as administration of the city with the most advanced communication technologies. To make emergency response real time, IoT enhances the way first responders and provides emergency managers with the necessary up-to-date information and communication to make use of those assets. IoT mitigates many of the challenges to emergency response including present problems, like a weak communication network and information lag. In this paper, it is proposed that an emergency response system for fire hazards is designed by using IoT standardized structure. To implement this proposed scheme a low-cost Expressive wi-fi module ESP-32, Flame detection sensor, Smoke detection sensor (MQ-5), Flammable gas detection sensor and one GPS module are used. The sensors detects the hazard and alerts the local emergency rescue organizations like fire departments and police by sending the hazard location to the cloud-service through which all are connected. The overall network utilizes a light weighted data oriented publish-subscribe message protocol MQTT services for fast and reliable communication. Thus, an intelligent integrated system is designed with the help of IoT.

**Keywords**—Internet of Things (IoT); Arduino IDE; GPS

## I. INTRODUCTION

The proposed system is capable of detecting smoke, different flammable gases and fire. This system is capable of providing hazard location coordinates to the nearby fire department. This fire hazard sensing system with systematic IoT framework emphasises an application innovation to the public safety and livelihood service sector [12], [13].

The overall billing of the proposed system is given in Table I.

TABLE I. BILLING OF THE OVERALL SYSTEM

Component	Quantity	Cost
ESP32 microcontroller	1	\$8.55
MQ-2 sensor	1	\$6.90
MQ-5 sensor	1	\$7.90
Flame sensor	1	\$4.19
GPS Module	1	\$6.19

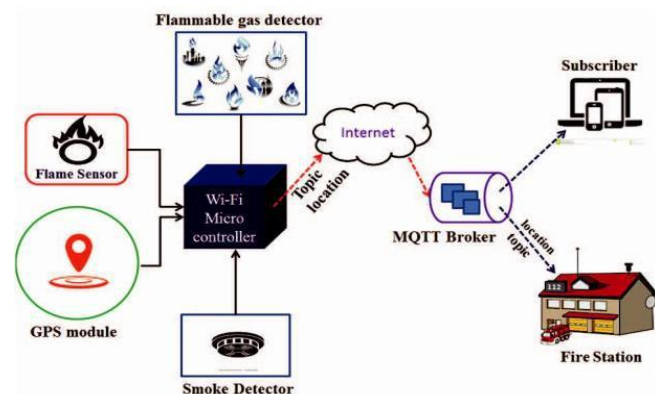


Fig. 1. Block diagram.

The fire hazard sensing system with IoT standardized design methods is shown in Fig. 1. The smoke detection sensor MQ-2 is used to detect the smoke, the Flame detection sensor is used to sense the flame, the flammable gas sensor MQ-5 is used to detect the gases like LPG/LNG and the GPS module is to obtain device location. These sensors along with Wi-Fi micro-controller are connected to a MQTT broker via Internet through which it communicates hazard status to the nearest fire-fighting organizations [14].

1) *Internet of Things*: The Internet of Things (IoT) is an umbrella of smart electronic devices like sensors and

intelligent software applications to build an effective data exchange system.

In IoT the devices can communicate with each other and independently configure themselves in a network of multiple Internet connected devices. To unleash smart cities development agenda, many existing systems with the specific application domain serve a greater good in urban areas adopting this modern IoT technology [3]. After adopting this emerging technology, machine to machine communication transforms the existing human-human or human-machine forms of communication [6]. IoT possess high resource sharing capabilities, high degree of intelligence, high scalability and other main characteristics. Internet of Things in the fire-fighting safety management field has more importance [2] in providing secured lifestyle in smart cities. This work gives the idea that designing an emergency response system for taking precautionary measures and rescue operations for fire hazards.

2) *Smart City Vision in INDIA*: The evolution from the emerging technologies has lead Government of India to launch Smart city Mission on 25th June 2015 to improve the quality of life the citizens with smart solutions. These smart solutions are to improve the services for livability of the entire city. Under the smart cities development plan around 60 cities are chosen and US \$21756.717 million is the total investment to implement smart solutions. In the whole investment US \$15786.554 million is for revamping area based projects, remaining US \$5970.163 million for smart city initiative investments [4]. The key importance has given to some of the core infrastructure elements like efficient public transport and urban mobility, digitalization and robust IT connectivity, affordable housing, sustainable environment and same importance has been given to safety and security of citizens [9]. In implementing the smart solutions and building life safety in urban landscape IoT has its own importance. IoT incorporates transparent, seamless heterogeneous end systems. In regarding fire safety and management, to safeguard the city's assets like local departments information systems, transportation systems, schools, libraries, hospitals, power plants and other community services Internet of Things (IoT) provides proper solutions in a secure fashion. The real time monitoring systems with IoT design structure with sensors integrated collects the leveraged data from the devices. The data from the devices processed and analyzed to take necessary precautionary actions [11].

The rest of the paper is organized as follows: Section II deals about experimental setup and working; Section III deals about required hardware aspects; Section IV deals about required software aspects; Section V deals about experimental results followed by Section VI as Conclusion.

## II. EXPERIMENTAL SET UP AND WORKING

The circuit connections for the proposed system are shown in Fig. 2. The Wi-Fi micro-controller board (ESP-32) is powered up by using USB cable. Different sensors for different measurements are used and interfaced to the micro-controller board using connecting wires.

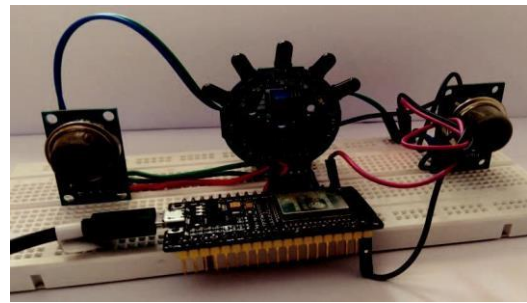


Fig. 2. The overall circuit connections of the System.

Flame sensor have 5 output pins which are connected to analog read general purpose I/O interface pins (GPIO pins) 36, 39, 34, 35 and 32 respectively. The MQ-2 gas sensor, MQ-5 gas sensor are connected to GPIO 25, GPIO 26 pins of the board respectively. And GPS module has both transmitter and receiver pins which are connected to GPIO17, GPIO18 pins of ESP-32 board respectively. After that, the logic is structured as required to operate the whole system as desired. For the desired system programming part is done in Arduino IDE. In the part of initialization pin configurations for respective connections are necessary.

## III. HARDWARE ASPECTS

### A. ESP32

ESP32 is the most advanced Espressif Wi-Fi micro-controller board. It is integrated with built in antenna switches, power amplifier and RF balun. Its compact design includes Flash memory and it has ESP32SoC and PCB antenna for better RF performance. ESP32 is well known for its hybrid functionality which consists of Bluetooth and Wi-Fi is shown in Fig. 3. It supports WPA/WPA and WEP for security aspects. For industrial environments it can give more reliability because it can adopt to environmental changes.

1) Its operating temperature range is  $-40^{\circ}\text{C}$  to  $+120^{\circ}\text{C}$ . It can be interfaced with other devices using I2C/UART or SPI/SIDO interfaces. It has some built in sensors like Hall sensor, Ultra low noise analog amplifier and touch interface. As compared to other Espressif models its performance is better. It's receiver sensitivity up to  $-98\text{dBm}$  and transmits power range up to  $19.20\text{dBm}$ . ESP32 is mainly designed for Low power applications like IoT based electronic industrial appliances [1].

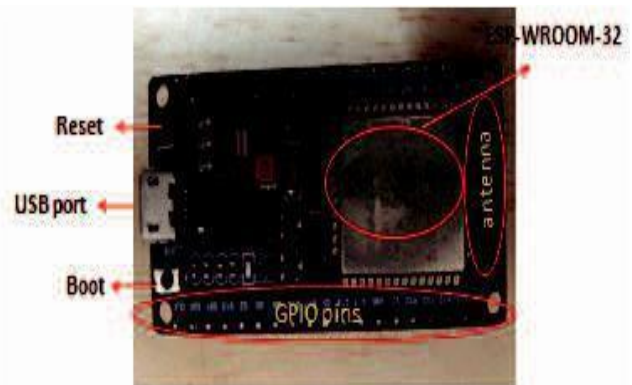


Fig. 3. ESP32.

TABLE II. ESP32 FEATURES

Specifications	Features
Micro-controller CPU	Xtensa Dual-core 32-bit LX6 600 DMIPS
Wi-Fi 802.11 b/g/n	WiFi MAC, WiFi base-band
Frequency	2.4Ghz
Network Protocols	IPv6, IPv4, TCP/UDP, HTTP/FTP, iWLAN, MAC Protocol
GPIOs	36
ADC	12-bit
ROM memory	448 KB
Instruction RAM	520 KB
Operating current	80mA
Operating voltage	3.3v

TABLE III. DIFFERENT TYPES OF WI-FI MICRO-CONTROLLER BOARDS

Board type	Special features	Cost
ESP32 NANO IOT development board	4MB Flash, 3.3V 0.5A Regulator, Xtensa Dual-Core 32-bit LX6 microprocessors, 32 gpio pins, Hall sensor, 10x capacitive touch interface, SD card interface support	\$8.36
WEMOS ESP32	32 gpio pins, USB to UART Chip CP2102, 16mb flash memory, One Tensilica LX6 micro-controller	\$9.8
Noduino Quantum	SPI Flash 16MB, 16 GPIO pins, 5V-12V Power supply, FreeRTOS	\$25.90
Fipy	8mb flash memory, 8x12 bit ADCs, Tensilica LX6 micro-controller with dual processor, 22 GPIO pins, U.FL LoRa/Sigfox antenna connector, WiFi, BLE, cellular LTE-CAT M1/NB1, LoRa, and Sigfox Micropython enabled	\$55.90

Table II is taken with reference to the documentation [1] the other ESP-32 On board wi-Fi micro-controller boards available along with their own specifications are listed in Table III.

2) *Flame sensor*: The flame sensor consists of emitter, detector with an associative circuitry. The emitter consists of an Infrared Light emitting Diode and the detector consists of an Infrared Photo diode which senses the Infrared light which is having same wavelength as that of emitted wave wavelength by IR LED. The basic principle that involved in working of the sensor is photon energy strikes out the electrons so that circuit resistance will change accordingly. Whenever Photo diode senses the IR light, the Resistance and corresponding Output voltage will be changed in proportion to the received IR light magnitude. Because of this flame detection sensor can often responds very quickly and give accurate measurement. This sensor is designed such that it ignores constant background IR radiation because it present in

all environments. Instead it is designed to sense sudden changes in the IR radiation. So that it avoids false detection.



Fig. 4. Flame sensor.

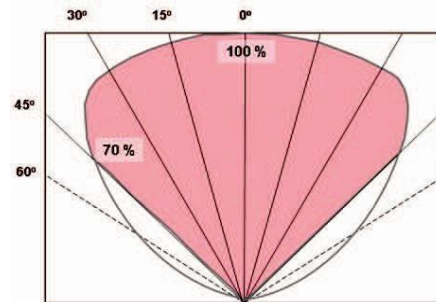


Fig. 5. View of the flame sensor.

The flame sensor needs to be aligned precisely in a Particular position to take care that it should not sense potential background radiation sources is shown in Fig. 4 and 5. The flame sensor detects flames in 3-D view and this cone of vision is not necessarily round. For the cone of vision vertical and the horizontal angles are often different. This sensor has advantage for projection that it has a high sensitivity on edges of its angle of vision. Flame sensor detection range is depends on mounting.

Location, so when making projection it is important to know what it sees. The flame detector is mounted in a height as twice the height of highest object in the view. While projecting the sensor it is preferred that it should be accessible to maintenance and repairs.

By mounting a second flame sensor in the opposite of the first sensor the shadow effect can be reduced. In general several flame sensors can be mounted such that sensors look each other but not to the walls so that blind spots can be avoided as shown in Fig. 6 [19].

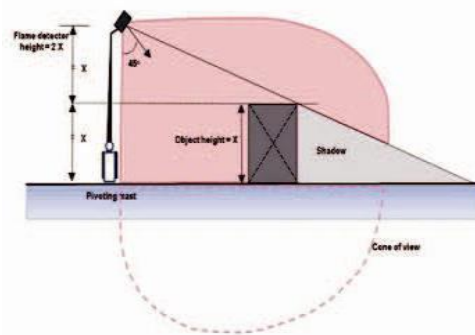


Fig. 6. Alignment of the flame sensor.

3) *MQ-Gas sensors*: In designing of these two sensors, SnO<sub>2</sub> is used as gas sensing layer, Au, Pt are used as electrodes, Ni-Cr alloy is used as heater coil, SUS36 100-mesh (Stainless steel gauze) is used as anti-explosion network and Bakelite is used as resin base is shown in Table IV. In the clean air, the sensitivity material SnO<sub>2</sub> of MQ-gas sensor initially has lower conductivity. If the sensor detects the target gas, the sensing material conductivity raises along with the concentration of the gas is shown in Fig. 7 and 8. The variation in conductivity in turn converted into variation in voltage over load resistance. The resistance of sensor can be measured from the below given formula:

Resistance of sensor R<sub>s</sub>:

$$R_s = (V_c / V_{RL} - 1)R_L \tag{1}$$

where, R<sub>L</sub> is adjustable.

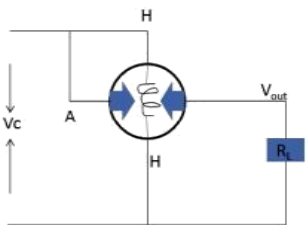


Fig. 7. Internal circuitry of MQ-gas sensor.



Fig. 8. MQ-gas sensors.

TABLE IV. MQ-GAS SENSORS FEATURES

S.NO	Specifications	MQ-5	MQ-2
1	Sensing resistance	10KΩ-60KΩ	2KΩ-20KΩ
2	Range	200-1000 ppm	300-1000 ppm
3	Gases to be detected	LNG, LPG Natural gas, propane isobutane	Smoke
4	Applications	portable gas detector, industrial combustible gas detector, Domestic gas leakage detector	Smoke detectors, fire alarms
5	Temp condition Humidity condition	18°C to 20°C 60% to 70%	18°C to 20°C 60% to 70%
6	Power consumption	≤ 800mW	≤ 900mW

4) *GPS Module*: GPS receivers use constellation of ground stations and satellites to track position on the earth. At any time above 12,000 miles over the object location there are 12 satellites orbiting and transmitting information back to the earth RF range 1.1GHz. GPS receiver uses that information and math to provide its orbital position [17]. It uses NMEA data format to display in sentences and sent out using serial Tx pin. Satellite position, weather, signal to noise ratio and obstructions such as mountains and buildings are primary variables that effects GPS accuracy is shown in Fig. 9. A GPS receiver must be able to get a lock on 4 satellites to be able to solve for a position.

*Features:*

- 1) 66 acquisition/22 tracking-channel receiver
- 2) WAAS/GAGAN/MSAS/EGNOS support
- 3) NMEA protocols (speed: 9600bps)
- 4) Ultra high sensitivity: -165dBm
- 5) Temperature range: -40 C to 85 C
- 6) RoHS compliant (Lead-free)
- 7) Form factor 20.5mm x 12.8mm x 7.8mm
- 8) Embedded patch antenna 12\*12\*4 mm
- 9) One serial port

*Applications:*

- 1) Location Based Service
- 2) Portable Navigation Device
- 3) GPS mouse and Bluetooth GPS receiver
- 4) Vehicle navigation system
- 5) Timing application



Fig. 9. GPS Module.



IV. SOFTWARE ASPECTS

A. Arduino IDE

Arduino IDE software is used to write programs, and programs can be uploaded directly to the board. It is available for many operating systems like Windows, Linux, Mac OS X, Portable IDE (Linux & Windows). It is an open source platform for electronics design, and very easy to use for both hardware and software. Arduino ide comes with few advantages like fast prototyping and also helps the students who don't have any prior knowledge in electronics and software programming. It provides flexible, simple and clear programming environment for beginners [5].

1) *Adafruit.io*: Adafruit IO is a cloud-service that makes sensed data useful. It is well known for ease of use, and allows simple data connections with little programming. The client libraries that wrap MQTT APIs and available to receive and send data with Adafruit IO. It can be built on Node.js and Ruby on Rails. Adafruit MQTT Client Library, PubSubClient MQTT Library are very popular MQTT client libraries used for Arduino IDE to access Adafruit IO [15]. The main Idea that data can be sent or receive by defining feed. The data can be published or subscribed to the feed. The MQTT client is connected to Adafruit Io with port number 1883, Adafruit account username and Adafruit IO key. The important features of MQTT are the ability to specify a QoS and impose a rate limit to prevent excessive load.

B. MQTT

Message Queue Telemetry Transport (MQTT) is extremely light weight, simple and publish/subscribe messaging protocol is shown in Fig. 11. This is specifically designed for low-bandwidth, high latency networks which are unreliable and constrained devices. It satisfies the design principles like minimum network bandwidth and meets the device resource requirements and allows simple way of telemetry transport. MQTT assures fast delivery and ensures reliability is shown in Fig. 10. These principles makes this protocol ideal for the emerging technologies like Internet of Things and M2M technologies with multiple connected devices where bandwidth and battery power are concerned [18]. A paradigm shift steering away from request-response protocols to the leading publish-subscribe protocols because it is needed for easy implementation of human web and communication among machines at a large scale [7], [8].

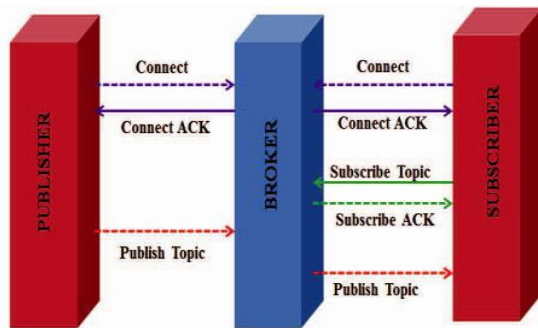


Fig. 10. MQTT publisher-broker-subscriber communication.

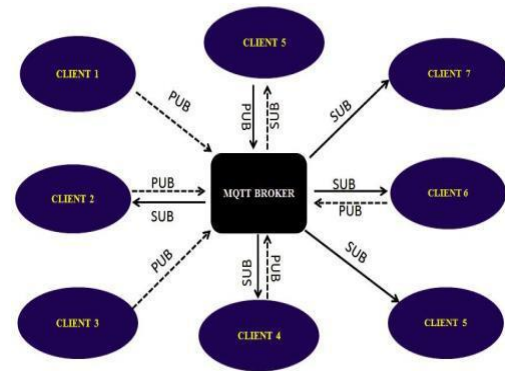


Fig. 11. Architecture for MQTT connection.

MQTT is standardized by OASIS. MQTT utilizes many features of the TCP transport. It requires minimal working TCP stack protocol overhead where small code footprint is used which is now available for even the smallest micro-controllers. MQTT follows the brokered publish/subscribe pattern which decouples the clients. The publisher (client1) which sends particular information broker and the subscriber (client2) which receives that information from the broker doesn't know each other [16]. The broker does the decoupling, by filtering all incoming messages from publishers and it distributes them to the correct subscribers. Decoupling can be done in three ways like space decoupling in which publisher and subscriber don't know each other. In Time decoupling, publisher and subscriber need not be connected at the same interval. In synchronization decoupling, both are not halted during receiving and publishing [10]. In comparison with the other networking protocols like HTTP/S, MQTT is best suit for Internet of Things applications. The main differences and important features of MQTT over HTTP/S are listed in Table V.

TABLE V. DIFFERENCES BETWEEN MQTT AND HTTP/S

S. No	Feature	MQTT	HTTP/S
1	Architecture	Publish- Subscribe architecture	Request- Response Architecture
2	Design Methodology	Data oriented	Document Oriented
3	Data security	secure	Not secured
4	Upper layer protocol	TCP	UDP
5	Message Size	Small with 2Byte Header	Large ASCII for- mat
6	Data Distribution	One to Many	One to One

## V. EXPERIMENTAL RESULTS

The overall circuited system with different fire hazard detection sensors and Wi-Fi micro-controller is shown in Fig. 2 and results given in snapshots. After connecting each and every device in the desired manner and making sure that each and every component is connected in accordance with the other components. An external power supply is given to ESP32 board with USB for current to flow. The sensor detects the hazard if any smoke or any flammable gas or any flame in the surroundings. Here the GPS module gives the hazard location in Decimal Degree format. The sensed data and the location co-ordinates are published on to the Arduino.Io cloud service. The dashboard shown in Fig. 12 indicates NO hazard status. The dashboard shown in Fig. 13 is when hazard detected. The Location obtained by the Decimal Degree co-ordinates is shown in Fig. 14.

The system utilizes the light weighted protocol MQTT services to provide fast responses so as to provide emergency rescue operations immediately. The information along with hazard location in decimal degrees is published on to Adafruit.IO along with hazard status which is subscribed by social organizations.

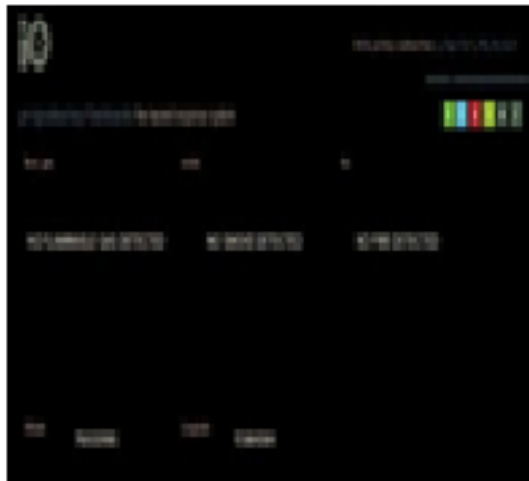


Fig. 12. Dashboard status when NO hazard detected.



Fig. 13. Dashboard status when hazard detected.



Fig. 14. Location with the obtained DD co-ordinates.

## VI. CONCLUSION

In this paper, it is mentioned that Internet of Things is an emerging technology which helps in providing smart solutions in Smart city development aspect. In providing a quality public safety and security services it is very important to adopt leveraged data driven emergency response systems with urban IoT design standards. A smart emergency response system for fire hazards is designed and implemented with required IoT standards which prioritize the immediate rescue operations by pushing relevant information to the public safety managements.

## REFERENCES

- [1] <http://espressif.com/en/products/esp32>
- [2] IOT based control of Appliances by Ravi Kishore Kodali, SreeRamya Soratkal and Lakshmi Boppana. International Conference on Computing, Communication and Automation (ICCCA2016)
- [3] IoT Based Smart Security and Home Automation M. Farooq, M. Waseem, S. Mazhar, A. Khairi, and T. Kamal, International Journal of Computer Applications, vol. 113, no. 1. [Online]. Available: <http://research.ijcaonline.org/volume113/number1/pxc3901571.pdf>
- [4] Internet of Things for Smart Cities by Andrea Zanella, Senior Member, IEEE, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, Senior Member, IEEE, and Michele Zorzi, Fellow, IEEE
- [5] <https://www.arduino.cc/en/main/software>.
- [6] R. K. Kodali, S. Soratkal and L. Boppana, "IOT based control of appliances," 2016 International Conference on Computing, Communication and Automation (ICCCA), Noida, 2016, pp. 1293-1297. doi: 10.1109/CCAA.2016.7813918
- [7] A low cost implementation of MQTT using ESP8266 Ravi Kishore Kodali and Kopolwar Shishir Mahesh Department of Electronics and Communication Engineering National Institute of Technology, Warangal WARANGAL 506004 INDIA Email: ravikkodali@gmail.com
- [8] Meena Singh, Rajan MA, Shivraj VL, and Balamuralidhar P, Secure MQTT for Internet of Things (IoT), 2015 Fifth International Conference on Communication Systems and Network Technologies
- [9] Ministry of Urban Development Government of India - Smart Cities [smartcities.gov.in/upload/uploadfiles/files/SmartCityGuidelines\(1\).pdf](http://smartcities.gov.in/upload/uploadfiles/files/SmartCityGuidelines(1).pdf)
- [10] Valerine Lampkin, Weng Tat Leong, Leonardo Olivera, Sweta Rawat, Nagesh Subrahmanyam, Rong Xiang, Building Smarter Planet Solutions with MQTT and IBM WebSphere MQ Telemetry, First Edition, September 2012.
- [11] An IoT based Fire Alarming and Authentication System for Workhouse using Raspberry Pi 3, International Conference on Electrical, Computer and Communication Engineering (ECCE), February 16-18, 2017, Cox's Bazar, Bangladesh

- [12] Discussion of Society Fire-fighting Safety Management Internet of Things Technology System WANG Jun, ZHANG Di, LIU Meng, XU Fang, SUI Hu-lin, YANG Shu-feng, 2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications
- [13] [14] Chen, Thou-Ho, et al. The smoke detection for early fire-alarming system base on video processing, in Proceedings of International Conference on Intelligent Information Hiding and Multimedia, 2006.
- [14] <https://learn.adafruit.com/>
- [15] <https://mqtt.org/>
- [16] A comparison of IoT application layer protocols through a smart parking implementation Paridhika Kayal; Harry Perros 2017 20th Conference on Innovations in Clouds, Internet and Networks (ICIN)
- [17] Implementation of a wireless sensor network using standardized IoT protocols Gustavo A. da Costa; Joo H. Kleinschmidt 2016 IEEE International Symposium on Consumer Electronics (ISCE)
- [18] [https://en.wikipedia.org/wiki/Flame\\_detector#IR.2FIRflame\\_detection](https://en.wikipedia.org/wiki/Flame_detector#IR.2FIRflame_detection)
- [19] [http://wiki.seeed.cc/Grove-Flame\\_Sensor/](http://wiki.seeed.cc/Grove-Flame_Sensor/)

# An Information Theoretic Analysis of Random Number Generator based on Cellular Automaton

Amirahmad Nayyeri

Division of Computer Science, Faculty of Science  
Salman Farsi University of Kazerun, Kazerun, Iran

Gholamhossein Dastghaibifard

Computer Engineering and IT Department  
Shiraz University, Shiraz, Iran

**Abstract**—Realization of Randomness had always been a controversial concept with great importance both from theoretical and practical Perspectives. This realization has been revolutionized in the light of recent studies especially in the realms of Chaos Theory, Algorithmic Information Theory and Emergent behavior in complex systems. We briefly discuss different definitions of Randomness and also different methods for generating it. The connection between all these approaches and the notion of Normality as the necessary condition of being unpredictable would be discussed. Then a complex-system-based Random Number Generator would be introduced. We will analyze its paradoxical features (Conservative Nature and reversibility in spite of having considerable variation) by using information theoretic measures in connection with other measures. The evolution of this Random Generator is equivalent to the evolution of its probabilistic description in terms of probability distribution over blocks of different lengths. By getting the aid of simulations we will show the ability of this system to preserve normality during the process of coarse graining.

**Keywords**—Random number generators; entropy; correlation information; elementary cellular automata; reversibility

## I. INTRODUCTION TO RANDOMNESS

Realization of randomness has great importance both from theoretical and practical perspective. The successful application of randomness for guiding search processes in spaces which scales exponentially to the input size shows the theoretical importance of having access to a proper source of randomness [1]. One can show generally that the existence of proof for theorem T with specific length  $n$  can be reduced to combinatorial problem of finding a tour of length  $\leq B$  that reaches all  $N$  cities while  $N = poly(n)$  [2].

Although the concept of Randomness has entered literarily by the theory of probability, this theory cannot define the randomness associated with individual objects whether finite or infinite. Probability theory is a theory about sets of objects not individual objects. Therefore it cannot distinguish between two sequences of length  $n$  generated on binary alphabet  $\{0,1\}$  in which one of them consists of just ones and the other corresponds to the tossing of a fair coin. It turns out that realization of randomness is very challenging and this concept resists theoretical investigation. Historically scientists have tried to define it in a specific domain. Fortunately one can observe the source of problem as a main thread in all of these domain specific approaches.

Von Mises [3] was the first person who tried to define randomness mathematically based on an intuitive aspect of unpredictability. He described randomness as an inability to predict the elements of an infinite binary sequence over  $\{0,1\}$  with probability better than  $\frac{1}{2}$  while the elements in the string are chosen randomly. Then he tried to improve his definition by replacing the random selection of elements with an acceptable selection rule. Evidently this change did not increase the mathematical clarity of his definition. Subsequently Wald [4, 5] introduced the notion of countability of selection function in order to make this definition more clear. Finally Mises's definition was refined in the light of computability of selection functions by Church [6]. Therefore acceptable selection rules were replaced by computable functions and as such the theoretical realization of randomness integrated with the notion of computability.

This type of evolved definition is known as Mises-Wald-Church definition [7]. Although Mises-Wald-Church definition of randomness was criticized by other thinkers like Ville [8], it kept its theoretical effect on subsequent works in this area.

The relation between randomness and computability has been deepened in the light of modern interpretation of defining and detecting randomness relatively to the amount of computational sources which have been used [9]. The notion of randomness is measured in its modern formulation for finite objects by Kolmogorov complexity [10], [11] which is the result of Solmonof, Kolmogorov and Chaitins theory [12] and for infinite objects by the Mrtin-Lof measure of randomness [13], [14].

These new definition of randomness are deeply connected to the theory of computation. The Kolmogorov complexity of a string  $x$  is defined as the size of the shortest program that produces it. Obviously random strings must have higher Kolmogorov complexity due to their incompressibility. Therefore the Kolmogorov complexity of random string  $x$ , denoted by  $C(x)$  would be approximately equals to length of string  $x$ . Formally  $C(x) \approx |x|$  and random strings cannot be compressed [10]. Introducing finite random objects as incompressible objects connects the notion of randomness to the philosophical interpretation of theory which was presented for the first time by Leibniz [15].

Philosophically, theory is recognized as a compression form of statements which can describe a set of large experimental data in the shortest way. Otherwise the theory with the same size of data set, which it explains would be

useless. In section V of Discourse de Metaphysique, Leibniz explains the comprehensibility of the world as the result of God's creation, in which the greatest possible diversity of phenomena are controlled by the smallest set of ideas. Today this fact can be rephrased in the language of algorithmic randomness. It means in spite of this apparent diversity in the world the set of rules which are responsible for all these phenomena is small. Therefore we can follow a scientific method to discover the theory which explains a set of wide natural phenomena. Based on this perspective the set of random data would be theory-less. Chaitin's  $\Omega$  number is recognized as an incompressible sequence of zeros and ones. This number is the probability of halting for program  $P$  generated by a successive independent tosses of a fair coin. Mathematically  $\Omega$  is defined by (1)[16].

$$\Omega = \sum_{P \text{ halts}} 2^{-(\text{size in bits of } P)} \quad (1)$$

$\Omega$  is an example of algorithmically irreducible or random string. The bits of  $\Omega$  cannot be compressed. In other words, its bits are true for a reason not simpler than itself. This breaking of Leibniz's famous principle of sufficient reason is one of the most controversial aspect of randomness in philosophy. It must be mentioned that Chaitin's result about halting probability is in fact another type of Godel's famous Incompleteness theorem [16]. Today we know that proving the randomness of a finite string  $x$ , is an example of incompleteness phenomena.

The computability of selection function in the Mises-Wald-Church definition of randomness for finite objects can be observed in terms of effective measure for continuous objects in Martin-Lof definition [14]. A real number  $x$  is considered random, if  $x$  is not contained in any event of effective measure zero [14].

The unpredictability as an intuitive notion of randomness plays a great role in other approaches for defining randomness. Ville [8] tried to define randomness by gambling approach. In his definition, it is impossible to win an infinite amount of money by betting on the bits of a random binary sequence. The amalgamation of the theory of randomness and the theory of computation has provided the opportunity of using theoretical apparatuses developed in the theory of computation to analyze the randomness in a deeper way.

Today we know the set of strings which are random in the sense of Kolmogorov complexity is not even computably enumerable and because of that, the statement of the form:  $x$  is a random string, is not provable [16].

In addition to the previous results, randomness in its modern formulation is considered relatively. In the light of the theory of computability, we interpret randomness in contrast to the amount of computational sources, used to detect it [17].

Although randomness has been always seen as a sign of complexity especially in the lack of causal model, in recent years, it has been realized that randomness is responsible for emergence of many complex phenomena by providing the opportunity of having interactions between many agents in systems. There is a tendency to disentangle randomness from complexity. For a recent work on this please refer to [18]. It is

believed that complexity in its own true form is the result of directed interactions between elements of system.

Randomness and its mysterious aspects have played a great role especially in fertilizing the multidisciplinary studies at the crossroad of mathematics, physics and computer science. Using random resources for solving hard problems in Randomized Algorithms has been very advantageous [19] both in substantial reduction of time complexity and also in deepening of our understanding of the nature of hard problems.

After the seminal work of Russel impagliazzo and Avi Wigderson about the tradeoff between hardness and randomness [20], today we know the importance of having randomized algorithm for solving hard problems as a theoretical key for designing deterministic algorithms.

There is no doubt that nature uses randomness extensively in the evolution. In recent years, Gregory Chaitin has started working on Metabiology [21]. He is studying the random evolution of artificial software in order to realize Busy Beaver function as a fitness function. The soul of his technique has been based on applying randomness as it has been applied in evolution in terms of mutation.

This paper has been organized into 6 sections. After this introduction about randomness, Random generators and their categorization and applications will be reviewed in Section 2. In Section 3, Cellular Automaton is formally defined and its application as Random generator will be discussed. Primary information Theoretic measures are explained in Section 4. These measures are used to analyze our inhomogeneous ECA60 as a Random Generator in Section 5 and finally in Section 6, concluding remarks will be presented.

## II. RANDOM GENERATOR

The difficulties of giving a complete definition of randomness were discussed briefly in the previous section. Randomness in its modern formulation is considered as a relative concept. Therefore Random Generators must satisfy criteria which guarantee their efficiency for the specific application.

The success of simulations which are based on Monte Carlo method is highly dependent on the quality of their random generators [22]-[25]. The security of information transfer on internet and the cryptography need random generators [26]. Randomness and Random generators are used extensively in solving hard problems in the framework of stochastic optimization and naturally inspired algorithm in order to bypass the problem of getting stuck in local extremes [27]-[29].

Random generators play a key role in many techniques of program validation [30] and Machine learning [31]. The rational behavior in strategic zero sum games needs to use randomness and Random Generators to mislead the opponent [32]. In recent years, Random Generators are used for Analyzing and simulation of interactions in complex social, economical and political systems at different scales [33]-[35].

Historically three approaches have been followed for generating randomness although there are many controversial

issues about the possibility of producing intrinsic randomness. True Random Number Generators (TRNG) use physical phenomena with inherent stochastic mechanism for generating random numbers [36]. There are many phenomena which can be used for TRNGs, for example unstable nuclear decay processes [37], cosmic background radiation [38], quantum based systems [39]-[41] and more exotic examples like superconducting nanowires and Josephson junctions near superconducting critical current [42].

The randomness of physical source can be amplified in some cases [43] provided that fundamental theoretic limits are preserved [44], although Santha and Vazirani proved that randomness amplification is impossible using classical resource [45].

The next family of generators is called Random Number Generators (RNG). In these generators randomness is transformed from a priori distribution in source to the desired posterior distribution [46], therefore these generators have an access to sources of randomness.

RNGs are categorized to three main groups: Von Neumann RNG [47] in which an identically independent priori distribution is transformed to the unbiased random numbers. Knuth and Yao RNG [48], in which an identically independent prior distribution is transformed to any desired distribution in output and finally Roche and Hoshi RNG [49], [50] which transform an arbitrary random distribution in source to an arbitrary distribution in output.

The third family of generators for generating randomness is called Pseudo-Random Number Generators (PRNG) in which arithmetical methods are used in order to produce randomness. Although John Von Neumann once said [47] "Anyone who considers arithmetical methods for producing random digits is, of course, in the state of sin", Chaos theory shows how randomness can emerge from deterministic systems, while we have a hypersensitive dynamic to the initial states [51], [52].

Therefore, in the light of Chaos Theory, there is a deep dichotomy between order and randomness. This fact provides the opportunity of applying deterministic algorithms for generating randomness. Hence, PRNGs try to generate randomness without having access to any source of randomness [53]-[56].

### III. CELLULAR AUTOMATA AND RANDOM GENERATORS

Cellular Automata were created by John Von Neumann, in his attempt to create a self-replicating machine [57]. He tried to show that, these machines are universal constructors and can generate even themselves. Cellular Automata have found a better place in theoretical studies when Stephen Wolfram published his book, called A New Kind of Science [58]. In this book he tried to present an extensive analysis of these systems and their effectiveness to realize a wide range of natural phenomena which are common to exhibit a particular type of behavior known as Emergent behavior.

Elementary Cellular Automata (ECA) is defined on alphabet set  $A = \{0,1\}$  as 1-dimensional cellular array of size  $N$ . The state of each cell  $i$  at time  $t$  is denoted by  $s_i^t \in A$ . The global state or lattice configuration of Cellular Automaton

at time  $t$  is represented by  $S^t$ , where  $S^t = (s_0^t, s_1^t, \dots, s_{N-1}^t) \in A^N$  ( $N$  is the size of lattice) [59].

All cells in the lattice are updated according to local update function  $f$  which generally has  $2r + 1$  arguments, where  $r$  is the radius of local function ( $r = 1$  in Elementary Cellular Automata). Formally  $s_i^{t+1}$  is defined by local function  $f$ .

$$s_i^{t+1} = f(s_{i-r}^t, \dots, s_i^t, \dots, s_{i+r}^t) \quad (2)$$

Applying  $f$  to all cells simultaneously, leads to the formation of next global state and maps  $S^t$  to  $S^{t+1}$  under the action of global function  $F: A^N \rightarrow A^N$ . It has been conjectured that applying very simple local function  $f$  at micro scale can give rise to a very complicated behavior in macro scale [58]. This highly interesting behavior is called Emergence and has inspired an extensive type of studies [60].

Cellular Automata were used for the first time as Pseudo-Random Generator by Wolfram [61], [62]. Actually he used the unpredictable emergent behavior of the global function in Cellular Automata resulted from simultaneous application of local function on different cells, as a main mechanism for generating randomness.

Afterwards, people tried to improve the quality of CA's random generator by using a combination of controllable cells [63], increasing the dimensionality of Cellular Automata [64]-[66], changing the neighborhood of cells [67], using Cellular Automata with additive rules [68], applying the evolutionary principles for designing Cellular Automata [69] and focusing on the parallelism associated with the evolution of global state in Cellular Automata for generating randomness [70].

Stephen wolfram in his well-known book "The New Kind of Science" categorized elementary Cellular Automata into four families [58]. The states of cells in elementary Cellular Automata are selected from simple binary alphabet set  $\{0,1\}$  and the radius of their local function  $r$  is equal to one. It is believed that in spite their simplicity, Elementary Cellular Automata (ECA) show all types of behaviors which can be observed in Cellular Automata in higher dimensions and with more complex local functions. Since the number of possible Elementary Cellular Automata is limited to  $2^{2^3} = 256$ , these systems have been examined from different perspectives [71], [72].

As an evidence for the power of ECA and the mysterious aspects of emergent behavior at macro scale in complex systems, like Cellular Automata, coordinated by simple interactions between cells at micro scales, please refer to [73] in which Emergent behavior has been used in combination with other complex system's properties for generating pseudo-randomness. It turns out that this strategy can generate randomness which does not have any dependency on the initial values of the system.

In this paper, a modified type of an Elementary Cellular Automaton (ECA60) would be used as random generator and its evolution from random initial state would be analyzed information theoretically. In the next section, primary Information measures will be introduced briefly.

#### IV. INFORMATION THEORETIC MEASURES

Our goal in this paper is to analyze the capability of a modified type of Elementary Cellular Automaton 60, the number of which is assigned according to Wolfram's rule, as a random generator. The system is initialized by a random binary sequence on alphabet set  $\{0, 1\}$ . During the evolution of Elementary Cellular Automaton, this random initial sequence of size  $N$  is transformed to other sequence of size  $N$ , while it preserves the initial randomness and simultaneously shows dynamical reversibility. It avoids building correlations in the evolution while all of the initial information at each time step is transformed to the next global state and because of this conservative behavior and dissipation-less dynamic the initial state of the system would be observed again. The periodicity of this conservative behavior is  $O(N)$ . In order to analyze this behavior, we use information theoretic measures.

Elementary Cellular Automaton with rule number 60 has been used as a random generator in which the central cell in local configurations 101,100, 011 and 010 is transformed to 1 and the central cell in the remaining four configurations 000,001,110 and 111 is transformed to 0, except for the first cell  $s_0^t$ , which is transferred without change to the next generation. As we will observe in the following section, this direct transfer of first cell is responsible for the reversibility of system. The evolution of our random generator from random initial state is shown in Fig. 1.

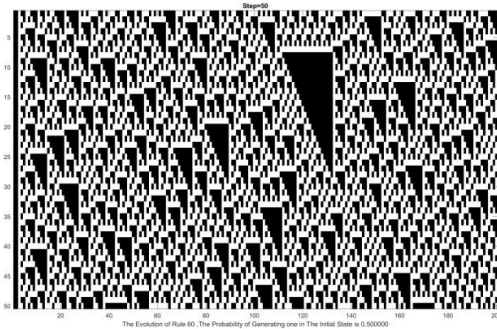


Fig. 1. The evolution of ECA60 from random initial configuration of size 200 during 50 generations.

The system is initialized by random global state, consisting of zeroes and ones (here our alphabet set is  $A = \{0,1\}$ ). Randomness implies that, starting from the beginning of system, one cannot predicate the next symbol in the sequence by observing its previous symbols. Statistically it means:

$$\lim_{i \rightarrow \infty, k \rightarrow \infty} P(x_i | x_{i-1}, \dots, x_{i-k+1}) = \frac{1}{|A|} \quad (3)$$

Here, randomness has been interpreted as a uniform distribution on the characters of alphabet which is led to the maximum possible entropy rate.

The entropy function which measures the average of uncertainty can be defined over substring of global state of length  $n$ , provided that conditions in (4) and (5) are satisfied [74].

$$P(\delta_n) \geq 0, \text{ where } \delta_n \in A^n \quad (4)$$

$$\sum_{\delta_n \in A^n} P(\delta_n) = 1 \quad (5)$$

Furthermore, it is assumed that sequences are generated by the stationary stochastic process. The Ergodicity of generator is preserved due to the nature of local function in Elementary Cellular Automata [74].

These substrings are in fact the microstates of the system upon which macro state  $S^t$  is built. Technically, the famous method of describing macro state in terms of probabilistic mass function over its constituent microstates in statistical physics has been applied. Therefore, the probabilistic description of the system's global state called  $P_n = \{p_i\}_{i=1}^n$ , defines the Probability Mass Function (PMF) over system's microstates.

Having in mind such probabilistic description of the system, the entropy of system is defined by (6) provided  $n \rightarrow \infty$  [75].

$$S_n = S[P_n] = \sum_{\delta_n} P(\delta_n) \log_2 \frac{1}{P(\delta_n)} \quad (6)$$

Actually  $S_n$  quantifies the disorder of  $n$ -length subsequences of the global string  $S^t$  at generation  $t$ . Usually  $S_n$  increases as  $n$  (the size of substrings or microstates) grows, but the action of local function  $f$  makes certain correlations, which mitigates the growth of disorder proportional to the length of microstates. The maximum entropy associated with each character which is selected from alphabet set  $A$ , is  $\log_2 |A|$  and consequently in the case of pure randomness, the maximum entropy of  $n \log_2 |A|$  for sequence  $\delta_n$  of length  $n$  over alphabet set  $A$  is expected to be observed.

The correlation can be detected as the result of  $S_n - S_{n-1}$  becomes less than  $\log_2 |A|$  (the maximum amount of entropy per symbol when the characters are chosen from set  $A$ ). But this reduction ( $S_n - S_{n-1}$ ) may be the result of correlations which have been built at much lower lengths. In order to detect the correlation specifically at length  $n$ , one can compare the consecutive gaps in entropy, computed on blocks of lengths  $n-2, n-1$  and  $n$ . Consequently, correlation at length  $n$  in (7) is calculated.

$$K_n = (S_{n-1} - S_{n-2}) - (S_n - S_{n-1}) = \Delta S_{n-1} - \Delta S_n \quad (7)$$

All correlations which have been formed at lengths less than  $n$  are considered in  $\Delta S_{n-1}$ , hence if there is correlation at length  $n$ ,  $\Delta S_n$  must be less than  $\Delta S_{n-1}$  and their difference is purely related to  $K_n$  (Correlation at length  $n$ ). Although there are different approaches for calculating correlations, one can observe the simple phenomenon which produces it at specific distance. Formally the emergence of correlation at length  $n$  requires (8) to be met.

$$P(x_n | x_{n-1}, \dots, x_1) > P(x_n | x_{n-1}, \dots, x_2) \quad (8)$$

The maximum amount of entropy per symbol is decomposed to the entropy rate of the system and all the

correlations which have formed at different lengths. Mathematically it means [74]:

$$S_{max} = \log_2 |A| = \Delta S_\infty + \sum_n K_n \quad (9)$$

#### V. ANALYZING THE RANDOM GENERATOR

The modified version of ECA60 has been used as a Random Generator which transforms an initial random sequence into another random sequence in the next generation. This transformation is done against our normal expectation to observe the increment of regularity in the sequence, due to the applying of deterministic local function. Fortunately the system is able to preserve the initial randomness during its evolution in spite of having a considerable hamming distance between consecutive global states of the system. It means that, this conservation of randomness would not decrease the activity of system.

As Lindgren [74] showed in Elementary Cellular Automata, due to the deterministic nature of local function, we expect the global entropy to be decreased during the evolution of system. Mathematically it means  $\Delta_t S(t) = S(t+1) - S(t) \leq 0$ . this fact can be realized intuitively, since applying local function  $f: \{0,1\}^3 \rightarrow \{0,1\}$  in Elementary Cellular Automata is usually accompanied with the omission of variations in the system's global state and would force system to converge to the very small subset of all possible global states. Obviously moving toward regularity and reduction of variations would lead to decrementing of the initial randomness. This behavior is not suitable for our purpose. On the other side it can be shown that  $\Delta_t S(t)$  is zero for systems with conservative nature [74].

It is not hard to realize that  $\Delta_t S(t)$  would be zero for almost reversible systems, i.e. entropy is constant during the evolution. A Cellular Automaton with rule R and range r is called almost reversible if R can be decomposed in the following way [74]:

$$R(x_1, x_2, \dots, x_{2r+1}) = f(x_1, x_2, \dots, x_{2r}) + x_{2r+1} \text{ mod } 2 \quad (10)$$

Or

$$R(x_1, x_2, \dots, x_{2r+1}) = x_1 + f(x_2, x_3, \dots, x_{2r+1}) \text{ mod } 2 \quad (11)$$

It means R is one to one if one can support it by giving it the information about its first or last argument. It is easy to show that our random generator is an almost reversible rule since its local function can be written as:

$$f(x_{i-1}^t, x_i^t, x_{i+1}^t) = x_i^t \text{ XOR } x_{i-1}^t = (x_{i-1}^t + x_i^t) \text{ mod } 2 = x_i^{t+1} \quad (12)$$

Therefore  $f$  is reversible if we have an access to the value of  $x_{i-1}^t$  or  $x_i^t$ . In order to recover the global state at time  $t$  from the global state at time  $t+1$ , we can use the following equation inductively for  $= 1, 2, \dots, n-1$ , when  $n$  is the size of system, if we have an access to the first bit of global state at time  $t(x_0^t)$ .

$$x_i^t = (x_{i-1}^t + x_i^{t+1}) \text{ mod } 2 \quad (13)$$

Please remember that our Random Number Generator is an inhomogeneous type of ECA60 in which the state of first cell in the system is transformed to the next generation without change. In fact, having access to this single bit from the previous generation makes us able to recover all the bits of previous state by applying (13), iteratively. Generally in Almost Reversible Rule with range  $r$ , one can recover the previous state by having access to the  $2r$ -bits of the previous generation [74].

As a matter of fact we are taking the advantage of emergence to produce next random sequence from the current one. This emergence here manifests itself as complex global function induced by simultaneous applying of local function  $f$  to all cells in the system. Generally there is no mathematical method to find the relation between global function  $F$  and local function  $f$  in such systems.

Conservation of information in this Random Generator which is led to the conservation of initial randomness due to the (9) is related to the global reversibility of Cellular Automata.

Studies by Hedlund [76] and Richardson [77] have shown that for Cellular Automaton A, A is injective if it is invertible (Reversible). Therefore reversibility in CA is equivalent to the injectivity of its global function. Furthermore A is injective if it is surjective and there is a close relation between injectivity and surjectivity of the Cellular Automaton A.

Injectivity and Surjectivity for global functions of CA's were studied by Moore [78] and Myhill [79] for the first time in the way of analyzing Garden of Eden (a global state without predecessor). Studies about the reversibility of dynamical systems have attracted many attentions on the part of scientists. We know that physical world at micro scale is governed by reversible rules. Surprisingly what can be observed at macro scale is irreversible. It is believed that this irreversibility at macro scale motivated by reversible rules at micro scale can be interpreted as a sign of Emergence in the system [60].

In addition to that, studies about Reversibility has found an extra importance after the seminal work of Landauer [80]. He found a relation between consuming energy and irreversibility in dynamical systems. In other words, he showed that dissipation of heat in the system which is the main source of consuming energy has its root in erasing information during the process in the system. Erasing information happens in irreversible systems, in which one cannot recover the information of previous state by having the information of current state. Although it is not hard to imagine that every irreversible process would be accompanied by erasing information, Landauer [80] predicted the minimum amount of energy dissipation associated with erasing one bit. Recently his prediction has been verified experimentally [81]. It has been shown by Bennett [82] that it is possible to do any computation reversibly. It means we have a reversible analogue for Universal Turing Machine. Then people started to analyze the advantages of reversibility in computing system [83]. It is exciting to know that in principle, reversible computation can be done with zero energy consumption. Many fundamental



limits to the computation process were realized better in the light of these studies [84]. Due to the importance of reversibility, many algorithms have been developed in order to detect it in the variety of dynamical systems [85], [86]. It has been shown that deciding the reversibility or its equivalent property in systems with dimension higher than one is impossible [87], [88].

In spite of conservative nature of this Random Generator, the initial random configuration is transformed into another random configuration with considerable hamming distance with previous global state. Furthermore the normality of random binary sequence is kept during the evolution of system. Let's look at the definition of normality and normal sequence.

Normality demands the balanced form of appearance for all patterns which means every block of digits of the same length occurs with the same frequency when all digits in the expansion are considered. Normality can be interpreted in the light of information theoretic measures. Although there are other similar concepts which are related to notion of normality, one has to know that random sequences must be normal in order to satisfy their expected unpredictability.

It is expected that unpredictability demonstrates itself as our inability to forecast the dominant frequency of observing specific pattern. Normality can be described with the aid of other measures like block complexity, block entropy, etc.

Definition [89]: the block complexity of a sequence with values in a finite alphabet is the function  $k \rightarrow P(k)$ , where  $P(k)$  is the number of different blocks of length  $k$  that occur in the sequence.

Clearly for a sequence over alphabet set  $A$  with length  $k$ , block complexity satisfies the (14).

$$1 \leq P(k) \leq |A|^k \tag{14}$$

For normal sequences, maximum block complexity is expected. Obviously low block complexity would not be seen in random sequence and can be responsible for generating periodic behavior. It is interesting to realize how block complexity as a local measure for sub-sequences can be used to predict the global behavior of the sequence. Morse and Hedlund proved the following theorem about this relation.

Theorem [90]: if the complexity of a sequence satisfies  $\exists k \geq 1 P(k) \leq k$ , then the sequence is ultimately periodic.

The block complexity can easily be related to block entropy. In section 4, the block entropy for the global state of a Cellular Automaton was defined when the process responsible for generating it is ergodic. When the block complexity for blocks of length  $k$  over alphabet set  $A$  is equal to  $|A|^k$ , it can be concluded that every possible pattern of length  $k$  over this alphabet set has been generated; therefore, maximum entropy would be expected. There is a simple relation between block complexity and block entropy. Considering their definition we can simply reach to (15).

$$H(B_n) = S_n = \frac{\log P(B_n)}{\log |A|} \tag{15}$$

Here  $A$  is a binary set therefore (14) can be simplified into  $H(B_n) = \log P(B_n)$ . As we discussed before, the entropy rate would be calculated when the length of block tends to infinity and all correlations are considered. Thus we have:

$$h = \lim_{n \rightarrow \infty} \frac{H(B_n)}{n} = \lim_{n \rightarrow \infty} \frac{\log P(B_n)}{n \log(|A|)} = \lim_{n \rightarrow \infty} \frac{\log P(B_n)}{n} \tag{16}$$

When the length of block tends to infinity, the correlations among symbols at different lengths exhibit themselves as a restriction of freedom for choosing among the possible  $|A|$  characters of alphabet. Therefore in the case of having correlations, the block complexity is not increased proportionally to the length of the block and the normality is broken. In other words for a sequence  $\delta_n$  of length  $n$  with correlations among its characters we have  $P(\delta_n) < |A|^n$ .

Since the correlation at specific length  $n$  is accompanied by the reduction of growth in block complexity at length  $n$  in contrast to its growth at length  $n - 1$  it can be detected when the entropy rate is decreased at some specific length. In Fig. 2, the entropy rate over blocks of size 10 for a system of size  $10^6$ , over 100 generation has been shown.

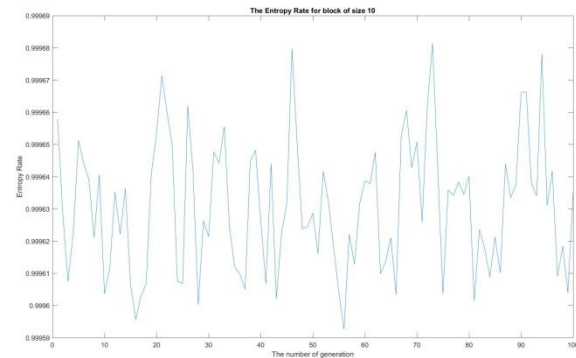


Fig. 2. The entropy rate for system of size  $10^6$  over blocks of size 10 for 100 consecutive generations.

The entropy rate shows fluctuations in the scale of  $10^{-5}$  over generations. Furthermore the lower value of entropy rate for blocks of size 10 is bigger than 0.99 which approves of the random nature of sequences generated by this random generator since the maximum entropy rate for our system is  $\log |A| = \log 2 = 1$ .

In this Random Generator, the block complexity is much more than the lower threshold predicated by Morse & Hedlund [90], [91] and it is around  $2^{nh_\mu} \approx 2^{9.9} \gg 10$ . Surprisingly here in spite of keeping considerable local variations and nearly maximum block complexity, the global state of system shows periodicity due to its conservative nature of system and the trajectory of global state starting from initial configuration meets the initial state again. Fortunately in its entire route from initial state, the normality at different lengths is preserved because the entropy rate does not depict considerable difference calculated on blocks of different sizes. In Fig. 3, the entropy rate for system of size  $10^6$  over 200 generation has been displayed when the entropy rate is calculated over blocks of different lengths ranging from 5 to 10.

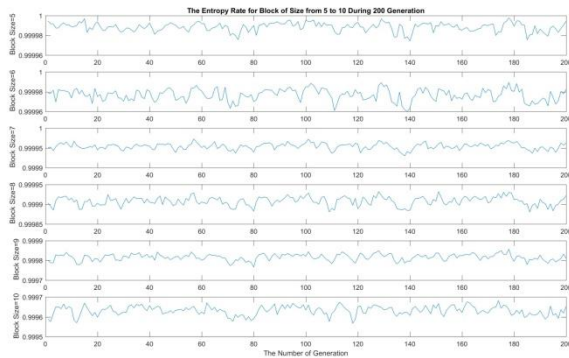


Fig. 3. The entropy rate for system of size  $10^6$  over blocks of different lengths ranging from 5 to 10 during 200 generations.

Fortunately the conservation of information does not impose restriction on the micro-dynamic of system and consecutive global states have considerable hamming distance. Please see the Fig. 4.

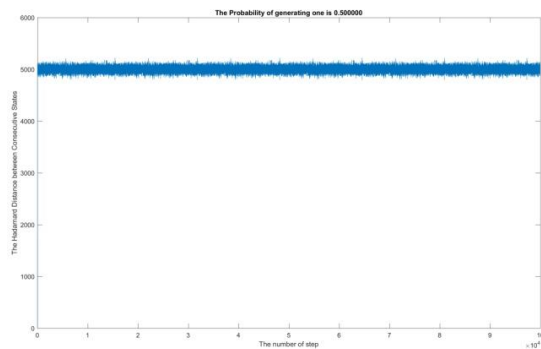


Fig. 4. the hamming distance between consecutive generations for system of size 10000 during  $10^5$  generations.

## VI. DISCUSSION AND CONCLUSION

The work of this Random Generator can be analyzed from different perspectives such as the Conservation of initial information during the evolution leads to reversible dynamic, the intrinsic parallelism (as the result of simultaneous updates of cells in the system), the efficient use of initial randomness and the ability of generating acceptable number of sequences which are equivalent to the initial configuration considering their randomness or their entropy rate. In other words this Random Generator is able to show varied behavior while it keeps its initial information during its evolution.

Taking into account (9) as the Hamiltonian of this Random Generator, the maximum amount of entropy rate of the system can be decomposed into the entropy of the system or its random parts plus its regular parts which are the result of summing all of its correlations at different lengths. Actually in this system, the random part of the current global state is mapped into the random part of the next global state. The fluctuations in the entropy rate calculated over blocks of different lengths are negligible and can be the result of encoding randomness from one length into another. Furthermore due to its avoidance of building correlations or

empowering them, the Normality of generated sequences is preserved during the evolution at all scales.

Preservation of Normality at all scales and the conservative nature of this Random Generator, in spite of having considerable micro-activity turn this system into remarkable Random Generator. In addition, this Random Generator gets the benefits of intrinsic parallelism and proves to be technically easy to implement due to its simple logical local function.

## REFERENCES

- [1] H. H. Holger, S. Thomas, Stochastic Local Search: Foundations and Applications, Elsevier, 2005.
- [2] J. Garey, M. R. Garey, D. S. Johnson, Computers and Interactability, Freeman Press, 1979.
- [3] M. R. Von, "Grundlagen der Wahrscheinlich Keit Srechung," Mathematische Zeit schrift, 5, pp. 52-99, 1919.
- [4] A. Wald, "Sur La Notion de Collectif Dans La Calcul Des Probabilities," Comptes Rendus Des Seances De l'Scademie Des Sciences, 202, pp. 180-183, 1936.
- [5] A. Wald, "Die Widerspruchsfreiheit Des Kollektiv Begriffes Der Wahrscheinlich keitsrechnung," Ergebnisse Eines Mathematischen Kolloquiums, 8, pp. 38-72, 1937.
- [6] A. Church, "On the Concept of Random Sequence," Bulletin of the American Mathematical Society, 46, pp. 130-135, 1940.
- [7] S. A. Terwijn, "The Mathematical Foundations of Randomness," In the Challenge of Chance, K. Landsman and E. Van Wolde(Editors), pp. 49-66, Springer, 2016.
- [8] J. Ville, "Etude Critique De La Notion De Collectif, Monographies Des Probabilities," Calcul Des Probabilites et Ses Applications, Gauthier-Villars, 1939.
- [9] A. Nies, Computability and Randomness, Oxford University press, 2009.
- [10] M. Li, P. Vitanyi, An Introduction to Kolmogorov Complexity and Its Applications (3<sup>rd</sup> Edition), Springer, 2008.
- [11] G. Chaitin, "On the Length of Programs for Computing Finite Binary Sequences," Journal of ACM, 13(145), 1966.
- [12] H. Zenil, Randomness through Computation Some Answers More Question, World Scientific, 2011.
- [13] R. G. Downey, D. R. Hirschfeldt, A. Nies, S. A. Terwijn, "Calibrating Randomness," Bulletin of Symbolic Logic, vol. 12(3), pp. 411-491, 2006.
- [14] P. Martin-Lof, "the Definition of Random Sequences, Information and Control," vol. 9, pp. 602-619, 1966.
- [15] G. W. Leibniz, Discourse de Metaphysique, Svide Monadologie, Gallimard, 1995.
- [16] G. Chaitin, Thinking about Godel and Turing: Essays on Complexity, World scientific, 2007.
- [17] S. Arora, B. Boaz, Computational Complexity: A Modern Approach, Cambridge University Press, 2009.
- [18] L. Zuchowski, "Disentangling Complexity From Randomness and Chaos," Entropy, 14, pp. 177-212, 2012.
- [19] J. Hromkovic, Design and Analysis of Randomized Algorithms: Introduction to Design Paradigms, Springer, 2005.
- [20] R Impagliazzo, A. Wigderson, "P=BPP if E Requires Exponential Circuits: Derandomizing the Xor Lemma," in Proceeding of the 29<sup>th</sup> Annual ACM Symposium on the Theory of Computing, pp. 220-9, 1997.
- [21] G. Chaitin, Proving Darwin: Making Biology Mathematical, Pantheon Book, 2012.
- [22] J. Gentle, Random Number Generation and Monte Carlo Methods, 2<sup>nd</sup> Edition, Springer, 2003.
- [23] PL. Ecuyer, "Random Numbers for Simulation," Communication of ACM, vol. 33, pp. 85-97, 1990.
- [24] R. Y. Rubinstein, D. P. Kroese, Simulation and the Monte Carlo Method, John Wiley & Sons, 2011.

- [25] J. E. Gentle, *Random Number Generation and Monte Carlo Methods*, Springer, 2013.
- [26] D. R. Stinson, *Cryptography: Theory and Practice*, CRC Press, 2005.
- [27] J. Carmelo, A. Bastos-Filho, D. Jaulio, D. Andrade, R. Marcelo, S. Pita et al. "Impact of the Quality of Random Numbers Generations on the Performance of Particle Swarm Optimization," In *IEEE International Conference on Systems, Man and Cybernetics*, pp. 4988-93, 2009.
- [28] A. Reese, "Random Number Generators in Genetic Algorithms for Unconstrained and Constrained Optimization," *Nonlinear Analysis Theory Methods & Application*, vol. 71, pp. 679-92, 2009.
- [29] A. Brabazon, M. O'Neill, S. McGarraphy, *Natural Computing Algorithms*, Springer, 2015.
- [30] R. G. Sargent, "Verification and Validation of Simulation Models," In *Proceeding of the 37<sup>th</sup> Conference on Winter Simulation*, pp. 130-143, 2005.
- [31] E. Alpaydin, *Introduction to Machine Learning*, MIT Press, 2014.
- [32] J. H. Conway, *On Numbers and Games*, Volume 6, IMA, 1976.
- [33] J. M. Epstein, *Generative Social Science: Studies in Agent-Based Computational Medeling*, Princeton University Press, 2006.
- [34] J. M. Epstein, *Agent-Zero: Toward Neurocognitive Foundation for Generative Social Science*, Princeton University Press, 2013.
- [35] O. Dowlen, *The Political Potential of Sortition: A Study of the Random Selection of Citizens for Public Office*, Volume 4, Andrews UK Limited, 2015.
- [36] C. D. Motchenbacher, F. C. Fitchen, *Low-Noise Electronic Design*, John Wiley & Sons, 1973.
- [37] G. F. Knoll, *Radiation Detection and Measurement*, John Wiley & Sons, 2010.
- [38] N. Yoshida, R. K. Sheth, A. Diaferio, "Non-Gaussian Cosmic Microwave Background Temperature Fluctuation from Peculiar Velocities of Clusters," *Monthly Royal Astronomy Society*, vol. 328(2), pp. 669-677, 2001.
- [39] T. Jennewein, U. Achleitner, G. Weihs, H. Weinfurter, A. Zeilinger, "A Fast and Compact Quantum Random Number Generator," *Rev. Sci. Instr.*, vol. 71(4), pp. 1675-1680, 2000.
- [40] A. Stefanov, N. Gisin, O. Guinnard, L. Guinnard, H. Zbinden, "Optical Quantum Random Number Generator," *Journal of Modern Optics*, vol. 47(4), pp. 595-598, 2000.
- [41] A. Acin, L. Masanes, "Certified Randomness in Quantum Physics," *Nature*, vol. 540, pp. 213-219, 2016.
- [42] M. Foltyn, M. Zgirski, "Gambling with Super-Conducting Fluctuations," *Physical Review Application*, vol. 4(2), 024002, 2015.
- [43] R. Colbeck, R. Runner, "Free Randomness Can be Amplified," *Nature Physics*, vol. 8, pp. 450-454, 2012.
- [44] S. Pironi et al., "Random Number Certified by Bell's Theorem," *Nature*, vol. 464, pp. 1021-1024, 2010.
- [45] M. Santa, U. V. Vazirani, "Generating Quasi-Random Sequences from Semi-Random Sources," *Journal of Computing System Science*, vol. 33, pp. 75-87, 1986.
- [46] L. Devroye, "Sample-Based Non-Uniform Random Variable Generation," In *Proceeding of the 18<sup>th</sup> Conference on Winter Simulation*, ACM, pp. 260-265, 1986.
- [47] J. V. Neumann, "Various Techniques Used in Connection with Random Digits," *National Bureau of Standards Applied Math Series*, pp. 36-38, 1963.
- [48] D. E. Knuth, A. C. Yao, *the Complexity of Non-Uniform Random Number Generation*, Academic Press, 1976.
- [49] J. R. Roche, "Efficient Generation of Random Variables from Biased Coins," *Proceeding of IEEE international Symposium in Information Theory*, pp. 169, 1991.
- [50] M. Hoshi, "Interval algorithm for Random Number Generation," *IEEE Transaction on Information Theory*, vol. 43(2), pp. 599-611, 1997.
- [51] A. Kanso, "Search-Based Chaotic Pseudo-Random Bit Generator," *International Journal of Bifurcation and Chaos*, vol. 19, no. 12, pp. 4227-4235, 2009.
- [52] R. Lozi, "Emerging of Randomness from Chaos," *International Journal of Bifurcation and Chaos*, vol. 22, no. 2, pp. 1250021, 2012.
- [53] B. A. Wichmann, I. D. Hill, "Algorithm AS 183: An Efficient and Portable Pseudo-Random Number Generator," *Journal of Royal Statistic Society Series C*, vol. 31(2), pp. 188-190, 1982.
- [54] L. Blum, M. Blum, M. Shub, "A Simple Unpredictable Pseudo-Random Number Genertor," *SIAM Journal of Computing*, vol. 15(2), pp. 364-383, 1986.
- [55] M. Mascagni, S. A. Cuccaro, D. V. Pryor, M. L. Robinson, "A Fast High Quality and Reproducible Parallel Lagged Fibonacci Pseudo-Random Number Generator," *Journal of Computational Physics*, vol. 119(2), pp. 211-219, 1995.
- [56] J. K. Salmon, M. A. Moraes, R. O. Dror, D. E. Shaw, "Parallel Random Numbers: As Easy As 1 2 3," *International Conference for High Performance Computing Networking Storage and Analysis*, IEEE, pp.1-12, 2011.
- [57] J. V. Neumann, *Theory of Self-Replicating Automata*, Edited and Completed by A. W. Burks, University of Illinois press, 1966.
- [58] S. Wolfram, *A New Kind of Science*, Wolfram Media, 2002.
- [59] N. Boccaro, *Modeling Complex Systems*, Second Edition, Springer, 2010.
- [60] B. Falkenburg, M. Morrison, *Why More is Different: Philosophical Issues in Condensed Matter Physics and Complex Systems*, Springer, 2015.
- [61] S. wolfram, "Cryptography with Cellular Automata," In *Proceeding of the CRYPTO 85, Advances in Cryptography*, vol. 218, pp. 429-32, 1985.
- [62] S. Wolfram, *Theory and Applications of Cellular Automata*, River Edge, NJ:World Scientific, pp. 1983-6, 1986.
- [63] S. U. Guan, S. Zhang, "a Family of Controllable Cellular Automata for Pseudo-Random Generation," *International Journal of Modern Physics C*, vol. 13(8), pp. 1047-73, 2002.
- [64] M. S. Tomassini, M. Sipper, M. Perrenoud, "On the Generation of High Quality Random Numbers by Two-Dimensional Cellular Automata," *IEEE Transactions on Computer*, vol. 49, pp. 1146-51, 2000.
- [65] B. Kang, D. Lee, C. Hong, "Pseudorandom Number Generation Using Cellular Automata," In *Novel Algorithms and Techniques in Telecommunication Automata and Industrial Electronics*, pp. 401-4, 2008.
- [66] S. Shin, G. Park, K. Yoo, "A Virtual Three-Dimensional Cellular Automata Pseudorandom Number Generator Based on More Neighborhood Method," In *4<sup>th</sup> International Conference on Intelligent Computing*, ICIC 2008, pp. 174-81, 2008.
- [67] S. Shin, K. Yoo, "Analysis of 2-state 3-neighborhood Cellular Automata Rules for Cryptographic Pseudorandom Number Generation," In *International Conference on Computational Science and Engineering*, CSE 2009, pp. 399-404, 2009.
- [68] P. Chaudhuri, D. Chowdhury, S. Nardi, S. Chattopadhyay, *Additive Cellular Automata: Theory and Application*, vol. 1, IEEE Computer Society Press, 1997.
- [69] S. Guan, S. Zhang, "An Evolutionary Approach to the Design of Controllable Cellular Automata Structure for Random Generation," *IEEE Transaction on Evolutionary Computing*, vol. 7, pp. 3-6, 2003.
- [70] P. Hortensius, R. Mcleod, H. Card, "Parallel Random Number Generation for VLSI system Using Cellular Automata," *IEEE Transaction on Computing*, vol. 38, pp. 1466-73, 1989.
- [71] R. Dogaru, I. Dogaru, H. Kim, "Synchronization in Elementary Cellular Automata," In *Proceedings of the 10<sup>th</sup> International Workshop on Multimedia Signal processing and Transmission*, CMSPT 2008, pp. 35-40, Jeonju, Korea, July 21-22, 2008.
- [72] R. Dogaru, I. Dogaru, H. Kim, "Binary Chaos Synchronization in Elementary Cellular Automata," *International Journal of Bifurcation & Chaos*, vol. 19, pp. 2871, 2009.
- [73] S. M. Hossseini, H. Karimi, M. Vafaei Jahan, "Generating Pseudorandom Numbers by Combining two Systems with Complex Behaviors," *Journal of Information Security and Applications*, vol. 9, pp. 149-162, 2014.

- [74] K. Lindgren, *Information Theory for Complex Systems*, Chalmers University Press, 2014.
- [75] J. T. Cover, J. A. Thomas, *Elements of Information Theory*, Wiley, 2006.
- [76] G. Hedlund, "Endomorphisms and Automorphisms of the Shift Dynamical System," *Mathematical System Theory*, vol. 3, pp. 320-375, 1969.
- [77] D. Richardson, "Tesselations with Local Transformations," *Journal of Computing System Society*, vol. 6, pp. 373-388, 1972.
- [78] E. Moore, "Machine Models of Self-Reproduction," *Proceeding of Syposia in Applied Mathematics*, American Mathematical Society, vol. 14, pp. 17-33, 1962.
- [79] J. Myhill, "The Converse of Moore's Garden of Eden Theorem," *Proceeding of American Mathematical Society*, vol. 14, pp. 658-686, 1963.
- [80] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM Journal of Research & Development*, vol. 5, pp. 183-191, 1961.
- [81] A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, E. Lutz, "Experimental Verification of Landauer's Principle Linking Information and Thermodynamics," *Nature*, vol. 483, pp. 187, 2012.
- [82] C. H. Bennett, "Logical Reversibility of Computation," *IBM Journal of Research & Development*, vol. 17, pp. 525-532, 1973.
- [83] C. H. Bennett, "The Thermodynamics of Computation," *International Journal of Theoretical Physics*, vol. 21, pp. 905-940, 1982.
- [84] C. H. Bennett, "The Fundamental Physical Limits of Computation," *Scientific American*, vol. 253, pp. 38-46, 1985.
- [85] K. Sutner, "De Bruijn Graphs and Linear Cellular Automata," *Complex Systems*, vol. 5(1), pp. 19-30, 1991.
- [86] T. Head, "Linear CA: Injectivity for Ambiguity," *Complex Systems*, vol. 3(4), pp. 343-348, 1989.
- [87] J. Kari, "Reversibility of 2D Cellular Automata is Undecidable," *Physica D*, vol. 45, pp. 397-385, 1990.
- [88] J. Kari, "Reversibility and Surjectivity Problems of Cellular Automata," *Journal of Computing System Science*, vol. 48, pp. 149-182, 1994.
- [89] J. P. Allouche, *Algebraic and Analytic Randomness*, In *Noise Oscillation and Algebraic Randomness* Edited by M. Planet, Springer, pp 345-356, 2000.
- [90] M. Morse, G. A. Hedlund, "Symbolic Dynamics," *American Journal of Mathematics*, vol. 60, pp. 815-866, 1938.
- [91] M. Morse, G. A. Hedlund, "Symbolic Dynamics II, Sturmian Trajectories," *American Journal of Mathematics*, vol. 62, pp. 1-42, 1940.

# Requirement Elicitation Techniques for Open Source Systems: A Review

Hafiza Maria Kiran

Department of Computer Science and IT  
The University of Lahore, 1Km, Defence Road  
Lahore, Pakistan

Zulfiqar Ali<sup>1,2</sup>

<sup>1</sup>Department of Computer Science and IT  
The University of Lahore, 1Km, Defence Road  
Lahore, Pakistan

<sup>2</sup>Department of Computer Science  
National University of Computer and Emerging Sciences,  
A.K Brohi Road, H-10/4,  
Islamabad, Pakistan

**Abstract**—The trend of Open Source Software development has been increased from the past few years. It has gained much attention of developers in the industry. The development of open source software systems is slightly different from traditional software development. In open source software development, requirement elicitation is a very complex and critical process as developers from different regions of the world develop the system so it's really difficult to gather requirements for such systems. A variety of available tools, techniques, and approaches are used to perform the process of requirement elicitation. The purpose of this study is to focus on how the process of requirement elicitation is carried out for open source software and the different ways which are used to simplify the process of requirement elicitation. This paper comprehensively describes the techniques which are available and are used for requirement elicitation in open source software development. To do so, a literature survey of the existing techniques of requirement elicitation is conducted and different techniques are found that can be used for requirement elicitation in open source software systems.

**Keywords**—Requirements engineering; requirement elicitation; open source system; requirement elicitation techniques

## I. INTRODUCTION

Requirement elicitation is very first and important step in the process of the development of a software system [1]. In requirement elicitation, all the requirements related to the system which is going to be developed are gathered from stakeholders. Although it's a very critical phase to gather requirements, so it also effects on the quality of software. Most of the times, the cause of software system's failure is due to poor requirement elicitation [2].

The open source concept was started in the 1980s with a general public license model of Richard Stallman [3]. According to this model, software should be flexible enough so that it can be modifiable. In many cases, open source systems become distributed systems as the people from all around the world are involved in development. In such cases, the task of requirement elicitation is really crucial. Open source software has gained much attention in the industry. Open source software has changed the scenario in which

millions of line of code are accessible to developers and they can read, enhance and improve the source code [4].

The reputation of open source software development has been increased from past few years and the companies and organizations are focusing on the development of such projects frequently. The importance of software requirement engineering rises expressively with the rapid growth of open source software development practices, more time and required resources for development purpose. As the requirements are the critical and important base of software, it required much time and fruitful effort so that the process can be as right as possible.

The purpose of this study is to focus on how the process of requirement elicitation is carried out for open source software and the different ways which are used to simplify the process of requirement elicitation. This paper comprehensively describes the techniques which are available and can be used for requirement elicitation in open source software development. To do so, a literature survey of the existing techniques of requirement elicitation is conducted and different techniques are found that can be used for requirement elicitation in open source software systems.

Section II introduces the requirement elicitation, Section III explains the Open Source Systems, Process of Requirement Elicitation in Open Source Systems is provided in Section IV. Section V provides the review of requirement elicitation techniques for open source systems. Section VI provides the comparative analysis of the various techniques with respected open source software categories and the last Section concludes the survey study.

## II. REQUIREMENT ELICITATION

The term 'elicitation' means to obtain, gather, collect or identify. In software engineering, requirement elicitation is the process of gathering requirements from stakeholders for software development. The process of the development of every software initiates with the phase of requirement elicitation [5]. Different elicitation techniques are used for the process of requirement elicitation in which analysts note down stakeholder's wants, needs, and desires [6]. In requirement elicitation process, the analyst focus on the understanding of

requirements, vision, and constraints of the system which is going to develop [7]. The scope of the requirements is determined in requirement elicitation phase and broader requirements are defined [8].

The phase of requirement gathering is the earliest and continuous process in software development. These requirements are gathered from different sources which include existing systems, stakeholders, documentation and the problem owners [9]. In requirement engineering process, the activity of requirement elicitations is examined the most critical activity. It is a very complex process and it also involves some other activities. Multiple techniques are available to perform these activities. The process of requirement elicitation is divided into five types: to understand the application area, find out the sources of requirements, exploring stakeholders, selection of appropriate tool and technique, elicitation of the requirements from different sources.

Before starting the activity of requirement gathering, the sources of requirements are identified. These sources include documentation, existing systems interviews, etc. [10], [11].

### III. OPEN SOURCE SYSTEMS

The open source concept was started in the 1980s with a general public license model of Richard Stallman [3]. According to this model, software should be flexible enough so that it can be modifiable. In many cases, open source systems become distributed systems as the people from all around the world are involved in development. In such cases, the task of requirement elicitation is really crucial. From the developer's point of view, open source is a blend of two imperative properties [4]. The first property is the visibility and access to the source code. And the second one is the authority to make changes or enhancements to the source code. In open source software development, organizations do not pay much attention to software engineering activities like detailed requirement gathering, testing of system etc. The design of software which is going to be enhanced is totally based on the elements of existing software. Open source software development is a new, unique and different from traditional software development and the resulting product of open source software development is not the property of a single organization [6]. The importance of why to adopt open source practices is independence in the context of the price of the product and the licensing conditions. The advantages of adopting open source software development are that it ensures a high level of security, maximum stability, independence from vendors of major software. The reason of why open source software is different from traditional software is because of licensing. In open source software systems, the license fee is not required. Its lower cost is the key factor in its adoption in small businesses [3]. Open source software is flexible enough to modify it for enhancement and improvements. In terms of reliability and quality, open source software is much better than traditional software. But there are some limitations of open source software. Open source software has lack of personalized support and this software do not come with a warranty. The types of open source systems include office automation, web design, communications, E-

Commerce, content management systems and operating systems.

### IV. PROCESS OF REQUIREMENT ELICITATION IN OPEN SOURCE SYSTEMS

This is absolutely true that the process of requirement elicitation is applicable in OSS. But in the context of open source software development, the activity of requirement elicitation carried out in a different way as in traditional software development process. In OSS, requirements are presented in natural language text format rather than in a formal template [12]. Why the process of requirement elicitation is different in OSS because of the nature of the project that is distributed and a huge amount of participants like users, stakeholders, developers, and customers. Moreover, it's also different because of the informal nature of documentation and communication.

In traditional software development, there are so many contexts in which requirements can be exposed. For example, there is face-to-face requirement elicitation in the form of interviews and workshops. Moreover, there is a concept of recycling old manuals and specifications. In case of open source software development, use of these requirement elicitation techniques is slightly different. In this scenario, requirements are gathered by discussions, through email or messaging or by communication over the internet. This is the reason that these requirements are usually informal and unstructured.

### V. REQUIREMENT ELICITATION TECHNIQUES FOR OPEN SOURCE SYSTEMS

Requirements for open source systems come from multiple sources and these requirements are of different nature in some aspects as compared to traditional software systems [13]. As discussed in the previous sections that there multiple techniques for requirement elicitation in developing a software system [10], [11]. There are some requirement elicitation techniques which can be as it is used for open source systems as they are used for traditional systems for example [13]. But which requirement elicitation techniques we can use for open source software development, is a question. The elicitation techniques which are commonly used are mentioned in Fig. 1 and described briefly in further sections. Fig. 1 is shows the requirement elicitation techniques which are used for requirement elicitation of open source software systems.

#### A. Groupware Tools

In the 1980s, the term collaborative software is being initiated to selected as groupware [14]. Groupware is basically a software application which is designed with the help of those people who are engaged in a common task for the achievement of their desired goal. In 2002, a study was conducted by researchers Rosson and Lloyd in which they tried to pay attention to find the importance and effectiveness of the process and activities of requirement elicitation in open source software development [15]. According to this study, the method used for the process of requirement elicitation was Groupware Tools. Requirements for OSS were gathered or elicited by communication among groups of several stakeholders using groupware tool.

### B. Web Survey

In the 1930s, the survey was a typical and standard way for the research purpose in different fields like marketing, social sciences, etc. [16]. These surveys are used for the purpose of data collection from a sample of persons. A number of methods are involved in data collection via surveys [17]. The first method in survey data collection which was paper and pencil interviewing has been changed into computer-assisted interviewing. And the other methods like face to face survey, telephonic survey, and mail survey are progressively replaced by web surveys. In 2015, a study was conducted by researchers named as Kuriakose and Parsons in which the technique which they used for the purpose of requirement elicitation was Web Survey. This study also focused to figure out the requirement engineering practices which the developers are following the process of requirement elicitation

in open source software development. The study concluded that although the practice of requirement elicitation is much important and helpful for development process according to the survey which they conducted, the usage of RE practices in the development of OSS is low. This technique is effectively used when the target population is spread over the large geographical area [18]. Surveys should be designed in such a way that they must be clear and contain domain knowledge. Proper attention and pre-planning are required in order to make this technique successful and quick. There are some pros and cons of adopting this technique. For example, this technique is easy to gather requirement because there are multiple choice questions in the survey and same questions are asked of multiple people. This technique can only be used for general purpose software. In this technique, sometimes useful feedback is not received because of ambiguities in questions.

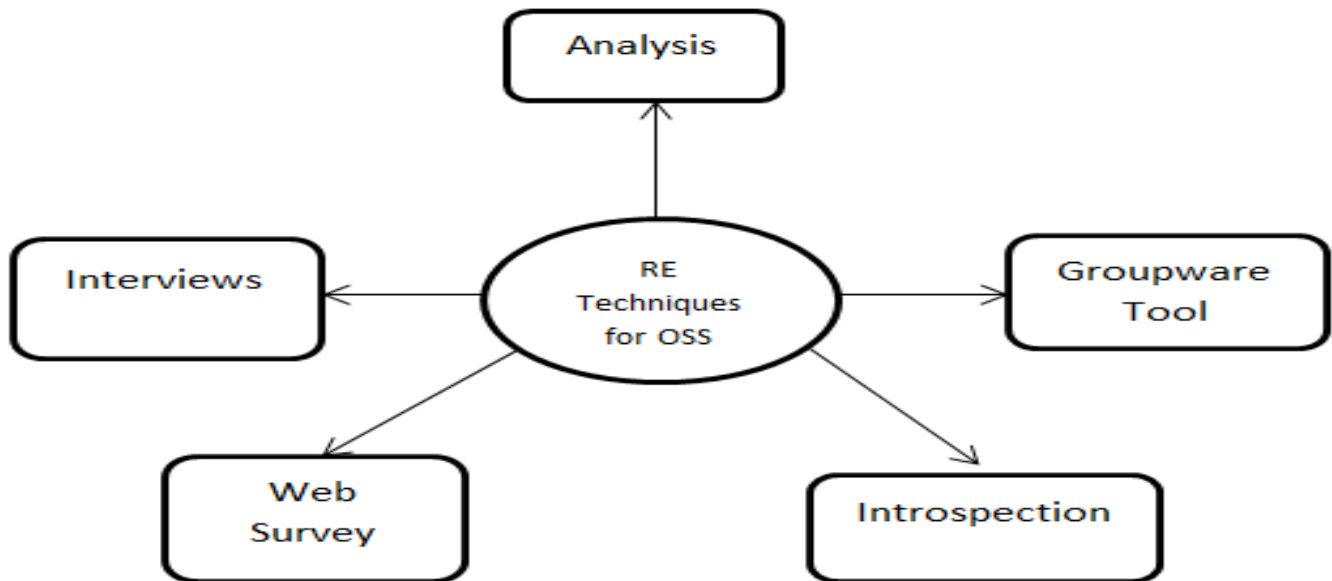


Fig. 1. RE Techniques for OSS.

### C. Interviews

In requirement elicitation process, the most common and traditional technique used is interview [19]. With the help of this technique, a huge amount of data can be collected quickly and efficiently. The feedback of using this technique highly depends on interviewer's skills. The three primary types of interviews are structured interviews, unstructured interviews, and semi-structured interviews.

Structured interviews include a collection of pre-defined questions which are asked to stakeholders [9]. This technique results in quantitative data and considered as an effective technique in requirement elicitation process. On the other hand, a semi-structured interview is a mixture of some pre-defined and some unplanned questions. It is actually a combination of structured and unstructured interviews. Whereas un-structured interviews aka open interviews include totally unplanned questions [18]. Qualitative data is produced as a result of this technique. In this technique, stakeholders and analysts discuss software system and finalize requirements. This technique is more effective when a specific

issue is going to be resolved and focus is on a deep understanding of the particular issue.

There are some pros and cons of interview technique, discussed in [20]. For example, this technique is really effective for complex topics and provides the overview of the whole system. The chances of no responsiveness are much low. A lesser number of people involved in this technique but still it is much time-consuming.

### D. Introspection

In this technique, the requirement analyst tries to develop all the requirements on the basis of needs and wants of stakeholders for a particular system [10]. This introspection technique for requirement elicitation is more efficient when the analyst is well aware of the domain and outcome of the project also has expertise in business processes which are performed by the user. This technique is used with a combination of other techniques to initiate the process of requirement elicitation [20]. This technique is basically the practice of observing the thoughts of stakeholders. In this

technique, the system analyst should be experienced and domain expert.

As they focus on how the design of the system should be. This requirement elicitation technique can be a useful technique but stakeholders and experts are from diverse fields and they may not easily understand one another. There are some pros and cons of introspection technique. For example, this technique is quite easy to implement and there is no implementation cost required for this technique. This technique can be an initial step to start the process of requirement elicitation. Stakeholders and analyst should be aware of the domain. There is no chance of discussion between stakeholders and other experts. It is difficult for the requirement analyst to understand or imagine the situation in which new system will work.

#### E. Analysis

The process of dividing a complex task into smaller chunks so that it can be easily understandable in a better way is called analysis. In software requirement engineering, the term requirement analysis is used to determine the needs and wants of stakeholders for a new system.

In this technique, information which comes from existing documents is gathered and then analyzed [20]. This requirement elicitation technique is successfully used in order to initiate the process of requirement elicitation. The collected information in this technique may diverge because of availability of documents and interaction with humans. This technique is mostly used when there is a need to have the detailed domain information studied by a domain expert. In this technique, experts usually analyze design documents, manuals and templates of existing systems. When there is a need to enhance the existing system or to replace it with another system, then this technique is mostly adopted.

There are some pros and cons of adopting this technique. For example, this technique is more helpful in a scenario when there is no availability of users and stakeholders. This technique helps us by providing some previous historical data about the system. This technique is also used when there is need for requirement reuse and it is an inexpensive technique. This technique is little bit time consuming as it is a difficult and time taking process to find out information from a big amount of documents and the information gathered from

existing documents may be incomplete. The information which is gathered can be invalid.

## VI. DISCUSSION

In 2002, a researcher claimed that the process of requirement elicitation in open source software development is not same as in traditional software development [21]. The requirement elicitation process in open source software development is actually a combination of the social and technical process which involves the positive social relationships development and social agreements or contracts which are negotiate informally. The techniques which are used in the process of requirement elicitation for open source software systems are mentioned in this paper. There are following five techniques mentioned in the paper which are groupware tools, web surveys, interviews, introspection, and analysis. These requirement elicitation techniques are used in different scenarios for the development of different open source software depending upon the nature and category of the software.

A comparison of these mentioned techniques is given in the paper which clearly shows that which requirement elicitation technique best matches with which type of software.

Table I show the relationship between requirement elicitation techniques and different scenarios in which they are used for open source software development. For the development of very complex software, the requirement elicitation technique which suits the process of requirement gathering is an *interview*. This technique is really an easy and effective technique for stating requirements, needs and wants of the user and stakeholder. The requirement analyst can have detailed specifications with the help of using different variants of this technique like structured, unstructured and semi-structured interviews regarding particular software which is going to be developed. The more questions asked by the interviewer from the stakeholder, the more it will be clear to state requirements.

If there is a scenario in which the availability of stakeholder or user is not possible then the requirement elicitation technique which suits best is *Analysis*. Requirement analysts study previous documentation and manuals regarding specific software so that they can understand the system and gather requirements for system development.

TABLE. I. REQUIREMENT ELICITATION TECHNIQUES WHICH SUIT BEST FOR MAJOR CATEGORIES OF OSS

Type of S/W RE Techniques	Complex Software	Stakeholder & User not available	Large no. of responses required within short time	No cost to RE process	Engagement of multiple stakeholders in common task
Groupware Tool					*
Interviews	*				
Introspection				*	
Analysis		*			
Web Survey			*		



For the development of that software in which the requirements for a system are gathered from a large population, *Web Survey* technique of requirement elicitation is best for this scenario. Requirement analyst simply prepares an online survey and spread it to the intended audience for a response. Same questions are asked in web survey from multiple people.

For the development of software in which there is a low budget for development and the requirement analysts do not want to spend cost for requirement elicitation, the technique which best matches for this is *Introspection*. This technique is very cheap and no implementation cost is required for this technique. It is a simple and easy technique and can be adopted as an initial step to start the process of requirement elicitation.

All the discussion and suggestion of requirement elicitation techniques are based on the literature and previous studies in this domain. So according to different researchers, following techniques which are mentioned in the paper are used for the process of requirement elicitation in open source software development.

## VII. CONCLUSION

In this paper, a detailed review of the techniques for the process of requirement elicitation in open source software development is presented. The paper described what open source software is and how it is different from traditional software. According to literature survey, all the techniques which are mentioned in this paper have some plus points and some negative points. Some techniques are applied at early stages of requirement gathering, and some techniques are applied later according to need. Every technique is specifically suitable for a particular software system in particular scenario.

## ACKNOWLEDGMENT

We are thankful to the Higher Education Commission (HEC) of Pakistan as this research work is sponsored by the HEC of Pakistan in the form of Scholarship for the Ph.D. program.

## REFERENCES

[1] H. Kitapci, and B. W. Boehm, "Formalizing informal stakeholder decisions--A hybrid method approach." pp. 283c-283c, 2007.  
[2] B. Davey, and K. R. Parker, "Requirements elicitation problems: a literature analysis," *Issues in Informing Science and Information Technology*, vol. 12, pp. 71-82, 2015.

[3] T. Jaeger, and A. Metzger, "Open Source Software," *Rechtliche Rahmenbedingungen der Freien Software*, vol. 2, pp. 51, 2011.  
[4] D. Spinellis, and C. Szyperski, "How is open source affecting software development?," *IEEE Software*, vol. 21, no. 1, pp. 28, 2004.  
[5] M. Christel, and K. Kang, *Issues in Requirements Elicitation*. Software Engineering Institute Technical Report CMU, SEI-92-TR-12. Carnegie Mellon University, 1992.  
[6] A. Davis, O. Dieste, A. Hickey, N. Juristo, and A. M. Moreno, "Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review." pp. 179-188, 2006.  
[7] S. A. Fricker, R. Grau, and A. Zwingly, "Requirements engineering: best practice," *Requirements Engineering for Digital Health*, pp. 25-46: Springer, 2015.  
[8] M. K. Niazi, "Improving the requirements engineering process through the application of key process areas approach," *AWRE02*, 2002.  
[9] D. Zowghi, and C. Coulin, "Requirements elicitation: A survey of techniques, approaches, and tools," *Engineering and managing software requirements*, pp. 19-46: Springer, 2005.  
[10] J. A. Goguen, and C. Linde, "Techniques for requirements elicitation." pp. 152-164, 1993.  
[11] A. M. Hickey, and A. M. Davis, "Requirements elicitation and elicitation technique selection: model for two knowledge-intensive software development processes." p. 10 pp., 2003.  
[12] I. Alexander, "Does requirements elicitation apply to open source development?," *OSSG*, 2009.  
[13] B. Massey, "Where do open source requirements come from (and what should we do about it)," 2002.  
[14] L. Richman, and J. Slovak, "SOFTWARE CATCHES THE TEAM SPIRIT New computer programs may soon change the way groups of people work together--and start delivering the long-awaited payoff from office automation. founttoun,," *FORTUNE Magazine*, 1987.  
[15] W. J. Lloyd, M. B. Rosson, and J. D. Arthur, "Effectiveness of elicitation techniques in distributed requirements engineering." pp. 311-318, 2002.  
[16] V. Vehovar, and K. L. Manfreda, "Overview: online surveys," *The SAGE handbook of online research methods*, pp. 177-194, 2008.  
[17] J. Bethlehem, and S. Biffignandi, "Handbook of Web Surveys (Wiley Handbooks in Survey Methodology)," 2011.  
[18] Y. Zhang, and B. M. Wildemuth, "Unstructured interviews," *Applications of Social Research Methods to Questions in Information and Library Science*, pp. 000-060, 2016.  
[19] R. Agarwal, and M. R. Tanniru, "Knowledge acquisition using structured interviewing: an empirical investigation," *Journal of Management Information Systems*, vol. 7, no. 1, pp. 123-140, 1990.  
[20] M. Yousuf, and M. Asger, "Comparison of various requirements elicitation techniques," *International Journal of Computer Applications*, vol. 116, no. 4, 2015.  
[21] W. Scacchi, "Understanding the requirements for developing open source software systems," *IEE Proceedings-Software*, vol. 149, no. 1, pp. 24-39, 2002.

# A Parallel Community Detection Algorithm for Big Social Networks

Yathrib AlQahtani

College of Computer and Information Sciences  
King Saud University  
Collage of Computing and Informatics  
Saudi Electronic University, Riyadh, Saudi Arabia

Mourad Ykhlef

College of Computer and Information Sciences  
King Saud University  
Riyadh,  
Saudi Arabia

**Abstract**—Mining social networks has become an important task in data mining field, which describes users and their roles and relationships in social networks. Processing social networks with graph algorithms is the source for discovering many features. The most important algorithms applied to social networks are community detection algorithms. Communities of social networks are groups of people sharing common interests or activities. DenGraph is one of the density-based algorithms that used to find clusters of arbitrary shapes based on users' interactions in social networks. However, because of the rapidly growing size of social networks, it is impossible to process a huge graph on a single machine in an acceptable level of execution. In this article, DenGraph algorithm has been redesigned to work in distributed computing environment. We proposed ParaDengraph Algorithm based on Pregel parallel model for large graph processing.

**Keywords**—Data mining; social networks; community detection; distributed computing; Pregel

## I. INTRODUCTION

Social networks have become extremely popular in the last years, and they have important roles in the dissemination of information and innovation. The analysis of such networks attracted more attention in the research area. Social networks are modeled as graphs, also called social graphs. An important property of social networks is that they have communities of entities with strong connections. Communities of social networks are groups of people sharing common interests or activities [1]. The typical way to identify communities is graph clustering.

The main techniques of graph clustering are hierarchical clustering, partitioning and density-based clustering. Hierarchical clustering algorithms, such as Newman-Girvan algorithm [2], detect several levels of clusters, where small clusters are included within the large ones. Partitioning algorithms, such as K-mean [3], divide the graph into  $k$  clusters, where  $k$  is predefined to the algorithm. Density-based algorithms consider a graph as areas of high density (clusters), surrounded by some areas of low density (noise). Not only clusters of arbitrary shape can be discovered, but also outliers and noise [4]. This capability makes density-based clustering more appropriate for social networks analysis, since usually there are a very high number of active users, but there is also a

high number of users that do not contribute and can result in noise.

DenGraph [5] is a density-based clustering algorithm for community detection in social networks, inspired by the well-known clustering algorithm for spatial data, DBSCAN [6]. The main idea of DenGraph is to find clusters and outliers of weighted social networks, based on the interaction. It requires two parameters: epsilon  $\epsilon$ , which is the maximum distance threshold; and  $\eta$ , the minimum number of nodes in the  $\epsilon$ -neighborhood.

However, processing big data such as social networks with millions of vertices and edges by using conventional computation is infeasible, since it is impossible to process a huge graph on a single machine in an acceptable time. Therefore, adapting parallel computing for mining social networks has become an urgent need to address processing massive data.

In this research article, we perform DenGraph algorithm for mining implicit social graphs in distributed environment. Therefore, we propose a parallel density based clustering algorithm for social networks (ParaDengraph) in Pregel [7] model as follows. First, compute the  $\epsilon$ -neighborhood to determine core and non-core nodes. Then, generate a new graph of the core nodes. Then, clusters are identified by finding connected components in the core graph. Finally, expand clusters with the non-core nodes.

The article is organized as follows. Section II reviews the related work. ParaDengraph algorithm is then presented in Section III. Finally, ParaDengraph was tested using real social networks, where experiments and evaluation are presented in Sections IV and V.

## II. BACKGROUND AND RELATED WORK

Density-based algorithms consider a graph as areas of high density (clusters), surrounded by some areas of low density (noise). According to the graph, the algorithm reveals the number of clusters. Not only clusters of arbitrary shape can be discovered, but also outliers and noise. Density-based clustering requires some parameters, and generates clusters such that each cluster is a maximal set of density-connected points. Points that are not contained in any cluster are considered as noise.

A. DenGraph

Given a graph  $G = (V, E)$  consisting of a set of nodes  $V$  and a set of weighted undirected edges  $E$ , DenGraph [5] algorithm produces clusters  $\{C_1, \dots, C_k\}$  and noise nodes, that are not part of any cluster. Other non-noise nodes are either core nodes or border nodes. A node  $u \in V$  is considered as core node if it has an  $\epsilon$ -neighborhood  $N_\epsilon(u) = \{v \in V \mid \exists(u, v) \in E \wedge \text{dist}(u, v) \leq \epsilon\}$  of at least  $\eta$  neighbors ( $|N_\epsilon(u)| \geq \eta$ ). A Node that is non-core and connected to at least one core node is considered as border node. A core node along with its border nodes begin a cluster that can be expanded later.

For undirected and weighted graph  $G = (V, E)$ , the number of interactions between two actors reflects the closeness of them, so the distance between two actors,  $p$  and  $q$ , is defined as:

$$\text{dist}(p, q) = \begin{cases} 0, & p = q \\ \min(I_{pq}, I_{qp})^{-1}, & (I_{pq} > 1) \wedge (I_{qp} > 1) \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

Where  $I_{pq}$ ,  $I_{qp}$  are the numbers of interactions between actors  $p$  and  $q$  initiated by  $p$  and  $q$ , respectively. The actual cluster criterion is based on the concepts: directly density-reachable, density-reachable and density-connected, which are shown in Fig. 1. These three concepts are defined as follows:

- **Definition 1.** Let  $u, v \in V$  be two nodes.  $u$  is directly density-reachable from  $v$  within  $V$  with respect to  $\epsilon$  and  $\eta$  if and only if  $v$  is a core node and  $u$  is in its  $\epsilon$ -neighborhood, i.e.  $u \in N_\epsilon(v)$ .
- **Definition 2.** Let  $u, v \in V$  be two nodes.  $u$  is density-reachable from  $v$  within  $V$  with respect to  $\epsilon$  and  $\eta$  if there is a chain of nodes  $p_1, \dots, p_n$  such that  $p_1 = v$ ,  $p_n = u$  and for each  $i = 2, \dots, n$  it holds that  $p_i$  is directly density-reachable from  $p_{i-1}$  within  $V$  with respect to  $\epsilon$  and  $\eta$ .
- **Definition 3.** Let  $u, v \in V$  be two nodes.  $u$  is density-connected to  $v$  within  $V$  with respect to  $\epsilon$  and  $\eta$  if and only if there is a node  $m \in V$  such that  $u$  is density-reachable from  $m$  and  $v$  is density-reachable from  $m$ .

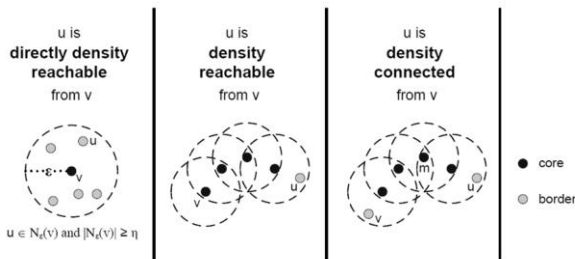


Fig. 1. Density reachability concepts.

In general, a set of core and border nodes  $V_C$  forms a cluster  $C$  if each node  $u \in V_C$  is density-connected to each node  $v \in V_C$ . It is usually that an actor in a social network might be part of more than one community. This overlapping was not allowed in the first version of DenGraph algorithm. An extended version, called DenGraph-O, has addressed this

drawback by allowing border nodes to belong to more than one cluster and it is described in Table I.

B. Graph Parallel Module: Pregel

Graphs are completely data-driven computations, dictated by vertices and edges structure rather than directly expressed in code. In addition, because of the irregular structure of graph data, scalability can be quite limited by unbalanced computational loads [8].

The Bulk Synchronous Parallel (BSP) model [9] provides the means to design parallel processing algorithms. It addresses the problem of parallelizing tasks across multiple workers by using a message-passing interface (MPI) instead of a shared memory.

Pregel introduced originally by Google [7], addresses distributed processing of large-scale graphs. It is a vertex-centric approach, where user focuses on the local action, processes each item independently, and then the system processes all actions on the large dataset.

TABLE I. DENGRAPH ALGORITHM

<b>DenGraph</b>	
	<b>input :</b> Graph, $\epsilon$ , $\mu$
	<b>output:</b> Overlapped clusters, noise.
1.	Begin
2.	Repeat
3.	Select a $u \in V$ that is not yet labeled
4.	Compute $\epsilon$ -neighborhood( $u$ )
5.	If $u$ is core vertex then
6.	A new cluster id is generated
7.	$u$ is assigned to the cluster and labeled as "core vertex"
8.	All $v \in \epsilon$ -neighborhood( $u$ ) are labeled as "border vertex"
9.	The new id is added to list of cluster-ids for all $v$
10.	All $v$ are pushed on a stack
11.	Repeat
12.	Pop the top vertex $v$ of the stack
13.	Compute the $\epsilon$ -neighborhood of $v$
14.	If $v$ is core vertex then
15.	Label $v$ as "core vertex"
16.	For each $n \in \epsilon$ -neighborhood( $v$ ) do
17.	Add new cluster-id to list of $n$
18.	Label $n$ as "border vertex"
19.	Push $n$ on the stack
20.	End
21.	End
22.	Until (the stack is empty)
23.	End
24.	If $u$ is not labeled then
25.	Label $u$ as "noise vertex"
26.	End
27.	Until (all vertices in $V$ are labeled)
28.	End

The basic idea is that each node of the graph corresponds to a task. The node generates output messages that are destined for other nodes, and then each node processes the inputs it receives. This computation consists of a sequence of iterations,

called super-steps. During each super-step, the framework in parallel invokes a user-defined function for each vertex.

### III. PARADENGRAPH

In this section, we present a parallel density based community detection algorithm in social networks based on Pregel parallel model (ParaDengraph). Given two real parameter  $\epsilon$  and  $\eta$ , and undirected weighted graph represented in adjacency lists, ParaDengraph finds clusters and any possible overlapping or noise.

ParaDengraph works as follows: Firstly, cut off the edges with distance more than  $\epsilon$  to compute the  $\epsilon$ -neighborhood  $N_\epsilon(u)$  for each node  $u$ , and classify it as core or non-core node. Then, generate a new graph of the core nodes only, and use this graph in Pregel model to find all the connected components in the core graph, each of which is a cluster. Finally, process non-core nodes by assigning each non-core node to the cluster(s) of its adjacent core nodes if exist, or mark it as a noise if no core node is adjacent to it. Notice that all steps are executed in parallel. ParaDengraph is divided into the following three stages:

#### A. Computing $\epsilon$ -neighborhood

ParaDengraph uses a graph parallel model and generate a graph structure from the input adjacency lists. Therefore, many graph functions can simplify the process, such as filtering feature. This step is accomplished by filtering the graph from all edges with distance  $> \epsilon$ . The number of the remaining edges that are adjacent to each node is used to classify it.

#### B. Core Graph Connected Components

Core graph is generated by filtering the graph of the previous step from all nodes with adjacent edges  $< \eta$ . Then, the core graph is processed in Pregel model to find connected components. The connected core nodes form a cluster. Solving such problem requires unknown number of iterations (super-steps) to find all the connected core nodes by passing messages.

#### C. Expanding Clusters

After discovering all connected core nodes and generating all initial clusters, the next step is to process non-core nodes, assigning each to the same cluster of its neighbor core nodes, and mark it as border. Overlapping must be considered when the border node connects to more than one cluster. Note that when a node does not adjacent to any core node (cluster), it must be marked as noise. The proposed algorithm of ParaDengraph is shown in Table II.

### IV. EXPERIMENTS

ParaDengraph was tested with two real networks. The whole workflow is shown in Fig. 2. ParaDengraph was implemented with Apache Spark GraphX engine for large-scale graph processing and Hadoop distributed file system

(HDFS), and was run on a 64-bit PC with Intel® Core™ i5-3337U CPU @ 1.80GHz  $\times$  4, 5.7 GB RAM and 732.0 GB HD.

#### A. Enron Emails Network

The Enron email network [10] consists of 1,148,072 emails sent between 87,273 employees of Enron. Nodes in the network are individual employees, and edges are individual emails. We run ParaDengraph for many times, changing both parameters: epsilon  $\epsilon$  and minimum point  $\eta$ , as shown in Table III. Fig. 3 shows the relation between the values of ParaDengraph parameters and the number of generated clusters of Enron network.

TABLE II. PARADENGRAPH ALGORITHM

<b>ParaDengraph</b>	
<b>input</b>	: Graph adjacency lists, $\epsilon$ , $\mu$
<b>output</b>	: Overlapped clusters, noise.
1.	Begin
2.	Generate the graph (G)
3.	Filter G from any edge with distance $> \epsilon$ .
4.	CoreGraph=Filter G from nodes with neighbors $< \eta$ .
5.	Mark each $u$ in CoreGraph as "CORE"
6.	NoneCore_nodes=Filter G from nodes with neighbors $> \eta$ .
7.	<b>Start Pregel(CoreGraph): vertex-program</b>
8.	Receive messages from the connected vertices.
9.	Compute label = min(u ID, u label, IDs of all source vertices of the received messages).
10.	If label is changed
11.	send messages to connected vertices with the new one
12.	End
13.	<b>Until (no new messages).</b>
14.	Join CoreGraph and NoneCore_nodes
15.	Set for each non-core $u$ label = list of labels of all directly reachable core.
16.	If $u$ label is empty
17.	Mark $u$ as "NOISE"
	Else
	Mark $u$ as "BORDER"
18.	End
19.	End

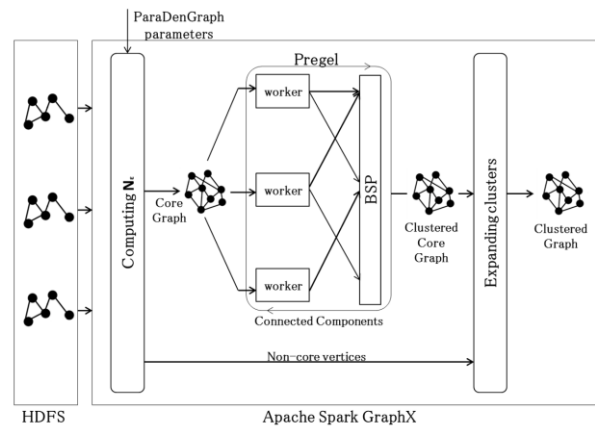


Fig. 2. ParaDengraph workflow.

TABLE III. ENRON NETWORK RUN STATISTICS

Epsilon	Min points	Clusters	Noise	Cores	Borders
0.1	1	31	86346	927	0
	2	17	86374	387	512
	3	10	86408	237	628
	4	10	86445	168	660
	5	9	86466	131	676
	6	9	86494	102	677
	7	7	86515	87	671
	8	9	86521	81	671
	9	8	86544	69	660
	10	8	86559	60	654
0.2	1	31	85494	1779	0
	2	15	85526	830	917
	3	10	85568	557	1148
	4	7	85596	409	1268
	5	5	85618	320	1335
	6	5	85647	254	1372
	7	6	85663	215	1395
	8	7	85679	191	1403
	9	6	85715	166	1392
	10	5	85750	144	1379
0.3	1	25	85068	2205	0
	2	15	85088	1028	1157
	3	11	85124	707	1442
	4	12	85149	534	1590
	5	10	85188	426	1659
	6	7	85226	342	1705
	7	5	85250	287	1736
	8	4	85279	245	1749
	9	6	85309	216	1748
	10	6	85318	193	1762
0.4	1	26	84476	2797	0
	2	13	84502	1336	1435
	3	11	84533	963	1777
	4	8	84573	740	1960
	5	7	84607	598	2068
	6	8	84630	487	2156
	7	7	84656	417	2200
	8	5	84691	358	2224
	9	4	84714	310	2249
	10	4	84725	272	2276
0.5	1	40	83327	3946	0
	2	12	83383	1902	1988
	3	11	83426	1406	2441
	4	8	83471	1103	2699
	5	6	83509	908	2856
	6	6	83538	748	2987
	7	6	83568	648	3057
	8	5	83599	567	3107
	9	2	83647	491	3135
	10	2	83663	441	3169

Enron Clustering

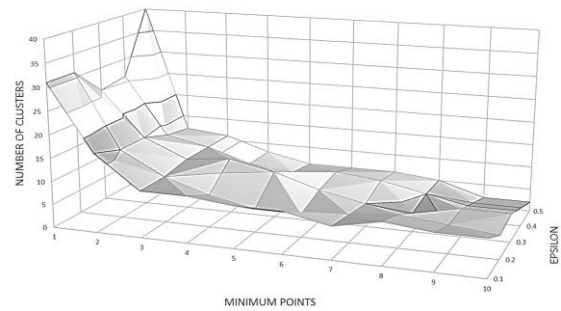


Fig. 3. Enron clusters and parameters values.

TABLE IV. TWITTER NETWORK RUN STATISTICS

Epsilon	Min points	Clusters	Noise	Cores	Borders
0.1	1	9	8792	395	0
	2	7	8796	36	355
	3	9	8796	16	375
	4	9	8796	15	376
	5	9	8801	13	373
	6	8	8807	12	368
	7	8	8807	12	368
	8	8	8807	12	368
	9	8	8807	12	368
	10	8	8807	12	368
0.2	1	5	8377	810	0
	2	4	8379	70	738
	3	8	8379	29	779
	4	7	8385	19	783
	5	7	8385	18	784
	6	7	8385	17	785
	7	7	8385	16	786
	8	7	8385	16	786
	9	7	8385	16	786
	10	7	8385	16	786
0.3	1	6	8176	1011	0
	2	3	8182	85	920
	3	8	8182	31	974
	4	8	8182	21	984
	5	7	8187	20	980
	6	6	8191	18	978
	7	7	8191	16	980
	8	7	8191	16	980
	9	7	8191	16	980
	10	7	8191	16	980
0.4	1	4	7756	1431	0
	2	3	7758	111	1318
	3	6	7759	44	1384
	4	8	7763	26	1398
	5	8	7763	22	1402
	6	6	7772	19	1396
	7	6	7772	18	1397
	8	5	7780	17	1390
	9	6	7780	16	1391
	10	6	7780	16	1391
0.5	1	3	6787	2400	0
	2	3	6787	182	2218
	3	5	6787	61	2339
	4	5	6787	36	2364
	5	9	6792	25	2370
	6	9	6792	23	2372
	7	9	6792	22	2373
	8	8	6799	20	2368
	9	8	6799	20	2368
	10	8	6799	20	2368

B. Twitter Interactions Network

The second data set was Twitter interaction graph, where nodes represent users and edges represent interactions, such as retweets or replays. ParaDengraph was applied many times to the graph of 9187 nodes as shown in Table IV. The relation between the generated clusters of Twitter graph and the parameters values of ParaDengraph is shown in Fig. 4.

From both Fig. 3 and 4, it can be seen that the resulted clusters were not related directly to the parameters values, since the number of clusters decreased in Enron network with increasing the values, while it increased in the Twitter network. Therefore, the result depends mainly on the graph nature and the core nodes locations to each other's.

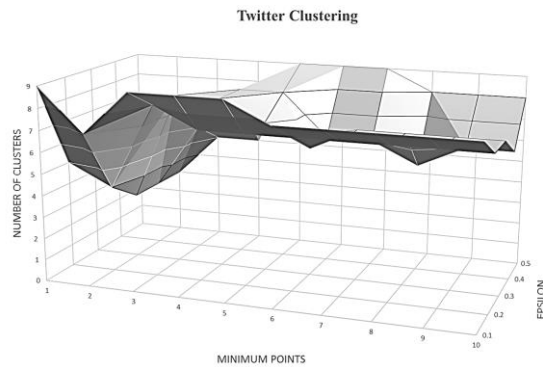


Fig. 4. Twitter clusters and parameters values.

However, the clear result was the relation between the number of core nodes and the parameters values. The number of core nodes increased when the epsilon value increased, but decreased with increasing the value of minimum points.

## V. EVALUATION

For evaluation purpose, both ParaDengraph and Newman Edge Betweenness Clustering (EBC) [2] algorithms were applied to the same social network graph to compare both results and obtain the characteristics of both clustering methods. The previous Enron and Twitter graphs, with 87,273 and 9,187 nodes respectively, are too large to fit in the low capacity of EBC. So, for evaluation, a smaller Twitter data set (539 nodes) was used to run both algorithms. ParaDengraph was applied on the graph with different collections of parameters values, epsilon ( $\epsilon$ ) and minimum points ( $\eta$ ). Since the average noise percentage for all runs (140 runs) was 54.24 %, we chose the closer epsilon value ( $\epsilon=0.09$ ). The result with  $\epsilon=0.09$  and  $\eta=5$  was 8 clusters, and 294 nodes as noise.

Then, Newman EBC was applied to the same graph of 539 nodes. Notice that EBC is a hierarchical clustering algorithm. EBC gave the best modularity at level 9 with 17 clusters. While EBC generated 17 clusters, ParaDengraph with the ability of density-based clustering to discover noise generated only 8 clusters. In order to test the effect of noise in the result, a total of 294 noise nodes that discovered by ParaDengraph ( $\epsilon=0.09$  and  $\eta=5$ ) were removed from the original graph, and EBC was applied again. On the graph of 245 nodes (after removing noise), the best modularity was found at level 6 with 10 clusters. The number of clusters decreased from 17 to 10, which was closer to ParaDengraph result (8 clusters).

As a result, we found that ParaDengraph outperformed EBC mainly in two parts. First, EBC failed to deal with large social networks due to its high time complexity. However, ParaDengraph was capable to run with better performance, even in the sequential version. This result proves that density-based clustering algorithms are more appropriate for large social networks than hierarchical clustering algorithms. Second, notice that EBC at the first run on the complete graph (539 nodes) gave a result of 17 clusters. However, the result of the second run on the same graph after excluding all noise (245

nodes) was 10 clusters. In the other hand, ParaDengraph gave both 8 clusters and some noise. We can notice that this result (8 clusters) was too close to the second result of EBC (10 clusters) after removing all noise.

## VI. DISCUSSION

In addition to the suitability for large-scale graph computing, which is most critical in dealing with massive networks, ParaDengraph is also able to give reasonable clusters by discovering and excluding noise from the clustering process. This feature is essential in social networks, since there is usually a large number of actors, but there is also a high number of them do not contribute (outliers). While EBC found clusters, it was unable to detect outliers, resulting in more clusters where some may contained only one or very few nodes. However, outliers did not affect the outcome of ParaDengraph.

Moving to the limitations of this proposed algorithm, the most noticeable limitation is that ParaDengraph was proposed based on the static version of DenGraph, where communities are observed as static. However, they are evolving continuously and such dynamic networks require considering the time and changes over time.

## VII. CONCLUSION

This article proposed ParaDengraph, a graph-parallel algorithm for community detection in large social networks based on the original sequential algorithm called DenGraph. The suggested algorithm is suitable for large-scale graph computing. ParaDengraph has been applied to two real social networks: Enron emails and Twitter. For evaluation, ParaDengraph was compared with Newman Edge Betweenness Clustering (EBC) algorithm. ParaDengraph outperformed EBC in terms of performance and the ability to generate clusters that are more reasonable by excluding noise from the clustering process. For future work, ParaDengraph can be improved for studying communities on dynamic social networks.

## ACKNOWLEDGMENTS

This work was supported by the Deanship of the Scientific Research and Research Center of the Collage of Computer and Information Sciences, King Saud University.

## REFERENCES

- [1] Zafaran, R., Abbasi, M. A., & Liu, H. (2014). "Social media mining, an introduction" Cambridge University Press.
- [2] Newman, M., & Girvan, M. (2004). "Finding and evaluating community structure in networks". *Phys. Rev. E*, pp. 69, 026113.
- [3] MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations". *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press.
- [4] Subramani, K., Velkov, A., Ntoutsis, I., Kroger, P., & Kriegel. (2011). "Density-based community detection in social networks". *Internet Multimedia Systems Architecture and Application (IMSAA)* (pp. 1-8). IEEE 5th International Conference.
- [5] Falkowski, T., Barth, A., & Spiliopoulou, M. (2007). "DENGRAPH: a density-based community detection algorithm". *Web Intelligence* (pp. 112-115). IEEE/WIC/ACM International Conference.

- [6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", In: Proc. Of KDD, 226-231, 1996.
- [7] G. Malewicz, M. H. (2010). "Pregel: a system for large-scale graph processing". ACM SIGMOD International Conference on Management of data (pp. 135-146). New York, NY, USA: ACM.
- [8] Lumsdaine, A., Gregor, D., Hendrickson, B., & Berry, J. (2007, 5 17). "Challenges in parallel graph processing". *Parallel Processing Letters*, pp. 5–20.
- [9] Leslie G. Valiant, "A bridging model for parallel computation". *Comm. ACM* 33(8), 1990, 103–111.
- [10] W. Cohen, Enron network dataset, KONECT. (2017)

# Comparison Between Two Adaptive Controllers Applied to Greenhouse Climate Monitoring

Mohamed Essahafi

Preparing a PhD Thesis in Adaptive Control at the Faculty  
of Sciences and Technology, University of Sultan  
Moulay Slimane, Morocco

Mustapha Ait Lafkih

Professor at the Electrical Engineering Department  
Faculty of Sciences and Technology, University of Sultan  
Moulay Slimane, BniMellal Morocco

**Abstract**—This paper presents a study of a multivariable Adaptive Generalized Predictive Controller and its application to control the thermal behaviour of an agricultural greenhouse, which is composed of a number of different elements (cover, internal air, plants, soil, actuators and sensors). The thermal model was obtained after the study of energy balances reacting the physical behavior of the greenhouse. For this reason, we opted to estimate the dynamic model of the greenhouse with algorithm based on recursive least squares (RLS) method. Simulation results are exposed to show the controller's performances in terms of response time, stability and the rejection of disturbances.

**Keywords**—Generalized predictive control; greenhouse; multi-variable control; identification; recursive least square

## I. INTRODUCTION

The role of the agricultural greenhouse is to produce a crop while avoiding the local climate. It helps to improve the yield of plants [1], and to grow plants that would not survive the natural climate. The most general objective of the producer is to place on the market quantities of agricultural products in relation to the economic demand. For this, he must determine the favorable conditions according to the biological needs of the plant.

The "climate" inside a greenhouse depends on its ventilation. The aeration process is complex, it participates in most of the heat and mass exchanges with the outside, and its control allows to control the physical parameters such as temperature, humidity, or gas concentrations, like CO<sub>2</sub> for example. This control is essential to maintain the plants in favorable metabolic conditions (respiration, photosynthesis, transpiration) and in a satisfactory biological state. Adaptive control is necessary to control the greenhouse throughout the functional life of production [2].

In this article we will compare two very famous control strategies, namely, the generalized predictive auto-tuning control (GPC) and the generalized minimum variance (GMV) [3], developed respectively by D. Clark 1988, and Astrom and Wittenmark 1973. The GPC control is an extension with an extended horizon of the GMV. This method of control is an approach which has proven its performances in industry. Its algorithm is easier, flexible and robust with respect to other methods. It is applicable to all types of processes, be it variable delay, long or unknown processes, non-minimal phase shift processes, as well as unstable processes (open loop). However, to effectively control the microclimate in the greenhouse, we chose an adaptive controller that allows online identification

of greenhouse parameters and at every moment it calculates the control law that allows to follow in real time the needs of the plant. Although this is based on the recursive least-squares identification (RLS).

To clarify this, we propose in this paper in Section 2, the proposed control approaches GPC and GMV. In Section 3, we present the adaptive control based on online RLS identification method, in Section 4 the simulation results are discussed. Finally in Section 5, a conclusion with future work prospects.

## II. DESIGN OF MULTIVARIABLE ADAPTIVE CONTROLLERS

### A. Greenhouse design

The control structure [4] chosen for the greenhouse of Fig. 1 shows that the system has two control inputs: heating ( $R_c$ ) and ventilation ( $V_t$ ), and two outputs to be measured: temperature ( $T_i$ ) and humidity ( $H_i$ ), with

$R_c$ : heating energy applied to the plant (KW).

$V_t$ : ventilation angle outside the greenhouse (C).

$T_o, H_o$ : air temperature and relative humidity outside the greenhouse (C, /100).

$S_r$ : Solar radiation (W / m).

$T_i, H_i$ : air temperature and relative humidity inside the greenhouse (C, /100).

$S_w$ : wind speed outside the greenhouse (km/h).

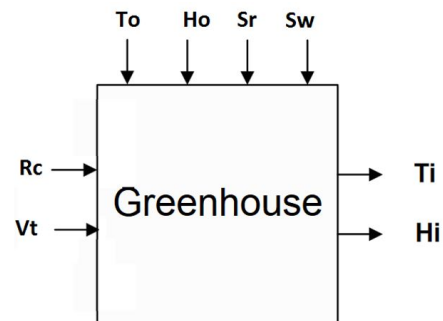


Fig. 1. Schematic diagram of controlled greenhouse.

### B. CARIMA Representation

To describe the discrete behaviour of the greenhouse consider a CARIMA [5] (Controlled Auto-Regressive Integrated



Media Moving) model

$$A(q^{-1})y(t) = B(q^{-1})U(t-1) + \Upsilon(t) \frac{C(q^{-1})}{\Delta} \quad (1)$$

Where  $A$ ,  $B$  and  $C$  are polynomials in the back-ward shift operator  $q^{-1}$ :

$$A(q^{-1}) = 1 + A_1q^{-1} + \dots + A_{n_a}q^{-n_a} \quad (2)$$

$$B(q^{-1}) = B_0 + B_1q^{-1} + \dots + B_{n_b}q^{-n_b} \quad (3)$$

$$C(q^{-1}) = 1 + C_1q^{-1} + \dots + C_{n_c}q^{-n_c} \quad (4)$$

$\Upsilon$  is an uncorrelated random sequence,  $\Delta$  is the difference operator  $1 - q^{-1}$ . For simplicity,  $C(q^{-1})$  is chosen to be 1 to give the model:

### C. Generalized Adaptive Controller GPC

To simplify the implementation of the adaptive GPC controller [6], the Diophantine resolution is necessary

$$A(q^{-1})Y(t) = B(q^{-1})U(t-1) + \frac{\Upsilon(t)}{\Delta} \quad (5)$$

Then the predicted output  $Y(t+j)$  for the prediction step  $j$  is

$$Y(t+j) = E_j B \Delta U(t+j-1) + F_j y(t) + E_j \Upsilon(t+j) \quad (6)$$

Since  $E_j(q^{-1})$  is a polynomial of degree  $j$ , the noise components are all at the next discretization steps, so that the optimal predictor, given the measured output data, is explicitly

$$\hat{Y}(t+j|t) = G_j \Delta U(t+j-1) + F_j Y(t) \quad (7)$$

Where,  $G_j(q^{-1}) = E_j B \hat{Y}(t+j|t)$

$$I = E_j A \Delta q^{-j} F_j(q^{-1}) \quad (8)$$

Where,  $E_j$ ,  $F_j$ ,  $A(q^{-1})$  Given  $A(q^{-1})$  denote  $A$  as  $\tilde{A} = A\Delta$ , such that  $\tilde{A} = 1 + \tilde{A}_1q^{-1} + \dots + \tilde{A}_{n_a}q^{-n_a+1}$

$$B = E_j B \tilde{A} + B q^{-j} F_j \quad (9)$$

$$E_j B = \frac{B[1 - q^{-j} F_j]}{\tilde{A}} \quad (10)$$

$$G_j = \frac{B[-q^{-j} F_j]}{\tilde{A}} \quad (11)$$

Consequently, using the recursion of the diophantine equation, so as to obtain the polynomials  $E_{j+1}$  and  $F_{j+1}$  Considering the values of  $E_j$  and  $F_j$  [7]

$$F_1 = -(\tilde{A}_1 + \dots + \tilde{A}_{n_a+1}q^{-n_a}) \quad (12)$$

$$F_j = F_{j,0} + F_{j,1}q^{-1} + \dots + f_{j,n_a}q^{-n_a} \quad (13)$$

with  $i = 0, 1, \dots, n_a - 1$

$$F_{j+1,i} = F_{j,i+1} - \tilde{a}_{i+1} F_{j,0} \quad (14)$$

and for  $i = 0, 1, \dots, n_a - 1$

$$F_{j+1,n_a} = -\tilde{a}_{n_a+1} F_{j,0} \quad (15)$$

Where,

$$G_1 = B = B_0 + \dots + B_{n_b}q^{-n_b} \quad (16)$$

for  $n_G j = n_b + j - 1$

$$G_j = G_{j,0} + G_{j,1}q^{-1} + \dots + G_{j,n_g}q^{-n_g} \quad (17)$$

and for  $i = 0, 1, \dots, j - 1$

$$G_{j+1,i} = G_{j,i} \quad (18)$$

Also we can write for  $i = j, 1, \dots, j + n_b + 1$

$$F_{j+1,i} = G_{j+i} + B_{i-j} F_{j,0} \quad (19)$$

$$G_{j+i,n_b+j} = B_{n_b} F_{j,0} \quad (20)$$

The term E may be separated by a second Diophantine equation in  $G_j$  and  $H_j$  as follows:

$$E_j B = G_j + q^{-j} H_j \quad (21)$$

$$G_j = G_0^j + G_1^j q^{-1} + \dots + G_{j-1}^j q^{-j+1} \quad (22)$$

$$H_j = H_0^j + H_1^j q^{-1} + \dots + H_{n_h}^j q^{-n_h} \quad (23)$$

Equation (7) can be rewritten as

$$\hat{Y}(k+j) = G_j \Delta U(k+j-1) + H_j \Delta U(k-1) + F_j Y(k) \quad (24)$$

And which considers the following quantities:

$$\begin{pmatrix} H_{N_1} \Delta U(k-1) + F_{N_1} Y(k) \\ H_{N_1+1} \Delta U(k-1) + F_{N_1+1} Y(k) \\ \vdots \\ H_{N_2} \Delta U(k-1) + F_{N_2} Y(k) \end{pmatrix} \quad (25)$$

Adopt a reference sequence  $W$  is available. In most cases  $W$  will be a constant  $w$  equal to the current set-point  $W(t)$ . The purpose of the predictive control law is to drive the future outputs  $Y$  close to future set-point  $W$  in some sense [8].

The basic cost function used in GPC has the form [9]

$$J = (\hat{Y} - W)^T (\hat{Y} - W) + \tilde{U}^T \Lambda \tilde{U} \quad (26)$$

with

$$\Lambda = \begin{pmatrix} \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \dots & \vdots \\ \vdots & \dots & \vdots \\ 0 & \dots & \lambda_m \end{pmatrix} & 0 \\ 0 & \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \dots & \vdots \\ \vdots & \dots & \vdots \\ 0 & \dots & \lambda_m \end{pmatrix} \end{pmatrix} \quad (27)$$

$$\tilde{U} = [\Delta U(t), \Delta U(t+1) + \dots + \Delta U(t+N_2-1)]^T \quad (28)$$

and

$$F = [F(t+1), F(t+1), \dots, F(t+N_2)]^T \quad (29)$$

As cited previously, the first  $j$  terms in  $G_j q^{-1}$  are the parameters of the step-response and therefore  $G_{ij} = G_j$  for  $j < i$ .

$G$  is then a matrix of dimension  $(m(N_2 - N_1 + 1) \times mN_u)$

$$\begin{bmatrix} G_{N_1-1} & \Lambda & G_0 & \Lambda & \Lambda & 0 \\ G_{N_1} & G_{N_1-1} & \Lambda & G_0 & \Lambda & 0 \\ M & M & 0 & 0 & 0 & M \\ G_{N_u-1} & G_{N_u-2} & G_{N_u-3} & \Lambda & \Lambda & G_0 \\ M & M & M & M & M & M \\ G_{N_2-1} & G_{N_2-2} & G_{N_2-3} & \Lambda & G_{N_2-N_u+1} & G_{N_2-N_u} \end{bmatrix} \quad (30)$$

Where,  $N_1$  and  $N_2$  represent minimum and maximum prediction horizons, respectively.  $N_u$  represents a control horizon [9].

Given that the first element of  $\tilde{u}$  is  $\Delta u(t)$  so that the current control  $u(t)$  is given by:

$$U_{opt}(t) = U(t-1) + \Delta U_{opt}(t) \quad (31)$$

with

$$\Delta U_{opt}(t) = M1(W - Fc) \quad (32)$$

The benefit of the GPC algorithm [10] is the expectations made about future control actions. As an alternative of allowing them to be free as for the above improvement, GPC uses the idea that after an interval  $N_U < N_2$  predictable control steps are supposed to be zero, so we have

$$\Delta U(t+j-1) = 0 \quad , \quad j > N_U \quad (33)$$

#### D. Generalized Minimum Variance Controller GMVC

After seeing the following steps for the design of the GPC controller [11], in the following section we also detail the GMV controller.

Consider the stochastic matrix polynomial model:

$$A(q^{-1})y(t) = B(q^{-1})q^{-1}u(t) + D(q^{-1})v(t) \quad (34)$$

is assumed, with

$$D(q^{-1})y(t) = D_0 + D_0q^{-1} + \dots + D_mq^{-m} \quad (35)$$

A generalized minimum variance controller [12] is obtained by minimizing the criterion.

$$I(k+d+1) =$$

$$E[y(k+d+1) - w(k)]^T [y(k+d+1) - w(k)] + [u(k) - u_w(k)]^T R [u(k) - u_w(k)] \quad (36)$$

With  $R = R^T$  positive semi-definite  $u_w(k)$  is the offset steady state value of  $u(k)$

$$u_w(k) = B^{-1}A(1)w(k) \quad (37)$$

Corresponding to (36) the process and signal model is split up into

$$\begin{aligned} q^{d+1}y(t) &= A(q^{-1})[B(q^{-1})qu(t) \\ &L(q^{-1})v(t)] + F(q^{-1})q^{d+1}v(t) \end{aligned} \quad (38)$$

Where, the new matrix polynomials are defined by:

$$F(q^{-1}) = I + F_1q^{-1} + \dots + F_dq^{-d} \quad (39)$$

$$L(q^{-1}) = L_0 + L_1q^{-1} + \dots + L_{m-1}q^{m-1} \quad (40)$$

Their parameters are determined by:

$$D(q^{-1}) = A(q^{-1})F(q^{-1}) + q^{-(d+1)}L(q^{-1}) \quad (41)$$

The term  $I(k+d+1)$  is rewritten in the time domain [13], knowing that  $[\partial I(k+d+1)/\partial u(k)] = 0$  we obtain:

$$\begin{aligned} B_1^T(q^{-1})[A(q^{-1})[B(q^{-1})qu(t) + L(q^{-1})v(t)] \\ - w(q^{-1})] + R[u(k) - u_w(k)] = 0 \end{aligned} \quad (42)$$

Where,  $v(t)$  can be reconstructed by:

$$v(t) = D^{-1}(q^{-1})[A(q^{-1})y(t) - B(q^{-1})q^{-1}u(t)] \quad (43)$$

The control vector resulting from the GMV algorithm [14] is written as:

$$\begin{aligned} u(t) &= [F(q^{-1})D(q^{-1})B(q^{-1})q + (B_1^T)^{-1}R]^{-1} \\ &(I + (B_1^T)^{-1}RB^{-1}(1)A(1))w(q) \\ &- A^{-1}(q^{-1})L(q^{-1})D^{-1}(q^{-1})A(q^{-1})y(t) \end{aligned} \quad (44)$$

If  $R = 0$  is set, The minimum variance controller result is:

$$\begin{aligned} u(t) &= B^{-1}(q^{-1}) \frac{q^{-1}}{1 + q^{-(d+1)}} [A(q^{-1})[w(t) - y(t)]] \\ &+ [D(q^{-1}) - L(q^{-1})]v(t) \end{aligned} \quad (45)$$

Where,  $v(t)$  must reconstructed from (43). This controller yields for the closed-loop system

$$y(t) = F(q^{-1})v(t) + q^{-(d+1)}w(t) \quad (46)$$

### III. ADAPTIVE CONTROL ALGORITHM

The general structure of the multivariable adaptive control applied to the agricultural greenhouse is shown in Fig. 2. The online identification of the parameters of the greenhouse is made in such a way as to converge towards the real values using the least squares recursive method [15].

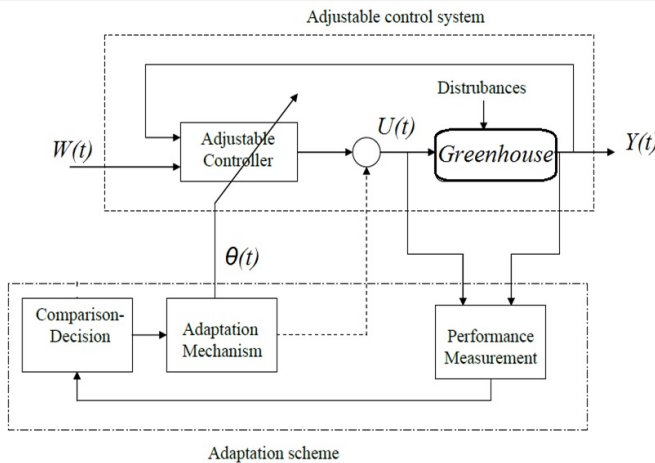


Fig. 2. Adaptive control strategy.

$$Y(k) = \theta^T(k)\varphi(k) \quad (47)$$

Where,  $\theta$  is the vector of the unknown parameters defined as

$$\theta^T(k) = [a_{11}(k), \dots, d_{24}(k)] \quad (48)$$

In (47),  $\varphi(k)$  is a regression vector partly consisting of measured input/output variables and is defined as:

$$\varphi(k) = [Ti \ Hi \ Rc \ Vt \ To \ Ho \ Sr \ Sw] \quad (49)$$

The parameter vector is calculated by the following recursive algorithm [11]:

$$\theta(k) = \underset{i=1}{\text{Argmin}} \sum^k \lambda^k ((y(k) - \varphi^T \hat{\theta}(k))^2) \quad (50)$$

$$\hat{\theta}(k) = \hat{\theta}(k-1) + K(k)[(y(k) - \varphi^T \hat{\theta}(k-1))] \quad (51)$$

$$K(k) = \frac{P(k-1)\varphi(k)}{\lambda + \varphi^T(k)P(k-1)\varphi(k)} \quad (52)$$

$$K(k) = \frac{1}{\lambda} (P(k-1) - K(k)\varphi^T(k)) \quad (53)$$

### IV. SIMULATIONS AND RESULTS

After identifying the agricultural greenhouse by a mathematical model in space state, the validation phase of the model by the simulations made on the greenhouse is necessary. The purpose of this paragraph is to test in simulation mode the control law multivariable predictive GPC on the identified discrete model of the greenhouse and compare it with the multivariable GMV control.

Then, the discrete model of the greenhouse is written by:

$$A1 = \begin{bmatrix} -0.1 & 0.2 \\ 0.33 & 0.4 \end{bmatrix} ; \quad A2 = \begin{bmatrix} -0.5 & 0.66 \\ -0.77 & 0.8 \end{bmatrix}$$

$$B0 = \begin{bmatrix} -0.1 & 0.2 \\ -0.3 & 0.4 \end{bmatrix} ; \quad B1 = \begin{bmatrix} 0.5 & -0.6 \\ -0.7 & 0.8 \end{bmatrix} ;$$

$$B2 = \begin{bmatrix} 0.9 & -0.1 \\ 0.11 & -0.12 \end{bmatrix}$$

The subsequent experiments represent variations of humidity and temperature set points in the greenhouse process. In general, each trial track can be separated into three sequential phases. Through the first, start-up phase the system plant is achieving steady state around the operation points. The second one is a suitably chosen identifying phase using acquired data, which gives us an initial estimation of CARIMA model parameters. The third phase finally displays the results of the adjusted multivariable adaptive GPC controller. Several of real-time tests had been simulated in order to choice parameters that would offer the desired controller performance. As a final point, the succeeding values of design parameters were set.

- For adaptive GPC control sampling times  $T = 1.5s$ , horizons  $N1 = 1, N2 = 10, Nu = 4$ , dead time  $d = 3$  and weights  $[\lambda1, \lambda2] = [0.97, 0.95]$
- For the cases using GMV as the control algorithm,  $R = \begin{bmatrix} 10 & 1 \\ 5 & 10 \end{bmatrix}$  is used.

In the following figures we present the simulations made for the two adaptive controllers. For the GMV controller the influence of the multivariable coupling of the parameters of the greenhouse is clearly felt on the first control at the time of the change of the setpoint applied to the second output (Fig. 3 and 4). The first control reacts so that the air temperature is disturbed by the change in behavior that affects the airflow.

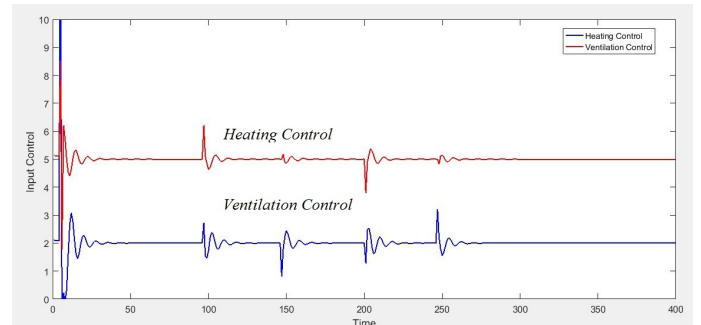


Fig. 3. The input control of the greenhouse with the AGMV controller.

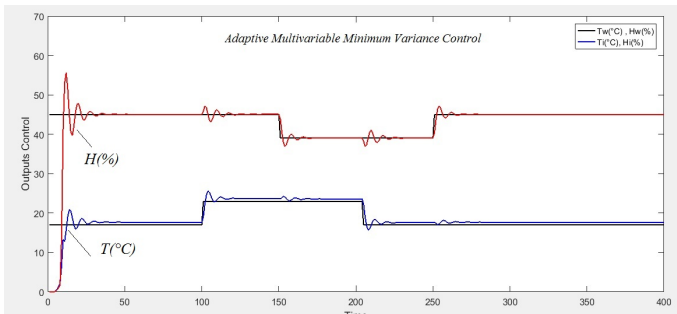


Fig. 4. Responses of greenhouse air temperatures and relative humidity using the AGMV controller.

With the observation of Fig. 3 and 4 for the GPC controller, we note despite the presence of the coupling between the variables of the thermal process, the set point change in real time on both the first output (top) and the second output (bottom) has less impact on both outputs. The inputs controls applied to the system are given in Fig. 5.

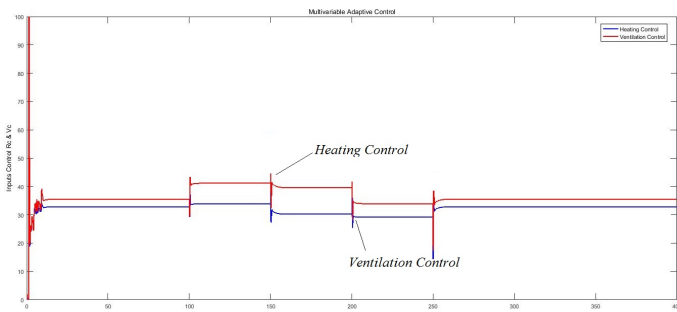


Fig. 5. The input control of the greenhouse with the AGPC controller.

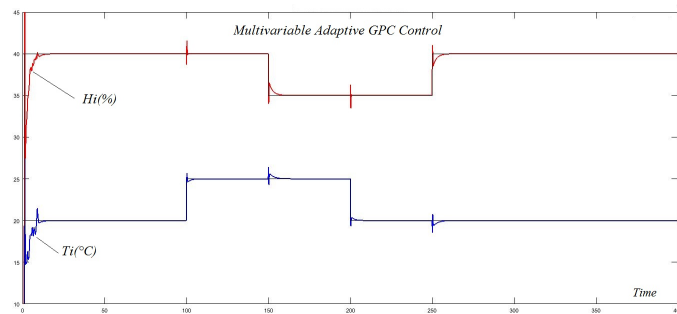


Fig. 6. Responses of greenhouse air temperatures and relative humidity using the AGPC controller.

The evolution over time of the dynamic parameters of the greenhouse is illustrated in Fig. 6. These parameters are initialized to values less than unity, but they are then recalculated taking into account the thermal behavior of the greenhouse. When the disturbed load is applied, the parameters are again re-estimated.

The aim purpose of this study was to compare the performance of the two adaptive controllers AGPC and AGMV for disturbance rejection as well as set-point tracking. The various simulation tests show that the GPC adaptive controller is more powerful than the GMV adaptive controller. In fact, the output response of the greenhouse follows the fixed setpoint for both controllers as shown in the figures Fig. 4 and 7. The only difference is the way in which the output arrives at the setpoint.

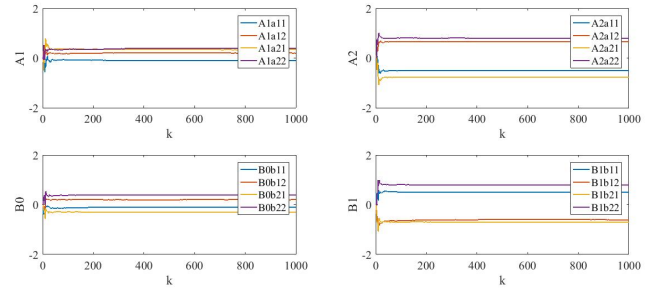


Fig. 7. The online estimation of the greenhouse's parameters A1, A2, B0, B1.

Moreover, for the GPC controller the output is characterized by a better stability and a lack of the oscillations of overtaking by comparing to the GMV controller, while the control effort is better for the GMV controller.

## V. CONCLUSION

A Summary Comparison of GPC Adaptive Controllers and GMV is rigorously implemented for controlling the temperature and humidity of the air inside the environment of a greenhouse, based on least-square estimation technique. The control objective is to enforce the air temperature set point within the greenhouse to track its desired value. It appears, using real results and simulation that the GPC adaptive controller retains so much stability and robustness in comparison with the GMV adaptive controller.

As future work, the global microclimate of the greenhouse taking into account the external and internal disturbance will be developed, in order to meet the basic needs of greenhouse cultivation.

## REFERENCES

- [1] N. Dinesh Kumar, S. Pramod, C. H. Sraboni, *Intelligent irrigation system*, International Journal of Agricultural Science, vol. 3, pp. 23-30, August, 2013.
- [2] Li. XH, X Cheng, K Yan, P. Gong, *An monitoring system for vegetable greenhouse based on a wireless sensor network*, PubMed journal on Sensors (Basel), vol. 10, pp. 8963-8980, October 2010.
- [3] Astrom, K. J., and B. Wittenmark, *Adaptive Control* Addison-Wesley Publishing Company Second Edition., 1995.
- [4] Lauri, J. V. Salcedo, S. Garcia-Nieto and M. Martinez, *Model predictive control relevant identification: multiple input multiple output against multiple input single output* in IET Control Theory and Applications, vol. 4, no. 9, pp. 1756-1766, September 2010. doi: 10.1049/iet-cta.2009.0482
- [5] K. P. Lam, *On Using the Filtered CARMA and CARIMA Models for State-Space Self-Tuning Control* 1987 American Control Conference, Minneapolis, MN, USA, 1987, pp. 1286-1290.
- [6] Clarke, D. W., C. Mohtadi, and P.S. Tuffs, *Generalized Predictive Control Part I. The Basic Algorithm; Part II. Extensions and Interpretations*, Automatica, Vol. 23, No. 2, pp. 137-160, 1987.
- [7] M. J. Grimble, P. Majecki and L. Giovanini, *Polynomial approach to nonlinear predictive GMV control* 2007 European Control Conference (ECC), Kos, 2007, pp. 4546-4553.
- [8] D. W. Clarke, P. J. Gawthrop (1975), *Self-tuning control*, Proc.IEE, vol. 126, no. 6, pp. Part D: Control Theory and Applications 122(9), 929-934 [Seminal paper on GMV-based self-tuning control]
- [9] D. W. Clarke, P. J. Gawthrop, *Self-Tuning controller*, Proc.IEE, vol. 122, no. 9, pp. Control Theory and Applications 126(6), 633-640, 1979.[A review of GMV-based self-tuning control.]
- [10] Z. Zidane, M. Ait Lafkih and M. Ramzi, *Application of Multivariable Linear Quadratic Gaussian Control and Generalized Predictive Control in a Hydropower Plant*, International Journal on Sciences and Techniques of Automatic control and computer engineering IJ-STA, vol. 7 no. 1, April (2013), pp. 1890-1906.
- [11] S. C Chai, G. P. Liu and D. Rees, *Design and implementation of networked predictive control systems*, 16 IFAC World Congress, Prague, 2005

- [12] G. A. Montague, A. J. Morris and M. T. Tham, *Performance evaluation of three multivariable self-tuning controller design techniques* 1986 25th IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 1564-1569. doi: 10.1109/CDC.1986.267148
- [13] G. A. Montague, M. T. Tham and A. J. Morris, *A Comparison of Multivariable Long Range Predictive Control with GMV Control in a Highly Nonlinear Environment* 1986 American Control Conference, Seattle, WA, USA, 1986, pp. 721-727.
- [14] M. J. Grimble and P. Majecki, *Nonlinear Predictive GMV control* 2008 American Control Conference, Seattle, WA, 2008, pp. 1190-1195. doi: 10.1109/ACC.2008.4586654
- [15] J. E. Bobrow and W. Murray, *An algorithm for RLS identification parameters that vary quickly with time* in IEEE Transactions on Automatic Control, vol. 38, no. 2, pp. 351-354, Feb 1993. doi: 10.1109/9.250491

# A Predictive Model for Solar Photovoltaic Power using the Levenberg-Marquardt and Bayesian Regularization Algorithms and Real-Time Weather Data

Mohammad H. Alomari\*, Ola Younis<sup>†</sup> and Sofyan M. A. Hayajneh<sup>‡</sup>

\*Electrical Engineering Department

Applied Science Private University, Amman, Jordan

<sup>†</sup>School of Electrical Engineering, Electronics and Computer Science

University of Liverpool, Liverpool, United Kingdom

<sup>‡</sup>Electrical and Computer Engineering Department

Isra University, Amman, Jordan

**Abstract**—The stability of power production in photovoltaics (PV) power plants is an important issue for large-scale grid-connected systems. This is because it affects the control and operation of the electrical grid. An efficient forecasting model is proposed in this paper to predict the next-day solar photovoltaic power using the Levenberg-Marquardt (LM) and Bayesian Regularization (BR) algorithms and real-time weather data. The correlations between the global solar irradiance, temperature, solar photovoltaic power, and the time of the year were studied to extract the knowledge from the available historical data for the purpose of developing a real-time prediction system. The solar PV generated power data were extracted from the power plant installed on-top of the faculty of engineering building at Applied Science Private University (ASU), Amman, Jordan and weather data with real-time records were measured by ASU weather station at the same university campus. Huge amounts of training, validation, and testing experiments were carried out on the available records to optimize the Neural Networks (NN) configurations and compare the performance of the LM and BR algorithms with different sets and combinations of weather data. Promising results were obtained with an excellent real-time overall performance for next-day forecasting with a Root Mean Square Error (RMSE) value of 0.0706 using the Bayesian regularization algorithm with 28 hidden layers and all weather inputs. The Levenberg-Marquardt algorithm provided a 0.0753 RMSE using 23 hidden layers for the same set of learning inputs. This research shows that the Bayesian regularization algorithm outperforms the reported real-time prediction systems for the PV power production.

**Keywords**—Solar photovoltaic; solar irradiance; PV power forecasting; machine learning; artificial neural networks; Levenberg-Marquardt; Bayesian regularization

## I. INTRODUCTION

The rising fuel costs and increasing energy demands with the ongoing industrial growth and environmental awareness have engaged to the importance of new renewable energy sources such as the solar Photovoltaic (PV) systems [1, 2]. As one of the most important renewable energy sources, PV energy is becoming the dominant clean and reliable energy source that is widely used around the world without caus-

ing any damage to the environment. Mentioning the light-electricity process, the term “Photovoltaic” is first used by Alfred [3], as the light conversion process into electricity. There are two modes of installation for solar PV power plants: grid-tied and off-grid systems [4]. The first mode is widely used and proven to be hugely beneficial. It depends on the variable weather conditions according to the geographical area of the system which is the reason why it was known as uncertain, uncontrollable, and non-scheduling power source [5]. The second mode, off-grid systems, is used for isolated or remote areas that are normally on a smaller scale.

Many studies were reported in the literature suggesting different modeling, simulation, and prediction methods for the expected power production of solar PV plants for the purpose of improving the investment feasibility and maintaining a stable power quality and scheduling [6, 7]. Fonseca [8], compared the accuracy of one-day ahead prediction for the power produced by 1MW PV System using two methods: Support Vector Machines (SVM) and Multilayer Perceptron (MP) Artificial Neural Networks (ANNs). It was found that the two algorithms approximately obtained almost the same accuracy with 0.07 KWh/m<sup>2</sup> and 0.11 KWh/m<sup>2</sup> Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), respectively.

Various forecasting methods of PV power output were reviewed in [9]. It was demonstrated that any model uses numerically predicted weather data will not take into account the effect of cloud cover and cloud formation when initializing, therefore sky imaging and satellite data methods used to predict the PV power output with higher accuracy. The article also outlined some key factors affecting the accuracy of prediction, such as forecast horizon, forecasting interval width, system size and PV panels mounting method (fixed or tracking). A model using multilayer perceptron-based ANN was proposed in [5] for one day ahead forecasting. The daily solar power output and atmospheric temperature for 70 days used for training the ANN. For the different settings of the ANN model (number of hidden layers, activation function, and learning rule), the minimum MAPE achieved was 0.855%.

The aim of the work published in [10] was to study the effect of forecast horizon on the accuracy of the method used to predict the PV power production, which was Support Vector Regression (SVR) using numerically predicted weather data. Two forecast horizons studied: up to 2 and 25 hours ahead. As expected, the forecasting of up to 2 hours ahead was more accurate with RMSE and MAE increased 13% and 17%, respectively, when the forecast horizon was up to 25 hours ahead. Cococcioni [11], developed and validated a model that adapted an ANN with tapped delay lines and built for one day ahead forecasting. The inputs were the irradiation and the sampling hours. The model achieved seasonal MAE ranging from 12.2% to 26% in spring and autumn, respectively. Monteiro [12], compared two short-term forecasting models: the analytical PV power forecasting model (APVF) and the MP PV forecasting model (MPVF), with both of the models using numerically predicted weather data and past hourly values for PV electric power production. The two models achieved similar results (RMSE varying between 11.95% and 12.10%) with forecast horizons covering all daylight hours of one day ahead, thus the models demonstrated their applicability for PV electric power prediction.

Leva [13], proposed a new Physical Hybrid ANN (PHANN) method to improve the accuracy of the standard ANN method. The hybrid method is based on ANN and clear sky curves for a PV plant. The PHANN method reduced the Normalized MAE (NMAE) and the Weighted MAE (WMAE) by almost 50% in many days compared to the standard ANN method. In [14], the PV energy production for the next day with 15-minutes intervals was accurately predicted with an SVM model that uses historical data for solar irradiance, ambient temperature, and past energy production. The method demonstrated very good accuracy with  $R^2$  correlation coefficients of more than 90%, and the coefficient was strongly dependent on the quality of the weather forecast.

In our previous work [15], we proposed an initial real-time forecasting model for the PV power production using ANNs based on the available solar irradiation records for the last few days. In this research work, ANNs were optimized comparing the Levenberg-Marquardt (LM) and Bayesian Regularization (BR) algorithms to analyze and correlate the available data of temperature, solar irradiance, timing, and the generated solar PV power. The suggested system provides real-time PV energy forecasts for the next 24 hours based on real-time weather data for the last week.

## II. PV AND WEATHER DATA

### A. PV Systems

There are four separate PV systems installed at the university campus for a total generation capacity of 550KWp:

- PV ASU00 (The Test Field): This system was installed in 2013 with a capacity of 56.4KWp including a CPV tracker, a Polycrystalline tracker, Poly-crystalline and Mono-crystalline panels (South and East/West oriented), and thin film panels.
- PV ASU08 (The Library): A rooftop-mounted 130.1KWp system of Yingli Solar panels and SMA sunny tripower inverters.



Fig. 1. Rooftop mounted Solar panels on top of the engineering building.

- PV ASU09 (Faculty of Engineering): This is the largest rooftop-mounted PV system at ASU that is installed on top of the faculty of engineering building with a capacity of 264KWp [16]. It consists of 14 SMA sunny tripower inverters (17KW and 10KW) connected with Yingli Solar (YL 245P-29b-PC) panels that are tilted by 11° and oriented 36° (S to E) (see Fig. 1).
- PV ASU10 (Deanship of Student Affairs): A rooftop mounted 117.4KWp system of Yingli Solar panels and SMA sunny tripower inverters.

### B. ASU Weather Station

ASU's weather station was installed in 2015 to be the first of its kind in Jordan providing a wide range of weather data measured by the latest sensors and devices [17]. It is located about 175m from the engineering building as shown on the map of Fig. 2. The station is equipped with many instruments to measure:

- The wind speed and direction (4 altitudes between 10-36m).
- Ambient temperature (3 altitudes between 1-35m).
- Relative humidity (at altitudes 1m and 35m).
- Global solar irradiance (including separate direct radiation and diffuse radiation records).
- Subsoil and soil surface temperature.
- Barometric pressure.
- Precipitation amounts.

The real-time weather measurements obtained by the station are updated continuously and published on the website: <http://energy.asu.edu.jo> as depicted in Fig. 3.

In this research work, we created a two years dataset using the available hourly records of weather and PV energy data for the duration between 16 May 2015 and 15 May 2017. This dataset for 731 days includes 17544 weather records and 17544 PV power values. PV power records were obtained only from the PV ASU09 system as it is the largest and most stable system on campus.



Fig. 2. A map showing part of ASU's campus.

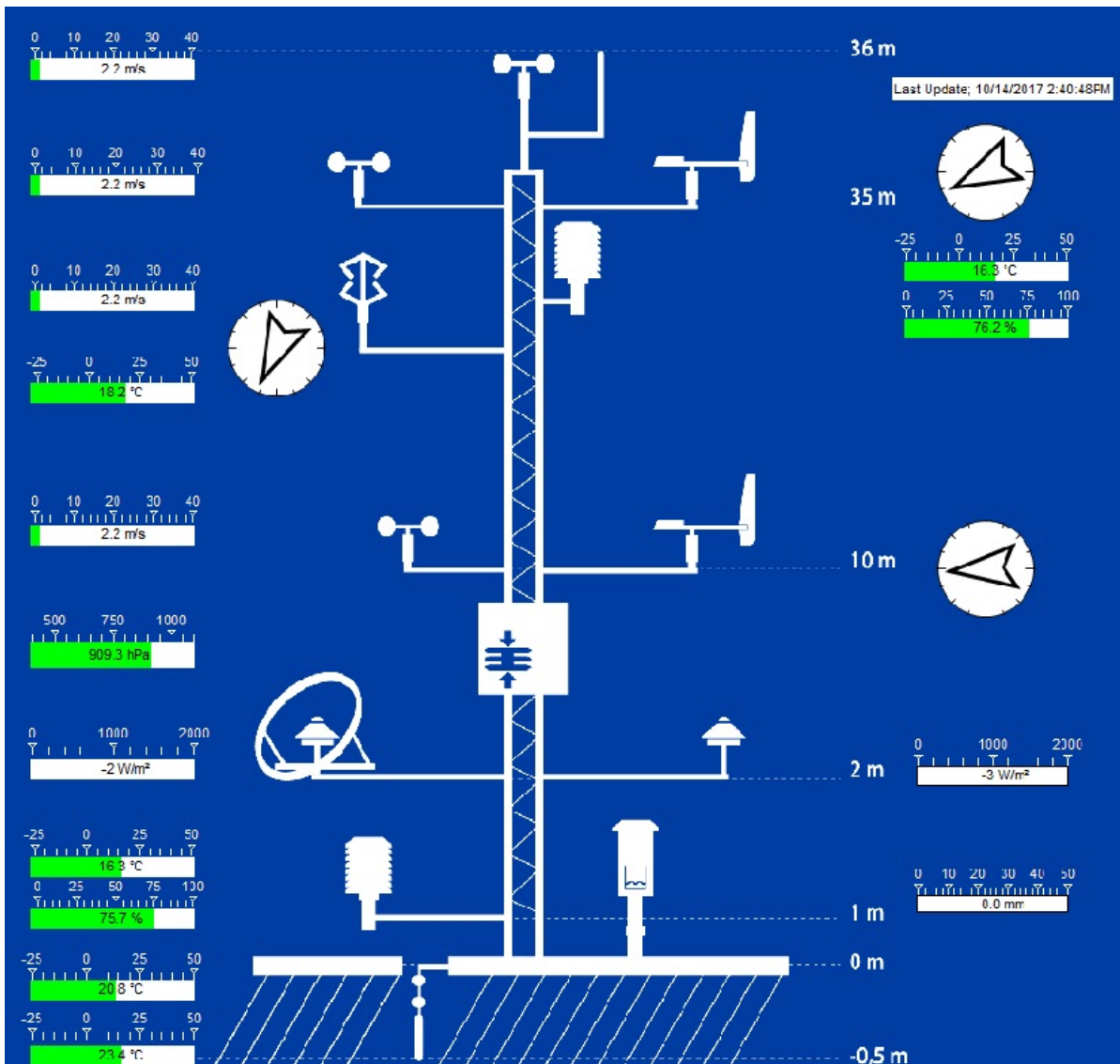


Fig. 3. Real-time weather station data available at [17].



### III. FORECASTING SYSTEM

The suggested forecasting system can be represented by the block diagram of Fig. 4 and is described in the following subsections.

#### A. Data Filtering and Association

The collected weather and PV power data are first filtered, as shown in Fig. 4, to filter out any missing records which guarantee the consistency of the dataset. This includes any weather information with no PV power values associated at the same time or any PV power records with a missing weather data.

Based on the timing data, associations were found by matching solar PV power records with weather records including the record time, temperature, and global solar irradiance. To obtain homogeneous data and reliable machine learning experiments, the final dataset was normalized between 0 and 1.

#### B. LM and BR ANNs

Compared with similar algorithms [18, 19, 20], ANNs are known as one of the most powerful machine learning techniques with a wide range of applications [21, 22, 23]. ANNs map non-linear inputs through adjustable weights into the desired targets. The network is created by three layers: the input, hidden, and output layers [24] as illustrated in the example of Fig. 5 for a network of 11 inputs, 23 hidden layers, and one output.

ANNs showed excellent learning and classification performances while dealing with real-world sensor data [25, 26, 27].

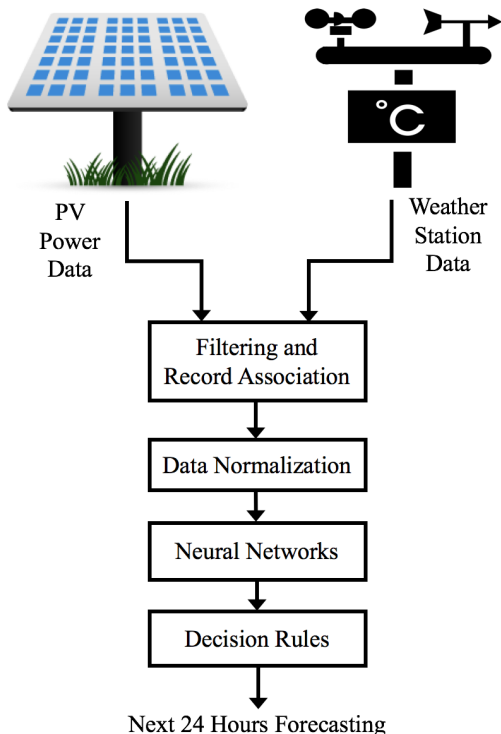


Fig. 4. Implementation flowchart for the proposed forecasting system.

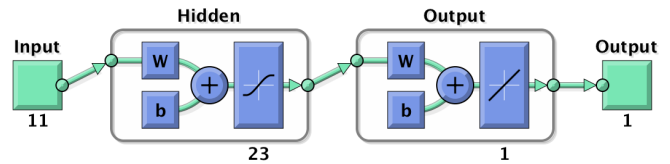


Fig. 5. An example for the structure of ANNs with 23 hidden layers.

In this work, we applied the LM and BR to neural networks and compared the learning performances using different training configurations.

The LM backpropagation optimization algorithm has been initially reported in [28] and has been applied later to neural networks in [29, 30]. The LM ANNs algorithm is implemented in MATLAB and it is known as the fastest backpropagation supervised algorithm especially while training feedforward ANNs with moderate sizes [31]. The BR backpropagation algorithm has been introduced in [32] and [33] and it is implemented in MATLAB [34]. Both of the LM and BR ANNs calculate the neural network errors' derivative functions with respect to weights and biases to obtain a Jacobian matrix that is used for calculations which means that the performance can only be measured by the mean squared errors [29].

#### C. Training and Testing Experiments

Data of the global solar irradiance ( $Rad_d(t)$ ) and the temperature ( $Temp_d(t)$ ) at the altitude of 1m are used to form the weather information vector  $W_d(t)$  at time  $t$  of day  $d$ . Two neural network models were created based on the LM and BR algorithms with the target function of the mean PV power  $P_d(t)$ . The inputs to these models are the current time stamp from the beginning of the current year ( $T_d(t)$ ) and the available weather vectors  $W_d(t)$  at the same time  $t$  over the previous five days before day  $d$ . So, as depicted in Fig. 6, each input sample of the training dataset consists of:

$$T_d(t), W_{d-1}(t), W_{d-2}(t), W_{d-3}(t), W_{d-4}(t), W_{d-5}(t) \quad (1)$$

and one output value  $P_d(t)$  with

$$W_{d-i}(t) = \{Rad_{d-i}(t), Temp_{d-i}(t)\}, \quad i = 1, 2, \dots, 5 \quad (2)$$

In this work, the MATLAB Neural Networks toolbox was used for huge amounts of training, validation, and testing experiments while using different input combinations and varying the number of hidden layers from 1 to 30. The used set of inputs are:

- ALL inputs (1 time value, 5 radiation values, and 5 temperature values).
- Only 5 radiation values.
- Only 5 temperature values.
- 6 inputs (1 time value and 5 radiation values).
- 6 inputs (1 time value and 5 temperature values).
- 3 inputs (1 time value, 1 radiation value, and 1 temperature value).

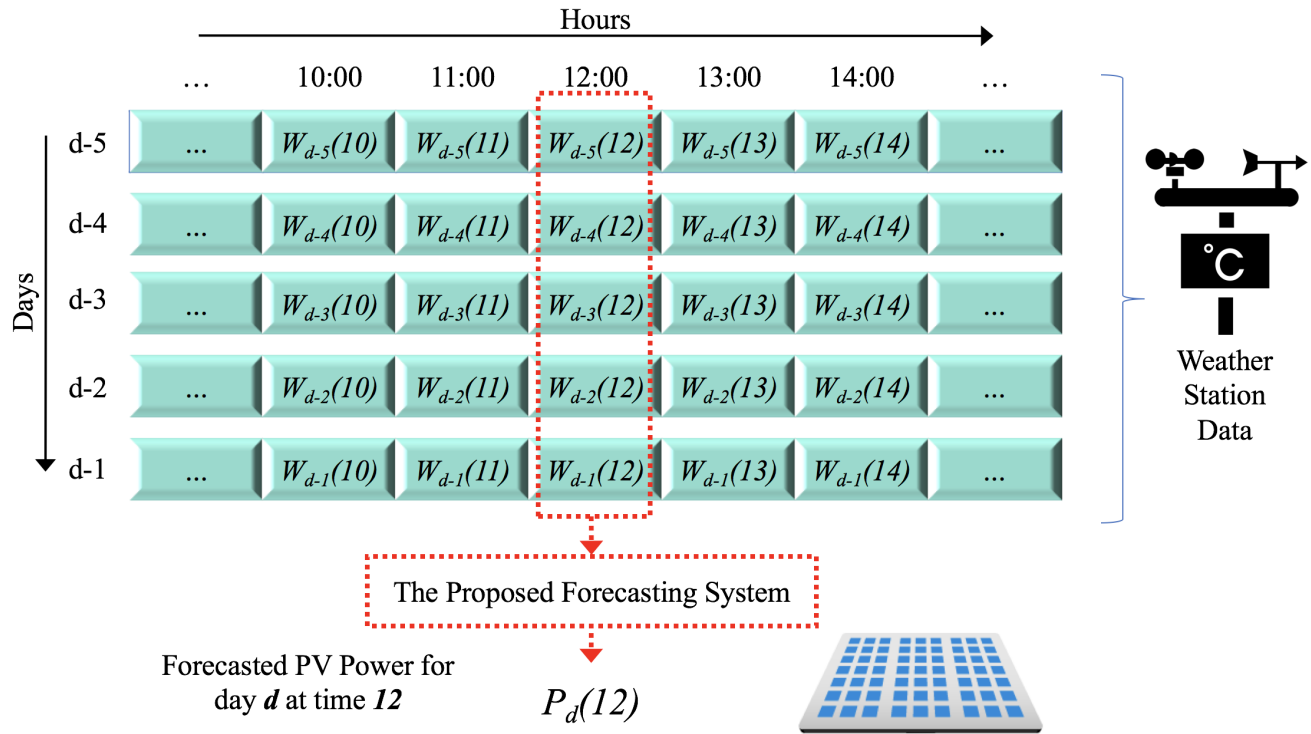


Fig. 6. Next-day PV forecasting based on the weather data of the previous five consecutive days.

- 10 inputs (5 radiation values and 5 temperature values).

Ten experiments were handled at each value for the number of hidden layers. At each experiment, the samples of the dataset were randomly mixed to generate the sub-datasets: 80% for training, 5% for validation, and 15% for testing. Then, the performance was evaluated by calculating the average RMSE for each of ten experiments using:

$$RMSE = \sqrt{\sum_{n=1}^N D_n^2} \quad (3)$$

The network configurations that provided the best performances are listed in Table I. It can be concluded from the results that the best training/testing experiments provided an average RMSE of 0.0706 and a best testing correlation coefficient of  $R=0.9660$  and mean square error of 0.00485 while using all inputs to the BR ANNs with 28 hidden layers for the testing performance illustrated in Fig. 7 and 8.

A histogram of 20 Bins is depicted in Fig. 9 for the overall errors. These results are very low compared to the methods and measures reported in the literature and related to the current research as summarized in Table II.

#### IV. CONCLUSION

In this research, a predictive forecasting model is proposed by applying the Levenberg-Marquardt and Bayesian Regularization algorithms to neural networks for the purpose of correlating historical weather data to photovoltaic outputs. Two years of hourly data were processed to associate the available

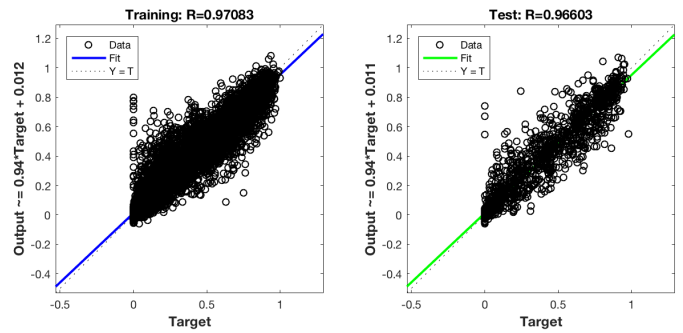


Fig. 7. Correlation coefficients calculations for the best performance.

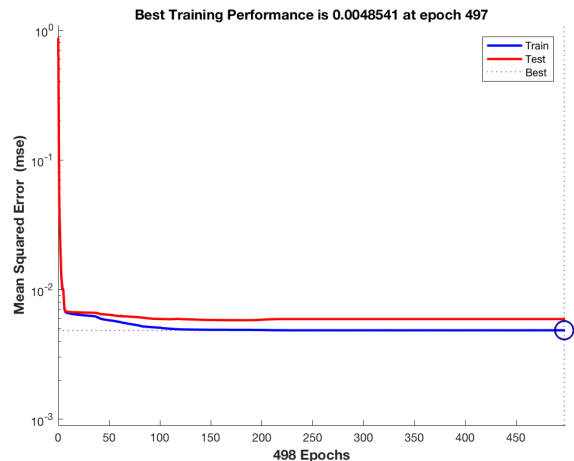


Fig. 8. Best ANN training performance using 28 hidden layers.

TABLE I. RESULTS OBTAINED USING DIFFERENT SETS OF INPUTS

Inputs	Levenberg-Marquardt			Bayesian Regularization		
	Hidden Neurons	Average RMSE	Testing R	Hidden Neurons	Average RMSE	Testing R
ALL	23	0.0753	0.9532	28	0.0706	0.9660
Rad	23	0.0828	0.9417	30	0.0797	0.9514
Temp	23	0.2340	0.9034	28	0.2285	0.8911
Time, Rad	24	0.0819	0.9420	26	0.0774	0.9501
Time, Temp	27	0.2107	0.9041	30	0.2061	0.8921
Time, Rad, Temp	23	0.0935	0.9271	14	0.0941	0.9384
Rad, Temp	27	0.0969	0.9385	15	0.0966	0.9295

TABLE II. A COMPARISON BETWEEN THE FORECASTING PERFORMANCE FOR DIFFERENT METHODS AND MEASURES RELATED TO THE CURRENT RESEARCH

Reference	Forecasts	Method	Measure	Result
[5]	One day ahead	MP-ANNs	MAPE	0.855%
[8]	One day ahead	SVM	MAE, RMSE	0.07 KWh/m <sup>2</sup> , 0.12 KWh/m <sup>2</sup>
[8]	One day ahead	MP-ANNs	MAE, RMSE	0.11 KWh/m <sup>2</sup> , 0.12 KWh/m <sup>2</sup>
[10]	2 hours ahead	SVR	MAE	0.065 MWh
[10]	25 hours ahead	SVR	MAE	0.076 MWh
[11]	One day ahead	ANN	MAE	0.122
[12]	One day ahead	APVF	RMSE, MAE	0.121, 0.0597
[12]	One day ahead	MPVF	RMSE, MAE	0.1195, 0.0646
[13]	One day ahead	PHANN	NMAE, WMAE	50% error reduction
[14]	One day ahead	SVM	R <sup>2</sup> correlation coefficients	90%
[15]	One day ahead	ANN	RMSE, R <sup>2</sup> correlation coefficients	0.0721, 96.7%
This work	One day ahead	LM-ANN	RMSE, R <sup>2</sup> correlation coefficients	0.0753, 95.32%
This work	One day ahead	BR-ANN	RMSE, R <sup>2</sup> correlation coefficients	0.0706, 96.60%

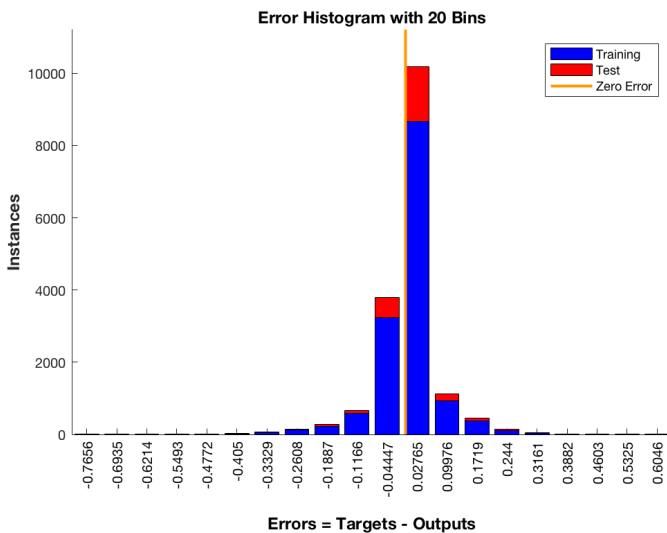


Fig. 9. Error histogram of 20 Bins.

temperature and global radiation records to the generated PV power. The associated datasets were used as a source of learning for a neural network model that use real-time weather data to provide PV power forecasts for the next 24 hours.

After a vast amount of training/testing experiments, excellent prediction results were obtained using the BR ANNs based on time, temperature, and radiation inputs. These predictions can be used by many energy management systems and power control systems of grid-tied PV plants. The proposed model is being developed into a real-time online application in our near future work.

#### ACKNOWLEDGEMENT

The authors would like to acknowledge the financial support received from Applied Science Private University that helped in accomplishing the work of this article.

#### REFERENCES

- [1] P. Ramsami and V. Oree, "A hybrid method for forecasting the energy output of photovoltaic systems," *Energy Conversion and Management*, vol. 95, no. Supplement C, pp. 406 – 413, 2015.
- [2] I. E. Agency, "Technology roadmap: Solar photovoltaic energy 2014 edition," Paris, France, 2014, last accessed: 2017-10-2. [Online]. Available: <https://goo.gl/6opzZ8>
- [3] A. Smee, *Elements of electro-biology: or the voltaic mechanism of man; of electro-pathology, especially of the nervous system; and of electro-therapeutics*. London, UK: Longman, Brown, Green, and Longmans, 1849.
- [4] V. Karthikeyan, S. Rajasekar, V. Das, P. Karuppanan, and A. K. Singh, *Grid-Connected and Off-Grid Solar Photovoltaic System*. Cham: Springer International Publishing, 2017, pp. 125–157.
- [5] R. Muhammad Ehsan, S. P. Simon, and P. R. Venkateswaran, "Day-ahead forecasting of solar photovoltaic output power using multilayer perceptron," *Neural Computing and Applications*, vol. 28, no. 12, pp. 3981–3992, Dec 2017.
- [6] M. Elyaqouti, L. Bouhouch, and A. Ihlal, "Modelling and predicting of the characteristics of a photovoltaic generator on a horizontal and tilted surface," *International Journal of Electrical and Computer Engineering*, vol. 6, no. 6, pp. 2557–2576, 2016.
- [7] S. Mankour, A. Belarbi, and M. Benmessaoud, "Modeling and simulation of a photovoltaic field for 13 kw," *International Journal of Electrical and Computer Engineering*, vol. 7, no. 6, pp. 3271–3281, 2017.
- [8] J. G. da Silva Fonseca Junior, T. Oozekia, T. Takashimaa, G. Koshimizub, Y. Uchidab, and K. Ogimoto, "Forecast of power production of a photovoltaic power plant in japan with multilayer perceptron artificial neural networks and support vector machines," in *26th European Photovoltaic Solar Energy Conference and Exhibition*. WIP-Renewable Energies, Sep. 2011, pp. 4237–4240.
- [9] A. Tuohy, J. Zack, S. Haupt, J. Sharp, M. Ahlstrom, S. Dise, E. Gritmit, C. Moehrlen, M. Lange, M. Garcia Casado, J. Black, M. Marquis, and C. Collier, "Solar forecasting: Methods, challenges, and performance," *IEEE Power and Energy Magazine*, vol. 13, pp. 50–59, 11 2015.
- [10] J. G. da Silva Fonseca, T. Oozeki, T. Takashima, G. Koshimizu, Y. Uchida, and K. Ogimoto, "Photovoltaic power production forecasts with support vector regression: A study on the forecast horizon," in *2011 37th IEEE Photovoltaic Specialists Conference*, June 2011, pp. 002 579–002 583.
- [11] M. Cococcioni, E. D'Andrea, and B. Lazzerini, "One day-ahead forecasting of energy production in solar photovoltaic installations: An empirical study," *Intelligent Decision Technologies*, vol. 6, no. 3, pp. 197–210, Aug. 2012.

- [12] C. Monteiro, L. A. Fernandez-Jimenez, I. J. Ramirez-Rosado, and P. M. Lara-Santillan, "Short-term forecasting models for photovoltaic plants: Analytical versus soft-computing techniques," *Mathematical Problems in Engineering*, vol. 2013, pp. 1–9, 2013.
- [13] A. Dolara, F. Grimaccia, S. Leva, M. Mussetta, and E. Ogliaari, "A physical hybrid artificial neural network for short term forecasting of pv plant power output," *Energies*, vol. 8, no. 2, pp. 1138–1153, 2015.
- [14] R. Leone, M. Pietrini, and A. Giovannelli, "Photovoltaic energy production forecast using support vector regression," *Neural Computing and Applications*, vol. 26, no. 8, pp. 1955–1962, Nov. 2015.
- [15] M. H. Alomari, J. Adeeb, and O. Younis, "Solar photovoltaic power forecasting in jordan using artificial neural networks," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 1, pp. 497–504, February 2018.
- [16] ASU, "Pv system asu09: Faculty of engineering," 2017, last accessed: 2017-10-14. [Online]. Available: <https://goo.gl/cxGYVb>
- [17] U. ASU, "Weather station," 2017, last accessed: 2017-10-14. [Online]. Available: <http://web.asu.edu.jo/wsp>
- [18] R. Qahwaji, M. Al-Omari, T. Colak, and S. Ipson, "Using the real, gentle and modest adaboost learning algorithms to investigate the computerised associations between coronal mass ejections and filaments," in *Mosharaka International Conference on Communications, Computers and Applications*, 2008, pp. 37–42.
- [19] M. AL-Omari, R. Qahwaji, T. Colak, S. Ipson, and C. Balch, "Next-day prediction of sunspots area and mcintosh classifications using hidden markov models," in *2009 International Conference on CyberWorlds*, Sept 2009, pp. 253–256.
- [20] M. Al-Omari, R. Qahwaji, T. Colak, and S. Ipson, "Machine leaning-based investigation of the associations between cmes and filaments," *Solar Physics*, vol. 262, no. 2, pp. 511–539, Apr 2010.
- [21] M. AL-Omari, R. Qahwaji, T. Colak, S. Ipson, and C. Balch, "Next-day prediction of sunspots area and mcintosh classifications using hidden markov models," in *2009 International Conference on CyberWorlds*, Sept 2009, pp. 253–256.
- [22] M. H. Alomari, R. S. Qahwaji, and S. S. Ipson, *Applied Machine Learning for Solar Data Processing: Developing Automated Technologies for Knowledge Extraction and Prediction of Solar Activities Using Machine Learning*. Germany: LAP Lambert Academic Publishing, 2011.
- [23] M. H. Alomari, "Mind controlled universal tv remote control," in *2016 International Conference on Image Processing, Production and Computer Science*, London (UK), Mar 2016, pp. 85–91.
- [24] N. J. Nilsson, "Introduction to machine learning. an early draft of a proposed textbook," 1996.
- [25] T. M. Mitchell, *Machine Learning*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 1997.
- [26] M. Alomari, E. Awada, and O. Younis, "Subject-independent eeg-based discrimination between imagined and executed, right and left fists movements," *European Journal of Scientific Research*, vol. 118, no. 3, pp. 364–373, 02 2014.
- [27] R. Qahwaji, T. Colak, M. Al-Omari, and S. Ipson, "Automated prediction of cmes using machine learning of cme–flare associations," *Solar Physics*, vol. 248, no. 2, pp. 471–483, Apr 2008.
- [28] D. W. Marquardt, "An algorithm for least-squares estimation of non-linear parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.
- [29] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the marquardt algorithm," *IEEE Transactions on Neural Networks*, vol. 5, no. 6, pp. 989–993, Nov 1994.
- [30] M. Hagan, H. Demuth, and M. Beale, *Neural Network Design*. Boston, MA, USA: PWS Publishing, 1996.
- [31] MathWorks, "trainlm: Levenberg-marquardt backpropagation," 2017, last accessed: 2017-12-14. [Online]. Available: <https://www.mathworks.com/help/nnet/ref/trainlm.html>
- [32] D. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.
- [33] F. D. Foresee and M. T. Hagan, "Gauss-newton approximation to bayesian learning," in *Proceedings of the International Joint Conference on Neural Networks*, June 1997, pp. 1930–1935.
- [34] MathWorks, "trainbr: Bayesian regularization backpropagation," 2017, last accessed: 2017-12-14. [Online]. Available: <https://www.mathworks.com/help/nnet/ref/trainbr.html>

# Improving Energy Conservation in Wireless Sensor Networks using Energy Harvesting System

Abdul Rashid, Faheem Khan, Toor Gul,  
Fakhr-e-Alam, and Shujaat Ali  
Department of Computer Science  
Bacha Khan University Charsadda,  
Pakistan

Samiullah Khan,  
Fahim Khan Khalil  
Institute of Business Management Sciences  
The University of Agriculture Peshawar  
Pakistan

**Abstract**—Wireless Sensor Networks assume an imperative part to monitor and gather information from complex geological ranges. Energy conservation plays a fundamental role in WSNs since such sensor networks are designed to be located in dangerous and non-accessible areas and has gained popularity since the last decade. The main issue of Wireless Sensor Network is energy consumption. Therefore, management of energy consumption of the sensor node is the main area of our research. Sensor nodes use non-changeable batteries for power supply and the lifetime of Sensor node greatly depends on these batteries. The replacement of these batteries is very difficult in many applications, such as an alternative solution to this problem is to use Energy Harvesting system in Wireless Sensor Network to provide a permanent power supply to sensor nodes. This process of extracting energies from nature and converting it into electrical energy is called energy harvesting. Energy can be harvested from the environment for sensor nodes. There are many sources of energies in nature like solar, wind and thermal which can be harvested and used for WSNs. In this research, we suggest to use energy harvesting system for Cluster Heads in a clustering based Wireless Sensor Networks. We will compare our proposed technique to a well-known clustering algorithm Low Energy Adaptive Cluster Hierarchy (LEACH).

**Keywords**—Wireless sensor network; Low Energy Adaptive Cluster Hierarchy (LEACH); clustering; cluster head; energy harvesting; energy conservation

## I. INTRODUCTION

A Wireless Sensor Network (WSN) is composed of a large number of small dispersed devices known as sensor nodes that are closely deployed in the environment sense changes. The location of nodes in a sensor network may not be predetermined and usually have a unique sensor node called Base Station (BS). All member nodes will forward data to BS either directly or through multi hop transmission. A base station may be either static or dynamic sensor node and provides wireless connectivity to its users. It is usually more capable than other sensor nodes in the WSN [1], [2] (see Fig. 1).

When sensor nodes in a WSN [3] run out of energy, they stop working which causes the whole sensor network to fail. Therefore, the main issue of WSN is energy conservation. Hence, such protocols should be designed which use minimum energy and should be utilized during sensing, processing and transmission. In Wireless Sensor Network, energy consumption is a vital challenge for sensor node for sensing and transmitting the data to the closest SN or the base station. The

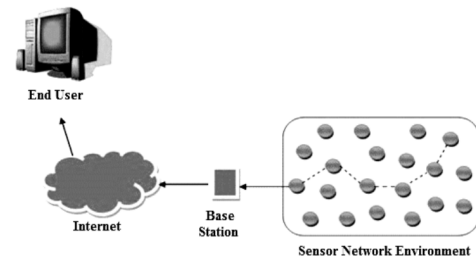


Fig. 1. Simulation results.

key requirement of a WSN is to minimize energy consumption and prolong network lifetime. Communication among sensor nodes stops when these sensor nodes lose their battery power [4].

### A. Challenges to Wireless Sensor Network

Several challenges still need to be faced by WSNs. The main challenges and essential design limitation which affect the performance of Wireless Sensor Network are discussed here [5]:

1) *Resource limitations*: Sensing nodes are bounded by energy, data processing abilities, memory and the data transmission rate to be gained.

2) *Security*: The Wireless Sensor Network must be secure to control illegal entities spreading false data to the sensor node or giving wrong information to the other sensor nodes and possibly causing significant damage to the sensor nodes.

3) *Self-Management*: After deployment of the sensor nodes, the user has less interaction with these sensor nodes with no infrastructure support or the capability to keep and repair itself. Therefore, the sensor nodes must be self-organised in a way that must be configured, work together with other sensor nodes.

4) *Heterogeneity*: It is a group of sensor nodes which are not identical and do not have similar capabilities, i.e. some sensor nodes are more powerful than others. Heterogeneity arises when two different sensor networks need to communicate with each other. There will be some mechanism which is required to enable efficient information exchange among these networks.

5) *Other Challenges*: Several other challenges may affect the design of Wireless Sensor Networks. For example, a group of sensor nodes is combined into a portable object, like a robot or automobile. This results in constant sensor network topologies being altered, which require repetitive changes. There is a need for routing (e.g., modifying neighbour lists), Media Access Control (e.g., modifying density), and data gathering.

### B. Energy Harvesting in WSN

Energy harvesting [6] refers to collecting energy from surrounding (sun, wind) or other sources of energies (body heat, finger stroke, foot strikes) and converting these energies to electrical energy. Energies from external sources can be harvested to power the nodes to increase their lifetime. Energy harvesting (EH) has significantly improved the lifetime of a WSN and enabling new devices much more reliable.

Thanks to Energy Harvesting, the general model of getting the longest possible lifetime while still giving a good sufficient result has moved in support of providing best possible outcomes with the amount of available of energy.

Energy Harvesting Wireless Sensor Networks are normal WSNs, where additional power sources assist the primary battery or entirely replaced by them. The main idea is to retrieve the power present in the nearby surroundings of a sensor node, transform it to a usable form using appropriate types of transducers and ultimately utilise it to power the sensor node themselves. If the power source used to stimulate the sensor network have enough energy level and always available then, EH supply a network with unlimited power and an approximately infinite lifetime.

Lastly, security for Wireless Sensor Networks is the critical concern. Given the level of confidence that we place into these networks, making sure that the data that a sensor node transmits should not be accessible to outside unauthorised users, or that these similar data can be altered only by the authorised users. These are the objectives of computer security [7].

There are some energy-harvesting sensor nodes which are commercially available in the market such as Crossbow MICAZ node with solar energy-harvesting sample circuit. Currently, available energy harvesting design for sensor nodes is classified into the following types:

- *Harvest-Use architecture*
- *Harvest-Store-Use architecture.*

In the harvest-use architecture [8], the harvesting architecture directly provides energy to the sensors. The harvest-store-use architecture [9] consists of a unit for storing harvested energy and also powers the SNs. When the harvested energies are more than the needs of the SNs, then the extra energies are stored for future use. Secondary storage is backup storage for situations when the Primary storage is running out of energy. In day time, harvested energies are used directly as well as stored for future use. In night time, the stored power is utilized to provide energies to the SNs.

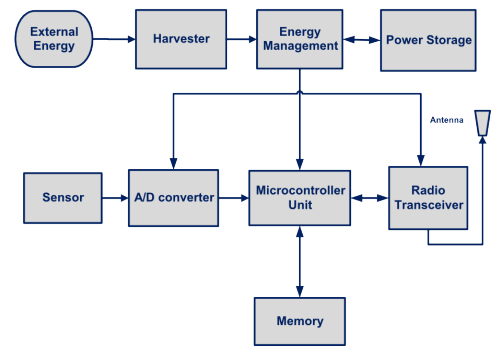


Fig. 2. System architecture of a wireless node with energy harvesters.

### C. Sensor Node Architecture with Harvesting System

The structure of harvesting based sensor node contains the following components (see Fig. 2):

- The harvester, which converts external energy or human-created energy to electrical energy.
- An energy management unit, that gathers electrical energy from the harvester for storage or distributes it to the other components for direct use.
- Power storage, store energy for future use.
- A microcontroller unit.
- A transceiver, for receiving and transmitting information.
- Sensor.
- An Analog/Digital converter, to convert the analog signal into digital signal and provides it available to the microcontroller unit for further processing.
- Memory unit for data storage.

### D. Application of Energy Harvesting

Several daily life applications are based on harvesting system. Energy harvesting from external sources, where a distant application is installed, and where such external energy source is unlimited, is the best alternative to expensive batteries. This is free power source available during the lifespan of the application [10].

Energy harvesting concept is used in a variety of applications in WSN. Some ad-hoc networks use minimum energy path to improve energy consumption at an SN such that the limited resources at SNs can be used more efficiently. In the meantime, if a low energy route is regularly used, this will cause decreasing in the SN energy along that route and may even cause network partition. Hence optimum routes based on energy efficient routing protocols and energy harvesting may be used to enhance the performance of the sensor network. Energy harvesting eliminates the need of battery replacement, which results to extend the lifetime of SNs.

Energy harvesting also has some industrial applications, e.g. the development of TEG using an AC Condenser, which uses measurements from thermometer which is placed in the condenser unit is the latest application [11]. The TEG had an

energy producing capability of 20Watt. Piezoelectric shoes are a new application of energy harvesting in which piezoelectric material is implanted in the sole of the shoe. A sensor inside the shoes senses the running, walking, or other vibrations occurring on the piezoelectric material and transforms it into electrical energy which is then used to power small electronic devices. An average of  $331 \mu\text{W}/\text{cm}^2$  was generated while walking. Another application of energy harvesting is in a club atmosphere in which energy harvesting conversion technique is placed on the dance floor in which electrical energy is produced from the dancing motion. The produced energy was utilised to provide energy to LED lighting systems in the club.

The energy harvesting also has applications in portable medical devices. These devices are likely to be small in size, lightweight and either to be wearable or inserted into the body. The dependence on batteries needs to be reduced in this field by using energy harvesting system.

Further research is going on, and more applications are introduced. These applications are now focusing on the grouping of many sources rather using a single source as can be used in case of Powerball, piezoelectric shoe and human-powered dance floor. Here the mechanical energy in the shape of vibrations is converted into electrical energy.

#### E. Nature of Energy Sources

Let us discuss the nature of energy (power) availability in everyday application situations.

1) *Photovoltaic energy*: Whereas panels that are able of energy harvesting from internal lighting have been developed, most photovoltaic sheets are used for energy harvesting from the sunlight. The exchange efficiency and electrical properties of current photovoltaic sheets are well studied [12]. Depends upon the season, environmental location and the existence or nonexistence of shadow creating entities such as buildings and trees, solar energy follows a tendency such that it raised as the day improvement to the point of time where it is maximum and then decreases gradually. After this, a stage of no energy availability follows at night.

2) *Vibration energy*: Vibration energy is taking out by piezoelectric fibres that are usually made of lead zirconate titanate that shows a noticeable piezoelectric effect. Roundy and some others researchers have confirmed that these devices can extract up to  $101 \text{ W}/\text{cm}^3$  [13]. If such an energy output were to maintained for only a one second, the sensor could monitor and forward thirty bytes of data at a transmission power of +5dBm which is a usual application requirement in WSNs. Also, the difference is seen at different periods, for instance, depending on the traffic on a bridge or the movement of people in a room.

3) *Mechanical energy*: Mechanical energy is harvested by converting mechanical energy into electrical energy by using vibrations, pressure, and strain from high-pressure motors and force.

4) *Electromagnetic energy*: Electromagnetic energy is harvested by using Faradays law of electromagnetic induction. Here, an inductive spring-mass system is used for converting mechanical energy into electrical. It generates voltage by the movement of a mass of magnetic material through a magnetic eld.

5) *Thermoelectric energy*: Thermoelectric energy is harvested from temperature difference (thermal gradients) using Thermoelectric Power Generators (TEGs).

6) *Radio frequency energy*: Radio Frequency energy harvesting is the process of converting electromagnetic energy into electrical energy. There are several Radio Frequency power sources such as television and radio broadcasting, mobile phone, Wi-Fi communications, microwaves and EM signals.

7) *Wind energy*: Wind energy is harvested by converting air ow (e.g., wind) energy into electricity. The wind turbine is used for linear motion coming from wind for generating electricity.

8) *Biochemical energy*: Biochemical energy is harvested from converting oxygen and endogenous substances into electricity through electrochemical reactions.

In the rest of the paper, Section II gives an overview about the stat of art schemes related energy harvesting systems. The proposed technique thoroughly explained in Section III. Realistic simulation scenarios and their results are discussed in Section IV. Finally, Section V gives the concluding remarks about the finding of this research article.

## II. LITERATURE REVIEW

In this section, a portion of the topology control calculations in Energy Harvesting-Wireless Sensor Networks and also in battery-fuelled Wireless Sensor Networks are exhibited. We give an outline thought of these calculations. Energy-aware routing protocols for WSN are a very demanding research area. Due to the restrictions of wireless sensor nodes, conventional routing protocols are not appropriate for WSN. Data aggregation is a simple mechanism for decreasing the number of packets forwarded over the network because data processing for excessive aggregation amount of power as compared to transmitting multiple packets having identical data [14], [15].

LEACH is a well-known clustering based protocol [16]. In LEACH sensor nodes are organised into the cluster. Each cluster has cluster head and member nodes. Cluster heads in each cluster are selected randomly. The main disadvantage of LEACH is that if a sensor node with less residual energy is selected as cluster head would die quickly; ultimately the whole cluster would become non-functional. LEACH performs local processing to reduce the amount of data being transmitted to the BS, therefore reducing energy consumption and improving network lifetime.

In this study, a game theory-based dispersed Energy-Harvesting-Aware (EHA) algorithm is proposed [17], which represents the behaviours of sensors as a game. This effort analyses the energy expenditure rate and energy-harvesting rate of every sensor node at different times. In this approach, the high harvesting energy sensor nodes assist with the low harvesting energy sensor nodes to keep the connectivity of the sensor network. The proposed algorithm first builds a beginning topology based on the Directed Local Spanning Sub-graph (DLSS) algorithm. Then every sensor node tries to and an adjacent node that covers up the remote neighbour of sensor node by adjusting the communication power stepwise.

In this approach, the authors proposed a protocol in EH-WSN with a hybrid storage model. The best proportion that

reduces outage probability is extracted, and some essential guiding principle is given [18]. The development of Carbon dioxide (CO<sup>2</sup>) sensor nodes that are powered by artificial light was proposed in [19]. For wireless communication, these nodes use IEEE 802.11. Which is the protocol commonly used in wireless LAN. The primary objective of developing IEEE 802.11-based sensing applications is the compatibility with current networks and infrastructures. The utilisation of the body heat was proposed in [20]. Any application that uses human body heat will work well as long as the external temperature is significantly below the temperature of the standard body, i.e. 98.6 °F. When the temperature of the environment is closer to that of the human body, the capability to harvest energy reduces. However, a human being living in an air-conditioned room can use the high-temperature difference. Even throughout winter seasons at Raleigh, NC, it was observed that the ranges of temperature lie from sub-zero to 15°C. Just moving a side of the TEG (Thermoelectric Power Generators) gives 79mV. This is enough for starting DC-DC converters.

The idea of Radio Frequency energy and thermal energy for harvesting, transmitted through a DC-DC converter, and is used for charging the battery was presented in [21]. This approach has a high probability of functioning when environmental conditions are continuously changing, as two sources are utilized. When a Television or Radio station stops broadcasting transmission, Radio Frequency energy sources are not present. Still, an existence of a temperature gradient can be harvested using a TEG.

Harvesting energy from movement and vibrations have become capable of providing power to sensor nodes. In fact, several urban environments, like highways railways, bridges and human bodies are focused on vibrations. To provide power to sensor nodes from these environmental vibrations, several researchers have built efficiently and tested many models based on piezoelectric resources [22].

In this literature, different techniques about WSN are studied. Earlier research has tried up to a certain extent to overcome the problem of energy consumption and network stability using energy efficient techniques. However, still, energy consumption and network stability is the primary challenging issue in Wireless Sensor Network. Therefore, we propose a technique of energy harvesting in clustering based Wireless Sensor Network to prolong network lifetime and network stability.

### III. PROPOSED TECHNIQUE

As per the literature review, a lot of energy efficient algorithms are used for prolonging network lifetime in WSN. One solution to reduce energy consumption in WSN is to use clustering technique as shown in Fig. 4. In clustering, the sensor nodes are divided into common nodes (member nodes) and some special nodes called Cluster Heads (CH). Member nodes sense data and transmit to the cluster head. Cluster head performs data aggregation on the received data from all member nodes. The CH then transmit the aggregated data to base station. However, there are still problems in clustering technique, i.e. frequent re-clustering or failure of the whole cluster in the case where the energy level of the CH depleted.

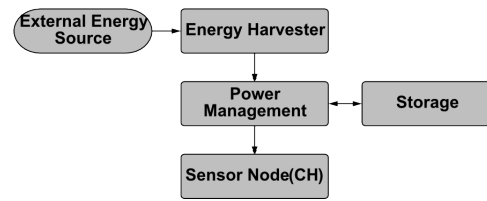


Fig. 3. Proposed architecture of cluster head with harvesting system.

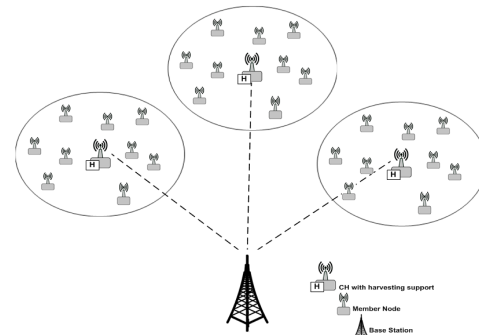


Fig. 4. Energy harvesting based clustering (EH Clustering) in WSN.

So, we suggest that there should be an additional source of energy with CH, i.e. Energy Harvesting system which provides continuous power to CH. So, in this way, the energy of the CH will always be at higher level.

The proposed technique is illustrated in Fig. 3. Here energy is harvested from an external source by energy harvester and converts into electrical energy through power management unit. This energy is stored for future use or directly supplied to CHs.

When energy harvested based sensor nodes are selected as CHs, then CH broadcast an advertisement message to other sensor nodes for cluster formation. Member nodes will transmit data to these selected CHs, CHs perform data aggregation and forward fused data to the base station as shown in Fig. 4 and 5.

#### A. Working Mechanism of Proposed Technique

The proposed approach consists of the following steps:

- Setup phase
- Operational phase

In the setup phase, all sensor nodes will find the proper CH for a cluster. In the start of setup phase, the energy-harvesting nodes in the sensor network broadcast themselves as CHs. Each node determines its CH by measuring the strength of the received signal message and notify itself as member node by sending a join request message (Join\_REQ) back to its chosen CH using a CSMA MAC protocol. CH create a TDMA schedule for data transmission coordination within the cluster and send to member nodes.

In Operational phase, sensors nodes detect events and then transmit the data to their respective CHs in assigned time slot defined in the TDMA schedule. CH performs aggregation on the received data and then transmit fused data to the Base



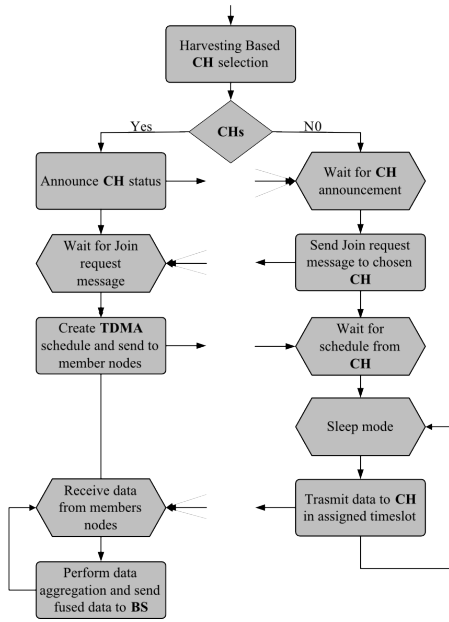


Fig. 5. Flow chart of proposed EH-Clustering technique.

Station via CSMA protocol. Thus, the CH, battery power is utilized in two ways: (1) energy consumption while receiving data from member nodes and then performing aggregation on the data. (2) The power consumption of transmitting the aggregated data to BS.

This way, the energy consumption in common SNs will be controlled through sleep mode defined in TDMA schedule. This way the lifetime of the SNs will be prolonged.

**Algorithm 1** EH-Clustering Algorithm

```

Input: [ Cluster member node]
Output: [Cluster Head ]
Let CM is the Cluster member node and NCH is Non-cluster head. → is used for Unicast while ⇒ is used for Broadcast.
CH = Harvested based SN
if CHi = True then
    CHi(MSG) ⇒ NCH
    CHi (wait-for-req)
    CHi (TDMA_Schd) → CMj
    Receive (CMj, DataPCK)
    Aggregate (CMj, DataPCK)
    CHi (CMj, agg_DataPCK) → BS
else
    NCHj (wait_MSG)
    NCHj (Join-REQ) → CHi
    CMj (in-sleep-mode)
    CMj (DataPCK) → CHi
end if
    
```

IV. RESULTS AND DISCUSSION

In this section, the performance of our proposed technique EH-Clustering is evaluated regarding energy consumption, throughput and lifetime of the sensor network and the results are compared with LEACH protocol. We will perform simulation on small scale as well as large-scale networks.

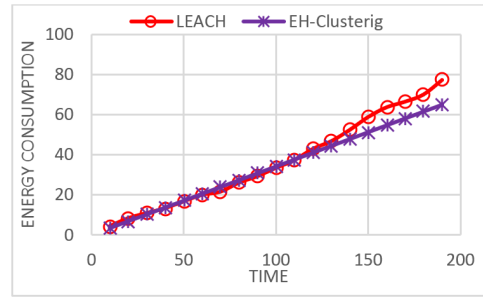


Fig. 6. Energy consumption vs Time interval for 20 nodes

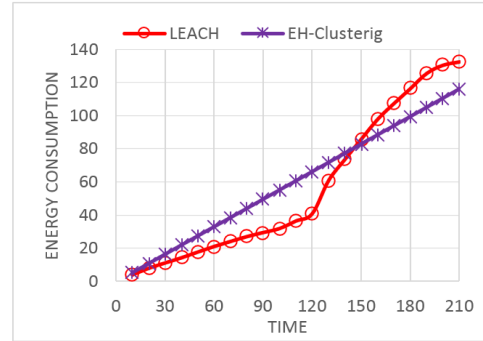


Fig. 7. Energy consumption vs time interval for 40 nodes.

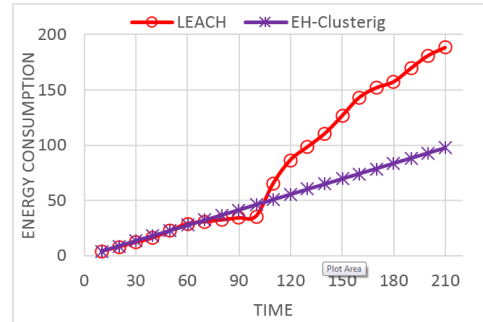


Fig. 8. Energy consumption vs Time interval for 60 nodes.

A. Simulation Result Scenario for Energy Consumption

Fig. 6 shows a comparison of energy consumption of LEACH protocol and our proposed algorithm for 20 nodes. It is clear that energy level of our proposed approach (EH-Clustering) is constant throughout the simulation time because regular power is supplied to CH from the harvesting system. The energy level of sensor nodes decreases with the passage of time in case of LEACH. So, we can say that our proposed algorithm works efficiently in this scenario.

Fig. 7 shows a comparison of energy consumption of LEACH protocol and our proposed algorithm for 40 nodes. It is clear that energy level of sensor nodes in EH-Clustering is constant throughout the simulation time because regular power is supplied to CH from the energy harvesting system, while the energy level of sensor nodes decreases with the passage of time in case of LEACH. So we can say that our proposed algorithm also works efficiently for 40 nodes in term of energy consumption.

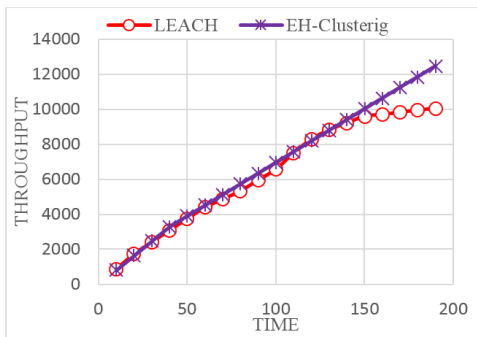


Fig. 9. Throughput vs Time interval for 20 nodes.

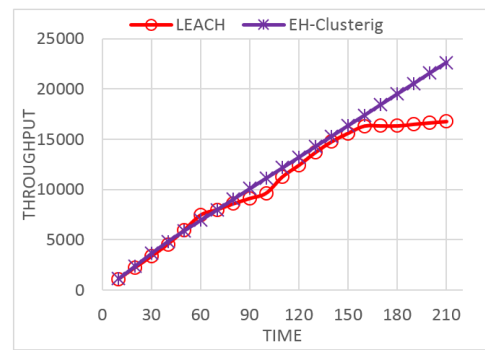


Fig. 11. Throughput vs time interval for 60 nodes.

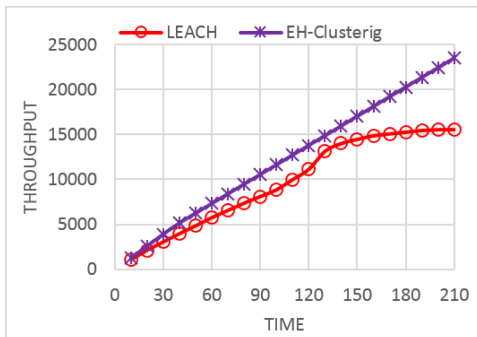


Fig. 10. Throughput vs Time interval for 40 nodes.

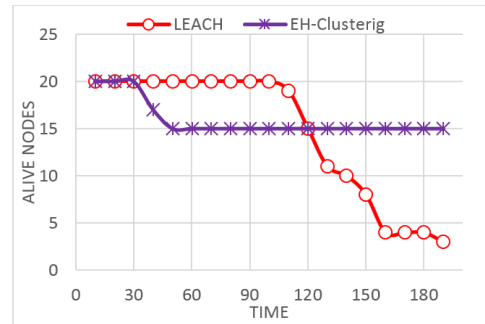


Fig. 12. Alive nodes vs Time for 20 nodes.

Fig. 8 shows a comparison of energy consumption of LEACH protocol and EH-Clustering for 60 nodes. It is clear that energy level of the whole Wireless Sensor Network in our proposed algorithm is constant throughout the simulation time, as regular power is supplied to CH from the energy harvesting system. The energy level of sensor nodes decreases in case of LEACH. So, we can say that our proposed algorithm also works efficiently for a network of size 60 nodes. From the above results of energy consumption for nodes 20, 40 and 60, it is clear that our proposed algorithm works efficiently for small as well as large-scale networks.

### B. Simulation Result Scenario of Throughput

Fig. 9 shows the throughput comparison of LEACH and EH-Clustering in case of 20 nodes. The throughput of our proposed approach is higher as compared to LEACH for the entire simulation time because regular power is supplied to CH from energy harvesting system. This will result in minimum re-clustering, and the battery power of the CH will always be at the optimum level. As a result, the throughput will also be high. So our proposed algorithm is efficient than LEACH protocol in term of throughput for 20 nodes.

Fig. 10 shows the throughput comparison of LEACH and EH-Clustering in case of 40 nodes. From above result, it is clear that the throughput of our proposed approach is higher as compared to LEACH for the entire simulation time, as regular power is supplied to CH from the energy harvesting system. This will result in minimum re-clustering, and the battery power of the CH will always be at the optimum level. As a result, the throughput will also be high. So, our proposed algorithm also works efficiently in term of throughput for 40

nodes than LEACH protocol.

Fig. 11 shows the throughput comparison of LEACH and EH-Clustering in case of 60 nodes. From above result, it is clear that the throughput of our proposed algorithm is higher as compared to LEACH for the entire simulation time, as regular power is supplied to CH from the energy harvesting system. This will result in minimum re-clustering, and the battery power of the CH will always be at the optimum level. As a result, the throughput will also be high. Hence our proposed algorithm also works efficiently in term of throughput for 60 nodes than LEACH protocol. From the above results of throughput comparisons of EH-Clustering and LEACH, we conclude that our proposed algorithm gives an efficient result for the small and large-scale network in term of throughput.

### C. Simulation Result Scenario of Alive Nodes

Fig. 12 shows a comparison of LEACH and EH-Clustering in term of Alive nodes for 20 nodes. The alive nodes in LEACH decrease as time passes because of frequent re-clustering, whereas in case of our proposed approach (EH-Clustering), the alive nodes are higher throughout the entire simulation time, as regular power is supplied to the CH from energy harvesting system, this will prolong network lifetime. So, we can say that our proposed algorithm gives the efficient result as compared to LEACH in term of Alive nodes for a network of 20 nodes.

Fig. 13 shows a comparison of LEACH and EH-Clustering in term of Alive nodes for 40 nodes. The alive nodes in LEACH decrease as time passes because of frequent re-clustering, whereas in case of our proposed approach (EH-Clustering), the alive nodes are higher throughout the entire

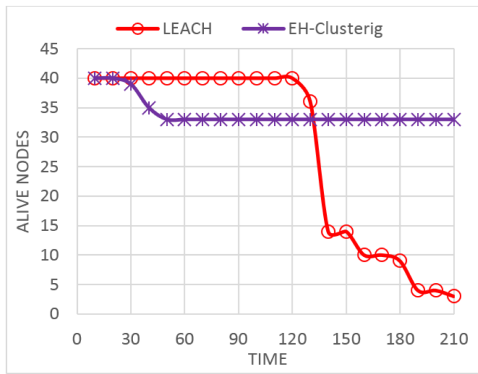


Fig. 13. Alive nodes vs Time interval for 40 nodes.

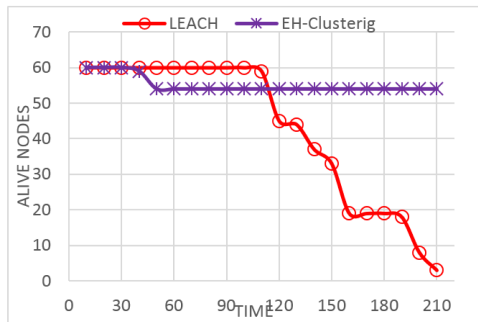


Fig. 14. Alive nodes vs Time interval for 60 nodes.

simulation time, as regular power is supplied to the CH from the energy harvesting system, this will prolong network lifetime. So, we can say that the lifetime of the sensor network in case of our proposed algorithm is also higher than LEACH in term of alive nodes for 40 nodes.

Fig. 14 shows a comparison of energy consumption of LEACH protocol and EH-Clustering for 60 nodes. It is clear that energy level of the whole WSN in our proposed algorithm is constant throughout the simulation time, as regular power is supplied to CH from the energy harvesting system. The energy level of sensor nodes decreases in case of LEACH. So, we can say that our proposed algorithm also works efficiently for a network of size 60 nodes. From the above results of energy consumption, it is clear that our proposed algorithm works efficiently for small as well as large-scale networks.

## V. CONCLUSION

In this research, energy harvesting based clustering approach is proposed. Our simulation results have shown that by using energy harvesting in clustering based WSNs, the lifetime of a WSN significantly enhanced. From this research, it is clear that energy harvesting is a best alternate source of energy for Wireless Sensor Networks. Using energy-aware clustering concepts, the battery usage and computation overhead will be decreased. Energy harvesting based clustering approach enhance the energy conservation of the sensor nodes. The energy harvesting based clustering approach enhances the performance and lifetime of Wireless Sensor Networks as compared to other algorithms. The deployment of low-cost energy harvesting based sensors nodes is the main cause for

the reliable communication and prolonging lifetime in Wireless Sensor Network.

## VI. FUTURE RECOMMENDATION

For future work, it is recommended to improve security in Wireless Sensor Network by using authenticated protocols to handle the malicious sensor node attack along with energy harvesting to get more robust systems.

## REFERENCES

- [1] A. Rathee, R. Singh and A. Nandini, *Wireless Sensor Network- Challenges and Possibilities*, International Journal of Computer Applications, vol. 140, no. 2, pp. 1-15, 2016.
- [2] S. Zin, N. B. Anuar, M. L. Kiah and A.-S. Pathan, *Routing protocol design for secure WSN : Review and open research issues*, Journal of Network and Computer Applications, February 2014.
- [3] S. Kumar, B. Prabhu, Rajkumar and D. S. Sophia, *A Methodology For Reducing Energy Utilization In Dense Wireless Sensor Networks*, International Journal of Research Granthaalayah, vol. 4, no. 1, pp. 125-130., 2016.
- [4] K. Vijeta, D. Kavita and D. B. Jangra, *Extended LEACH-Based Clustering Routing Protocols For WSN: A Survey*, International Journal of Engineering Development and Research, vol. 5, no. 1, pp. 362-367, 2017.
- [5] D. N. A. Shiltagh and A. H. Wheeb, *Priority Based Transmission Rate Control with Neural Network Controller in WMSNs*, Journal of Engineering, vol. 20, pp. 66-81, 2014.
- [6] S. Akbari, *Energy Harvesting for Wireless Sensor Networks Review*, in Federated Conference on Computer Science and Information Systems, pp. 987992, 2014.
- [7] A. D. Mauro and N. Dragoni, *On the Impact of Energy Harvesting on Wireless Sensor Network Security*, Technical University of Denmark (DTU), (DTU Compute PHD-2014; No. 349), 2015.
- [8] C. Anbarasan, B. Anantharaj and N. Balaji, *Cooperative Energy Allocation for Sensing and Transmission in Rechargeable WSN*, International Research Journal of Latest Trends in Engineering and Technology, vol. 3, no. 1, pp. 9-18, 2016.
- [9] M. Singh and T. Singh, *Energy Efficient, Distributed Clustering Approach for Ad Hoc Wireless Sensor Network*, International Journal of Science and Research, vol. 6, no. 4, pp. 2415-2421, 2017.
- [10] S. Sojan and D. R. Kulkarni, *A Comprehensive Review of Energy Harvesting Techniques and its Potential Applications*, International Journal of Computer Applications (0975 8887), vol. 139, no. 3, pp. 14-19, April 2016.
- [11] D. F. Yildiz and K. L. Coogler, *Low Power Energy Harvesting with a Thermoelectric Generator through an Air Conditioning Condenser in 121st ASEE Annual Conference & Exposition*, Indianapolis, 2014.
- [12] M. A. Green, K. Emery, Y. Hishikawa, Warta and E. D. Dunlop, *Solar cell efficiency tables (version 39)*, Progress in Photovoltaics: Research and Applications, vol. 20, no. 1, p. 1220, 2012.
- [13] S. Roundy, D. Steingart, L. Frechette, P. Wright and J. Rabaey, *Power sources for wireless sensor networks*, Springer Berlin Heidelberg, vol. 29, no. 20, pp. 1-17.2004
- [14] Anuj, *Wireless Sensor Network: A Review On Data Aggregation* International Journal of Innovations in Applied Sciences & Engineering, ISSN: 2454-9258, vol. 2, pp. 11-17, 2016.
- [15] Faheem khan, Sohail abbas, Samiullah khan, *An Efficient and Reliable Core-Assisted Multicast Routing Protocol in Mobile Ad-Hoc Network*, International Journal of Advanced Computer Science and Applications, vol:7:5, 2016.
- [16] M. J. Usman, Z. Xing, H. Chiroma, A. Y. U. Gital, A. I. Abubakar, A. M. Usman and Herawan, *Modified Low-Energy Adaptive Clustering Hierarchy Protocol for Efficient Energy Consumption in Wireless Sensor Networks for Healthcare Applications*, International Review on Computers and Software, vol. 9, no. 11, pp. 1904-1915, 2014.
- [17] T. Qian, A. Wei, Y. Han, Y. Liu and S. C., *Energy Harvesting Aware Topology Control with Power Adaptation in Wireless Sensor Networks*, Ad Hoc Network, vol. 27, no. C, pp. 44-56, April 2015.

- [18] S. Luo, R. Zhang and T. J. Lim, *Optimal save-then-transmit protocol for energy harvesting wireless transmitters*, IEEE Transactions on Wireless Communications, vol. 12, no. 3, p. 11961207, Mar. 2013.
- [19] X. Fafoutisa, T. Sorensenb and J. Madsena, *Energy Harvesting - Wireless Sensor Networks for Indoors Applications using IEEE 802.11*, Procedia Computer Science, vol. 32, p. 991-996, 2014.
- [20] L. C. Mateu, N. P. Lucas and P. Spies, *Human Body Energy Harvesting Thermogenerator for Sensing Applications*, International Conference on Sensor Technologies and Applications, p. 366372, Oct. 2007.
- [21] D. Har, S. Min and T. M. Mladenov, *Radio frequency energy harvesting for wireless sensor networks*, International Journal of Distributed Sensor Networks, vol. 13, no. 6, 2017.
- [22] A. Almusallam, R. N. Torah, D. Zhu, M. J. Tudor and S. P. Beeby, *Screen-printed piezoelectric shoe-insole energy harvester using an improved flexible PZT-polymer composites*, Journal of Physics: Conference, Vols. 476 (1742-6596/476/1/012108), 2013.

# A New Motion Planning Framework based on the Quantized LQR Method for Autonomous Robots

Onur Sencan

Department of Control and  
Automation Engineering  
Istanbul Technical University  
Istanbul, Turkey 34398

Hakan Temeltas

Department of Control and  
Automation Engineering  
Istanbul Technical University  
Istanbul, Turkey 34398

**Abstract**—This study addresses an argument on the disconnection between the computational side of the robot navigation problem with the control problem including concerns on stability. We aim to constitute a framework that includes a novel approach of using quantizers for occupancy grids and vehicle control systems concurrently. This representation allows stability concerned with the navigation structure through input and output quantizers in the framework. We have given the theoretical proofs of qLQR in the sense of Lyapunov stability alongside with the implementation details. The experimental results demonstrate the effectiveness of the qLQR controller and quantizers in the framework with real-time data and offline simulations.

**Keywords**—Robot motion; mobile robotics; hybrid systems; optimal control; quantization

## I. INTRODUCTION

Systems that contain both continuous dynamics and discrete events are called hybrid or discontinuous dynamical systems. The discontinuity in the system can be based on the control system transitions like the operation of gearbox shift pattern in vehicles or the steep system dynamics like a change of direction or final stop of a bouncing ball. In the early studies of control theory, this well-known phenomenon in the system dynamics interpreted as a predictable disturbance [1] or the noise on the signal [2]. After the widespread usage of the pulse width modulation signals in electric drives at the beginning of the 90s, sliding mode control techniques [3,4] use this nonlinear switching behavior as a control method for stabilizing nonlinear systems with an on-off(bang-bang) controller. Following the broad range of industrial practices, usage areas of sliding mode control reached many application areas in robotics like stabilization of autonomous surface vessels in rough open seas [5] and reactive position control of quadrotors [6]. Sliding mode control frameworks define the switching regions as discontinuity surfaces and aim to design feedback controllers that direct the optimal solutions of system states to settle around these surfaces. The near-optimal solutions around discontinuity surfaces induce chattering problem, which is high-frequency switching between various controllers that designed for different set points originated by neglected effects of actuator and system dynamics [7].

Projection of this chattering situation into robot navigation problem can be seen in a seesaw movement of a vehicle among attractive goal point and different repellent sources like the obstacles around the vehicle. At this point, derivative

studies of artificial potential fields[8–10] and vector fields [11,12] for navigation also tend to behave in the same manner around this instability boundaries. Different approaches like assigning coefficients for cost functions to stay away from these regions as in the dynamical window approaches [13,14] do not guarantee a stabilizing controller and works only the predefined areas where the parameters are tuned for these special cases. Decentralized cooperative control of swarm robots as in [15] has related approaches but in a different context with a switched control system. This phenomenon of using existing control methods with different notation can be seen very often in the robotics literature. Similarly, optimal control theory literature dismisses antecedent study of Carathéodory's [16] and this leads up to related works on discontinuous dynamics with Carathéodory's solutions [17–19] and its variant Filippov solutions [20]. These studies suggest the equivalent solutions of Pontryagin's minimum principle [21] by inspecting the vector fields of a specific operating point or neighborhood of an operating point respectively. Cortés [22] contributes a comprehensive review paper on discontinuous dynamics systems with Carathéodory's and Filippov's solutions. However, these approaches take a rather different way by focusing on the understanding of complex dynamics systems then discussing the stability or design control laws on discontinuous dynamics. The solutions of piecewise continuous vector fields are used in the area of state-dependent switching dynamical systems. In [23], the hybrid systems with sensor and actuator constraints are studied with a different notion called quantizers. It is evident that robotics applications are a perfect match for a hybrid system both with sensor and actuator constraints.

Quantization term is used to describe the levels or sets that are separated by discontinuous events like the change of gear in a vehicle. Quantized sets have their unique continuous dynamics, which may have similar properties like the shift change between second and third gears or completely different ones like switching between first and rear gears. Using quantization effect not as an error source to be predicted with white noise models, but rather as a stabilizer of an unstable discrete-time system by introducing the usage of quantized states as an input to state feedback mechanism is based on the study of Delchamps [24]. This strategy changes the traditional view of quantization from a simple rounding operator to a system state to be utilized in control system design process. The characteristics of quantization effects in hybrid systems are extensively studied in [23,25]. The proposed study is influenced by the idea of using the measurement quantization

as an information coding of the real data, which is used directly by the controller. Also, in [26], quantization intervals are associated with the stability of the system in an explicit relation with the systems closed-loop unstable poles. This relation is exhibited by the minimum number of quantization levels and the system poles in the right-half plane (RHP). Therefore, it is essential to determine the minimum information that is needed for the higher levels of the control system after applying quantizer.

In this paper, we propose a navigation framework with input and output quantizers which are used to express sensor data in grid segmentations and generate reference inputs within motion planner to achieve asymptotic stability with quantized LQR(Linear Quadratic Regulator) respectively. A new LQR system namely qLQR is adapted from the quantizer definitions of [27] to a navigation framework with theoretical and simulation results. Using quantization on creating occupancy grids from sensor data and in control system architecture is a novel approach to robot navigation problem.

The paper organized as follows, first we give the quantization concept in control theoretical sense with some illustrations in Section II. In Section II-A, we introduce the general framework of input and output quantization for navigation. In Section II-B, the first quantizer in the framework, the input quantizer, is explained in detail how to handle path chunks from path planner with quantization. Following Section (II-C) is about the second quantizer in the framework, which is the output quantizer. In this section, we express how a quantizer is implemented into a control system with motion planner, which is responsible for generating reference control signals from the output of input quantizer block. In Section III, we derive the quantized LQR system for this framework, which is used in previous sections. In Section III-A, the quantized control structure is derived in classical control terms. In the next Section (III-B) we derive stability boundaries of control signals as the same structure in hybrid systems[23]. We show the effectiveness of qLQR compared to the traditional LQR system and qLQR usage in our framework both with real-time and simulation results.

## II. QUANTIZED FEEDBACK CONTROL

The standard method for sampling a continuous error signal  $e(t)$  to a sequence of impulse functions  $e_s(t)$  achieved by a sampling switch with appropriately chosen sampling time  $T$ . The area of this impulse sequence is the sampling instant, which is denoted as  $e(kT)$ . Through the property of the area of Dirac-delta impulse function equals to 1 (i.e.,  $\int \delta(t) dt = 1$ ), the sampled error signal  $e_s(t)$  is represented as,

$$e_s(t) = \sum_{k=-\infty}^{\infty} e(kT)\delta(t - kT) \quad (1)$$

To convert this sampling instants to a physical signal, there should be a holder which is engaged to hold the previous signal value until the next signal is sampled. Zero-order Holder is the simple and effective solution, denoted by  $h_0$ ,

$$h_0(kT) = e(kT) \quad (2)$$

This sample and hold mechanism obeys the superposition rule. Thus, the linearity of the sampled continuous signal is not

affected at this stage. In addition to that, a quantizer, which can also be interpreted as a nonlinear discontinuous sampler is added just before the system. The quantization operation is a mapping operator, that is directed from a Euclidean space  $\mathbb{R}^k$  to a finite subset of this space  $\mathcal{P}_{\text{fin}}(\mathbb{R}^k)$ ,

$$q : \mathbb{R}^k \mapsto \mathcal{P}_{\text{fin}}(\mathbb{R}^k) \quad (3)$$

The operation introduced with a quantization operator  $q$  sets the sampled values to corresponding quantized set values. In Fig. 1 sampling, hold and quantization operations are illustrated together. Here, the quantizer operation for this example defined as,

$$q(x) := \begin{cases} L1 & 0 \leq h_0(kT) \leq l1 \quad (\text{setA}) \\ L2 & l1 \leq h_0(kT) \leq l2 \quad (\text{setB}) \\ L3 & l2 \leq h_0(kT) \leq l3 \quad (\text{setC}) \end{cases}$$

Quantizer operator is arranged to switch between three different sets  $A, B$  and  $C$  depending on the value of signal against intervals  $l1, l2, l3$  and get corresponding set values, which are  $L1, L2, L3$ .

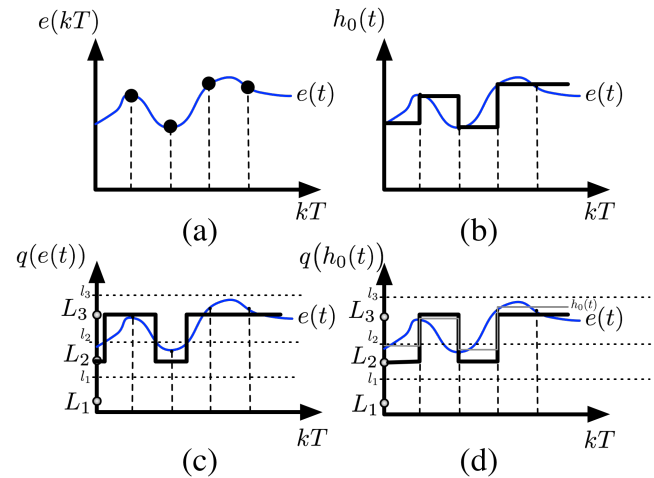


Fig. 1. Sampling, hold and quantization operations on a random error signal. (a) Discrete sampler (b) Zero order hold operator (c) Quantization operator applied on continuous signal (d) Quantizer operator applied on hold signal.

The unified framework considers quantization, time delay and disturbance to handle nonlinearities in system dynamics even the system is modeled or behaved like a linear system. Thus, the first step is modeling the additive nonlinear effects in deterministic error signals ( $e$ ), such as Gaussian white noise. The second step is to design a control law  $k$  disregarding the error signals with static feedback. Generalized dynamic equations for such a robotics system can be written as;

$$\dot{x} = f(x, u) \quad (4)$$

$$y = h(x) \quad (5)$$

where,  $x \in \mathbb{R}^n$  for system states,  $u \in \mathbb{R}^m$  is control signal,  $y \in \mathbb{R}^l$  is the system output,  $f$  is the nonlinear system characteristic function.  $h$  is the nonlinear measurement

function. We aim to design a control law  $k$  that asymptotically stabilizes the given system,  $u = k(x)$ .

Now suppose that the state feedback  $x$  is quantized, hence the control law becomes

$$u = k(q(x)) \quad (6)$$

$$u = k(x + e) \quad (7)$$

here  $e := q(x) - x$  is *quantization error*. The goal is to reduce the  $e$  with time goes to infinity;

$$\mathbf{G} : \lim_{t \rightarrow \infty} e(t) = 0 \quad (8)$$

### A. General Navigation Framework with Quantized Feedback Control System

The principal objective of using quantizers is the robustness of the controllers under measurement and modeling errors. This purpose is achieved via input to state stability(ISS), which uses Lyapunov functions and small-gain theorem. The quantized variable can be any signal like measurement output, control input or state variable. The quantizer maps these continuous variables to their quantized conjugates. Next sections explain that how we implement this quantization operator to describe grid maps in input quantizer and generate reference values and control a stabilizable control system with discontinuous dynamics of steering and throttle subsystems through output quantizer. The generalized framework is shown in Fig. 2.

Obstacle maps layer is used for pre-occupation of the 3D sensor data into initial grid layout in raw form or after some outlier elimination process [28,29]. Traffic planner block stands for the infrastructural decision logic and constraints like a temporary stop in traffic lights and in the intersection points or speed limits in different roads which are also resources of discontinuity in navigation problem. Input quantizer samples the input sensors as an input signal to path and motion planner by taking the static and dynamic obstacles and the current vehicle states into consideration. The output of the motion planner gives a maneuverable path, which is formed from grid cells. Each grid cell is a shaped reference input to the system. At this point, the second quantizer becomes a part of the system, which is called as output quantizer. Output quantizer converts the motion planner outputs like linear and angular velocities to the discrete levels available. Output quantizer provides reference inputs to the low-level controllers in the last layer. These controllers are typical PID controllers with protections like anti-windup, saturation delimiters.

### B. Input Quantizer

There are two stages here that need to be considered for stabilization of the system under quantization. One of them is the quantization levels, which are dynamic in the creation of occupancy grids in Region of Interest(ROI) and static in vehicle control systems. The second stage is to find a feedback control law that stabilizes the system. First, we derive the system model, which is the mathematical model of Ackermann type autonomous guided vehicle(AGV). The vehicle kinematics in Fig. 3 is given as a nonlinear function  $f$ , subject to state and control variables(or constraints)

$$\dot{x} = f(x, u) \quad (9)$$

The vehicle state variables are  $x, y$ , the locations in x-y coordinates,  $\theta$  orientation with linear velocity  $\nu$  and steering angle  $\varphi$ . Control variables of the vehicle are linear acceleration  $a$  and angular velocity  $\omega$ . In vector notation:

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ \theta \\ \nu \\ \varphi \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} a \\ \omega \end{bmatrix}$$

Eventually, the open form of the nonlinear system dynamics given in (9) can be derived as follows:

$$\dot{\mathbf{x}} = \frac{d}{dt} \begin{bmatrix} x \\ y \\ \theta \\ \nu \\ \varphi \end{bmatrix} = \begin{bmatrix} \nu \cos(\theta) \\ \nu \sin(\theta) \\ \frac{\nu}{\theta} \tan(\varphi) \\ a \\ \omega \end{bmatrix} \quad (10)$$

Quantization process on vehicle state variables expressed as follows:

$$\begin{aligned} x &\leftarrow G_x(x_c), \quad y \leftarrow G_y(y_c), \quad z \leftarrow G_z(z_c) \\ \psi &\leftarrow \arctan2(m_y, m_x), \quad \phi \leftarrow \arctan2(m_z, m_x), \\ \theta &\leftarrow \arctan2(m_z, m_y) \end{aligned} \quad (11)$$

In this assignment, state variables that hold the position for every axes, get values from the current vehicle position in corresponding Graph tree that is shaped from path and motion planners. Roll, pitch, yaw angles of the vehicle get values from the tangents of the path spline. This assignment visualized in Fig. 4.

Input planner may also return feedback to the path planner if the minimum gridding length is exceeded. The mission of the input quantizer is to divide the region into grid partitions, which are independent for each axis. The detailed input quantizer block of the framework is given in Fig. 5.

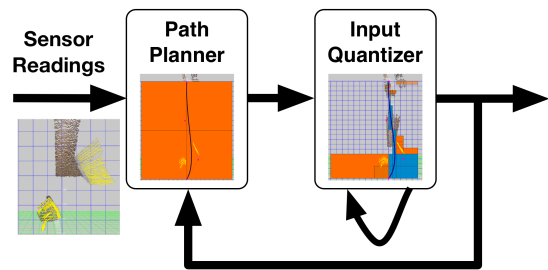


Fig. 5. Input quantizer and path planning blocks.

Division starts with coarse parent grids. Grids are assigned to the path (blue tetragons) if segmented path chunks (generated by using one of RRT planners in [30–33]) are located in grid regions. Then, the obstacles are initiated to the system with external sensors (in our case, a Kinect 3D sensor). Some of the grids are occupied (red tetragons) if there is an obstacle in grid region. In each step, the grids are become denser with splitting a coarse grid into a new grid with the half size (not necessary to be symmetrical in all axes). This division is

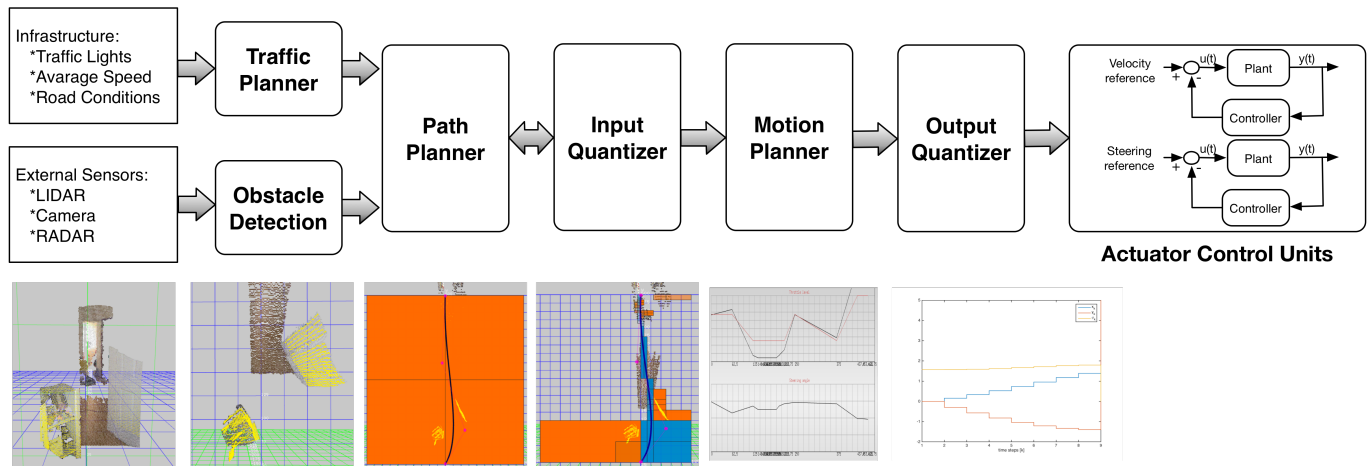


Fig. 2. General framework of the quantized feedback control system.

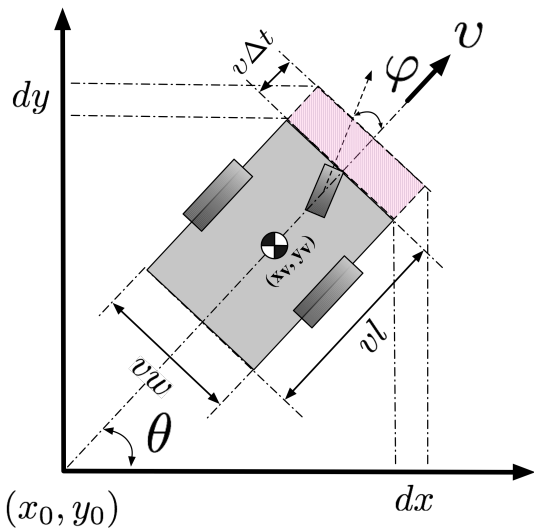


Fig. 3. Ackermann type steering model of AGV.

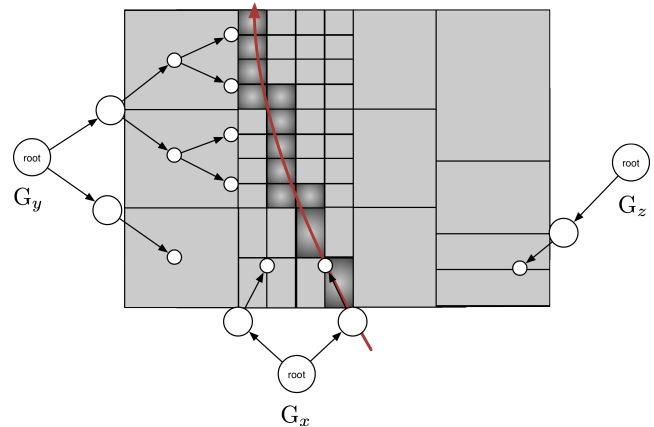


Fig. 4. Assignment of the sensor inputs in input quantization block.

repeated till the occupied obstacle grids, and grids associated to the path chunks are not overlapped (Fig. 6). All grids are symbolized with a graph tree data structure, which is independent in each axis. If the overlapping still exists to the minimum grid size allowed (which is determined by the steering and throttle actuator constraints, vehicle dimensions, etc.), then a feedback signal is generated to a new route among the alternative solutions.

The first path is initiated through the internal processes of the selected path planner algorithm. However, the afterward process is executed with the cooperation of the input quantizer. Grid segmentation logic is shown in Fig. 7.

### C. Output Quantizer

The output quantizer of the controller takes reshaped quantized reference inputs from motion planner and generates

outputs with regard to controllers sampling levels. After finding a non-overlapping grid segmentations through the selected path, then the grids are used in the motion planner section to generate velocity and steering angle references. Intuitively, if the path goes through in a close neighborhood of an obstacle, then the grids are denser in those areas. Also, the steering angle has a value between the tangent of grids and the tangent of the spline. Those references are selected from a suitable set of the actuator capabilities. After the motion planner step, we obtain the reference values to be tracked. However, we should also maintain the stability of the system while sweeping through the reference values. At this stage, the second quantizer type in the system, the *output quantizer* becomes a part of the framework. Connection between these blocks are illustrated in Fig. 8.



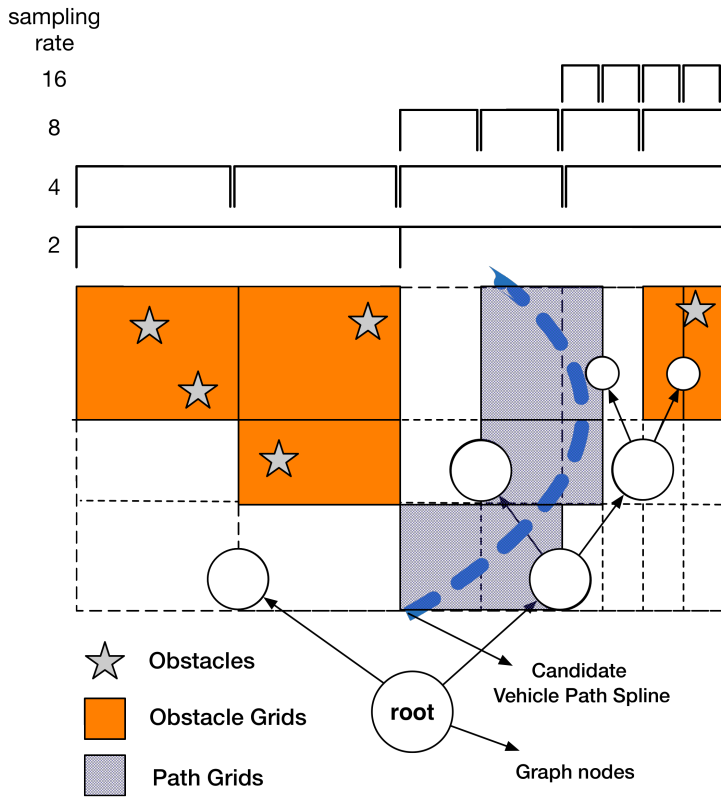


Fig. 6. Representation of the obstacle grids and path grids with node graphs in Input Quantizer.

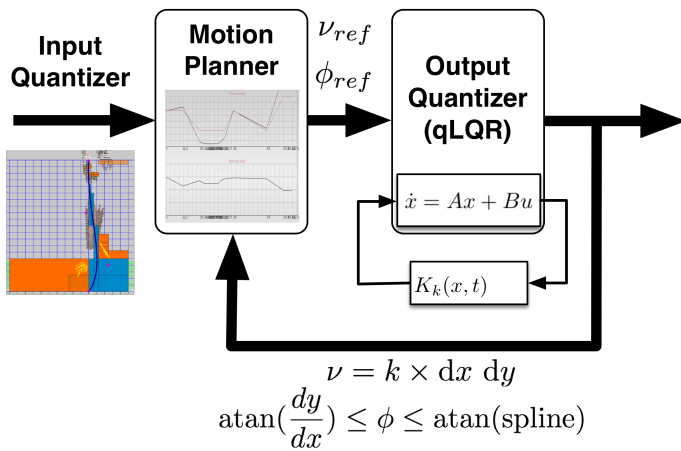


Fig. 8. Motion planner and output quantizer blocks.

In the simplest case, with an Ackerman type vehicle, there are two control variables  $\nu, \omega$  linear and angular velocities respectively. If we need a smooth ride, then we need to use the derivatives  $\dot{\nu}, \dot{\omega}$ . Linear velocity of the vehicle can be assigned as a fixed value like the many approaches in the literature. However, in this approach grid size of each axis promises a throughput that can be used to select the optimal velocity. To do this, we first define an expected mean value of the velocity

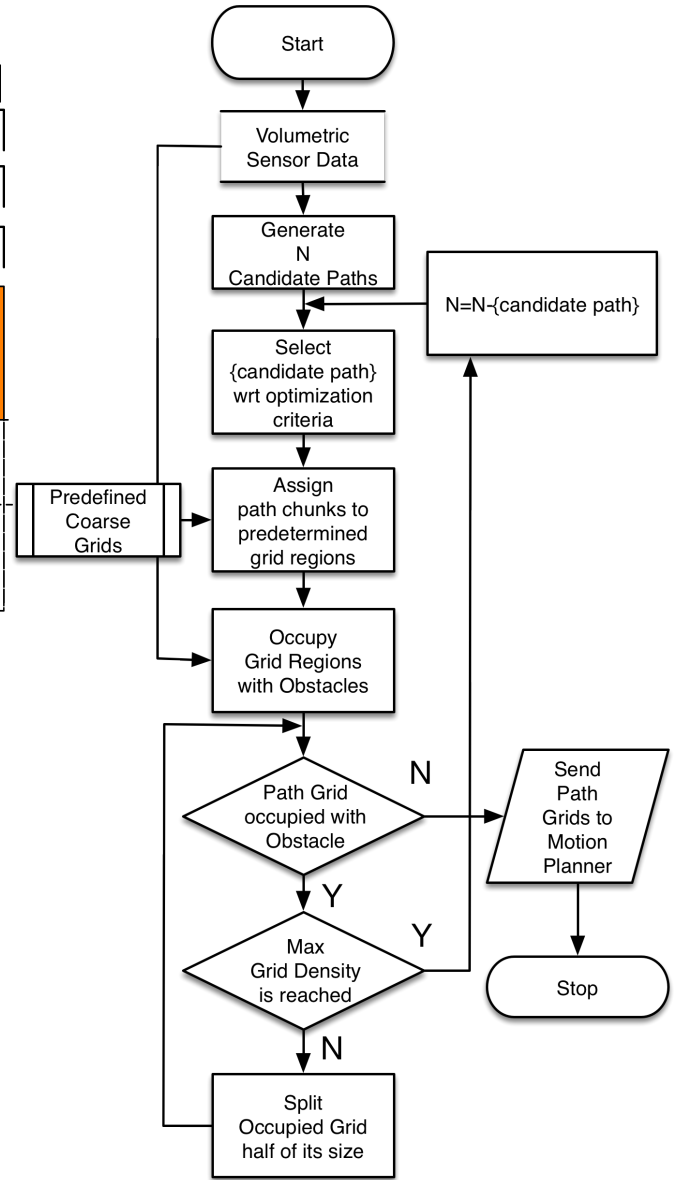


Fig. 7. Flow chart of the grid segmentation logic.

$\bar{\nu}$  in terms of grid structure,

$$\bar{\nu} = \frac{\sqrt{(dy_0 + dy^+)^2 + (dx_0 + dx^+)^2}}{\Delta t} \quad (12)$$

Here,  $dx_0, dy_0$  are the spatial samplers of the current grid and  $dx^+, dy^+$  are the spatial samplers of the next grid that is predicted in motion controller (Fig. 9).

Quantized value of the expected mean velocity is denoted as  $\lfloor \bar{\nu} \rfloor$ . This special floor operator is a direct consequence of the vehicle throttle levels which is shown with an example with a gear-shift mechanism in Fig. 10.

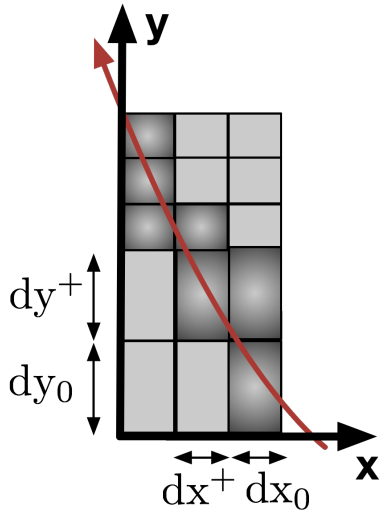


Fig. 9. Visualization of the parameters in expected mean velocity calculation.

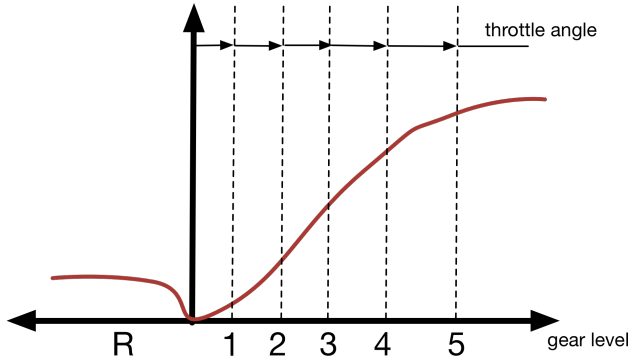


Fig. 10. Throttle with a gear-shifting is an example of switched system.

Heading angle of the vehicle is either can be found by using the tangent of the curvature (i.e.,  $\psi = \lfloor \arctan2(m_y, m_x) \rfloor$ ) or tangent of sequential grids (i.e.,  $\psi = \lfloor \arctan2\left(\frac{dy_0+dy^+}{dx_0+dx^+}\right) \rfloor$ ). The latter one is definitely a coarse estimation than the former one. State space of the closed-loop system with these control inputs can be found as follows:

$$\begin{aligned} x_{k+1} &= Ax_k + Bq_k \\ Q(y_k) &= Cx_k \end{aligned} \quad (13)$$

Here,  $q_k = Q(u_k)$  is quantized output of the output quantizer to the controller. Output quantizer signal with state feedback is found as,

$$u_k = r_k + Kx_k \quad (14)$$

$$r_k = \begin{bmatrix} \lfloor \bar{v} \rfloor_{V_t} \\ \lfloor \psi \rfloor_{V_s} \end{bmatrix} \quad (15)$$

K is gain matrix that gives us the flexibility in controller design. Special floor functions  $\lfloor \cdot \rfloor_{V_t}$  and  $\lfloor \cdot \rfloor_{V_s}$  quantize the reference output from motion planner to the system specific set levels. Here the optimization criteria is minimizing the

quantized values of the system output(i.e., state variables where  $C = I(n)$ ) which is,

$$\begin{aligned} &\underset{\text{error}}{\text{minimize}} && y_k - Q(y_k) \\ &\text{subject to} && \dot{x} = f(x, u). \end{aligned} \quad (16)$$

Starting from this point of view, state space equations can be written as follows:

$$\begin{aligned} x_{k+1} &= Ax_k + Bq_k \\ &= Ax_k + B(r_k + Kx_k) \\ &= (A + BK)x_k + Br_k \\ &= A_{cl}x_k + Br_k \end{aligned} \quad (17)$$

Here **A** is a constant valued system matrix and relates the state changes with previous states. The term **B** is the control matrix and has dependency on heading angle  $\psi$ . Apparently, with static quantization levels of the controller denoted by  $r_k$  the only alternative that we can control is **K** gain matrix. By changing values of the gain matrix, we can determine the eigenvalues of the closed-loop system. The system which is shown with linear state space model is *asymptotically stable* if all eigenvalues of the **A** is negative. This condition is equivalent to the following Lyapunov equation in terms of Lyapunov stability.

$$A^T P + PA = -Q \quad (18)$$

Here **P** is positive definite Hermitian matrix and **Q** is positive definite to make left hand side negative definite equation. The corresponding Lyapunov function is,

$$V(x) = x^T P x \quad (19)$$

Consider a linear system,

$$\dot{x} = Ax + Bu \quad (20)$$

where  $x \in \mathbb{R}^n$ ,  $u \in \mathbb{R}^m$ . Suppose this system is stabilizable with a gain matrix **K** and with a feedback control law  $u = Kx$ . State space model of the closed-loop system is now,

$$\begin{aligned} \dot{x} &= Ax + Bu \\ &= Ax + BKx \\ &= (A + BK)x \\ &:= A_c x \end{aligned} \quad (21)$$

Applying the Lyapunov stability equation in (18) to closed-loop stabilizable control system under the conditions  $B \neq 0$ ,  $K \neq 0$  gives,

$$\begin{aligned} A_c^T P + PA_c &= -Q \\ (A + BK)^T P + PA + BK &= -Q \end{aligned} \quad (22)$$

Parameters  $\lambda_{min}$ ,  $\lambda_{max}$  are the minimum and maximum eigenvalues of the system respectively. These are decisive parameters of the system boundaries and hold for following inequality.

$$\lambda_{min}(P) |x|^2 \leq x^T P x \leq \lambda_{max}(P) |x|^2 \quad (23)$$

The feedback law cannot be implemented directly because of the quantization takes place in our system. Quantizer  $q_\mu(\cdot)$  is defined for a variable  $z$  as,

$$q_\mu(z) := \mu q_\mu \left( \frac{z}{\mu} \right) \quad (24)$$

$\mu$  here is a strictly positive scalar value ( $\mu > 0$ ). We can think  $\mu$  as a zoom variable as in [23]. Increasing  $\mu$  yields coarser quantization, which increases the ROI and at the same quantization error. Conversely, decreasing  $\mu$  leads to denser quantization that decreases the quantization error but limits the range. The range of quantizer defined as  $ROI = R\mu$  and quantization error with  $\Delta\mu$ . Proof of asymptotic stability holds for zoom-in/out cases given in Appendix.

The output quantizer can be formed as a state feedback,

$$u = Kq_\mu(x) \quad (25)$$

For a closed-loop system, derivation of the quantized version of (21) is as follows:

$$\begin{aligned} \dot{x} &= \mathbf{A}x + \mathbf{B}u \\ &= \mathbf{A}x + \mathbf{B}\mathbf{K}\mu q_\mu \left( \frac{x}{\mu} \right) \\ &= (\mathbf{A} + \mathbf{B}\mathbf{K})x + \mathbf{B}\mathbf{K}\mu \left( q_\mu \left( \frac{x}{\mu} \right) - \frac{x}{\mu} \right) \end{aligned} \quad (26)$$

To inspect the behaviour of the system stability, we should check the trajectories of the system given in (26). Principally, we should define the overbounding polytope of stability regions of the state space. This ball  $B_1$  is associated with the state ranges which is defined with  $R\mu$  before.

$$B_1 := \{x : |x| \leq R\mu\} \quad (27)$$

Second polytope is defined to specify the lower boundary of the state space. To define this, first one should use the Lyapunov stability theorem. Suppose there is a *Lyapunov function*  $V$ , which has the following gradient,

$$\dot{V}(x) \leq 0, \forall x \neq 0, \dot{V}(0) = 0 \quad (28)$$

The general interpretation of the Lyapunov function  $V$  as a generalized energy function which is always loses energy(except at origin). For this system, Lyapunov function defined as in (19).

$$V(x) = x^\top \mathbf{P} x \quad (29)$$

Substituting closed-loop system in (26) in (28) satisfies,

$$\begin{aligned} \dot{V}(x) &= -x^\top \mathbf{Q} x + 2x^\top \mathbf{P}\mathbf{B}\mathbf{K}\mu \left( q_\mu \left( \frac{x}{\mu} \right) - \frac{x}{\mu} \right) \\ &\leq -\lambda_{\min}(\mathbf{Q}) |x|^2 + 2|x| \|\mathbf{P}\mathbf{B}\mathbf{K}\| \Delta \\ &\leq -|x| \lambda_{\min}(\mathbf{Q}) (|x| - \Theta_x \Delta) \end{aligned}$$

where,

$$\Theta_x := \frac{2\|\mathbf{P}\mathbf{B}\mathbf{K}\|}{\lambda_{\min}(\mathbf{Q})} \quad (30)$$

For a small  $\epsilon > 0$ , we can set boundaries for the system states,

$$\Theta_x \Delta (1 + \epsilon) \leq |x| \leq R\mu \quad (31)$$

Therefore, rate of change of the Lyapunov function in (30) has a lower limit which is bounded by the lower thresholds of the state variables. This lower limit constitutes the second polytope  $B_2$ ,

$$B_2 := \{x : |x| \leq \Theta_x \Delta (1 + \epsilon)\} \quad (32)$$

Ellipsoid for the operating point is constructed as,

$$R_1(\mu) := \left\{ x : x^\top \mathbf{P} x \leq \frac{\lambda_{\min}(\mathbf{P}) M^2 \mu^2}{\|\mathbf{K}\|^2} \right\} \quad (33)$$

Ellipsoids for attraction regions are defined as,

$$R_2(\mu) := \left\{ x : x^\top \mathbf{P} x \leq \lambda_{\max}(\mathbf{P}) \mathbf{Q}^2 \Delta^2 (1 + \epsilon)^2 \mu^2 \right\} \quad (34)$$

These regions are illustrated in Fig. 11

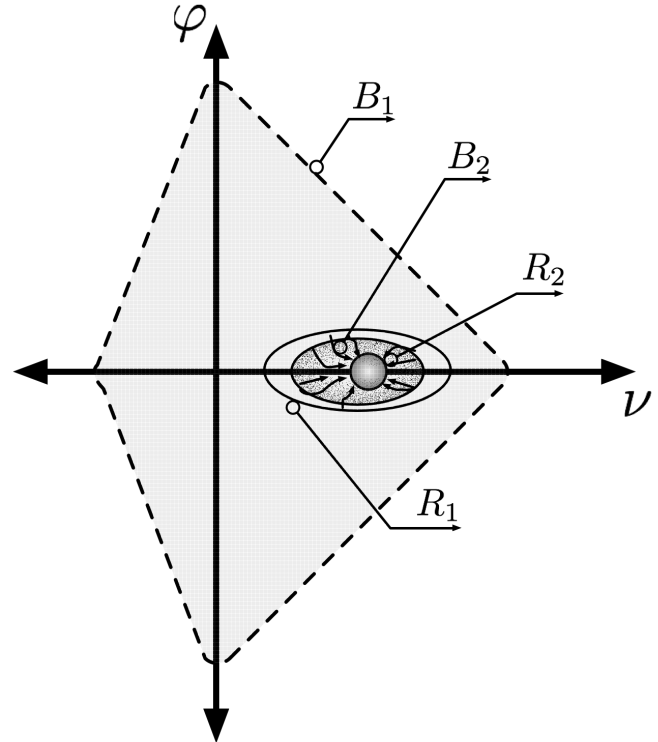


Fig. 11. Boundary polytopes of control variables for Lyapunov stability.  $B_1$  region denotes the state minimum and maximum values(max velocity and steering angles).  $R_1$  is the selected operating or linearization point of vehicle dynamics.  $R_2$  is the attraction(both appear as discontinuity or stability surface) region.  $B_2$  shows the rate of change constraints of the state variables, because speed and steering angle are dependent variables in terms of stability analysis.

All solutions initialized in region  $R_1(\mu)$  will be entered to inner region  $R_2(\mu)$  in finite dwell time given by the following formula in [23]

$$T := \frac{\lambda_{\min}(\mathbf{P}) M^2 - \lambda_{\max}(\mathbf{P}) \mathbf{Q}^2 \|\mathbf{K}\|^2 \Delta^2 (1 + \epsilon)^2}{\mathbf{Q}^2 \|\mathbf{K}\| \Delta^2 \epsilon (1 + \epsilon) \lambda_{\min}(\mathbf{Q})} \quad (35)$$

Asymptotic stability proof for a quantized feedback controller is given in [23]. Adaptation of this proofs to our framework is given in Appendix section.

Output quantizer layer is the most sophisticated part of the framework in control theoretical way. In Section III, an adaptation of hybrid control systems in [27] quantized LQR controller is introduced as a stabilizable controller for the framework, when there is traditional LQR controller can be found for the system (for a controllable and stabilizable system). The reason behind that we use a quantized LQR controller instead of a traditional LQR controller is to take account of the reference levels that can be able to follow by the low-level actuators. Without considering the low-level capability of the system, one is always to be mistaken on the assumption of a perfect low-level actuation system, and this causes undesirable results in real life scenarios. One of the most faced situation is switching systems, where from its nature, switching between two stable modes of control operations may cause an unstable behavior. Conversely, switching between two unstable modes may cause a stable operation contrary to the instincts.

### III. QUANTIZED LQR FOR OUTPUT QUANTIZER

In this section, we derive the control system which is denoted as K in previous sections for output quantizer in the framework. First, we explain the idea of stabilization for traditional Linear Quadratic Regulator(LQR) in discrete systems and expand it to the new approach of quantized LQR(q-LQR) structure. Both theoretical results and simulation results are given in order.

#### A. Optimal Control for Discrete System

Discrete time system with k sampling steps is defined as follows,

$$x_{k+1} = \mathbf{A}x_k + \mathbf{B}u_k \quad (36)$$

with given initial condition  $x_0$ . The design objective is to find an optimal control  $u_k^*$  so that the performance index  $J$ , which is designed for the control demands,

$$J = \frac{1}{2} x_N^T S_N x_N + \frac{1}{2} \sum_k^{N-1} x_k^T Q x_k + u_k^T R u_k \quad (37)$$

can be minimized. Here  $N$  is the final time,  $S$ ,  $Q$  and  $R$  are the weight coefficient matrices for the final state, run-time states and the control inputs. To solve this optimization problem, first we need to derive a Hamiltonian function  $H_k$  with Lagrange multiplier  $p$ ,

$$H_k(x_k, p_{k+1}, u_k) = \frac{1}{2} x_k^T Q x_k + \frac{1}{2} u_k^T R u_k + p_{k+1}^T A x_k + p_{k+1}^T B u_k \quad (38)$$

The necessary conditions where the performance index is to be a minima or maxima (i.e. optimum) are given as follows:

$$\frac{\partial H_k}{\partial u_k^*} = 0 = R u_k^* + B^T p_{k+1}^* \quad (39)$$

$$\frac{\partial H_k}{\partial p_k^*} = x_{k+1}^* = A x_k^* + B u_k^* \quad (40)$$

$$\frac{\partial H_k}{\partial x_k^*} = p_k^* = Q x_k^* + A^T p_{k+1}^* \quad (41)$$

Optimal control function for given discrete system can be found from (39) as,

$$u_k^* = -R^{-1} B^T p_{k+1}^* \quad (42)$$

Substituting the optimal control signal in (42) to (40) gives,

$$x_{k+1}^* = A x_k^* - B R^{-1} B^T p_{k+1}^* \quad (43)$$

Using (41) and (43), we can show the feature states and the Lagrange multipliers in left hand side,

$$x_{k+1}^* = A x_k^* - B R^{-1} B^T p_{k+1}^* \quad (44)$$

$$A^T p_{k+1}^* = p_k^* - Q x_k^* \quad (45)$$

Conjugated equations (44) and (45) are difficult to solve. However, there is a special solution for this case,

$$p_k = S_k x_k \quad (46)$$

Rearranging (44) with (46) gives,

$$x_{k+1}^* = (I - B R^{-1} B^T S_{k+1})^{-1} A x_k^* \quad (47)$$

It can be shown that the  $S_{k+1}$  is positive semi-definite and thus inverse will always exist. Substituting the  $p_k$  term in (46) into (45) brings us,

$$A^T S_{k+1} x_{k+1}^* = (S_k - Q) x_k^* \quad (48)$$

By using (48) with (47),

$$A^T S_{k+1} (I - B R^{-1} B^T S_{k+1})^{-1} A x_k^* = (S_k - Q) x_k^* \quad (49)$$

If one inspect (49), that can be seen the equation is independent from the states i.e.,

$$A^T S_{k+1} (I - B R^{-1} B^T S_{k+1})^{-1} A = (S_k - Q) \quad (50)$$

Equation (50) is called as *Discrete Riccati Equation*. Here,  $S_k$  is referred as *Riccati*. To find the optimal control, substituting (46) to (42) gives us,

$$u_k^* = -R^{-1} B^T S_{k+1} x_{k+1}^* \quad (51)$$

with substitution  $x_{k+1}$  term of (47) into (51),

$$\begin{aligned} u_k^* &= -R^{-1} B^T S_{k+1} (I - B R^{-1} B^T S_{k+1})^{-1} A x_k^* \\ &= -(R + B^T S_{k+1})^{-1} B^T S_{k+1} A x_k^* \\ &= -K_k x_k^* \end{aligned} \quad (52)$$

here  $K_k$  is the state feedback coefficient matrix for the optimal controller,

$$K_k = (R + B^T S_{k+1} B)^{-1} B^T S_{k+1} A \quad (53)$$

Also, the Riccati term can be extracted using (53) and (50),

$$S_k = Q + A^T S_{k+1} (A - B K_k) \quad (54)$$

Some rule of the thumb information for performance index are,

- Increasing the  $Q$ , increases the bandwidth.
- Increasing only the diagonals of the  $Q$ , increases the damping ratio.
- Response of the state  $x_j$  can be made faster by increasing the diagonal entry  $q_{jj}$  in the weight matrix  $Q$ .

### B. Quantized Linear Quadratic Regulator as Minimum Energy Controller

If a LTI system  $[A, B]$  is *stabilizable*, it is also quadratically stabilizable. Thus, there is a state-feedback control input  $u$  that shapes the states, performs a decreasing Lyapunov function, which is an indicator of the stabilization. These Lyapunov functions are called *control Lyapunov functions* (CLF). A quadratic CLF is constituted by,

$$V(x) = x_k^T S_k x_k \quad (55)$$

The gradient of Lyapunov function is denoted as  $\Delta V(x)$  and it should always have a negative sign if  $V(x)$  is a continuously decreasing function.

$$\Delta V(x) := V(x_{k+1}^*) - V(x_k^*) < 0 \quad (56)$$

Now, we convert our problem to find a control set  $U$  that minimizes the given CLF in (55),

$$U : \{u_i \in \mathbb{R}; i \in \mathbb{Z}\}$$

a quantizer is a function that maps the state set  $X$  to the control set  $U$  with one to one correspondence,

$$f : X \mapsto U$$

where,

$$X : \{x_i \in \mathbb{R} \mid f(x) = u_i, i \in \mathbb{Z}\}$$

Substituting the quantizer into (56) and using the discrete system equation  $x_{k+1} = Ax_k + Bu_k$  we have,

$$\Delta V(x) = V(Ax_k + Bf(x_k)) - V(x_k) < 0 \quad (57)$$

The problem of the quantization is to find the minimum coarsest control level that stabilizes the system. So, let  $\rho$  a multiplier on the unit control set  $U$ . Aim is to find the  $\rho$  value that minimizes the performance index,

$$f : X \mapsto \beta U, \quad X : \{x_i \in \mathbb{R} \mid f(x) = \beta u_i; i \in \mathbb{Z}, \beta \in \mathbb{R}, \beta > 0\} \quad (58)$$

Equation (57) shows there is no loss of generality with the scaling[26] and hereafter, the  $\beta$  value is assumed to be  $\beta = 1$  and CLF is a robust CLF for these given fixed control values. Like the expanded version of normal LQR controllers in (51), same equation can be written for the quantized LQR controller;

$$u_k^* = - (R + B^T S_{k+1} B)^{-1} B^T S_{k+1} A x_k^* = K_k x_k^* \quad (59)$$

Close loop system transition matrix can be written using new feedback matrix in (59),

$$A_c = A + BK_k \quad (60)$$

So, the new CLF gradient  $\Delta V(x)$  becomes,

$$\Delta V(x) = V(x_{k+1}) - V(x_k) \quad (61)$$

$$= x^T A_c S A_c x - x^T S x \quad (62)$$

$$= x^T (A_c S A_c - S) x \quad (63)$$

The middle term  $(A_c S A_c - S)$  corresponds to the  $Q$  term in the classical LQR performance index in (37). since  $\Delta V(x) < 0$ , a positive  $Q$  matrix is constructed as,

$$\begin{aligned} Q &= S - A_c S A_c \\ &= S - A^T S A - \\ &\quad - (R + B^T S_{k+1} B)^{-1} A^T (S_{k+1} B B^T S_{k+1}) A \end{aligned} \quad (64)$$

After we form the quantized LQR law, we define the control inputs with boundaries. Let  $V(x) = x_k^T S_k X_k$  a CLF with  $S_k > 0$  and positive semi-definite Riccati matrix. A quantizer  $f : X \mapsto U, f(x) = u$  is defined and the control law is,

$$U = \{\pm u_i : u_{i+1} = \rho u_i, i \in \mathbb{Z}\} \quad (65)$$

Control law satisfies an interval, which is the solution of the  $\Delta V(x) = 0$ . If we extend (57),

$$\begin{aligned} \Delta V(x) &= V(x_{k+1}) - V(x_k) \\ &= x_{k+1}^T S_{k+1} x_{k+1} - x_k^T S_k x_k \\ &= x^T (A^T S_k A - S_k) x + \\ &\quad + 2x^T A^T S_k B u_k + u_k^T B^T S_k B u_k \end{aligned} \quad (66)$$

$u^{(1)}$  and  $u^{(2)}$  are the roots of the Eq. 66 and one can found the solution as,

$$u^{(1),(2)} = \frac{B^T S_k A Q^{-1} A^T S_k B}{B^T S_k B} \pm \sqrt{\frac{B^T S_k A Q^{-1} A^T S_k B}{B^T S_k B}} \quad (67)$$

These roots are the control boundaries for the quantized controller. Now, our optimization problem can be expressed as follows. We are trying to find the *coarsest* quantizer that satisfies the control boundaries  $[u_k^{(1)}, u_k^{(2)}]$  and decreases the performance index in every step (i.e.  $\forall k \Delta V(x_k) < 0$ ). We define the performance criteria as the *minimum energy control* that stabilizes the system.

$$\min_{x_{k+1}=Ax_k+Bu_k} \text{stable} \sum_{k=0}^{\infty} u_k^2 \quad (68)$$

Positive semi-definite Riccati equation for this system is,

$$S_k^* = A^T S_k^* A - (B^T S_k^* B + 1)^{-1} A^T S_k^* B B^T S_k^* A \quad (69)$$

besides, (69) also gives the same solution of the following LQR problem,

$$\min_{x_{k+1}=Ax_k+Bu_k} \text{stable} \sum_{k=0}^{\infty} u_k^2 = x^T R x \quad (70)$$

where  $R = S_k^*$ . State feedback gain is also evaluated as,

$$K_{lqr} = (B^T R_k B)^{-1} B^T R_k A \quad (71)$$

#### IV. RESULTS AND DISCUSSION

First simulation results are aimed to show the superiority of the quantized LQR (qLQR) to the traditional LQR approach. In Fig. 12 generated control inputs for the same system in traditional LQR and qLQR are compared. While the control inputs nearly have the same values, qLQR control signal has an additional feature, which is the upper and lower boundary stability thresholds from (67).

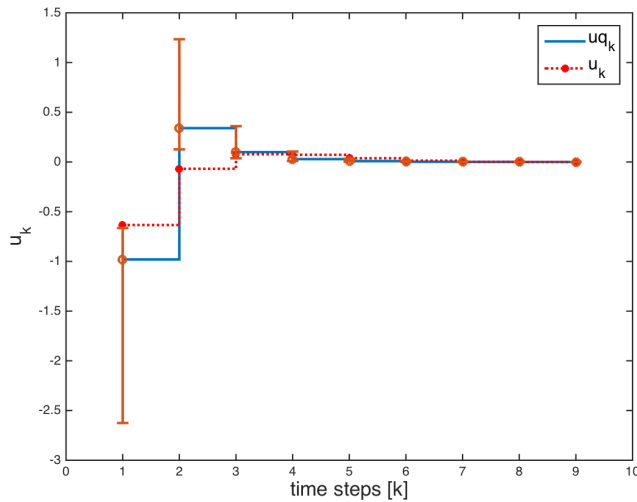


Fig. 12. Comparison between the LQR and qLQR control inputs. Control signal of the qLQR has additional upper and lower boundaries.

System outputs for the controller inputs in Fig. 12 is showed in Fig. 13. One can see that the system controlled with qLQR controller is stabilized faster than the traditional one.

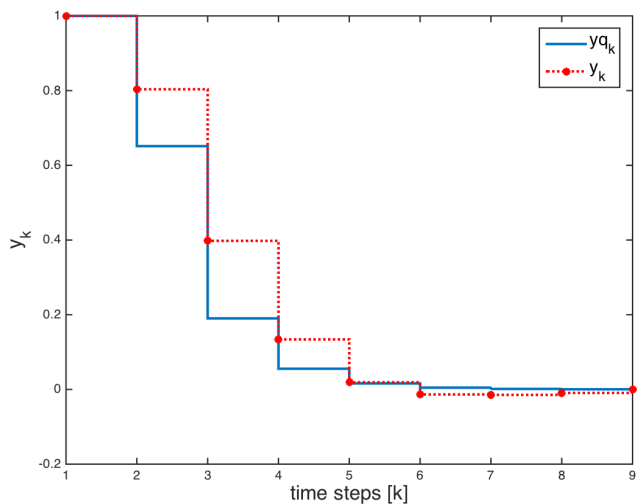


Fig. 13. Comparison between the LQR and qLQR system outputs. System controlled with qLQR is stabilized faster than the traditional one.

Red rectangles show the obstacle grids, and blue rectangles

show the path allocations like before. In the beginning, grid graphs of both dimensions ( $G_x, G_y$ ) have one root node with two siblings. Thus, the ROI is separated into two halves, and very coarse representation occurs for obstacles and the path which is shown in Fig. 14(a). Using the algorithms that are given in the proposed paper, dimensions are divided into different partitions on the existence of obstacles in the path route locations. This effect can be seen clearly in Fig. 14(b). While the left half of the obstacle grids remain coarser, the right half plane grids are denser due to the path route passes nearby. This procedure proceeds continuously in real-time, so it is convenient for representing both static and dynamic objects in the framework. This approach has a disadvantage when the ROI is selected with large scale with a long path. In this case, further objects that intersect with the path may affect the spatial sampling rate along their axes. This drawback can be solved by using cascade ROI areas or setting the path spline only in local waypoints. Final results are the given for the implementation of the qLQR to the given path in the Fig. 14. Linearized system states are specified as the x position, y position and the heading angle  $x_k = [x_i, y_i, \phi_i]$ . All states are gradually converged to the reference points. Gradual amplitude changes in the system states for a small route is given in Fig. 15. Besides, phase vector representation (position and the heading angle) of the states is shown in Fig. 16

#### V. CONCLUSION

In this paper, we proposed a new framework for robot navigation problem concerning the asymptotical stability of the system with input and output quantizers. In control theory, hybrid or switched systems literature produce multiple approaches to solve stability and control under nonlinearity for many industrial usages, including robotics. However, these application areas on robotics restricted with only control designs or optimizations on multi-robot swarm formations. Using quantization for representing the occupancy grid maps and control the discontinuous dynamics of robot systems are novel to this study.

First, we have given theoretical backgrounds for quantization as a projection to robot navigation problem. Quantized control framework is represented in general and in detail as input and output quantizers. We have explained the controller used in output quantizer, denoted as qLQR, in detail. Next, we have derived the upper and lower boundaries for the control signal that ensures asymptotical stability for the system. The results are given both in the simulation environment and real-time experiments including the comparison between traditional LQR and qLQR for an arbitrary run. In this particular application of the framework with LQR derived qLQR quantizers has limitations likely the same as the limitations in LQR controllers. Obtaining an analytical solution for the Riccati equation could be difficult in complex systems. The states that are used in the state feedback could not be observed in all situations. Thus, an observer design may have needed to be implemented to the system. However, unbounded input nature of traditional LQR methods is eliminated by using the quantizers.

In further studies, we will show the equivalence of Markov property with the state transitions between grids and redefine the structure in the probabilistic framework. Besides, we want

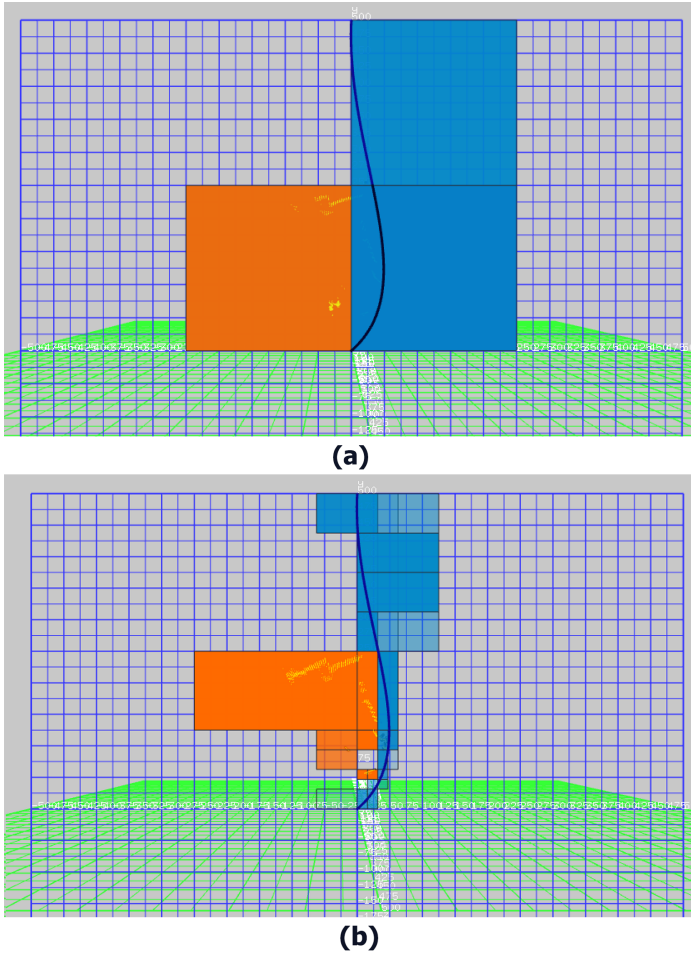


Fig. 14. Demonstration of the independent sampling rates in real life example.(a)In the beginning, both visible dimensions divide the ROI into two halves with two sibling nodes. (b) After algorithm runs, each visible dimensions are separated into different partitions due to the existence of the obstacles and intersected path splines.

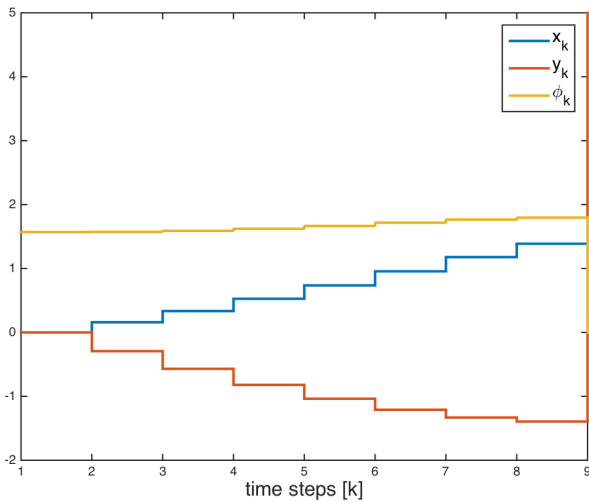


Fig. 15. System state changes with qLQR controller in vehicle control system.

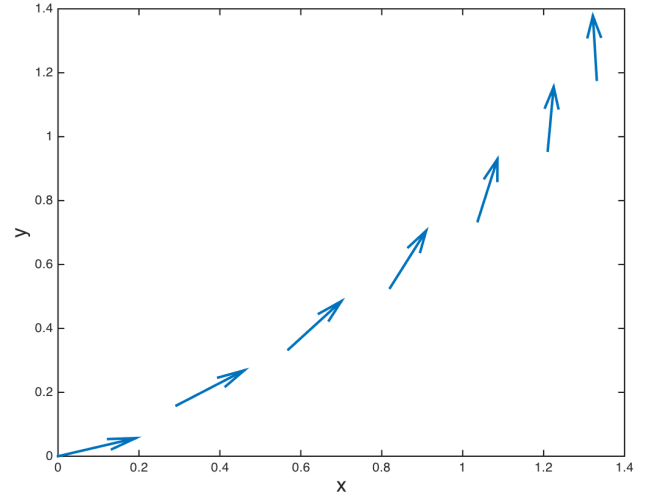


Fig. 16. Phase portrait of the qLQR controller system for x-y positions of the vehicle.

to apply different state feedback methods like eigenstructure assignment to examine the feasibility of the framework.

#### APPENDIX

##### Proof for Asymptotic Stability in Quantized Feedback Control for Zoom-in/out Cases

**Theorem V.1.** Assume that we have the following inequality.

$$\sqrt{\frac{\lambda_{\min}(\mathbf{P})}{\lambda_{\max}(\mathbf{P})}} > 2\Delta \frac{\|\mathbf{PB}\| \|\mathbf{K}\|}{\lambda_{\min}(\mathbf{Q})}$$

Then there exists a quantized feedback control solution that makes the system in (26) globally asymptotically stable.

*Proof: Zoom-out case:* Set  $u = 0$ , Let  $\mu(0) = 1$  And increase  $\mu$  fast enough to dominate the rate of growth of  $\|e^{At}\|$ . Then, there will be a time  $t_0 \geq 0$  such that

$$\|x(t_0)\| \leq \sqrt{\frac{\lambda_{\min}(\mathbf{P}) M \mu(t_0)}{\lambda_{\max}(\mathbf{P}) \|\mathbf{K}\|}}$$

which implies that  $x(t_0)$  belongs to ellipsoid  $R_1(\mu(t_0))$ .

*Zoom-in case:* Let's pick a  $\epsilon > 0$  for  $t \geq t_0$  that holds,

$$\sqrt{\lambda_{\min}(\mathbf{P}) M} > \sqrt{\lambda_{\max}(\mathbf{P}) Q_u \|\mathbf{K}\| \Delta (1 + \epsilon)}$$

where,

$$Q_u := \frac{2\|\mathbf{PB}\|}{\lambda_{\min}(\mathbf{Q})}$$

Let  $\mu(t) = \mu(0)$  for  $t \in [t_0, t_0 + T)$  where  $T$  is given by (35). Then, we can  $x(t_0 + T)$  belongs to the ellipsoid  $R_2(\mu(t_0))$ .

For  $t \in [t_0 + T, t_0 + 2T)$  let  $\mu(t) = \Omega \mu(t_0)$  where,

$$\Omega := \frac{\sqrt{\lambda_{\max}(\mathbf{P}) Q_u \|\mathbf{K}\| \Delta (1 + \epsilon)}}{\sqrt{\lambda_{\min}(\mathbf{P}) M}}$$

Here we have  $\mu(t_0 + T) < \mu(t_0)$  and  $R_2(\mu(t_0)) = R_1(\mu(t_0 + T))$

#### REFERENCES

- [1] J. Keenan and J. Lewis, "Estimation with quantized measurements," in *1976 IEEE Conference on Decision and Control including the 15th Symposium on Adaptive Processes*. IEEE, 1976, pp. 1284–1291.
- [2] R. Curry, P. Mirchandani, and C. Price, "State Estimation with Coarsely Quantized, High-Data-Rate Measurements," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-11, no. 4, pp. 613–621, 1975.
- [3] C. Edwards, "Sliding mode control using only output information," in *IEE Colloquium on Robust Control: Theory, Software and Applications*. IEE, 1997, pp. 10–10.
- [4] V. I. Utkin, "Sliding Modes in Control of Electric Motors," in *Sliding Modes in Control and Optimization*. Berlin, Heidelberg: Springer, Berlin, Heidelberg, 1992, pp. 250–264.
- [5] H. Ashrafiuon, S. Nersesov, F. Mahini, and G. Clayton, "Full state sliding mode trajectory tracking control for general planar vessel models," in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 5158–5163.
- [6] A. Sanchez, V. Parra-Vega, C. Tang, F. Oliva-Palomo, and C. Izaguirre-Espinosa, "Continuous reactive-based position-attitude control of quadrotors," in *American Control Conference - ACC 2012*. IEEE, 2012, pp. 4643–4648.
- [7] H. Lee and V. I. Utkin, "Chattering suppression methods in sliding mode control systems," *Annual Reviews in Control*, vol. 31, no. 2, pp. 179–188, Jan. 2007.
- [8] O. Khatib, "The Potential Field Approach And Operational Space Formulation In Robot Control," in *Adaptive and Learning Systems*. Boston, MA: Springer, Boston, MA, 1986, pp. 367–377.
- [9] M. C. Mora and J. Tornero, "Path planning and trajectory generation using multi-rate predictive Artificial Potential Fields," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 2990–2995.
- [10] U. Baroudi, G. Sallam, M. Al-Shaboti, and M. Younis, "GPS-free robots deployment technique for rescue operation based on landmark's criticality," in *International Wireless Communications and Mobile Computing Conference (IWCMC)*. IEEE, 2015, pp. 367–372.
- [11] W. Feiten, R. Bauer, and G. Lawitzky, "Robust obstacle avoidance in unknown and cramped environments," in *IEEE International Conference on Robotics and Automation*. IEEE Comput. Soc. Press, 1994, pp. 2412–2417.
- [12] D. Zhou and M. Schwager, "Vector field following for quadrotors using differential flatness," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 6567–6572.
- [13] O. Brock and O. Khatib, "High-speed navigation using the global dynamic window approach," in *International Conference on Robotics and Automation*. IEEE, 1999, pp. 341–346.
- [14] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 4, no. 1, pp. 23–33, Mar. 1997.
- [15] G. A. S. Pereira, M. F. M. Campos, and V. Kumar, "Decentralized algorithms for multi-robot manipulation via caging," *The International Journal of Robotics Research*, vol. 23, no. 7-8, pp. 783–795, 2004.
- [16] H. J. Pesch and R. Bulirsch, "The maximum principle, Bellman's equation, and Carathéodory's work," *Journal of Optimization Theory and Applications*, vol. 80, no. 2, pp. 199–225, Feb. 1994.
- [17] M. Bernardo, C. Budd, A. R. Champneys, and P. Kowalczyk, "Piecewise-smooth dynamical systems: theory and applications," 2008.
- [18] J.-M. Coron and L. Rosier, "A relation between continuous time-varying and discontinuous feedback stabilization," *Journal of Mathematical Systems, Estimation, and Control*, vol. 4, pp. 67–84, 01 1994.
- [19] A. Bacciotti and F. Ceragioli, "Nonpathological Lyapunov functions and discontinuous Carathéodory systems," *Automatica*, vol. 42, no. 3, pp. 453–458, Mar. 2006.
- [20] A. F. Filippov, "Existence and General Properties of Solutions of Discontinuous Systems," in *Differential Equations with Discontinuous Righthand Sides*. Dordrecht: Springer, 1988, pp. 48–122.
- [21] J. M. Blatt and J. D. Gray, "An elementary derivation of Pontryagin's maximum principle of optimal control theory," *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, vol. 20, no. 02, pp. 142–156, Apr. 1977.
- [22] J. Cortes, "Discontinuous dynamical systems," *IEEE Control Systems Magazine*, vol. 28, no. 3, pp. 36–73, May 2008.
- [23] D. Liberzon, *Switching in Systems and Control*. Springer, Jul. 2003.
- [24] D. F. Delchamps, "Stabilizing a linear system with quantized state feedback," *Automatic Control, IEEE Transactions on*, vol. 35, no. 8, pp. 916–924, 1990.
- [25] R. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Transactions on Robotics*, vol. 45, no. 7, pp. 1279–1289, 2000.
- [26] N. Elia and S. K. Mitter, "Stabilization of linear systems with limited information," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1384–1400, 2001.
- [27] D. Liberzon, "Hybrid feedback stabilization of systems with quantized signals," *Automatica*, vol. 39, no. 9, pp. 1543–1554, 2003.
- [28] S.-W. Yang, C.-C. Wang, and C.-H. Chang, "RANSAC matching: Simultaneous registration and segmentation," in *IEEE International Conference on Robotics and Automation (ICRA 2010)*. IEEE, 2010, pp. 1905–1912.
- [29] R. Raguram, J.-M. Frahm, and M. Pollefeys, "A Comparative Analysis of RANSAC Techniques Leading to Adaptive Real-Time Random Sample Consensus," in



- Computer Vision – ECCV 2008*. Berlin, Heidelberg: Springer, Berlin, Heidelberg, Oct. 2008, pp. 500–513.
- [30] J. J. Kuffner and S. M. LaValle, “RRT-connect: An efficient approach to single-query path planning,” *2000 ICRA. IEEE International Conference on Robotics and Automation*, vol. 2, pp. 995–1001 vol.2, 2000.
- [31] N. Mukai and N. Ishii, “R-Tree Based Path Representation for Vehicle Routing Problem,” in *2009 21st IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2009, pp. 758–761.
- [32] S. Karaman, M. R. Walter, A. Perez, E. Frazzoli, and S. Teller, “Anytime Motion Planning using the RRT\*,” in *2011 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2011, pp. 1478–1483.
- [33] M. Otte and E. Frazzoli, “RRTX: Asymptotically optimal single-query sampling-based motion planning with quick replanning,” *The International Journal of Robotics Research*, vol. 35, no. 7, pp. 797–822, 2015.

# Bearing Fault Classification based on the Adaptive Orthogonal Transform Method

Mohamed Azergui, Abdenbi Abenaou and Hassane Bouzahir

Laboratory of Systems Engineering and Information Technology (LISTI)

National School of Applied Science, Ibn Zohr University, PO Box 1136, 80000 Agadir, Morocco

**Abstract**—In this work, we propose an approach based on building an adaptive base which permits to make accurate decisions for diagnosis. The orthogonal adaptive transformation consists of calculating the adaptive operator and the standard spectrum for every state, using two sets of vibration signal records for each type of fault. To classify a new signal, we calculate the spectral vector of this signal in each base. Then, the similarity between this vector and other standard spectra is computed. The experimental results show that the proposed method is very useful for improving the fault detection.

**Keywords**—Condition monitoring; vibration analysis; adaptive orthogonal transformation; bearing fault

## I. INTRODUCTION

The rolling bearing is one of the most widely used elements in rotating machinery. As a critical component, it carries most of the load during the running of rotating machinery. If the rolling bearing fails, serious problems arise, which will, in turn, result in the decrease of production efficiency and large economic loss. Records show that faulty bearings contribute to about thirty percent of the failures in rotating machinery [1]. As a result, it is of great importance to study the effective fault diagnosis approaches for rolling bearings.

Various monitoring have been developed for bearing fault diagnosis and condition monitoring, such as vibration analysis, temperature and acoustic emission monitoring [2]. Vibration signal analysis is one of the most efficient techniques thanks to the useful information to severity and type of bearing damage [3], [4]. Various signal processing techniques have been proposed for mechanical fault diagnosis are time domain [5], frequency domain [6]–[8], time–frequency domain analysis [9], high frequency resonance technique (HFRT) [10], [11], wavelet transform methods [12], [13] and automatic diagnosis techniques [14]. In summary, such methods can be primarily categorized into two classes: frequency identification and features classification.

The basic idea of these methods is the decomposition of the vibration signal in a system of function of orthogonal base as those of Fourier, Walsh or Haar, [15]–[17] to obtain the vector (spectre) of the informative characteristics. However, the spectrum obtained by these frequency methods in the majority of cases will complicate the procedure of comparing the signals of various types of faults, since the vibration signal is a non-stationary process. Hence the need for a method of computing the vector of the informative characteristics with a minimum dimension.

In this paper, for the first time, we propose to use the adaptive orthogonal transformations for the extraction of the

informative characteristics of bearing vibration signal. This method was used for voice signals [18] and was recently employed for classification of breast masses in mammography [19].

The use of these transformations is favored by the ability to adapt the shape of their basic functions according to the character of the standard vector. The latter is formed from the vibration signals of each fault type. In other words, each class of defects is associated with a system of basic functions adaptive for the projection of the signals. The formed basic function system is expressed as a factorization orthogonal matrix operator, which allows making a transformation with a fast calculation algorithm.

This paper is organized as follows. The principles of adaptive orthogonal transforms are introduced in Section II. The proposed method is validated using the data collected from bearing run-to-failure tests in Section III. Finally, the main conclusions are outlined in Section IV.

## II. THEORETICAL BACKGROUND

In digital treatment, transformed shelf space orthogonal of a signal  $X$  can be represented by the matrix (1).

$$Y = \frac{1}{N}HX \quad (1)$$

Where,

- $X = [x_1, x_2, \dots, x_N]^T$  is the initial signal is to be transformed (of size  $N = 2^n$ ).
- $Y = [y_1, y_2, \dots, y_N]^T$  is the vector of the spectral coefficients calculated by the operator orthogonal  $H$  of dimension  $N \times N$ .

To avoid the problem of signals synchronizations, we mention that  $X$  is transformed to the Frequency domain.

Factorization of Good [20] showed a possibility of representing the matrix operator  $H$  as product  $G_i$  (2) sparse matrix with a higher proportion of zero which has allowed the construction of the quick transformation algorithms of Fourier, Haar, and Walsh. The matrices  $G_i$  ( $i = 1, \dots, n$ ) are constructed by blocks of matrices  $V_{i,j}$  of minimum dimension that is called spectral nuclei:

$$G_i = \begin{bmatrix} \begin{bmatrix} \alpha_{i1} & 0 & \dots & 0 \\ \beta_{i1} & 0 & \dots & 0 \end{bmatrix} & \begin{bmatrix} \gamma_{i1} \\ \delta_{i1} \end{bmatrix} & \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ 0 & \begin{bmatrix} \alpha_{i2} & 0 & \dots & 0 \\ \beta_{i2} & 0 & \dots & 0 \end{bmatrix} & \begin{bmatrix} \gamma_{i2} \\ \delta_{i2} \end{bmatrix} \\ \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \begin{bmatrix} \alpha_{iN/2} & 0 & \dots & 0 \\ \beta_{iN/2} & 0 & \dots & 0 \end{bmatrix} & \begin{bmatrix} \gamma_{iN/2} \\ \delta_{iN/2} \end{bmatrix} \end{bmatrix} \quad (2)$$

With

$$v_{i,j} = \begin{bmatrix} \alpha_{ij} & \dots & \gamma_{ij} \\ \beta_{ij} & \dots & \delta_{ij} \end{bmatrix} = \begin{bmatrix} \cos(\alpha_{ij}) & \dots & w_{i,j} \sin(\alpha_{ij}) \\ \sin(\alpha_{ij}) & \dots & -w_{i,j} \cos(\alpha_{ij}) \end{bmatrix},$$

$$w_{i,j} = \exp(j\theta_{i,j}), \varphi \in [0, 2\pi], \theta \in [0, 2\pi]$$

Hence ((1)) can be written as follows:

$$Y = \frac{1}{N}HX = \frac{1}{N}G_1G_2 \dots G_nX = \frac{1}{N} \prod_{i=1}^n G_iX \quad (3)$$

By defining the angular parameters,  $\varphi_{i,j}$  and  $\theta_{i,j}$ , the operators of orthogonal transformations H can be formed with basic functions complex, or with real functions when  $\theta_{i,j} = 0$ . The calculation of the parameters depends  $\varphi_{i,j}$  on the choice of the structures of the spectral nuclei  $V_{i,j}$ . What allows generating a system of basic functions adapted to a given class of signals.

Yet, to assure a fast calculation, in this work, the spectral nuclei in matrices  $G_i$  are established so that they contain a higher proportion of zeros, such as he is explained below.

Adapting operator H in (1) is provided by the condition:

$$\frac{1}{N}H_aZ_{cd} = Y_c = [y_{c,1}, 0, 0, \dots, 0]^T, y_{c,1} \neq 0 \quad (4)$$

Where,

- $Y_c$  is the target vector which builds the criterion of adaptation of the operator  $H_a$ .
- $Z_{cd}$  represents the vector standard of a class calculated by means of the statistical characteristics of several vibratory signals.
- $H_a$  is adaptable to synthesize operator.

The synthesis of the adaptable operator  $H_a$  based standard  $Z_{cd}$ (for a given class), consists in calculating the angular parameters  $\varphi_{i,j}$  matrices  $G_i$  according to the condition (4). The procedure of the calculation of the parameters is illustrated by Fig. 1 the principle of which is based on an iterative algorithm introduced by Fig. 2, which allows the calculation of the target vector  $Y_c$  is according to the equation:

$$Y_i = G_iY_{i-1} \quad (5)$$

The calculation of the vector  $Y_c$  allows the obtaining of the adapted operator H. For the classification of the vibration signals, we dispose two sets of the vibration signals. The first

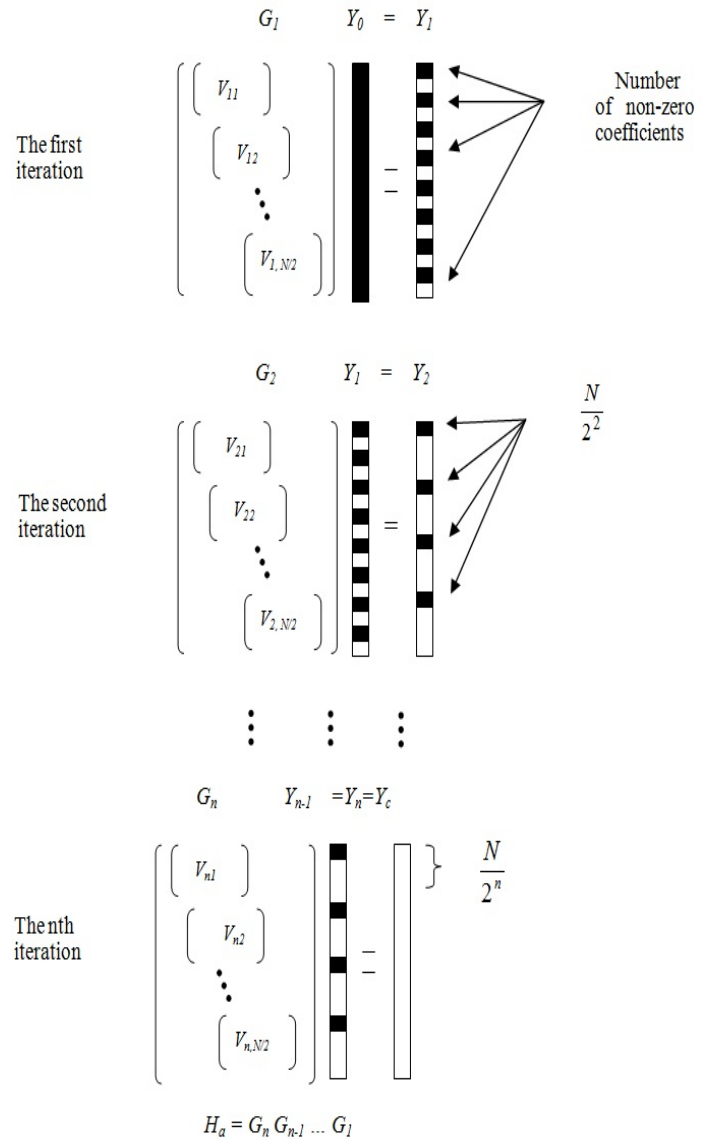


Fig. 1. The procedure of synthesis of the operator of the adaptive transformed.

one serves to calculate the standard  $Z_{cd}$  of  $i$  (class  $i$ ) and allows to generate the synthesis of the operator. Whereas the second set used to form the spectral standard  $Y_{sd,i}$  of  $i$ , which is obtained by the projection of the recordings of the second set in the adaptable base  $H_a$ .

To make the decision and classify vibration signal, we calculate each  $Y_i$  spectrum in each base  $H_{a,i}$ . To define the fault corresponding to the vector  $Y_i$  of the informative characteristics, we lean on a rule of decision formed by a combination of two criteria:

- The Euclidean distance  $\delta_i = \|Y_i - Y_{sd,i}\|$  and
- The distance of the energy concentrated in their first coefficients of the decomposition  $\varepsilon_i = |Y_{1,i}^2 - Y_{1,sd,i}^2|$ .

So, the vector  $Y_i$  will correspond to class  $i$  if  $\delta_i = \min(\delta_{k=1 \dots M})$  and  $\varepsilon_i = \min(\varepsilon_{k=1 \dots M})$ , with  $M$  is the number of classes. This procedure of classification is illustrated

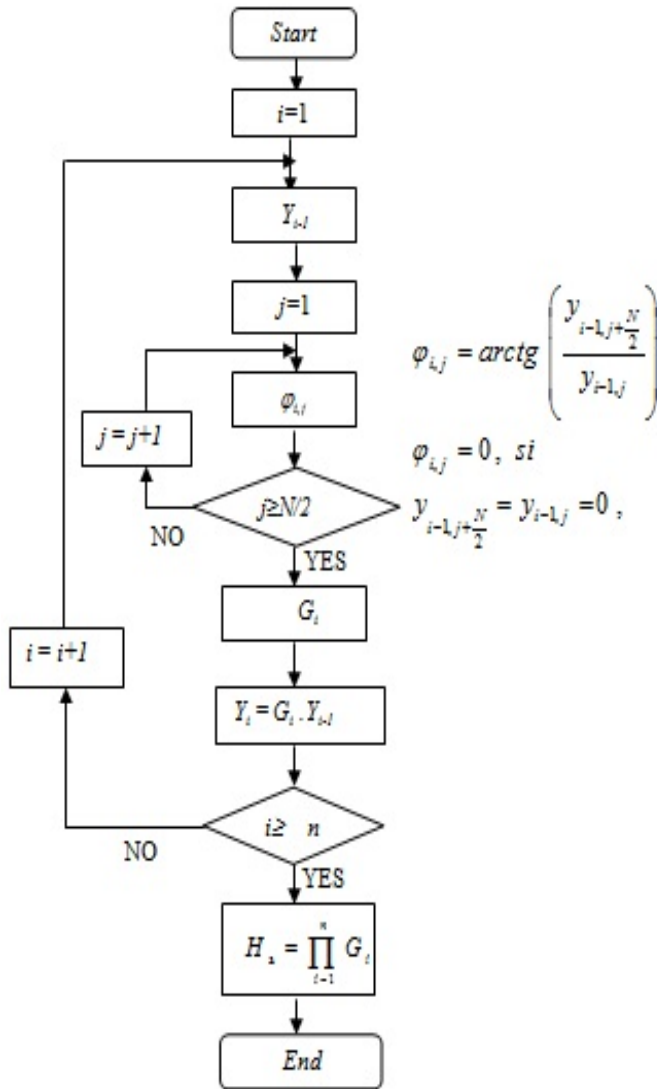


Fig. 2. The algorithm of synthesis of the operator of the adaptive transformed.

in the Fig. 3.

### III. APPLICATION TO EXPERIMENTAL SIGNAL

#### A. Experimental setup

The bearing test rig hosts four bearings were installed on a shaft. The rotation speed was kept constant at a rate of 2000 RPM by an alternative current motor coupled to the shaft via rub belts. A uniform radial load of 6000 lbs is applied onto the shaft and bearing. All bearings are lubricated.

Rexnord ZA-2115 double row bearings were installed on the shaft as shown in Fig. 4. A PCB 353B33 High Sensitivity Quartz ICP accelerometers were installed on the bearing housing. The test rig and sensors placement are also shown in Fig. 4. All failures occurred after exceeding designed lifetime of the bearing which is more than 100 million revolutions. Vibration data were collected every 10 minutes by NI DAQCard-6062E at the sample rate set at 20 KHz.

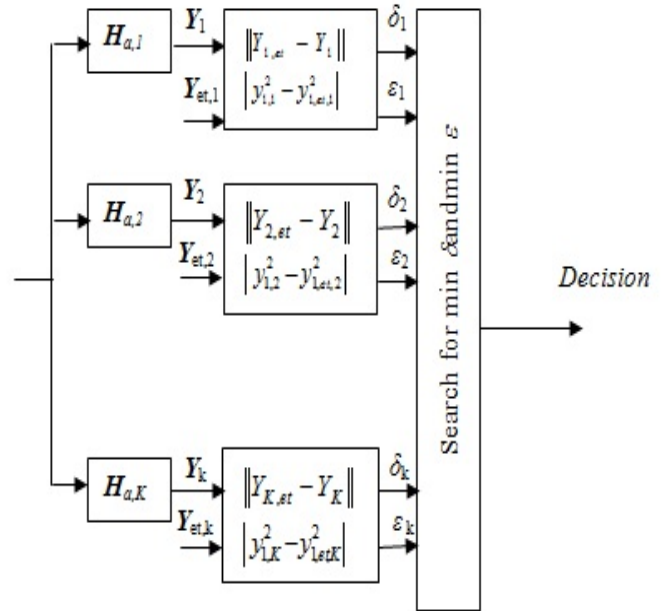


Fig. 3. Classification procedure.

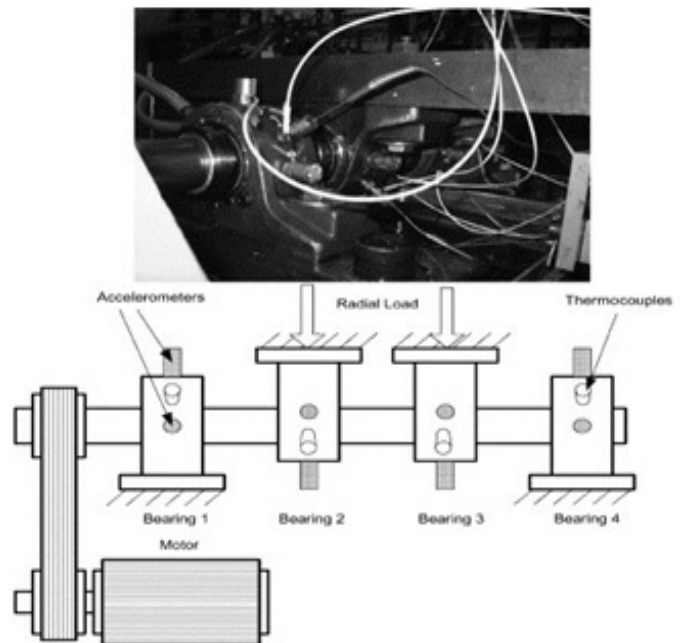
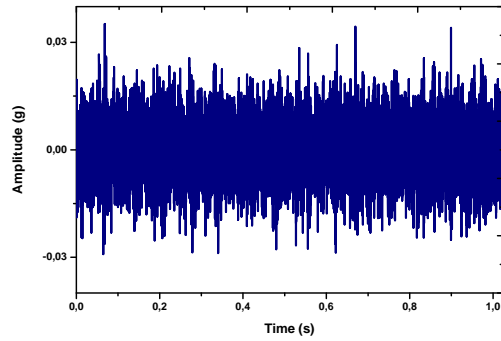


Fig. 4. Bearing test rig.

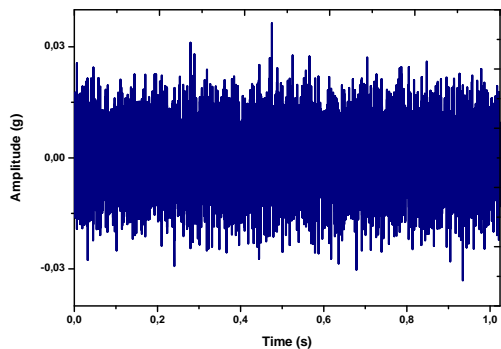
The test was carried out for 35 days until a significant amount of metal debris was found on the magnetic plug of the test bearing. An inner race defect was discovered in test bearing 1.

#### B. Experimental Results Analysis

The proposed method was applied to detect the bearing with outer race fault. The raw vibration signal of normal operating conditions and outer race failure occurred in bearing 1 are plotted in



(a)



(b)

Fig. 5. Vibration signal of: (a) normal operating conditions, (b) outer race failure occurred in bearing 1.

Fig. 5a and 5b, respectively.

Fig. 6a and 6b present the frequency spectrum of normal state and outer race. The characteristic defect frequencies cannot be obtained directly in FFT spectrum.

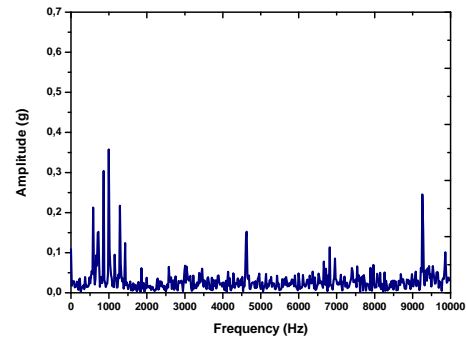
By using the elaborate method, the projection of normal signal in the normal class base and fault class base are plotted in Fig. 7a and 7b, respectively. We can notice that the energy of the projection of the normal signal in the adaptive base has a small spectral vector (Fig. 7a).

Fig. 8a and 8b illustrate the projection of fault signal in the normal class base and fault class base. It can be seen that during the projection of this signal at a normal base, we obtain rather a broad spectral vector (Fig. 8a).

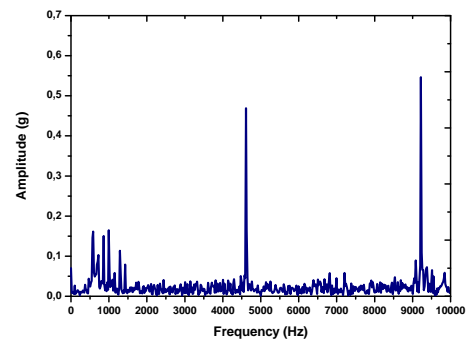
This result demonstrate that the first signal belongs to the class of the normal signal and the second signal belongs to the class of abnormal signal, respectively. The same conclusion also manifested by values of  $\delta$  and  $\varepsilon$ .

The results obtained by the developed method, illustrated in Fig. 7 and 8 indicate its effectiveness and show that it ensures a high distinction that will help to make the classification of bearing vibration signal.

The efficiency of the elaborate method is illustrated on Fig. 9 which reflects the certainty of classification according to the size of the interval of the analysis. The certainty of the



(a)



(b)

Fig. 6. FFT of: (a) normal operating conditions, (b) outer race failure occurred in bearing 1.

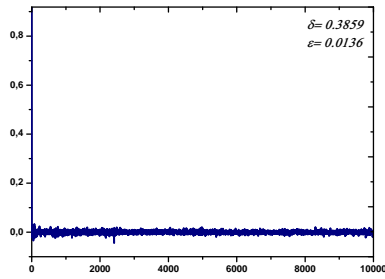
classification of the signals is much higher and can reach a 100 % value as the interval of analysis increases.

#### IV. CONCLUSION

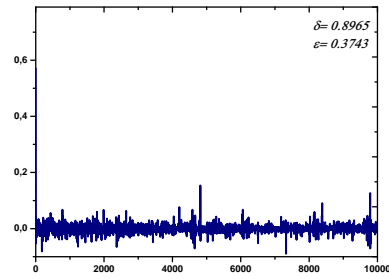
To improve the accuracy fault classification of bearings in rotating machines, a new method is developed based to calculate the informative characteristics of the vibration signal. The experimental results show that the method ensures a high distinction that will help to make the classification of bearing vibration signal. The developed software system according to this method will be beneficial for practical fault classification.

#### REFERENCES

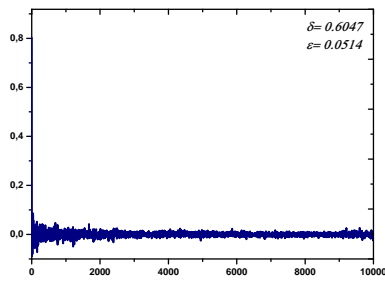
- [1] N. Tandon and A. Parey, "Condition monitoring of rotary machines," in *Condition Monitoring and Control for Intelligent Manufacturing*. Springer, 2006, pp. 109–136.
- [2] W. Zhou, T. G. Habetler, and R. G. Harley, "Bearing condition monitoring methods for electric machines: A general review," in *Diagnostics for Electric Machines, Power Electronics and Drives, 2007. SDEMPED 2007. IEEE International Symposium on*. IEEE, 2007, pp. 3–6.
- [3] N. Tandon and B. Nakra, "Comparison of vibration and acoustic measurement techniques for the condition monitoring of rolling element bearings," *Tribology International*, vol. 25, no. 3, pp. 205–212, 1992.
- [4] N. Tandon and A. Choudhury, "A review of vibration and acoustic measurement methods for the detection of defects in rolling element bearings," *Tribology international*, vol. 32, no. 8, pp. 469–480, 1999.
- [5] H. A. Khwaja, S. Gupta, and V. Kumar, "A statistical approach for fault diagnosis in electrical machines," *IETE Journal of Research*, vol. 56, no. 3, pp. 146–155, 2010.



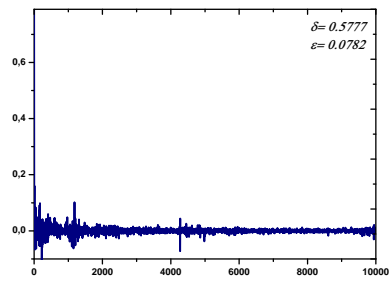
(a)



(a)



(b)



(b)

Fig. 7. Projection of normal signal in the: (a) normal class base, (b) fault class base.

Fig. 8. Projection of fault signal in the: (a) normal class base, (b) fault class base.

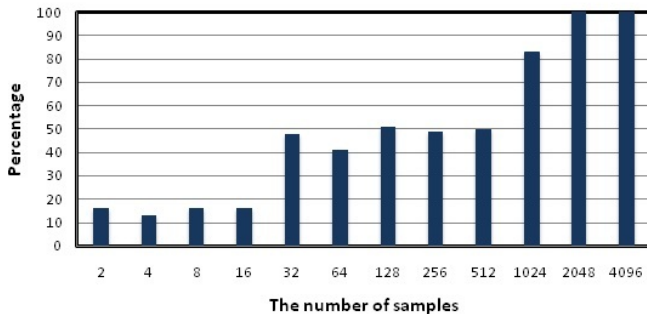


Fig. 9. Classification percentage according to the size of the interval of the analysis.

[6] E. Jantunen, "A summary of methods applied to tool condition monitoring in drilling," *International Journal of Machine Tools and Manufacturing*, vol. 42, no. 9, pp. 997–1010, 2002.

[7] F. P. G. Márquez, A. M. Tobias, J. M. P. Pérez, and M. Papaalias, "Condition monitoring of wind turbines: Techniques and methods," *Renewable Energy*, vol. 46, pp. 169–178, 2012.

[8] S. Gowid, R. Dixon, and S. Ghani, "A novel robust automated fft-based segmentation and features selection algorithm for acoustic emission condition based monitoring systems," *Applied Acoustics*, vol. 88, pp. 66–74, 2015.

[9] J.-H. Lee, J. Kim, and H.-J. Kim, "Development of enhanced wigner-ville distribution function," *Mechanical systems and signal processing*, vol. 15, no. 2, pp. 367–398, 2001.

[10] P. McFadden and J. Smith, "Vibration monitoring of rolling element bearings by the high-frequency resonance technique review," *Tribology international*, vol. 17, no. 1, pp. 3–10, 1984.

[11] T.-C. Liu and T.-Y. Wu, "Application of empirical mode decomposition and envelop analysis to fault diagnosis in roller bearing with single/double defect," *Smart Science*, vol. 5, no. 3, pp. 150–159, 2017.

[12] N. Nikolaou and I. Antoniadis, "Rolling element bearing fault diagnosis using wavelet packets," *Ndt & E International*, vol. 35, no. 3, pp. 197–205, 2002.

[13] Y. Jiang, B. Tang, Y. Qin, and W. Liu, "Feature extraction method of wind turbine based on adaptive morlet wavelet and svd," *Renewable energy*, vol. 36, no. 8, pp. 2146–2153, 2011.

[14] P. Jayaswal and A. Wadhvani, "Application of artificial neural networks, fuzzy logic and wavelet transform in fault diagnosis via vibration signal analysis: A review," *Australian Journal of Mechanical Engineering*, vol. 7, no. 2, pp. 157–171, 2009.

[15] K. Rao and N. Ahmed, "Orthogonal transforms for digital signal processing," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76.*, vol. 1. IEEE, 1976, pp. 136–140.

[16] H. Kekre, T. K. Sarode, P. Natu, and S. Natu, "Transform based face recognition with partial and full feature vector using dct and walsh transform," in *Proceedings of the International Conference & Workshop on Emerging Trends in Technology.* ACM, 2011, pp. 1295–1300.

[17] N. Ahmed and K. R. Rao, *Orthogonal transforms for digital signal processing.* Springer Science & Business Media, 2012.

[18] A. Abdenbi, A. A. Fadoua, and N. Benayad, "Vers un système de reconnaissance automatique de la parole en amazighe basé sur les transformations orthogonales paramétrables."

[19] K. El Fahssi, A. Elmoufidi, A. Abenaou, S. Jai-Andaloussi, and A. Sekkaki, "Feature extraction of the lesion in mammogram images using segmentation by minimizing the energy and orthogonal transformation adaptive," *WSEAS TRANSACTIONS on BIOLOGY and BIOMEDICINE*, vol. 11, 2014.

[20] I. J. Good, "The interaction algorithm and practical fourier analysis,"

*Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 361–372, 1958.

# Improving Security of the Telemedicine System for the Rural People of Bangladesh

Toufik Ahmed Emon  
Dept. of CSE  
Jahangirnagar University  
Savar, Dhaka 1212

Uzzal Kumar Prodhan  
Dept. of CSE  
Jatiya Kabi Kazi Nazrul Islam University  
Trishal, Mymensingh

Mohammad Zahidur Rahman, Israt Jahan  
Dept. of CSE  
Jahangirnagar University  
Savar, Dhaka 1212

**Abstract**—Telemedicine is a healthcare system where health-care professionals have the capability to observe, diagnose, evaluate and treat the patient from a remote location and the patient have the ability to easily access the medical expertise quickly and efficiently. Increasing popularity of Telemedicine increase the security intimidations. In this paper, a security framework is implemented for the developed cost-effective Telemedicine system. The proposed security framework secure all the sections of the model following the recommendations of Health Level 7, First Healthcare Interoperability Resources and Health Insurance Portability and Accountability Act. Implementation of this security framework including authenticating the different types of user, secure connection between mobile and sensors through authentication, protect the mobile application from hackers, ensures data security through encryption, as well as secure server, using secured socket layer called SSL. Finally, we can say that the developed Telemedicine model is more secure and it can be implemented in any remote areas of developing countries as like as Bangladesh.

**Keywords**—Telemedicine; security; encryption; hashing

## I. INTRODUCTION

About 400 million people around the world are deprived of the basic healthcare service. In Bangladesh, the number of physician per 10,000 people is only 3 and nurse per 10,000 is only 1.07 [1]. These data show a severe scarcity of healthcare services in Bangladesh. In this regards, we develop stored and forward telemedicine system. In our developed system, the expert and local doctors, the pharmacy admin and lab assistant need to register in the system. The system administrator does the registration. The local pharmacist registers the patients. The local doctors Login the mobile application which is connected to sensors to get sensors data. When the sensors data received by the mobile application, it shows in the mobile application. Then the data is encrypted as well as send to the remote server. These data decrypt in the server site and save on the server. Doctors examined all the data and send a prescription to the patients through proper channel. The expert doctors also registered in the system before they prescribe patients. The end user gets the prescription and gives the prescribed medicine as well as suggestions to the respective patients. To make the system reliable as well as trustworthy to the user, security becomes the main concern of this telemedicine system.

In this research, we improve the security and privacy features of the developed telemedicine system. The security framework is particularly designed for our developed

telemedicine system. Our developed telemedicine includes several modules such as sensors, mobile application, web application as well as the web server. There are some other modules as like Bluetooth connected devices, wifi or data transmission through other media. Therefore, to secure a telemedicine, we need to secure all of its modules. We divide the telemedicine framework into five different sections to improve the security, such as authentication and application security, client layer security, patient data security, web server security as well as database security. Furthermore, ensures the security of each module following the recommendations of Health Level 7 or HL7 and Fast Healthcare Interoperability Resources or FHIR. Not only that we also several security models, such as Access control list or ACL, Multi-level Security or MLS as well as Role-based access model or RBAC.

The rest of the paper is organized as follows: In Section II, we briefly discuss different methods currently being used to secure the telemedicine system as well as their advantages and disadvantages. In Section III, we describe the main module of our telemedicine system followed by a brief description of security module with its security advantages in Section IV. Section V discusses the drawbacks and future work of this proposed module. Finally, in Section VI, we conclude with the summary.

## II. LITERATURE REVIEW

Security of telemedicine is a growing concern as it becomes more complex day by day. There are several security models to secure the telemedicine system. M. Fahim Ferdous Khan and Ken Sakamura proposed a hybrid access control model for healthcare informatics considering the following issues such as network security, emergency access, the principle of minimum disclosure, user approval, access authorization, etc. Their authentication mechanism is based on the eTRON architecture [2]. On the other hand, Liu et al proposed a model named as Open and Trusted Health Information Systems or OTHIS targeting the Australian Health sector. Their proposed system is compatible with general Health Information System or HIS [3]. Prema T. Akkasaligar and Sumangala Biradar encrypt medical image using Chaos theory and DNA cryptography. They divided the image into odd and even DNA encoded image. Then they add both images to get the original encrypted image.

To decrypt the image they use reverse process [4]. M.J. Chang, J. K. Jung, M.W. Park and T.M. Chung find out the security holes in Telemedicine and suggest firewall to control



unauthorized access [5]. I. Chiuchisan, D.G. Balan, O. Geman and I. Chiuchisan and I. Gordin use signature verification, data encryption as well as secure network infrastructure to secure the telemedicine system [6]. C. Fu, Y. Lin, H. y. Jiang and H. f. Ma improves their encryption by extending the key length to 212 bits. They declared all the variables as 64-bit double precision type [7].

Basudev Halder and S.Mitra implement a watermarking based process in which they can recover the converted ECG images without any distortion [8]. C. Han, L. Sun, and Q. Du use fountain code and image segmentation to secure image transmission. They divide the image into two parts. The main advantages of their proposed method are lower complexity and lower cost [9]. R.M. Seepers, C. Strydis, I. Sourdis and C.I. De Zeeuw proposed a heartbeat based security module where they used IPI (Inter-Pulse-Interval) to generate the security code [10]. Uzzal Kumar Prodhana, Mohammad Zahidur Rahman, and Israt Jahan did a survey on Telemedicine status in Bangladesh and found most of the Bangladeshi people, doctors, and nurses, pharmacy, and hospital want telemedicine service [11]. A. Sudarsono, P. Kristalina, M.U.H. Al Rasyid and R. Hermawan encrypt 8-types of sensor data using AES128-bit and transmit these encrypt data using MD5 data sensor digest [12]. Ahmed Ibrahim et al. introduced a structure that permits the protected exchange of health information among different healthcare providers. Patients approve a particular type of medical information to be retrieved, which helps to prevent any undesired leakage of medical information [13].

Mamta Puppala et al. stated the METEOR framework which consists of two components: the enterprise data warehouse (EDW) and a software intelligence and analytics (SIA) layer which facilitates a wide range of clinical decision support (CDS) systems [14]. Role-based authorized method of access is proposed by T.W. Tseng, C.Y. Yang and C.T. Liu [15]. W.D. Yu, L. Davuluri, M. Radhakrishnan and M. Runiassy proposed a security-oriented design framework (SOD) which is a three-tier architecture. In their proposed system they use SHA1 algorithm to secure login data as well as use HTTPS secure web server [16]. Khan Zeb et al. introduced a U-Prove based security technique to authenticate Telemedicine users [17]. N. Jeyanthi et al. proposed a reputation based service where the users will be accepted by a proxy server which performs entry level authentication [18]. T. Vivas, A. Zambrano, and M. Huerta proposed a digital certificate based module to secure telemedicine [19]. Fatemeh Rezaeibagha and Yi Mu proposed a new protocol for telemedicine data security [20]. J. Singh and A.K. Patel proposed web late based watermarking for telemedicine security [21]. iMedic a four-tier based security model for telemedicine proposed by Amiya K. Maji et al. which includes an extra layer to make Telemedicine system more secure [22].

### III. DEVELOPED TELEMEDICINE SYSTEM

We develop a low cost, portable and secured telemedicine system for the rural and deprived people of Bangladesh. To make it more user-friendly and flexible we divide our telemedicine system into four main module. These are Local Administrator in Pharmacy, Local Doctors, Expert Doctors and Health System Administrator. The following business process

diagram shows working process of the four main module of our system (Fig. 1).

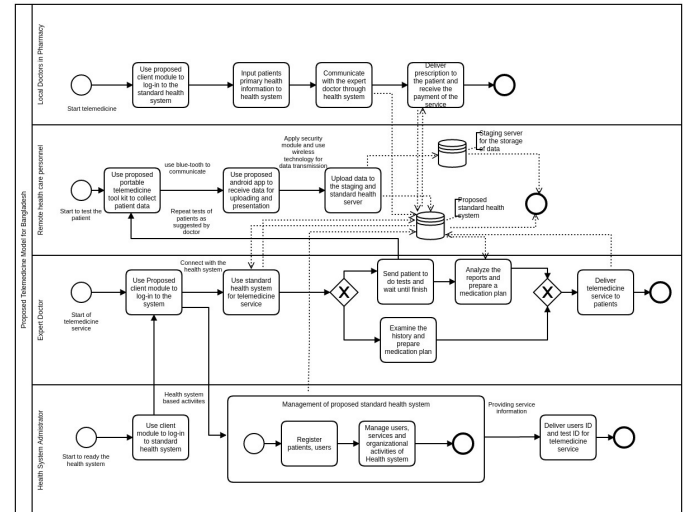


Fig. 1. Developed telemedicine system.

We have several components to complete the task for each module. The component diagram of our telemedicine module is given below (Fig. 2):

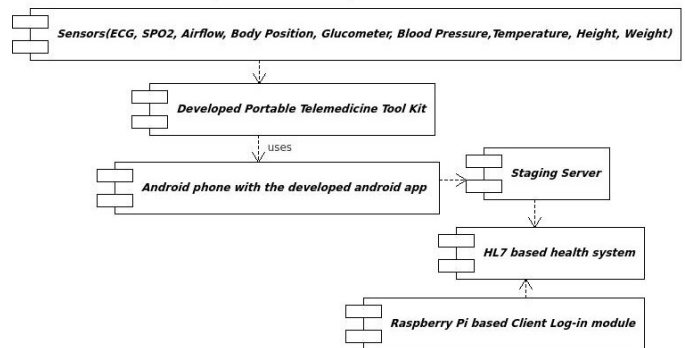


Fig. 2. Component diagram of developed telemedicine model.

In our developed Telemedicine system, every user must be registered. The health system administrator registers the expert doctors, the local pharmacy admin (a pharmacy admin is a person who works in the local pharmacy and responsible for all the local administrative work) as well as the local doctors. The pharmacy admin is responsible for the registration of the remote patients. The pharmacy admin also assigns a doctor for the patient at the time of registration. When the remote patients get registered, he/she have the patient id which is used for future correspondence. The completed registration in the system looks like Fig. 3:

When a patient needs Telemedicine service from the local pharmacy admin, he/she needs to describe his/her problems to the pharmacy admin. The pharmacy admin input all the patients data into the system using a Raspberry Pi based client login module against the respective patient ID. The expert doctor checks all the history of the patient. If needed then the expert doctor asked the pharmacy admin to do certain medical

Fig. 3. Patients registration form.

checkup for the patient. After getting the request from doctors to do medical checkup for the patients, the pharmacy admin instructs local doctors to complete the prescribed task. The local doctors do the prescribed task using the develop portable Telemedicine toolkit. The toolkit includes nine types of sensors including ECG, SP02, Airflow, Body Position, Glucometer, Blood Pressure, Temperature, Height and Weight sensors. The local doctors do the checkup using a mobile application which is connected with the developed toolkit through a secured Bluetooth connection. The local doctors need to Login the Android mobile application to get the sensors data. When the sensors data received by the Android application, it showed on the application interface as follows (Fig. 4 and 5):

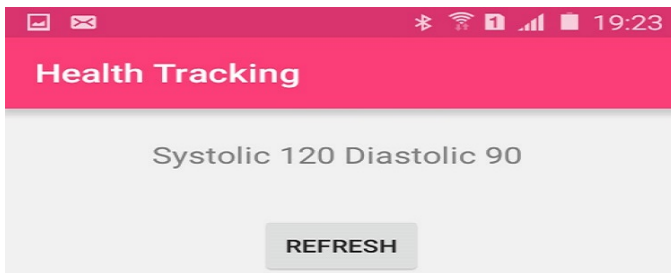


Fig. 4. Blood pressure data in mobile application interface.

The local doctors see the data and upload the data to the remote staging server against the respective patient id along with the test id. There needs to open the mobile internet connection to send the data. An encryption algorithm is implemented in the mobile application to encrypt the sensors data. The data sent to the staging server is originally an encrypted data. In the server site, there is a decryption algorithm to decrypt the data and store the original data. The expert doctors Login the HL7 based open health system called GNU health system and prescribe the patient by observing their medical test results. The remote pharmacy admin Login the system and get the prescription of the patient (Fig. 6).

After that the patients get the medicine or advice prescribed by the expert doctors from the pharmacy admin, giving a small amount of royalty fee for the prescription which is around 300 BDT. The prescription also stores in the system. Therefore, if needed then the patients get the prescription at any time. The pharmacy admin also generates invoice report, service report

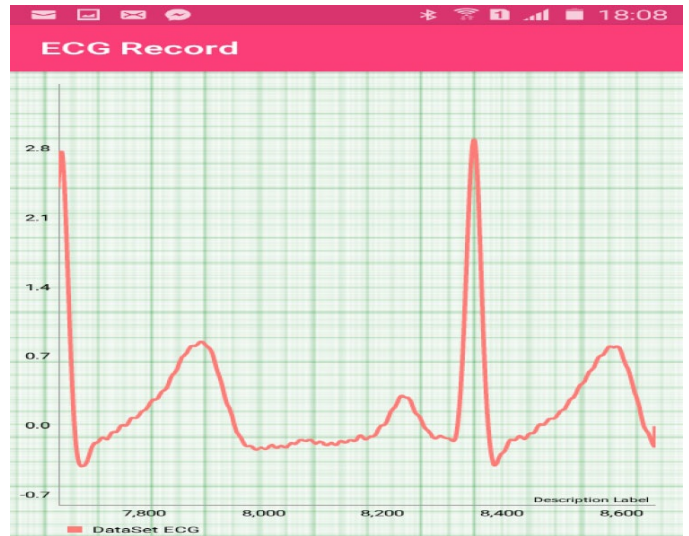


Fig. 5. ECG data in the mobile application interface.

Medication	Dose	Unit
Amoxicillin 500mg capsules	500	mg

Fig. 6. Prescription of a patient.

as well as fix an appointment for the patients.

#### IV. IMPLEMENTED SECURITY TECHNIQUE

We consider various security model to secure an authentication, authorization, transmission etc to secure our developed telemedicine system. Among them, Access Control List or ACL attach number permission to an object. Define the object accessibility as well as define which user access which section of the data [23]. Role-Based Access Control or RBAC works for a large number of people where ACL called the minimal RBAC model [24]. In Multi-Level Security, information flow between the authorized users only and only the privileged user can read the information. It also prevents unauthorized users to access the data. The most common Multi-Level Security model use for security purpose is called Bell-LaPadula model. Bell-LaPadula model checks the subject security model when a user tries to read or write on the subject as well as no object bypass any authorized users [25]. On the other hand, BIBA security model ensures the data security model by providing several access control rules. In BIBA security model, a lower level user is not permitted to request higher level user documents [26].

Not only security model but also some other organizations such as Health Insurance Portability and Accountability Act or HIPPA provide privacy, security, enforcement as well as breach notification rule to make a healthcare data to a Protected Health Information or PHI [27]. The National Authentication Service for Health or NASH provides PKI as well as Public Key Infrastructure certificate which helps users to know about their healthcare data authentication, integrity, non-repudiation as well as confidentiality [28].

Considering different security model as well different organizations security and privacy rules we divided our develop Telemedicine framework into five different sections such as application security, user authentication, web server security, database security as well as data security. The block diagram of our security measures are given below (Fig. 7):



Fig. 7. Block diagram of security framework.

#### A. Authentication and Application Security Layer

Authentication and Application Security Layer is designed to authenticate every user as well as provide security for the Android application. In our system, there are types of user, one is remote doctors or trained personnel, expert doctors and local pharmacy admin. There are three types of authentication needed to use this Telemedicine system.

**Firstly**, the Telemedicine system administrator registers the remote doctors, pharmacy admin as well as expert doctors. All these users get username and password to Login the system. The remote pharmacy admin registers the patients and assigned a doctor to their patient id.

**Secondly**, the remote doctors need to Login the mobile application with the same username and password which is provided when he/she register in the system before getting the patients data from sensors. **Thirdly**, a password is needed to establish the connection between toolkit Bluetooth module and mobile application. To get data from sensors we use HC-05 Bluetooth module which has a default password and username. To enrich security we set new password and username for this Bluetooth module.

To get the data from the sensors and send these data to the server we use an Android application. Therefore, reverse engineering made it possible for a hacker to find out the application data. To enrich security as well as prevent reverse engineering we will implement ProGuard. ProGuard compresses the application which saves a lot of space as well as encrypts the mobile application which makes the application code obfuscate to prevent any kind of security threats.

#### B. Client Layer

Client layer help users to interact with the system. It consists of both mobile application interface as well as web

application interface. Mobile application layer helps remote healthcare personnel to get sensors data. On the hand, web interface help doctors to prescribe patients as well as remote doctors to get the expert doctors prescription.

- We use HL7 based open source health system called GNU [29] healthcare system for our Telemedicine system which provides a web interface for both specialized doctors and remote users, also for pharmacy admin. GNU healthcare system provides a user interface called Tryton to Login the system. Python language used to create Tryton web interface. Whereas, Tryton is a three-tier high-level customary design computer application principles. Fig. 8 describes the security measures we implement to secure the web application: We develop an Android-based mobile application to

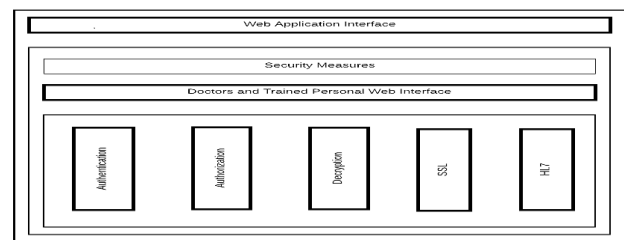


Fig. 8. Security steps implementation for web client.

get data from the sensor. In order to get data from sensors, we implement different security measures to get data without any kind of data distortion. The end user needs to Login the Android application using username and password given at the time of registration. To secure the Login data, we implement Message Digest or MD algorithm. There are MD2, MD4, MD5, and MD6 hash algorithm. The MD2 has 18 rounds with 512 bits digest size. This hash algorithm is optimized for 8 bits computer. The MD4 has 3 rounds with 128-bits digest size and 512 bits block size. The MD5 hashing algorithm improves its security feather by adding one more round. Therefore, it has 4 round with 128-bits digest size with 512-bits block size. There are some vulnerabilities in message digest hashing algorithm but it has a little effect on MD5, even though the MD5 algorithm is faster than SHA algorithm [30]. Fig. 9 shows the block diagram of the MD5 hashing algorithm.

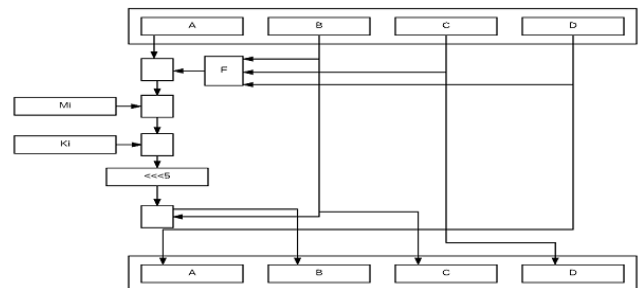


Fig. 9. MD5 hashing algorithm.

### C. Data Security Layer

There are various types of encryption algorithm including symmetric, asymmetric, shared or cryptographic hash function to encrypt data. Among them, we will implement symmetric encryption algorithm named as Advanced Encryption Algorithm. There are AES-128 bits, AES-192 bits, and AES-256 bits. The AES-128 bit has 10 rounds, AES-192 has 12 rounds and AES-256 has 14 rounds, here more rounds mean more security against security attack. United States National Security Agency reviewed all AES algorithm and recommend AES-256 and AES-192 to secure classified documents secure [31]. Not only that Fast Healthcare Interoperability Resources or FHIR also recommend AES algorithm to secure data. Therefore we will implement AES-256 to encrypt our sensors data.

We implement AES-256 bit encryption in mobile application and encrypt data as well as send it to the remote web server. In the server site, we decrypt the data with the same algorithm and save data in the database. The following block diagram shows AES algorithm (Fig. 10).

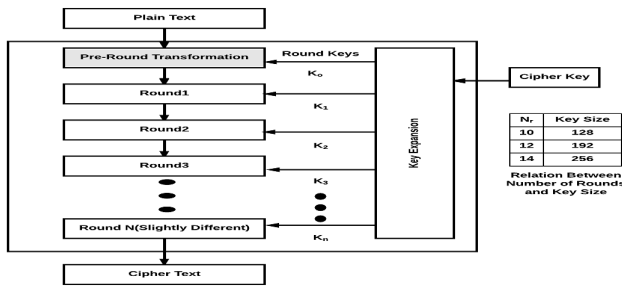


Fig. 10. AES algorithm block diagram.

### D. Middle Layer

Apache Tomcat server is the main part of the middle layer where the GNU healthcare system is installed. GNU health provides an interface called Tryton. These server responses for HTTPS request from the mobile application as well as web clients.

### E. Data Layer

The data layer consists of web database where the patient's data is stored. We use GNU healthcare which uses PostgreSQL database which an open source most secured database [32]. The remote healthcare personnel sends the data to the web server with a valid patients id. The patient's data store in the database with a valid and unique patients id with examination id.

There are several authentications procedure to get data from the sensors and send it to the server. Moreover, there are also some authentication steps to get the prescription from the doctors and make it available to the patients. The overall security steps with workflow from the Bluetooth connection between mobile application and toolkit to the data store in web server is represented in a flow chart as follows (Fig. 11):

After getting the data from remote doctors the expert doctors review the data. Therefore, he/she provide the prescription

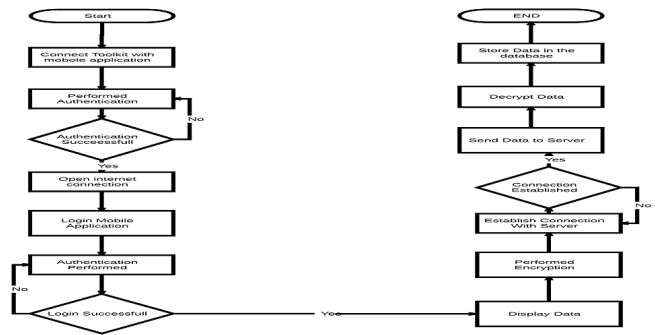


Fig. 11. Workflow from remote doctors to expert doctors.

for the respective patient. The remote doctor Login the system and find the prescription as well as provide medicine to the patients. The following Fig. 12 shows the complete workflow as well as proposed security steps.

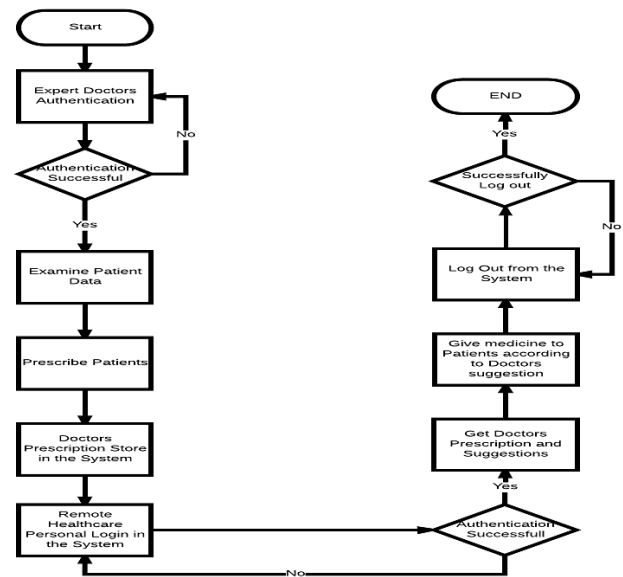


Fig. 12. Workflow from expert doctors to remote doctors.

## V. IMPLEMENTATION AND RESULT

The most common threats to data privacy and security include data theft, unauthorized access, improper disposal of data, data loss, hacking IT incidents and more. In this section, we have implemented the security measures to prevent unauthorized access, improper disposal of data and data loss. The list of section where the security measures are implemented are given below:

- Bluetooth Connection
- Mobile Application User Authentication
- Mobile Application Security
- Sensors Data Security
- Server Security

Therefore, Fig. 13 shows the sequence diagram of the Telemedicine system on which the security measures are implemented is as follows:

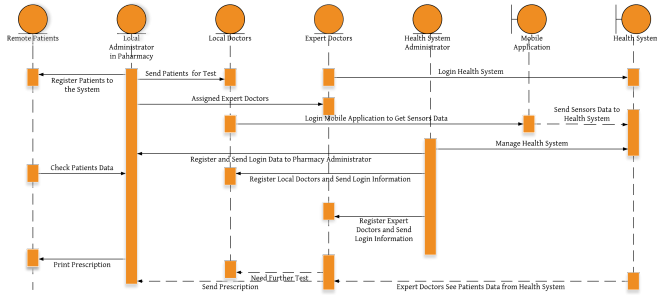


Fig. 13. Sequence diagram of the telemedicine system on which the security measures are implemented.

This chapter also describes the results of the implemented security technique. Fig. 14 shows the graphical representation of security implementation.

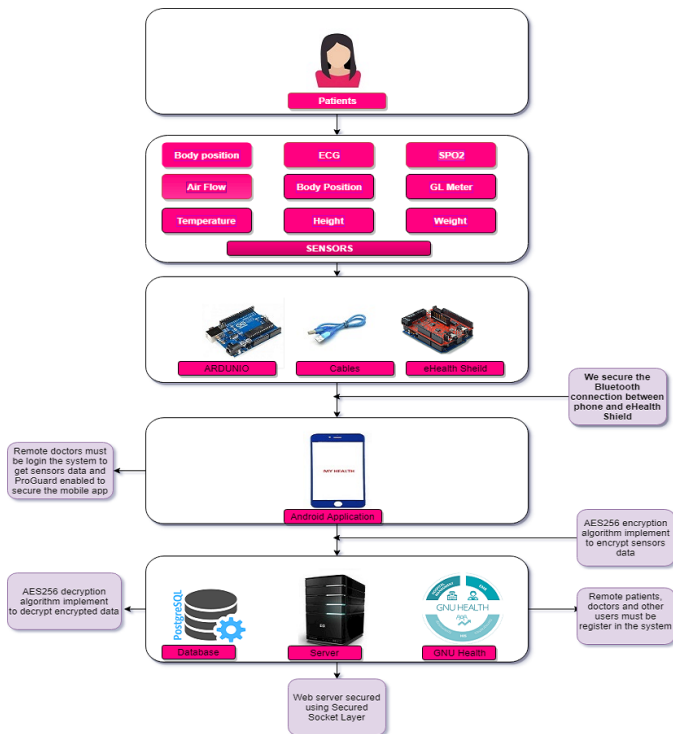


Fig. 14. Security techniques applied in different sections of the developed telemedicine system.

### A. Enhancing Bluetooth connection security

HC-05 has a default password which is '0000' or '1234' and default Bluetooth name 'HC-05'. To improve security we change the default name to "My Telehealth" and change default password to a 15 digit password with a combination of numbers, digits, uppercase letter, lowercase letter and special characters using AT command. Here, AT command stands for Attention Command used to change default Bluetooth setting. For a 4 digit number password, there is 10,000 possible combination. Therefore, for a 15 digit password with a

combination of number, character, special character, lowercase letter, and the uppercase letter, a total 70 character, the possible combination is 721480692460864. This password combination makes it unbreakable.

To set the password, first of all, we change the HC-05 mode to attention mode. Therefore, we press a enable button just on the opposite side of the HC-05 until the lid shows the indication. The indication shows that it is enabled for attention mode. Then, we upload a blank code to the module. To set the Bluetooth name we write the following code in Arduino serial monitor.

```
AT+NAME= JU-TELEMEDICINE-CENTER
```

And set password we write:

```
AT+PSWD= Telemedicineju1
```

After successfully configure the Bluetooth, the authentication process looks like as follows where Fig. 15 shows the available Bluetooth device, Fig. 16 shows the Bluetooth connection asked for a password, Fig. 17 shows the entering the password for connection and the last Fig. 18 shows the successful connection.

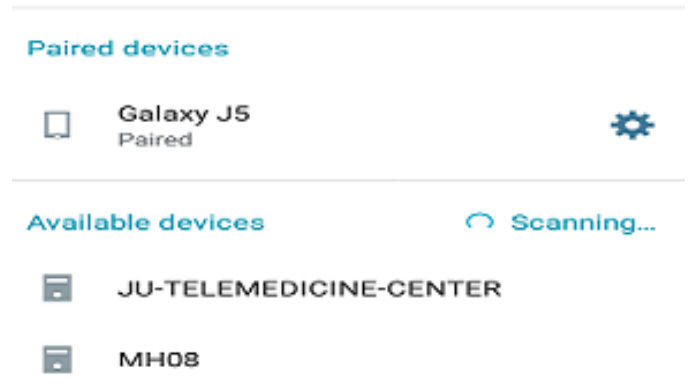


Fig. 15. Available Bluetooth devices.

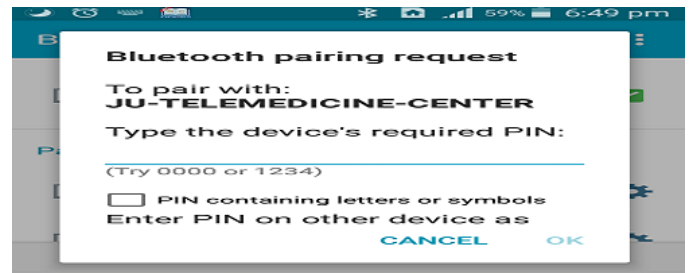


Fig. 16. Bluetooth connection asked for password.

### B. Authenticating Remote Doctors in the Mobile Application

We create a Login module for the remote doctors in the mobile application. Every time when remote doctors open the mobile application to get the sensors data, he/she must be login the mobile application with their username and password that

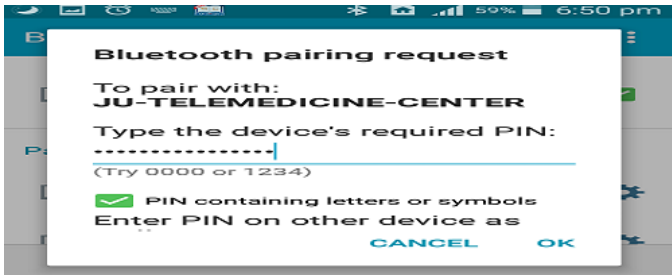


Fig. 17. Entering the password in the pop-up window.

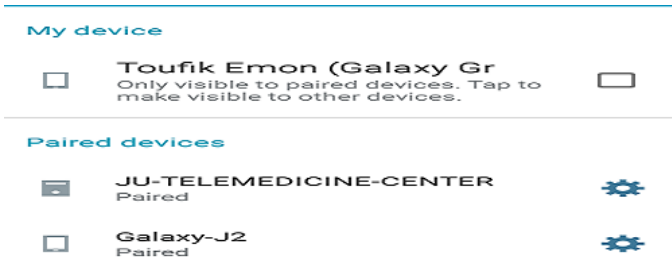


Fig. 18. Successful connection.

was given at the time of registration to use the app. Figures below shows the authentication process of the application (Fig. 19 to 21).

We also secure the username and password of the remote doctors by implementing MD5 hashing algorithm. There are several steps to implement the hashing algorithm. The implementation steps are:

- Step-1: Append Padding Bits

In this step, we extended the message so that the message length is similar to 448, modulo 512. The message is padded therefore it is just 64 bits which is a multiple of 512 bits. To complete the padding, first of all, a single "1" bit is added to the message, and then "0" bits are appended so that the length of bits of the full message becomes congruent to 448, modulo 512. At least one bit and at most 512 bits are appended.

- Step-2: Append Length

A 64-bit symbol of the message before the padding bits were added is appended to the result of the step-1. In the message, the bit is greater than  $2^{64}$ , then only the low-order 64 bits of the message is used.

- Step-3: Initialize MD5 Buffer

This step includes the initialization of 34 bits four-word buffer A, B, C, D with the hexadecimal number where A= 01 23 45 67, B= 89 ab cd ef, C= fe dc ba 98 and D= 76 45 32 10.



Fig. 19. Authentication interface.

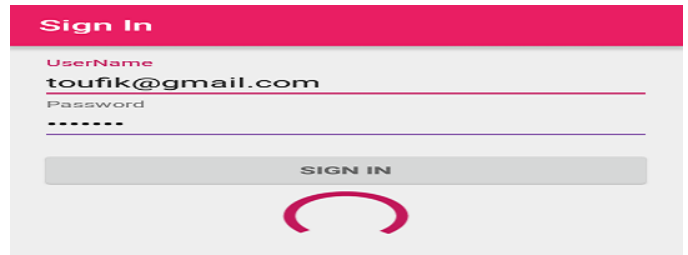


Fig. 20. Entering username and password.

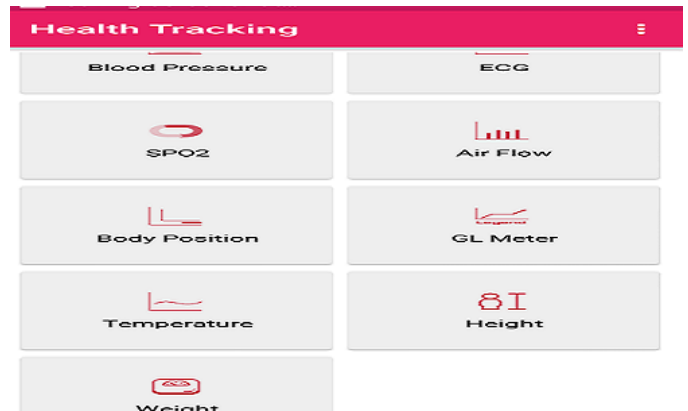


Fig. 21. Successful login.

- Step-4: Process Message in 16-Word Blocks

This step calculates the MD5 hashing.

- Step-5: Output

This step includes the implementation of MD5 hashing.

Fig. 22 and 23 shows the results of the MD5 implementation.

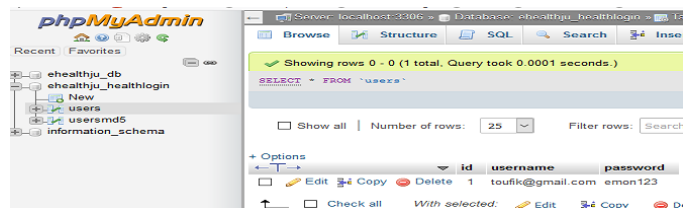


Fig. 22. Before implementation of MD5 hashing.

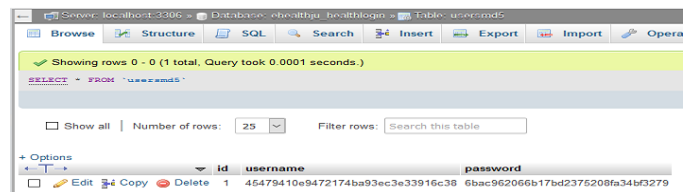


Fig. 23. After implementation of MD5 hashing.

### C. Mobile Application Security

Securing the mobile application is one of the most important tasks. We secure our Android-based mobile application

from hackers as well as prevent to struct data from our mobile application by enabling ProGuard. To enable ProGuard we have to open our mobile application in Android Studio. Then, find the file named build.gradle(Module: app) and open the file. In the buildTypes section, we found minifyEnabled false. Now to enable ProGuard, we change the state of minifyEnabled to true state which looks like minifyEnabled true. Now we generate a signed APK for our application. After successful generation of signed APK, there is some change in our application file. Though enabling ProGuard changes variable names as well as the class name, therefore, we need to write the following code in "proguard-rules.pro" file to enable the application working properly.

```
-ignore warnings
-keep class * { public private *; }
```

**D. Implementation of AES 256 bit Encryption Algorithm for Sensors Data Security**

To secure the sensors data we implement AES256 bit encryption algorithm with our modified vector size and key size. AES256 bit encryption secure the sensors data during transmission. The implementation steps of AES-256 encryption are:

- Step-1: Round keys are originated from the cipher key using Rijndael's schedule.
- Step-2: First Round. Each byte of the state is consolidated with the round key using bitwise XOR in add round key step. Each byte is substituted with another according to a lookup table in a non-linear replacement step called Sub Byte. In this step, each row of the state is shifted cyclically a certain number of steps. This called Shift Rows step. Mix Columns is a Mixing operation in Mix Columns step where the columns of the state, combining the four bytes in each column. Round Key is added in this step.
- Step-3: 2nd round to 13th Round

The following steps are repeated in this step: Sub Bytes. Shift Rows. Mix Columns. Add Round Key.

- Step-4: Final Round

In the final step, the following steps are repeated: Shift Rows. Mix Columns. Add Round Key.

On the server side, we implement AES256 decryption algorithm. Therefore, in the server the original data stored. Fig. 24, 25 and 26 shows the result of the implementation result of AES256 bit encryption.

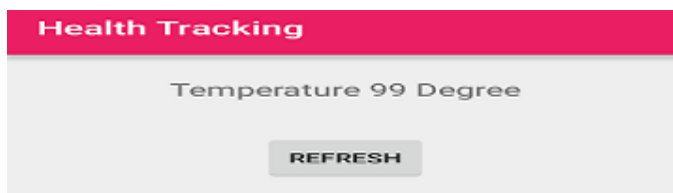


Fig. 24. Before AES 256 bit encryption.

296	566	T665	TEMPERATURE	SoYc3Aw6wWdLYmUo0i63PA==	1	2017-11-18 09:54:32	Update	delete
299	566	T667	TEMPERATURE	SoYc3Aw6wWdLYmUo0i63PA==	1	2017-11-18 10:03:16	Update	delete

Fig. 25. Encrypted data in transmission.

136	110	T220	TEMPERATURE	Temperature 99 Degree	1	2017-05-01 09:20:16	Update	delete
137	220	T440	BLOOD PRESSURE	Systemic: 120 Diastolic: 90	1	2017-05-01 09:22:34	Update	delete
138	330	T660	SPO	70.837204, 76.70445, 93.747734, 89.20929, 56.02 2762.92, 864095.63, 98386.77, 232796.76, 57673. 87, 123825.84, 14399.96, 94556.51, 959297.64, 79 313.66, 26843.95, 12451.52, 059944.84, 77905.54 83394.91, 04372.	1	2017-05-01 09:24:03	Update	delete

Fig. 26. Decrypted data in the server.

**E. Server Security**

All the sensors data stored in the remote staging server. Therefore, it is more important to secure the web server. Following the recommendation of Health Level 7, we implement the Secure Socket Layer or SSL-256 bit to secure our server. After implementation of SSL, our server URL link change from HTTP to https which means our server is secured. The implementation procedures of SSL is described below:

There are some prerequisites for SSL. These are certificates from the certificate authority (CA), registered domain name, web server (Apache HTTP, Nginx, HAProxy, or Varnish server).

There are three types of SSL server and these are single domain, wild card, and multiple domains.

1) *Generating private key:* For our system, we use single domain SSL certificate. To install SSL in our system, we buy SSL certificate from the certificate authority and get a.crt file bundles from them. After that, we generate a private key using OpenSSL which is called ehealthju.com.key and CSR file called ehealthju.com.csr. Therefore, we run the following command in the command line:

```
Openssl req -newkey rsa: 2048 -nodes -keyout ehealthju.com.key -out ehealthju.com.csr
```

After that, the following information is shown in the prompt. We should care about the common name field because common name field is the field that we put in our SSL certificate.

- Country Name (2 letter code) [AU]: BD
- State or Province Name (full name) Some-State: Dhaka
- Locality Name (eg, city) []: Dhaka
- Organization Name (eg, company) [Internet Widgits Pty Ltd]: ehealthju
- Organizational Unit Name (eg, section) []: Jahangir-nagar University
- Common Name (e.g. server FQDN or YOUR name) []:ehealthju.com
- Email Address []:uzzalcseju@gmail.com

This will generate a .key and .csr file. The .key file is the private key and should be kept secure. The .csr file will send to the CA to request SSL certificate. By using the generated private key and CSR file, we send to request your CA's to

provide the SSL certificate. They will validate our domain by sending an email.

2) *Certificate Installation:* We made a backup of our configuration file by copying it using these commands:

- `cd /etc/apache2/sites-available`
- `cp 000-default.conf 000-default.conf.orig`

Then open the file for editing:

- `sudo vi 000-default.conf`

We change the `<VirtualHost *:80>` entry to `<VirtualHost *:443>` Then add the `ServerName` directive. `ServerName ehealthju.com` Then add the following lines to specify certificate and key paths:

- `SSLEngine on`
- `SSLCertificateFile /home/user/ehealthju.com.crt`
- `SSLCertificateKeyFile /home/user/ehealthju.com.key`

Then we add the following code at the top of the file and then save the file.

- `VirtualHost *:80`
- `ServerName ehealthju.com`
- `Redirect permanent https://ehealthju.com/`
- `VirtualHost`

After that, we enable the Apache SSL module by running this command:

- `sudo a2enmod ssl`

Then we restart Apache to load the new configuration and enable TLS/SSL over HTTPS using the following command

- `sudo service apache2 restart`

After restart the Apache server, it converted to HTTPS instead of HTTP and our server is secured. Fig. 27 shows status of our server before implementation of SSL and Fig. 28 shows the status of our server after SSL implementation.

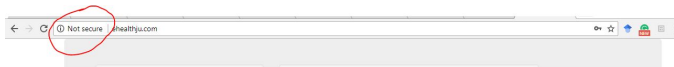


Fig. 27. Before SSL implementation.



Fig. 28. After implementation of SSL.

## VI. SECURITY ANALYSIS AND FUTURE WORK

In this paper, we discussed the security framework for our telemedicine system which is developed for the unprivileged people of Bangladesh. To secure this telemedicine system we find five section. Therefore, the entire security of our developed telemedicine system depends on the security improvement of this five section. We have applied the specific solution of these five security holes. First of all, a security technique must be needed to establish a secure connection between toolkit and mobile application. In this paper, we have introduced an authentication system for Bluetooth connection which improves the security. After that, we implemented an authentication system to use the mobile application, so that no unauthorized user use the system. It makes the mobile application more secure and prevents any type of data breaches.

Healthcare data is the more vulnerable because the medical record is more lucrative to hackers than any other data as like as credit card numbers [33]. Therefore, we improve the healthcare data security of our system by encrypting data with the AES-256 bit which the most advanced encryption algorithm. We decrypt the data in the server site using the same algorithm. Moreover, we have introduced registration for expert doctors, remote healthcare personnel as well as patients so that no unauthorized user use the system. This protects the system from any kind of security vulnerability. All these steps are taken following the instruction of Health Level 7 and FHIR.

Although there are many measures to secure our telemedicine system there are still some security threats. In our developed telemedicine system, we receive sensor data using mobile application. Before we send data to remote server, the remote user has the opportunity to observe data. Moreover, the end user gets the patients prescription before the patients get the prescription. In both cases, there are possibilities to breach data and prescription. There are some other problems as like as there are third parties like lab assistant who investigate the data. Sometimes there are different lab tests, therefore, lab assistant may change. In this case, there are also possibilities of data breaches.

Considering all these issues, in future, we improve our security model so that patients can easily get their prescription as well develop a module to authenticate the lab assistant. As well as our security model will play a vital role in the widespread use of developed telemedicine service so that a secured telemedicine service can be given to the remote poor people of our country at low cost.

## VII. CONCLUSION

In this paper, we implement a security framework to secure these principles and prevent the security breaches. The implemented security framework follows the recommendation of HL7 and FHIR, also consider the HIPPA and NEHTA recommendation for security and privacy of a health system. This paper also shows different advantages of our security framework. The implemented security framework is cost effective and efficient in performance. It can, therefore, be decided that the implemented security framework implements competent measures for real-time secure telemedicine data transmission.



#### ACKNOWLEDGMENT

The authors would like to thank all the faculty and employee of Computer Science and Engineering Department and ICT Division, Ministry of post, telecommunication, and information technology, Bangladesh for the full financial support granted to this research.

#### REFERENCES

- [1] U. K. P. Mohammad Zahidur Rahman, Israt Jahan, "Ardunio based telemedicine system for bangladesh," *Jahangirnagar University Journal of Science, Volume 40, No: 2*, 2017.
- [2] M. F. F. Khan and K. Sakamura, "Security in healthcare informatics: design and implementation of a robust authentication and a hybrid access control mechanism," in *Communications, Computers and Applications (MIC-CCA), 2012 Mosharaka International Conference on*. IEEE, 2012, pp. 159–164.
- [3] V. Liu, W. Caelli, L. May, and T. Sahama, "Privacy and security in open and trusted health information systems," in *Proceedings of the Third Australasian Workshop on Health Informatics and Knowledge Management-Volume 97*. Australian Computer Society, Inc., 2009, pp. 25–30.
- [4] P. T. Akkasaligar and S. Biradar, "Secure medical image encryption based on intensity level using chao's theory and dna cryptography," in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICIC)*, Dec 2016, pp. 1–6.
- [5] M. J. Chang, J. K. Jung, M. W. Park, and T. M. Chung, "Strategy to reinforce security in telemedicine services," in *2015 17th International Conference on Advanced Communication Technology (ICACT)*, July 2015, pp. 170–175.
- [6] I. Chiuchisan, D. G. Balan, O. Geman, I. Chiuchisan, and I. Gordin, "A security approach for health care information systems," in *2017 E-Health and Bioengineering Conference (EHB)*, June 2017, pp. 721–724.
- [7] C. Fu, Y. Lin, H. y. Jiang, and H. f. Ma, "Medical image protection using hyperchaos-based encryption," in *2015 9th International Symposium on Medical Information and Communication Technology (ISMICT)*, March 2015, pp. 103–107.
- [8] B. Halder and S. Mitra, "Modified watermarked ecg signals by using adaptive normalization factor," in *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*, July 2015, pp. 434–439.
- [9] C. Han, L. Sun, and Q. Du, "Securing image transmissions via fountain coding and adaptive resource allocation," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.
- [10] R. M. Seepers, C. Strydis, I. Sourdis, and C. I. D. Zeeuw, "Enhancing heart-beat-based security for mhealth applications," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 1, pp. 254–262, Jan 2017.
- [11] U. K. Prodhan, M. Z. Rahman, and I. Jahan, "Design and implementation of an advanced telemedicine model for the rural people of bangladesh," *Technology and Health Care*, no. Preprint, pp. 1–6, 2018.
- [12] A. Sudarsono, P. Kristalina, M. U. H. A. Rasyid, and R. Hermawan, "An implementation of secure data sensor transmission in wireless sensor network for monitoring environmental health," in *2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, Oct 2015, pp. 93–98.
- [13] A. Ibrahim, B. Mahmood, and M. Singhal, "A secure framework for medical information exchange (mi-x) between healthcare providers," in *Healthcare Informatics (ICHI), 2016 IEEE International Conference on*. IEEE, 2016, pp. 234–243.
- [14] M. Puppala, T. He, X. Yu, S. Chen, R. Ogunti, and S. T. Wong, "Data security and privacy management in healthcare applications and clinical data warehouse environment," in *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*. IEEE, 2016, pp. 5–8.
- [15] T. W. Tseng, C. Y. Yang, and C. T. Liu, "Designing privacy information protection of electronic medical records," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec 2016, pp. 75–80.
- [16] W. D. Yu, L. Davuluri, M. Radhakrishnan, and M. Runiassy, "A security oriented design (sod) framework for ehealth systems," in *2014 IEEE 38th International Computer Software and Applications Conference Workshops*, July 2014, pp. 122–127.
- [17] K. Zeb, K. Saleem, J. Al Muhtadi, and C. Thuemmler, "U-prove based security framework for mobile device authentication in health networks," in *e-Health Networking, Applications and Services (Healthcom), 2016 IEEE 18th International Conference on*. IEEE, 2016, pp. 1–6.
- [18] N. Jeyanthi, R. Thandeeswaran, and H. Mcheick, "Sct: Secured cloud based telemedicine," in *The 2014 International Symposium on Networks, Computers and Communications*, June 2014, pp. 1–4.
- [19] T. Vivas, A. Zambrano, and M. Huerta, "Mechanisms of security based on digital certificates applied in a telemedicine network," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2008, pp. 1817–1820.
- [20] F. Rezaeibagha and Y. Mu, "Practical and secure telemedicine systems for user mobility," *Journal of biomedical informatics*, 2017.
- [21] J. Singh and A. K. Patel, "An effective telemedicine security using wavelet based watermarking," in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICIC)*, Dec 2016, pp. 1–6.
- [22] A. K. Maji, A. Mukhoty, A. K. Majumdar, J. Mukhopadhyay, S. Sural, S. Paul, and B. Majumdar, "Security analysis and implementation of web-based telemedicine services with a four-tier architecture," in *2008 Second International Conference on Pervasive Computing Technologies for Healthcare*, Jan 2008, pp. 46–54.
- [23] K. C. Nelson and M. A. Noronha, "Facilitating ownership of access control lists by users or groups," Jul. 4 2017, uS Patent 9,697,373.
- [24] S. Li and L. Da Xu, *Securing the Internet of Things*. Syngress, 2017.
- [25] "Mls introduction — cryptosmith," <https://cryptosmith.com/mls/intro/>.
- [26] O. N. Ely, "Secure computing system," Sep. 19 2017, uS Patent 9,767,297.
- [27] L. M. Boyle and D. M. Mack, *HIPAA: a guide to health care privacy and security law*. Wolters Kluwer, 2017.
- [28] K. K. Htat, P. A. Williams, and V. McCauley, "Security of eprescriptions: data in transit comparison using existing and mobile device services," in *Proceedings of the Australasian Computer Science Week Multiconference*. ACM, 2017, p. 56.
- [29] L. F. Martín and G. Solidario, "Gnu health: A free/libre community-based health information system," in *OpenSym (Companion)*, 2016, pp. 11–1.
- [30] S. P. Dwivedi, "Message digestion," 2017.
- [31] D. Kraus and M. Welschenbach, "Advanced encryption standard," 2016.
- [32] Z. Shan, "A study on altering postgresql from multi-processes structure to multi-threads structure," *arXiv preprint arXiv:1609.09062*, 2016.
- [33] S. I. Khan and A. S. M. Latiful Hoque, "Digital health data: A comprehensive review of privacy and security risks and some recommendations," *Computer Science Journal of Moldova*, vol. 24, no. 2, 2016.

# Nonlinear Model Predictive Control for pH Neutralization Process based on SOMA Algorithm

Hajer Degachi

Tunis El Manar University  
National Engineering school of Tunis  
LR11ES20, Analysis, Conception  
and Control of Systems Laboratory

Wassila Chagra

Tunis El Manar University  
El Manar Preparatory Institute  
for Engineering Studies  
LR11ES20, Analysis, Conception  
and Control of Systems Laboratory

Moufida Ksouri

Tunis El Manar University  
National Engineering school of Tunis  
LR11ES20, Analysis, Conception  
and Control of Systems Laboratory

**Abstract**—In this work, the pH neutralization process is described by a neural network Wiener (NNW) model. A nonlinear Model Predictive Control (NMPC) is established for the considered process. The main difficulty that can be encountered in NMPC is solving the optimization problem at each sampling time to determine an optimal solution in finite time. The aim of this paper is the use of global optimization method to solve the NMPC minimization problem. Therefore, we propose in this work, to use the Self Organizing Migrating Algorithm (SOMA) to solve the presented optimization problem. This algorithm proves its efficiency to determine the optimal control sequence with a lower computation time. Then the NMPC is compared to adaptive PID controller, where we propose to use the SOMA algorithm to formulate the PID in order to determine the optimal parameters of the PID. The performances of the two controllers based on the SOMA algorithm are tested on the pH neutralization process.

**Keywords**—Nonlinear model predictive control; optimization; SOMA algorithm; adaptive PID; pH neutralization process

## I. INTRODUCTION

The pH neutralization process is characterized by a nonlinear behavior. The high nonlinear characteristic of this process makes the control of the pH a hard task. Since the necessity of maintaining the pH in a specific range value, many control strategies have been proposed. For this purpose, [1]–[3] designate a PID controller to control the pH value. An adaptive controller is developed in [4]. [5] proposed a model algorithmic control strategy. Other works [6]–[8] developed the Model predictive control (MPC) which is a more advanced control strategy. The MPC control strategy presents the major advantage to efficiently handle nonlinearity and constraints imposed on system input and output [10]. Indeed, for real processes, the control values are usually delimited by an upper and lower bound that should be respected.

The MPC is essentially based the choice of a suitable model and adequate optimization method to solve the minimization problem. Many structures was used to describe the nonlinear behavior of the pH neutralization process: NARX model [11], Fuzzy neural network model [12], Neural networks [13], Wiener model [8], [14]–[17].

Respecting the nonlinear nature of the model, the resulting NMPC minimization problem is nonlinear and nonconvex. The nonconvexity will complicate the implementation of the

NMPC. Added to that, solving a nonlinear optimization problem is a hard task in terms of computation time and burden. To overcome these difficulties a variety of solutions were proposed in literature to avoid solving a nonlinear optimization problem and ensure global convergence at each sampling time. In [18], [19], the minimization problem is converted into a linear one in the case of a simple polynomial description of the nonlinear block. In this case, the nonlinearity is removed by considering the inverse of the polynomial function. As a result, the input control will be linearly expressed in the predicted output leading to a quadratic optimization problem. Also, the nonlinear optimization problem is formulated in [8], [15], [16] as a linear problem by performing a linearization of the nonlinear model. To avoid solving the nonlinear optimization problem and reduce the implementation complexity of the NMPC, [20], proposed to describe the nonlinear process by a set of uncertain linear models instead of one nonlinear model. We can note that all these works avoid to solve the nonconvex optimization problem regarding the difficulty of implementation and the high computation burden necessary at each sampling time.

To obtain good control accuracy the NMPC optimization problem must be solved as nonlinear [21]. However, global optimization methods are generally not recommended since they are not able to ensure the real time feasibility of the NMPC. In fact, the sampling period of the process under consideration must be respected at each iteration when solving the NMPC problem. This constraint is difficult to be satisfied using global optimization method due to their slow convergence. Therefore, the main aim of this work is the use of an efficient algorithm to solve the NMPC optimization problem and ensure the real-time feasibility of the control algorithm. For this, we propose to use the Self Organizing Migrating algorithm (SOMA) in this work to solve the optimization problem. The SOMA is an evolutionary algorithm such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Artificial Bee Colony algorithm (ABC) and so on. The SOMA algorithm was successfully applied to solve a variety of engineering problems. The most interesting are: control issues [22], antenna design [23], system identification [24], [25], Aircraft wing design and Synthesis of robot control program [26]. This method presents the significant advantage of high convergence speed.

In This work, we will be interested to represent the pH neutralization process by a Wiener model. Wiener model

belongs to block oriented models that are defined by a dynamic linear block followed by a nonlinear static one. Due to the specific structure of the pH neutralization process, the Wiener model appears able to well reproduce the behavior of the process. Many works used the Wiener model to describe the nonlinear dynamic of the process.

The different representations vary in the way that linear and nonlinear blocks are described.

Many structures are used to describe the steady state block : Polynomial, [14], [27], support vector machine [16], neural network [8], [15], [19], cubic spline [18]. In this work, the nonlinear block is represented by a feed forward Neural Network (NN).

In this paper, the NMPC is integrated with SOMA algorithm to solve the optimization problem. We prove, in this work, the ability of the SOMA algorithm to ensure good control performances with a low computation time despite the large prediction and control horizons. In the sequel, The NMPC strategy integrated with SOMA algorithm is compared to adaptive PID controller. We propose, in this work, to adjust PID parameters using SOMA algorithm.

This paper is organized as follows: Section 2 describes the Wiener model. The identification of the considered process is detailed in Section 3. NMPC based on the Wiener model is presented in Section 4. The SOMA algorithm used to solve the minimization problem is described in Section 5. Simulation results are given in Section 6.

## II. WIENER MODEL

Wiener model belongs to block oriented models. It is described by a linear dynamic block followed by a nonlinear static one as shown in Fig. 1.



Fig. 1. Wiener model.

In this work, the linear dynamic block is described by a simple autoregressive model defined as:

$$A(q^{-1})s(k) = B(q^{-1})u(k) \quad (1)$$

Where polynomials A and B are given by:

$$A(q^{-1}) = 1 + a_1q^{-1} + a_2q^{-2} + \dots + a_{n_a}q^{-n_a} \quad (2)$$

$$B(q^{-1}) = b_1q^{-1} + b_2q^{-2} + \dots + b_{n_b}q^{-n_b} \quad (3)$$

$n_a$  and  $n_b$  are respectively the orders of the two polynomials A and B.

The nonlinear block is defined as:

$$y(k) = f(s(k)) \quad (4)$$

$f(\cdot)$  is a nonlinear function.

Different forms are used to describe the nonlinear block for Wiener model starting from the simple polynomial form

[19], [28] to more complex description Neural network [19], [15], support vector machine [16]. In this work the nonlinear static block of the Wiener model is given by a feed-forward neural network as adopted in [15].

## III. IDENTIFICATION OF THE WIENER MODEL

The aim of this paragraph is the determination of the parameters  $a_i$  and  $b_i$  of the linear block as well as the weights  $w_i$  of the network that present the parameters of the nonlinear block.

The structure of the Wiener model is depicted in Fig. 2.

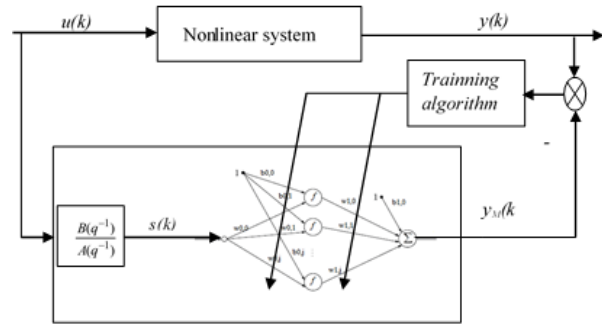


Fig. 2. Identification of the Wiener model.

### A. Identification of the Linear Block

Firstly, a small input signal is applied to the system to ensure linear perturbation of the nonlinear system [29]. Then, the recursive least square algorithm is firstly used to identify the parameters of linear dynamic block.

The output of the linear block can be expressed as:

$$y(k) = \psi^T(t)\theta(t-1) \quad (5)$$

$\psi$  is the data vector defined as:

$$\psi^T(t) = [-y(k-1), \dots, -y(k-n_a) u(k-1), \dots, u(k-n_b)] \quad (6)$$

$\theta$  is the parameter vector:

$$\theta^T = [a_1 a_2, \dots, a_{n_a} b_1 b_2, \dots, b_{n_b}] \quad (7)$$

The update of the parameter vector is ensured by the following equation:

$$\theta(t) = \theta(t-1) + P(t)\psi(t)\varepsilon(t) \quad (8)$$

P is weighting matrix given by:

$$P(t) = P(t-1) - \frac{P(t-1)\psi(t)\psi^T(t)P(t-1)}{1 + \psi^T(t)P(t-1)\psi(t)} \quad (9)$$

The initial parameter vector is  $\theta(0) = 0$ .

Once the parameters of the linear block are determined, the back-propagation algorithm is applied to train the feed-forward neural network such as in [29].

### B. Identification of the Nonlinear Block

The structure of the nonlinear block is illustrated in Fig. 3. The output of the nonlinear block can be computed from Fig. 3 as:

$$y_M(k) = b_{1,0} + \sum_{i=1}^j w_{1,i} f(e(k)) \quad (10)$$

where  $j$  is the number of hidden nodes,  $f$  is a nonlinear activation function taken as the hyperbolic tangent function and  $e(k)$  is the output of the hidden layer defined as:

$$e_i(k) = b_{0,i} + w_{0,i} s(k) \quad (11)$$

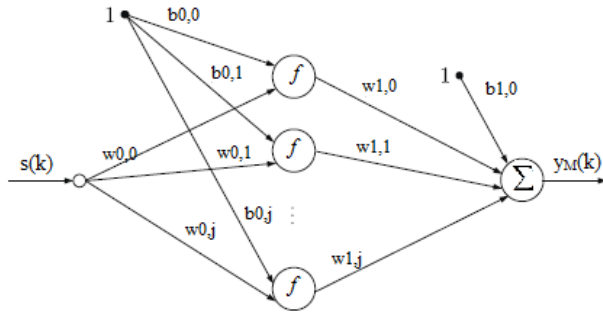


Fig. 3. Neural Network representation of the nonlinear Wiener block.

The weights  $w_{i,j}$  of the neural network are updated by minimizing the following criterion:

$$J_{iden}(k) = \sum_{i=1}^N (y(k) - y_M(k))^2 \quad (12)$$

where  $y(k)$  is the process output and  $y_M(k)$  is the model output defined as:

$$y_M(k) = b_{1,0} + \sum_{i=1}^j w_{1,i} f \left( b_{0,i} + w_{0,i} \left( \sum_{i=1}^{n_a} -a_i s(k-i) + \sum_{i=1}^{n_b} b_i u(k-i) \right) \right) \quad (13)$$

Applying the back-propagation training algorithm the optimization is carried out in order to minimize the criterion (12) with respect to the weights  $w_{i,j}$  of the network.

$$\frac{\partial J}{\partial w} = - (y(k) - y_M(k)) \frac{\partial y_M(k)}{\partial w} \quad (14)$$

The update equation of the different weights is defined as:

$$w(k+1) = w(k) + \Delta w(k) \quad (15)$$

where  $\Delta w(k)$  is defined as:  $\Delta w(k) = -\mu \frac{\partial J}{\partial w}$

$$w(k+1) = w(k) + \mu (y(k) - y_M(k)) \frac{\partial y_M(k)}{\partial w} \quad (16)$$

$\mu$  is the learning coefficient.

The Neural Network Wiener (NNW) model is used in this work to compute the  $j$ -step ahead predictions. Future prediction are used to define the cost function of the NMPC strategy.

### IV. MODEL PREDICTIVE CONTROL DESIGN FOR WIENER MODEL

The basic idea of the MPC strategy is the use of the process model to compute the  $j$ -step ahead prediction over a prediction horizon  $N_p$ . At each sampling time, a control sequence  $U = [u(k), u(k+1), \dots, u(k+N_u-1)]^T$  is computed, where  $N_u$  is the control horizon, by minimizing the cost function defined as the difference between the predicted output  $\hat{y}(k+i|k)$  and the future set-point  $y^{sp}(k+i)$  defined by:

$$J(k) = \sum_{i=1}^{N_p} (y^{sp}(k+i|k) - \hat{y}(k+i|k))^2 + \lambda \sum_{i=0}^{N_u-1} \Delta u(k+i|k)^2 \quad (17)$$

subject to

$$\Delta u^{\min} \leq \Delta u(k+i|k) \leq \Delta u^{\max}$$

$$u^{\min} \leq u(k+i|k) \leq u^{\max}$$

$\lambda$  represents a positive weighting coefficient,  $\Delta u(k+i|k)$  is defined as  $\Delta u(k+i|k) = u(k+i|k) - u(k+i-1|k)$ .

Based on the receding control horizon, only the first element of the control sequence  $U$  is applied to the system. Then, the whole procedure will be repeated at the next sampling time and the prediction horizon will be shifted one step forward.

In order to get efficient control results, predictions must be accurately computed. It is so important to take into account the effects of system and model mismatch coming from modeling errors and unmeasured disturbance, a correction term as in the dynamical matrix control, is added to model output defined as [30]:

$$d(k) = y(k) - y_M(k+i|k) \quad (18)$$

where  $y_M(k+i|k)$  is the model output and  $y(k)$  is the process output. The correction term is constant over the prediction horizon. Therefore, the model expression that will be used to compute predictions is defined as:

$$\hat{y}(k+i|k) = y_M(k+i|k) + d(k) \quad (19)$$

Due to the nonlinear nature of the NNW model the resulting optimization problem will be nonlinear and nonconvex. The convergence of this optimization problem is not guaranteed and the algorithm may be trapped in a local minimum that will lead necessarily to suboptimal control performances. Solving a nonlinear optimization problem is a high time demanding task. Added to that, real process requires generally a large prediction and control horizons that will raise the computation time. Moreover, the sampling period of the process presents an additional constraint that should be respected when solving the NMPC optimization problem.

Therefore, a fast convergence speed algorithm must be used to ensure the global convergence and the real-time feasibility of the control algorithm. Deterministic optimization methods are used in [9], [31] to solve the NMPC optimization problem. These methods are high time consuming also they can be

only applied to systems requiring small prediction and control horizons.

Many research papers propose to use stochastic optimization methods to solve the minimization problem. The Artificial Bee Colony (ABC) algorithm is combined with the NMPC in [32] to solve the minimization problem. This algorithm proves its simplicity of implementation and reduced computational complexity. The genetic algorithm (GE) is used in [33], [34] to determine the optimal control sequence. The particle swarm optimization (PSO) algorithm is integrated in [35] with NMPC to solve the resulting optimization problem. In [36], the neural network is used to determine the solution of the minimization problem. In [37], the Nelder Mead algorithm was applied which leads to global solution by using far initialization. The simulations gave optimal results with least computation time for SISO and MIMO models. However, it remains a local optimization method.

We propose in this work to use the SOMA algorithm to solve the presented optimization problem. SOMA presents an effective, robust and simple global optimization method.

## V. SELF ORGANIZING MIGRATING ALGORITHM

SOMA is a stochastic optimization method proposed first by Zelinka [25]. SOMA is based on the social group of individual not on the philosophy of evolution. The classification of SOMA as evolutionary algorithm is explicated by the fact that the obtained results after a migration loop is the same as the result of one generation of evolutionary algorithm [26]. The principle of SOMA algorithm can be summarized as a migration loops during which the position of each individual is enhanced in order to reach the leader position (individual with the best fitness).

Each individual will be randomly initialized in the search space described by the upper and the lower bounds of the variables. At each migration loop, individuals are evaluated, the one that has the less fitness will be the leader, the rest individuals will cross a trajectory (*pathlength*) with step *t* in the direction of the leader.

Similar to other evolutionary algorithms, the operation of SOMA is ensured by some control parameters. The recommended ranges of these parameters are fixed based on great number of empirical tests and they are defined and given by [24], [26]:

- **Dim:** It defines the problem dimension (number of decision variable).
- **Population size (*Popsiz*e):** It defines the number of individuals in population, *recommended value*  $\geq 10$ .
- **Migrations:** It defines the maximum number of iterations (migration loops). It is the stopping criterion in SOMA recommended range [10, up to user].
- **Pathlength:** It fixes how far the individual will stop its movement from the leader. If the *pathlength*=1 the individual will stop at the leader position, if *pathlength*=2 the individual will surpass the leader position by the same distance from the initial position, recommended value 3.

- **step:** It defines the step that uses individuals to cross the path.
- **PRT:** The *PRT* is used to determine the *PRTvec*. Individuals are allowed to change their position based on the *PRT*, recommended value 0.4.

SOMA is a population based algorithm. The initial population *P* is generated randomly in the search space defined by the lower  $x_{min}$  and upper  $x_{max}$  bound of the manipulated variables. So, *P* is defined as:  $P = \{X_1, X_2, \dots, X_{Popsiz$ e}\}. The *i*th individual  $X_i$  is defined by  $X_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,Dim}\}$ ,  $i = 1, \dots, Popsiz$ e,  $j = 1, \dots, Dim$ .

$$x_{i,j} = x_{min,j} + (x_{max,j} - x_{min,j}).rand \quad (20)$$

Like other evolutionary algorithm, SOMA performs a set of stochastic evolutionary operators. These operators are defined by mutation and crossover.

- **Mutation:** This operator defines in evolutionary algorithm the diversity in the population. In SOMA mutation is applied differently. It is performed by the *PRT* vector noted *PRTvec*. The *PRTvec* aims to perturb the path of individuals randomly in order to ensure diversification among them [38]. The perturbation in SOMA algorithm presents the mutation phase in the GA algorithm. Two values can be affected to this vector: 0 or 1 based on the SOMA *PRT* control parameter. Only individuals with *PRTvec* equal to 1 are allowed to change their positions. The *PRTvec* is created as follows:

$$\begin{aligned} & \text{if } rand < PRT \\ & \quad PRTvec_j = 1; \\ & \text{else } PRTvec_j = 0; \\ & \text{end} \end{aligned} \quad (21)$$

- **Crossover:** Crossover operator means the creation of new individual during the search. Since in SOMA algorithm no new individual is generated, the crossover is defined in this case by generating a new best position of individual across the search space to reach the leader. At each migration loop individuals explore a set of positions when mapping the path, memorize the best found one and move to this position at the end of the path. At the next migration loop, individuals start from this position. The movement of individuals to reach the position of the leader is given by the following equation:

$$x_{i,j}^{k+1} = x_{i,j}^k + (x_{L,j}^k - x_{i,j}^k).t.PRTvec_j \quad (22)$$

where  $x_{i,j}^{k+1}$  is the new individual position.,  $x_{i,j}^k$  is the position of individual at iteration *k*,  $x_{L,j}^k$  is the leader position and  $t \in [0 : step : pathlength]$

The SOMA algorithm can be summarized as follows:

- 01** Choose the control parameters :PRT,step,pathlengthandmigration
- 02** Evaluate the initial population P using equation (20),
- 03** Evaluate individuals of the population
- 04** Sort individuals and select the leader
- 05** iteration = 1;

```

while (iteration < migrations)
for i = 1 : Popsiz
Generate PRTvec using equation(21)
Move each individual toward the leader using
equation (22)
Evaluate the individual at the new position

    new fitness < fitness(individual (i))
Move individual(i) to the new best position.
end if
end for
sort individuals and select the new leader
iteration = iteration + 1
end while
    
```

## VI. SIMULATION RESULTS

The considered system is composed of the base stream  $q_1$ , the acid stream  $q_2$  and the buffer stream  $q_3$  that are mixed in continuous stirred tank reactor. The dynamic model of the reactor is derived from the conservation equation and equilibrium relations [4]. The dynamic of the pH process is given by two differential equations and a nonlinear one given by (23) and (24 ).

$$\begin{aligned}
 \dot{W}_a(t) &= \frac{q_1(t)}{V}(W_{a1} - W_a(t)) + \frac{q_2(t)}{V}(W_{a2} - W_a(t)) \\
 &+ \frac{q_3(t)}{V}(W_{a3} - W_a(t)) \\
 \dot{W}_b(t) &= \frac{q_1(t)}{V}(W_{b1} - W_b(t)) + \frac{q_2(t)}{V}(W_{b2} - W_b(t)) \\
 &+ \frac{q_3(t)}{V}(W_{b3} - W_b(t))
 \end{aligned} \tag{23}$$

$$\begin{aligned}
 W_a(t) + 10^{pH(t)-14} - 10^{-pH(t)} + \\
 W_b(t) \frac{1 + 2 \times 10^{pH(t)-k_2}}{1 + 10^{k_1-pH(t)} + 10^{pH(t)-k_2}} = 0
 \end{aligned} \tag{24}$$

where  $W_a$  and  $W_b$  represent the charge balance coefficients. The base stream  $q_1$  is manipulated to control the pH value, the acid stream  $q_2$ , the buffer stream  $q_3$  are maintained constant.

the different parameters of the reactor are listed in Table I.

In this section, we will consider the pH neutralization process.

TABLE I. PARAMETERS OF THE pH PROCESS

Parameter	value
$W_{a1}$	-3.05e-3
$W_{a2}$	-3e-2
$W_{a3}$	3e-3
$W_{b1}$	5e-5
$W_{b2}$	3e-2
$W_{b3}$	0
$q_2$	0.55
$q_3$	16.60
$V$	2900
$k_1$	6.35
$k_2$	10.25

### A. Identification of the pH Neutralization Process

The pH neutralization process will be described by a NNW model, where the linear block is described by an auto-regressive model and the nonlinear block is given by a multi-layer feed-forward neural network with one hidden layer. The plant is excited using two different input signals to get the identification and the validation data as shown respectively in Fig. 4 and 5. The bounds on the input are 0 and 30.

A feed-forward neural network with one hidden layer and 3 nodes appears sufficient to describe the nonlinear block of the Wiener model [15].

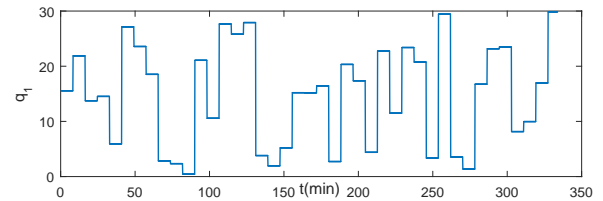


Fig. 4. Identification signal.

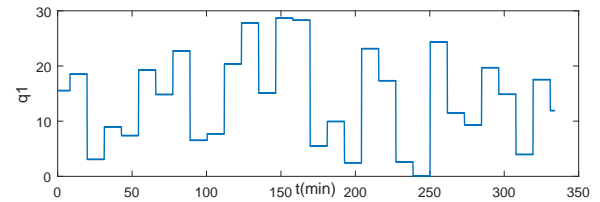


Fig. 5. Validation signal.

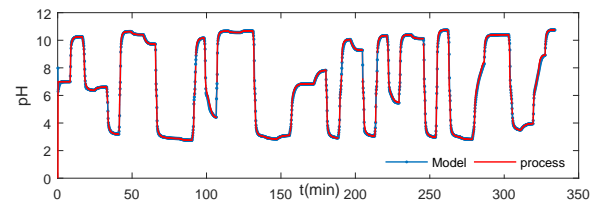


Fig. 6. Identification of the NNW model.

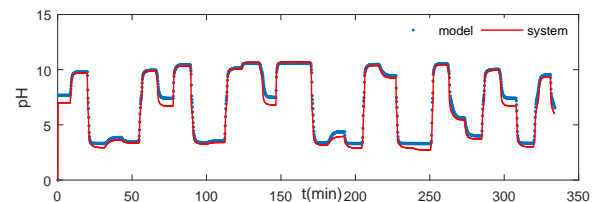


Fig. 7. Validation of the NNW model.

Fig. 6 and 7 illustrate the identification and the validation of the system. We can conclude that the neural Wiener model can reproduce the dynamic of the pH process with sufficient accuracy. This is well confirmed by the validation error depicted in Fig. 8 defined as the difference between the system and the model output  $y(k)$  and  $y_M(k)$ , respectively.

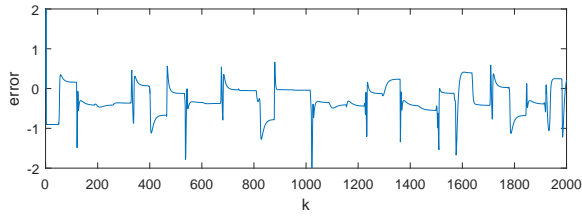


Fig. 8. Validation error.

**B. NMPC Control of the pH Neutralization Process**

The NMPC strategy aim to maintain the pH value at a desired value. For this the neural Wiener model is used to compute the j-step-ahead predictions for the NMPC. The parameters of the NMPC are fixed as  $N_p = 10$ ,  $N_u = 3$ ,  $\lambda = 0.08$ ,  $q^{\min} = 0$  and  $q^{\max} = 30$ .

The recommended values of the parameters of the SOMA algorithm are fixed as:  $pathlength = 3$ ,  $step = 0.31$ ,  $PRT = 0.4$ ,  $popsiz = 10$ , the number of migration loops is 30.

We can note that the SOMA algorithm proves its efficiency to ensure good output tracking as depicted in Fig. 9. The performance offered by the SOMA algorithm outperforms those offered by the GBM method. This can be proved in terms of less overshoot and best control quality as illustrated in Fig. 9.

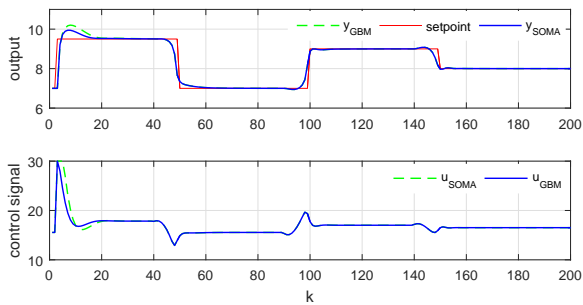


Fig. 9. Output tracking using the SOMA algorithm and the GBM.

In order to test the efficiency of the SOMA algorithm in presence of disturbance a constant  $v$  is added to the system output as:

$$v(k)=0.8 \quad 110 \leq k \leq 130$$

$$v(k)=0.6 \quad 160 \leq k \leq 180$$

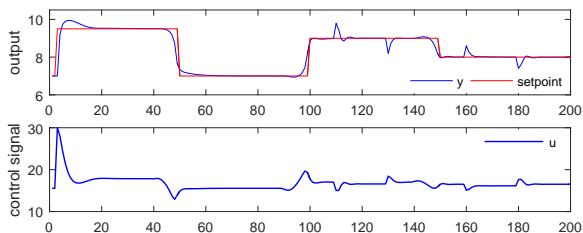


Fig. 10. Output tracking in the presence of disturbance.

We can note from Fig. 10 the good ability of the SOMA

algorithm to reject disturbance. This latter is rejected within 10 iterations that present a short time.

The NMPC strategy is compared to an adaptive PID controller. The Sum of square Error (SAE) defined by equation (25) is minimized in order to determine the proportional, integral and derivative coefficients of the PID controller:  $k_p$ ,  $k_i$  and  $k_d$ , respectively:

$$SAE(k) = \sum_{i=1}^k |e(k)| \quad (25)$$

Respecting the high nonlinear considered process, the SOMA algorithm is adopted to derive the coefficients of the adaptive PID controller at each iteration.

The control parameters of the SOMA algorithm are fixed for the PID controller as those of the NMPC.

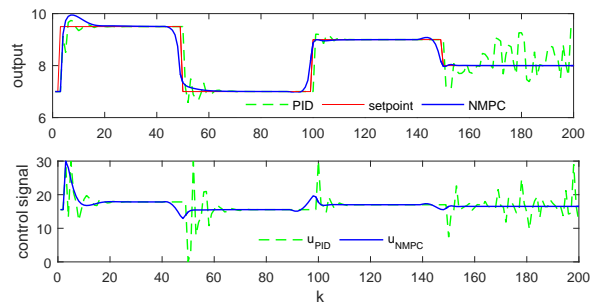


Fig. 11. Output tracking for NMPC and PID.

Simulation results for PID and NMPC are given in Fig. 11. The results show that both controllers ensure good output tracking in steady state for the first and second setpoint change. However, we can remark that the PID exhibits more overshoots as well as the deterioration of the control quality and consequently the output tracking in the last output change. This proves the superiority of the NMPC strategy to ensure good performances compared to PID.

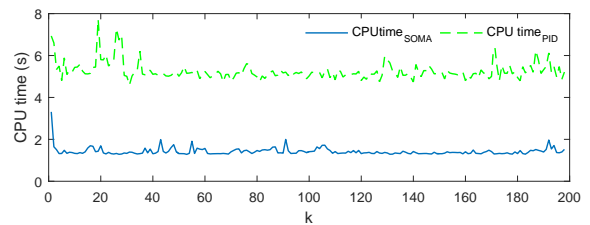


Fig. 12. Computation time of the NMPC and PID

Computation time present an important performance index in control problem. Fig. 12 shows the computation time for both NMPC and PID. The present results confirm the good ability of NMPC to give better performances with the low computation time. In fact, the determination of optimal PID parameters on line at each sampling time using the SOMA algorithm requires more migration loops this will directly affect the computation time of the implementation of the PID controller.

## VII. CONCLUSION

In this paper, a nonlinear model predictive control is combined with the SOMA algorithm to solve the NMPC optimization problem. The control performances of the SOMA algorithm are tested on a high nonlinear process. Control results prove the efficiency of this algorithm to ensure good output tracking and control accuracy with a low computation time. NMPC based on SOMA algorithm is compared to adaptive PID controller. We can conclude from simulation results that the NMPC outperforms the PID in terms of less overshoot and computation time. NMPC offers the best control results and the less computation time compared to PID. The difficulty of the determination of PID parameters for high nonlinear process limits the performance of this controller.

## REFERENCES

- [1] Rose, T.P., Devadhas, G.G., and Rex, S.R. "Invention of a suitable controller for a non linear Chemical process". In International Conference on Control, Instrumentation, Communication and Computational Technologies, pp. 1462-1467, 2014.
- [2] Ram, S.S., Kumar, D.D., Meenakshipriya, B., and Sundaravadivu, K. "Designing and comparison of controllers based on optimization techniques for pH neutralization process". In International Conference on Information Communication and Embedded Systems, pp. 1-5, 2016.
- [3] Puchalski, B., Rutkowski, T., Tarnawski, J., and Duzinkiewicz, K. "Comparison of tuning procedures based on evolutionary algorithm for multi-region fuzzy-logic PID controller for non-linear plant". In Methods and Models in 20th International Conference on Automation and Robotics (MMAR), pp. 897-902, 2015.
- [4] Henson, M.A., and Seborg, D.E. "Adaptive nonlinear control of a pH neutralization process". IEEE transactions on control systems technology, vol.2, pp. 169-182, 1994.
- [5] Zhiyun, Z.O.U., Meng, Y.U., Zhizhen, W.A.N.G., Xinghong, L.I.U., Yuqing, G.U.O., ZHANG, F., and Ning, G.U.O. "Nonlinear model algorithmic control of a pH neutralization process". Chinese Journal of Chemical Engineering, vol. 21, pp. 395-400, 2013.
- [6] Oblak, S., and Skrjanc, I. "Continuous-time Wiener-model predictive control of a pH process based on a PWL approximation". Chemical Engineering Science, vol. 65, pp. 1720-1728, 2010.
- [7] Waller, J.B., and Toivonen, H.T. "A neuro-fuzzy model predictive controller applied to a ph-neutralization process". IFAC Proceedings Volumes, vol. 35, pp. 495-500, 2002.
- [8] Lawrynczuk, M. "Computationally efficient nonlinear predictive control based on neural Wiener models". Neurocomputing, vol. 74, pp. 401-417, 2010.
- [9] Degachi, H., Chagra, W., Ksouri, M. (2015, March). Global optimization method for model predictive control based on Wiener model. In 12th International Multi-Conference on Systems, Signals & Devices (SSD), pp. 1-6, 2015.
- [10] Mayne, D.Q., Rawlings, J.B., Rao, C.V., Scokaert, P.O. "Constrained model predictive control: Stability and optimality". Automatica, vol. 36, pp. 789-814, 2000.
- [11] Bello, O., Hamam, Y., and Djouani, K. "Nonlinear model predictive control of a coagulation chemical dosing unit for water treatment plants". IFAC Proceedings Volumes, vol. 47, pp. 370-376, 2014.
- [12] Xue, A., Peng, D., and Guo, Y. "Modeling of pH neutralization process using fuzzy recurrent neural network and DNA based NSGA-II". Journal of the Franklin Institute, vol. 351, pp. 3847-3864, 2014.
- [13] Tharakan, L.G., Benny, A., Jaffar, N.E., and Jaleel, J.A. "Neural network based pH control of a weak acid Strong base system". In International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing, pp. 674-679, 2013.
- [14] Mahmoodi, S., Poshtan, J., Jahed-Motlagh, M.R., and Montazeri, A. "Nonlinear model predictive control of a pH neutralization process based on Wiener Laguerre model". Chemical Engineering Journal, vol. 146, pp. 328-337, 2009.
- [15] Lawrynczuk, M. "Practical nonlinear predictive control algorithms for neural Wiener models". Journal of Process Control, vol. 2, pp. 696-714, 2013.
- [16] Lawrynczuk, M. "Modelling and predictive control of a neutralisation reactor using sparse support vector machine Wiener models". Neurocomputing, vol. 205, pp. 311-328, 2016.
- [17] Gomez, J.C., Jutan, A., and Baeyens, E. "Wiener model identification and predictive control of a pH neutralisation process". IEE Proceedings-Control Theory and Applications, vol. 151, pp. 329-338, 2004.
- [18] Norquay, S.J., Palazoglu, A., and Romagnoli, J.A. "Application of Wiener model predictive control (WMPC) to a pH neutralization experiment". IEEE Transactions on Control Systems Technology, vol. 7, pp. 437-445, 1999.
- [19] Peng, J., Dubay, R., Hernandez, J.M., Abu-Ayyad, M.M. "A Wiener neural network-based identification and adaptive generalized predictive control for nonlinear SISO systems". Industrial & Engineering Chemistry Research, vol. 5, pp. 7388-739, 2011.
- [20] Khani, F., and Haeri, M. "Robust model predictive control of nonlinear processes represented by Wiener or Hammerstein models". Chemical Engineering Science, vol. 129, pp. 223-231, 2015.
- [21] Allgower, F., Findeisen, R., Nagy, Z.K. "Nonlinear model predictive control: From theory to application". J. Chin. Inst. Chem. Engrs, vol. 3, pp. 299-315, 2004.
- [22] David, N., and Lubomir, M. "Self-Organizing Migrating Algorithm used to control a semi-batch chemical reactor". In 13th International Conference on Control, Automation and Systems, pp. 1266-1269, 2013.
- [23] Kadlec, P., and Raida, Z. "Multi-objective self-organizing migrating algorithm applied to the design of electromagnetic components". IEEE Antennas and Propagation Magazine, vol. 55, pp. 50-68, 2013.
- [24] dos Santos Coelho, L., and Alotto, P. "Electromagnetic optimization using a cultural self-organizing migrating algorithm approach based on normative knowledge". IEEE Transactions on Magnetics, vol. 4, pp. 1446-1449, 2009.
- [25] Qi, H., Niu, C.Y., Jia, T., Wang, D.L., and Ruan, L.M. "Multiparameter estimation in nonhomogeneous participating slab by using self-organizing migrating algorithms". Journal of Quantitative Spectroscopy and Radiative Transfer, vol. 157, pp. 153-169, 2015.
- [26] Zelinka, I. "SOMA Self-organizing Migrating Algorithm". In Self-Organizing Migrating Algorithm Springer International Publishing, 2016.
- [27] Kazemi, M., and Arefi, M.M. "A fast iterative recursive least squares algorithm for Wiener model identification of highly nonlinear systems". ISA transactions, vol. 67, pp. 382-388, 2017.
- [28] Hagenblad, A., Ljung, L., Wills, A., "Maximum likelihood identification of Wiener models". Automatica, vol. 44, pp. 2697-2705, 2008.
- [29] Al-Duwaish, H., Karim, M. N., Chandrasekar, V. "Use of multilayer feedforward neural networks in identification and control of Wiener model". IEE Proceedings-Control Theory and Applications, vol. 143, pp. 255-258, 1996.
- [30] Tatjewski, P. "Advanced control of industrial processes: structures and algorithms". Springer Science & Business Media, 2007.
- [31] Kheriji, A., Bouani, F., and Ksouri, M. "A GGP approach to solve non convex min-max predictive controller for a class of constrained MIMO systems described by state-space models". International Journal of Control, Automation and Systems. vol. 9, pp. 452-460, 2011.
- [32] Sahed, O.A., Kara, K., and Benyoucef, A. "Artificial bee colony-based predictive control for non-linear systems". Transactions of the Institute of Measurement and Control, vol. 37, pp. 780-792, 2015.
- [33] Al-Duwaish, H., and Naem, W. "Nonlinear model predictive control of hammerstein and wiener models using genetic algorithms". In Proceedings of the 2001 IEEE International Conference On Control Applications, pp. 465-469, 2001.
- [34] Chen, W., Li, X., and Chen, M. "Suboptimal nonlinear model predictive control based on genetic algorithm". In Third International Symposium on Intelligent Information Technology Application Workshops, pp. 119-124, 2009.
- [35] Kaddah, S.S., Abo-Al-Ez, K.M., and Megahed, T.F. "Application of nonlinear model predictive control based on swarm optimization in power systems optimal operation with wind resources". Electric Power Systems Research. vol. 143, pp. 415-430, 2017.



- [36] Ramirez, D.R., Arahal, M.R., and Camacho, E.F. "Min-max predictive control of a heat exchanger using a neural network solver". IEEE transactions on control systems technology, vol. 12. pp. 776-786, 2004.
- [37] Chagra, W., Degachi, H., and Ksouri, M. "Nonlinear model predictive control based on Nelder Mead optimization method". Nonlinear Dynamics, pp. 1-12, 2017.
- [38] Agrawal, S., and Singh, D. "Modified Nelder-Mead self organizing migrating algorithm for function optimization and its application". Applied Soft Computing, vol. 51. pp. 341-350, 2017.

# An Efficient Participant's Selection Algorithm for Crowdsensing

\*Tariq Ali, Umar Draz, Sana Yasin, Javeria Noureen, Ahmad shaf  
CS. Department  
(CIIT) Sahiwal, Pakistan

Munwar Ali  
CS. Department  
(CIIT) Lahore, Pakistan

**Abstract**—With the advancement of mobile technology the use of Smartphone is greatly increased. Everyone has the mobile phones and it becomes the necessity of life. Today, smart devices are flooding the internet data at every time and in any form that cause the mobile crowdsensing (MCS). One of the key challenges in mobile crowd sensing system is how to effectively identify and select the well-suited participants in recruitments from a large user pool. This research work presents the concept of crowdsensing along with the selection process of participants from a large user pool. MCS provides the efficient selection process for participants that how well suited participant's selects/recruit from a large user pool. For this, the proposed selection algorithm plays our role in which the recruitment of participants takes place with the availability status from the large user pool. At the end, the graphical result presented with the suitable location of the participants and their time slot.

**Keywords**—Mobile crowdsensing (MCS); Mobile Sensing Platform (MSP); crowd sensing; participant; user pool; crowdsourcing

## I. INTRODUCTION

Today a mobile phone is the essential part of life. The use of mobile phones has greatly increased; the latest mobile phones now come with many embedded sensors. The capabilities of mobile phones have been greatly increased in the recent years, for instance, processing power, embedded sensors, storage capacities and network information rates [1]. This advancement of technologies combined with the huge number of client companioned cell phones empowers another and quickly developing sensing paradigm called Crowdsensing. Crowdsensing is the ability by which application developers can make tasks and recruit cell phone clients to give sensor information to be utilized towards a particular goal. Crowdsensing is also sometimes referred as a mobile Crowdsensing. A formal way to represent the mobile Crowdsensing (MCS) is: Mobile Crowdsensing (MCS) presents a new sensing model, which is based on the power of mobile devices. The absolute number of user companioned devices such as mobile phones, wearable devices, and smart vehicles so on [2], and their inherent mobility empowers a new and fast-growing.

Mobile Crowdsensing (MCS) awards a tremendous measure of wireless customers that offer neighborhood learning (e.g., nearby data, encompassing setting, calmer level, and activity conditions) accumulated by their sensor-improved contraptions. Mobile phone use for computation and acquires a richer functionality. It has a variety of sensors such as camera, microphone, Global Positioning System, accelerometer etc.

Health and pollution monitoring sensors will be intended in the coming future. Mobile sensors such as smartphones and vehicular systems represent a new type of geographically distributed sensing infrastructure that enables mobile people-centric sensing. Until recently mobile sensing research such as activity recognition, where peoples activity (e.g., walking, driving, sitting, talking) is classified and monitored, required specialized mobile devices (e.g., the Mobile Sensing Platform [MSP]) [3]. Crowdsensing have diverse applications which are separated into three classifications:

- 1) Framework checking.
- 2) Individual to individual correspondence checking.
- 3) Natural checking.

During a few years ago, a mobile device has been explored to contribute project-4 that continuously reported the total number of examined birds surrounding the US. Another measurement is Noise pollution [4] is the worst issue in the world. There are so many problems with noise pollution. These problems further affect the standard of health and life. Many diseases are being a part of noise pollution like blood pressure and so on [5]. The European Union commission is advised the other country to control the noisy areas through environment-friendly sensors [6]. To control the noise pollution in the city sensing nodes has been deploying by several governments through the noisy areas. With the help of noise map that visualizes the graphical view about those areas where the amount of sound level distribution is high. To measure the environmental noise some noise tube system also proposed [3]. Main architecture of mobile crowd sensing is in Fig. 1.

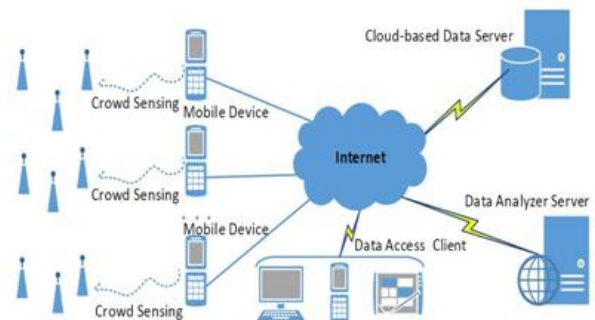


Fig. 1. Architecture of MCS.

This paper presents the selection algorithm of participants, that how participants select to perform a task from a large user

pool. Participants recruit on the base of availability to minimize the cost of sensing task. Section II describes the related work that was purposed on the crowd sensings parameters and its income challenges. In Sections III and IV, the selection criteria and selection algorithm are discussed. Results and conclusion are written in Sections V and VI, respectively.

## II. RELATED WORK

The growing sensing capabilities of smartphones have gone past the sensor systems concentrate on environmental and infrastructure monitoring. Individuals are currently the bearers of sensing devices and the sources and consumers about the sensed events [7]. Mobile Crowdsensing plays an analogous role with the one played through Amazons Mechanical Turk (MTurk) or ChaCha in crowdsourcing [8]. It permits individuals and organizations (clients) to access an absolute number of people (providers) ready to execute simple sensing tasks for which they are paid. Unlike the MTurks tasks which are executed on personal computers and always need human work. Mobile sensing tasks are accomplished on mobile devices that fulfill particular context/sensing requirements (e.g., location, time, particular sensors) and sometimes do not need human work (i.e., automatic sensing tasks [9]).

Smartphones already have numerous sensors like a camera; microphone, GPS, accelerometer and in the near future they are intended to include health and pollution monitoring sensors. Vehicular systems [10] have access to numerous hundred sensors embedded in cars, and latest vehicles come equipped with new types of sensors, for example, radar and camera. It has been believing that researchers in a number of fields of science and engineering as well as local state, and federal agencies can significantly benefit from this new sensing infrastructure as they will have access to valued data from the physical world. Moreover, commercial organizations might be very interested in gathering mobile sensing data to get more about customer behavior. The participants in mobile crowd sensing systems may need significant incentives to go out of their way and cover out of favor regions. In author situation, they give the complete detail in Biketastic that provide the incentives to participants through sharing bicycle ride. Another assorted quality of motivation is allowing information dealing to get extra data, for instance, deal chasing by means of value questions in Live Compare [11].

### A. Crowdsensing Challenges

Crowdsensing has many challenges here discussed some privacy and security challenges, issues and limitation for mobile crowd sensing [12]. Neighbourhood examination is entering the challenge of finding heuristics and outlining calculations that whole the pretend meaning. Similar examples of this function are reducing or eliminating the noise and cover the gaps of data. For example, GPS test can't have the capacity to get right or missing, in this time anomalies must be taking out excluded tests extrapolated. The 3-level structure building also have a couple of troubles are according to the accompanying: (a) simulate a computing, (b) arrangement and execution trial between virtual machines correspondences, (c) Correspondence execution is about near to the ground at what time stand out from between process correspondence [13]. Movement affected Reconfiguration is another challenge like

MoneyBee. By using different model and systems most of the MCS produce the same type of data. To make the MCS as an honest application the problem is what are the possible ways in which provide the true and essential information in which true contribution takes place. The arrangement is recent, various amusement hypothesis approaches have been proposed for versatile group detecting and registering to empower and compensate honest commitments. For a unique versatile group detecting and registering framework, there is still a requirement for new motivation and estimating components to draw in, move, and reward honest and excellent detecting information givers. For information conveyance, data conveyance in the transient system is additionally challenged in portable group detecting, how to dispatch the detected information from appropriated members. With the help of host, it can detect the account and its arrangement. versatile group detecting, figuring attributes, instance transfer speed, remote correspondence, repetitive system allotting because of human portability, and a colossal number of vitality obliged gadgets. Planning calculations can comprehend this inconvenience and utilized detecting servers to orchestrate detecting occasions of cell phones (a motivation instrument utilize selected). Note that shrewd detecting applications will just utilize the planning calculations.

1) *Crowd sensing privacy*: Assurance of privacy is basic for everyone. No one needs to reveal his/her security before anyone. In the current system use unmistakable systems to offer security to PDAs or center points. Privacy mechanism is responsible to provide security defending segments to data supporters. Some part, for example, assignment distribution, sensor doors, information anonymization, motivation component and huge information stockpiling are utilized as a part of this layer, which gathered information from the chose hubs [4], [14], [15]. Most famous privacy approaches are discussed in [16]:

- 1) Pseudonyms: It is the straightforward method that makes members unknown by supplanting their recognizable proof data with an assumed name.
- 2) Connection anonymization.

Utilizing this method, we can keep away from the system based following assaults utilizing IP addresses. One such procedure which is utilized as a part of Crowd detecting applications is onion steering. Another protection safeguarding approach in which some calligraphy techniques are used. Moreover, some essentialness usage also has been noticed in the mobile crowdsensing [17]. The k-namelessness procedure can be connected keeping in mind the end goal to give the area security of the members who transfer reports. The essential thought behind the k-namelessness method develop gatherings of 'N' members. Along these lines normal trait sharing, (like 'N' members arranged each district), translating then indistinguishable each other. To construct a gathering of 'k' clients it can utilize diverse strategies to locate the appropriate and basic property. So these techniques ordered into two fundamental segments, for example, speculation and bother. With the wide adoption of mobile Crowdsensing applications, task coverage and participant selection in MCS systems have captured the attention of researchers. First, there are several systems and experimental studies on either experimental study on MCS coverage or general framework of participant recruitment [18],

[19]. For example in [20] has performed a systematic study of the coverage and scaling properties of place-centric urban crowd sensing and shows promising results that MCS can provide relatively high coverage levels especially given area with large size. Then, there are also many theoretical studies on various task assignment and participant selection problems, playing tradeoffs among sensing cost, task coverage, energy efficiency [21], [22] and user privacy [23], and incentive. In offline study participant selection in the piggyback mechanism, in which MCS for probabilistic coverage at this situation so that task is easily performed. They aim to select a minimum number of participants to guarantee the selected participants will make enough number of calls at a certain percentage of the target locations over a long-fixed sensing period. Protection, security, difficulties, and hazard that uncovers the delicate information about members with respect to privacy is easily be solved if the well-suited participants need to select from the large user pool. We will concentrate on the social and specialized difficulties or dangers. Fig. 2 is represented a generic structure of task flow in MCS.

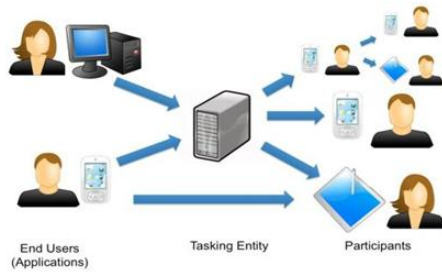


Fig. 2. Generic structure of task flow in MCS.

In this section, a different concept is presented such as Crowdsensing, Crowdsensing application, Crowdsensing challenges and privacy issues and Crowdsensing related work. MCS allows the extensive measure of mobile phone customers share local learning, for example; (nearby data, encompassing setting). Crowdsensing has many challenges some privacy and security challenges are mentioned in [16]. To get the privacy different privacy techniques can use to protect the user privacy such as Anonymization, encryption and data perturbation but still, there is some privacy threats need to be solved. The key challenge which identity in this research work is how to effectively identify the well-suited participant from the large user pool [17]. The recruitment process becomes more complicated when the sensing tasks are dynamic and heterogeneous.

### III. SELECTION CRITERIA FOR PARTICIPANTS

In selection algorithm, it is assumed that 'N' number of participants has been recruited. 'N' number of participants selects from the user pool and user pool is divided into two partitions. Selection is done by the parallel searching to check the availability of participants. Design algorithm uses two values '0' and '1' in two locations. If the value is 0-1 then the participant is selected same as in the 1-0. If 0-0 in both locations then participants will not be select. Table I shows the availability of participants.

TABLE I. PARTICIPANTS RAKING

Users	1	2	3	4	5	6	7	8	9	10
L1	0	0	1	1	0	0	1	1	0	0
L2	1	1	0	0	0	0	0	0	1	1

### IV. SELECTION ALGORITHM

In this algorithm, well-suited participants are selected from a large user pool. User pool divides into two segments that parallel check and selects the participants this increase the efficiency of the algorithm. For example, we have two locations L1 and L2. For the selection of participants, it is necessary that participants should be available either in L1 or L2. If participant available in any one location it will be selected otherwise it will not be selected.

#### A. Pseudo Code

In this algorithm, user pool is divided into two segments such as  $\frac{x}{2}$ . 'x' represents the rang of user pool.

#### Algorithm 1 Selection algorithm

```

1: Start
2: Input: N numbers of participants
3: Output: Selected participants.
4: a is user pool // takes input from user pool
5: selected = 0 // initialize the variable (selected) from zero
6: unselected = 0 // initialize the variable (unselected) from zero

```

---

**Segment 1**

```

7: for ( r in 1: mid) // this loop runs from 1 to mid for selecting users from the first segment of user pool
8: do
9:   for ( u in 1:2) // inner for loop is running from 1 to 2 times used for locations
10: do
11:   k=1
12:   if a [r, k] ==1 OR a [r, k = k + 1] == 1 then
13:     selected = selected + 1 // counts the selected participants from segment 1
14:     selected
15:   else
16:     unselected = unselected + 1 // counts the unselected participants from segment 1
17:   end if
18: end for
19: end for
20: Call Segment 2

```

#### B. Description of Pseudo Code

In this algorithm, participants are selected from a large user pool. To increase the efficiency of this algorithm user pool is divided into two partitions ( $x/2$ ) that efficiently selects the participants from the user pool. There are 'x' time slots and two locations L1 and L2.

**Segment 2**

```

21: for ( p in mid : x) // this loop runs from mid to x that
    checks the second segment of the user pool for selecting
    participants
22: do
23:   for ( t in 1:2) // inner for loop is running from 1 to 2
    times used for locations
24:   do
25:     q = 1
26:     if a [p, q] == 1 OR a [p, q = q + 1] == 1 then
27:       selected = selected + 1 // counts the selected
    participants from segment 2
28:       selected
29:     else
30:       unselected = unselected + 1 // counts the uns-
    elected participants from segment 2
31:     end if
32:   end for
33: end for
34: print ("overall selected and unselected users from user
    pool")
35: selected // counts the total number of selected users from
    pool
36: unselected // counts the total number of unselected users
    from pool
37: End

```

- 1) Variables initialization  
Selected =0  
Unselected = 0
- 2) Loops:  
Two loops are used in this algorithm. The first loop runs from 1 to mid for selecting users from the first segment of user pool. The second loop is running from 1 to 2 used for locations.
- 3) Conditions:  
Assign value to k=1 and use condition if (a [r, k] = =1 OR a [r, k=k+1] = =1). In this condition can specify as a(user pool), r(used for loop to check the participants), k(it checks the availability of participants).

- If else condition  
A conditional statement is used to check the availability of participants.
- If k finds 1 in any location it means that conditions are true then add 1 in a selected variable such as selected= selected +1. All selected participants value stored in a selected variable. This condition will run until a condition is false.  
Else if k finds zero in both locations, then the participant is not available and add 1 in the unselected variable such as unselected=unselected+1.

The above process will repeat for the second partition of user pool. At the end, print the total number of selected and unselected participants from the user pool.

**C. Working of Selection Algorithm**

In Fig. 3 flowchart of overall working of the selection, algorithm is shown.

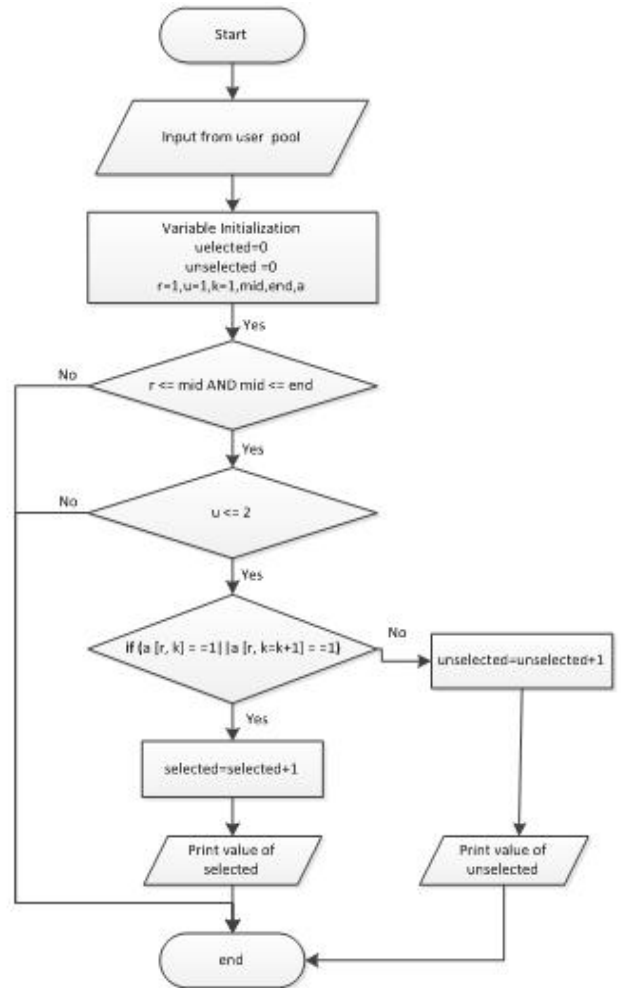


Fig. 3. Flowchart of selection algorithm.

**V. RESULTS**

The selection algorithm of participants from a large user pool on the base of their availability is successfully implemented. User pool divides into two segments that parallel check and selects the participants. If the participants are available in the location, they will be selected to participate in the user pool. Otherwise, they will not be selected. To participate in the user pool, the participants must be selected. In another case, participant will not be eligible to participate. So here we can see how participants selected for the participation. It is assumed that user pool range is 12 and 12/2=6. The design algorithm parallels check both partitions of user pool for the selection of participants. This Fig. 4 shows that 10 participants are willing to participate in the user pool.

**A. Graphical Representation of Results**

This section shows the graphical representations are as follows as in Fig. 5.

```

1 a6=matrix(c(0,0,1,1,0,1,1,0,0,0,1,0, 1,1,0,0,0,0,0,1,1,0,0),nrow=12,ncol=2)
2 selected = unselected = 0
3 for(r in 1:6)
4   for(u in 1:2)
5     k=1
6     if(a6[r,k]==1|a6[r,k-k+1]==1)
7       {
8         selected=selected+1
9         selected}
10      else
11        {unselected=unselected+1}
12      }
13    }
14  }
15 for(p in 7:12) # second segment of user pool
16   for(t in 1:2)
17     q=1
18     if(a6[p,q]==1|a6[p,q-q+1]==1)
19       {
20         selected=selected+1
21         selected}
22      else
23        {unselected=unselected+1}
24      }
25    }
26  }
27 print("overall selected and unselected users from user pool")
28 selected
29 unselected
30 }
63 (Top Level) :

```

```

[1] "overall selected and unselected users from user pool"
> selected
[1] 10
> unselected
[1] 2
>

```

Fig. 4. Selection of participants.

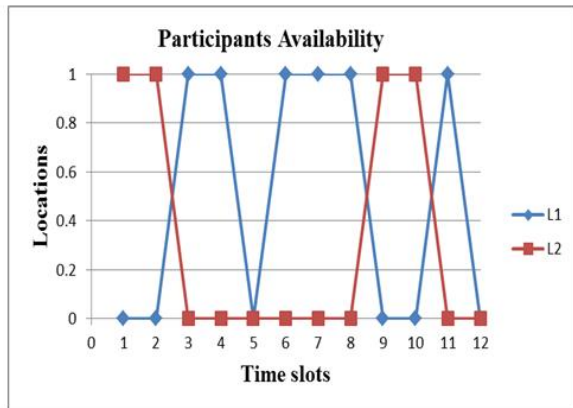


Fig. 5. Participant's availability.

Fig. 5 presents the availability of participants based on values as shown in Table I. From the figure at value 5 and 12, there is no participants are available and other points show the availability of participants. 'L' represents the locations.

## VI. CONCLUSION

In this research work, literature is reviewed to understand the dimensions related to ongoing and emerging issues in mobile crowd sensing. Different studies are analyzed to identify the focus of mobile crowd sensing, its applications domains, privacy and security challenges and limitation for mobile crowd sensing with their possible solution. There is existing participant recruitment process is analyzed. We have identified few shortcomings in the mobile crowd sensing system is how to effectively identify and select the well-suited participants in

recruitments from a large user pool and suggested selection algorithms that can improve the efficiency of the selection of the participants from the large user pool. The recommendations presented for the existing participant recruitment process that can improve the recruitment process, how we can recruit the well-suited participants from the large user pool.

Up to our best knowledge, this is the first work is done on the domain of MCS and the proposed designed modules to make the recruitment process efficient. These modules are a selection of 'N' numbers of participant's recruitment algorithm. This algorithm efficiently recruits the participants and improves the performance of the algorithm. The selection of 'N' number of participants recruitment algorithm with the goal of minimizing the sensing cost, while satisfying the certain level of coverage of mobile users. In future, we are interested to extend this research by developing energy efficient sensor device that will minimize the energy consumption and will increase the battery life of the smartphones.

## REFERENCES

- [1] Bellavista, P., et al. Human dynamics of mobile crowd sensing experimental datasets. in Communications (ICC), 2017 IEEE International Conference on. 2017. IEEE.
- [2] Du, R., et al. Predicting activity attendance in event-based social networks: Content, context and social influence. in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing. 2014. ACM.
- [3] Jaimes, L.G., I.J. Vergara-Laurens, and A. Raij, A survey of incentive techniques for mobile crowd sensing. IEEE Internet of Things Journal, 2015. 2(5): p. 370-380.
- [4] Ma, H., D. Zhao, and P. Yuan, Opportunities in mobile crowd sensing. IEEE Communications Magazine, 2014. 52(8): p. 29-35.
- [5] Stansfeld, S.A. and M.P. Matheson, Noise pollution: non-auditory effects on health. British medical bulletin, 2003. 68(1): p. 243-257.
- [6] Santini, S., B. Ostermaier, and A. Vitaletti. First experiences using wireless sensor networks for noise pollution monitoring. in Proceedings of the workshop on Real-world wireless sensor networks. 2008. ACM.
- [7] Liu, C.H., et al., Energy-aware participant selection for smartphone-enabled mobile crowd sensing. IEEE Systems Journal, 2015.
- [8] Narula, P., et al., MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid. Human Computation, 2011. 11: p. 11.
- [9] Kong, L., et al., Sustainable Incentive Mechanisms for Mobile Crowd-sensing: Part 1. IEEE Communications Magazine, 2017. 55(3): p. 60-61.
- [10] Pouryazdan, M., et al., Quantifying User Reputation Scores, Data Trustworthiness, and User Incentives in Mobile Crowd-Sensing. IEEE Access, 2017. 5: p. 1382-1397.
- [11] Ren, J., et al., Exploiting mobile crowdsourcing for pervasive cloud services: challenges and solutions. IEEE Communications Magazine, 2015. 53(3): p. 98-105.
- [12] Yang, K., et al., Security and privacy in mobile crowdsourcing networks: challenges and opportunities. IEEE Communications Magazine, 2015. 53(8): p. 75-81.
- [13] Govindaraj, D., et al., MoneyBee: Towards enabling a ubiquitous, efficient, and easyto use mobile crowdsourcing service in the emerging market. Bell Labs Technical Journal, 2011. 15(4): p. 79-92.
- [14] Ganti, R.K., F. Ye, and H. Lei, Mobile crowdsensing: current state and future challenges. IEEE Communications Magazine, 2011. 49(11).
- [15] Pournajaf, L., et al., A survey on privacy in mobile crowd sensing task management. Dept. Math. Comput. Sci., Emory Univ., Atlanta, GA, USA, Tech. Rep. TR-2014-002, 2014.
- [16] Krontiris, I., M. Langheinrich, and K. Shilton, Trust and privacy in mobile experience sharing: future challenges and avenues for research. IEEE Communications Magazine, 2014. 52(8): p. 50-55.

- [17] Yi, K., et al., Fast participant recruitment algorithm for large-scale Vehicle-based Mobile Crowd Sensing. *Pervasive and Mobile Computing*, 2017.
- [18] Reddy, S., D. Estrin, and M. Srivastava, Recruitment framework for participatory sensing data collections, in *Pervasive Computing*. 2010, Springer. p. 138-155.
- [19] Tuncay, G.S., G. Benincasa, and A. Helmy. Participant recruitment and data collection framework for opportunistic sensing: A comparative analysis. in *Proceedings of the 8th ACM MobiCom workshop on Challenged networks*. 2013. ACM.
- [20] Chon, Y., et al. Understanding the coverage and scalability of place-centric crowdsensing. in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. 2013. ACM.
- [21] Xiong, H., et al., EMC 3: Energy-efficient data transfer in mobile crowdsensing under full coverage constraint. *Mobile Computing, IEEE Transactions on*, 2015. 14(7): p. 1355-1368.
- [22] Zhao, D., H. Ma, and L. Liu, Energy-efficient opportunistic coverage for people-centric urban sensing. *Wireless networks*, 2014. 20(6): p. 1461-1476.
- [23] Pournajaf, L., et al. Spatial task assignment for crowd sensing with cloaked locations. in *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*. 2014. IEEE.

# An Energy-Efficient User-Centric Approach for High-Capacity 5G Heterogeneous Cellular Networks

Abdulziz M. Ghaleb, Ali Mohammed Mansoor and Rodina Ahmad

Dep. of Software Engineering, Faculty of Computer Science & IT, University of Malaya, Malaysia

**Abstract**—Today’s cellular networks (3G/4G) do not scale well in heterogeneous networks (HetNets) of multiple technologies that employ network-centric (NC) model. This destabilization is due to the need for coordination and management of multiple layers of the HetNets that the NC models cannot provide. User-centric (UC) approach is one of the key enablers of 5G wireless cellular networks for rapid recovering from network failures and ensuring certain communication capability for the users. In this paper, we present resource-aware energy-saving technique based on the UC model for LTE-A HetNets. We formulate an optimization problem for UC as a mixed linear integer programming (MILP) that minimizes the total power consumption (Energy Efficiency) while respecting the data rate per user and propose a low complexity iterative algorithm to user terminal (UE)-eNodeB association. In UC model, UE possessing terminal intelligence can establish the transmission and reception with different cells within the LTE-A HetNet assuming the existence of coordination between the different cells in the network. The performance is evaluated in terms of energy saving in the uplink and downlink and the added capacity to the network (data rate). The evaluation is carried out by comparing a UC model against a NC model with the same simulation setup. The results show significant percentage of energy saving at eNodeBs and UEs in a UC model. Also, system capacity is enhanced in the UC model in both the uplink and downlink due to utilizing best channel gain for transmission and reception.

**Keywords**—Energy efficiency; HetNets; green networks; user-centric; network-centric; 5G

## I. INTRODUCTION

Today’s 3G and 4G cellular networks are principally designed based on cell-centric or network-centric (NC) model with a focus on peak rate and spectral efficiency improvements. In the 5G era, dense deployment of heterogeneous network (HetNet) architecture will shift towards user-centric (UC) model to deliver a uniform connectivity experience. Therefore, 5G networks will require advanced source coding and advanced radio access networks. The objective is to significantly improve the flexibility of deployment and connectivity by making them more and more user-oriented [1]. The relationship between the downlink and the uplink in HetNets is different from that of the homogeneous ones. The transmit power of all transmitters in the uplink is roughly the same (independent of distance and amount of traffic) since all UEs are running off batteries. In contrast, there exist transmit power disparities between different eNodeB (eNB) types in the downlink (up to 20 dB) [2].

The efficient deployment of HetNets in 5G era calls for new disruptive technologies in a way that allows the corresponding information to flow in multiple data streams through different sets of heterogeneous nodes [4]. 5G networks should achieve

combined gains in three categories: extreme densification and offloading, increased bandwidth and increased spectral efficiency in order to support 1,000-fold gains in capacity and connections for at least 100 billion devices. The demand in capacity gain would increase the consumed energy by the network by a factor of 100 [1]-[5].

Therefore, the NC architecture should evolve into a UC one, and uplink and downlink could be considered as two separate networks. Each network will require different models for interference, cell association, and throughput [2]-[4]. In UC architecture, the UE has a crucial role in establishing the connectivity with the eNBs. The UE can decide whether to establish connectivity with the same cell or with different cells in the uplink and downlink communication. In this perspective, new carrier type was proposed in [5] where user/data and control planes can be separated in UEs by small cells at higher frequency bands (mmWave). This is expected to reduce the frequent handover between small cells and macrocell and among small cells. Hence, the connectivity can be maintained even when using small cells and higher frequency bands since connectivity and mobility is provided by the control plane [6].

### A. Motivation for this Work

The current works does not involve any performance evaluation of the UC model in term of power efficiency, capacity improvement or Quality of Service (QoS). The motivation for this work is to provide good insights of the performance of UC model deployment in future 5G networks. In this work, we have formulated an optimization problem for UC as mixed linear integer programming (MILP) that minimizes the total power consumption while respecting the data rate per user and proposed a low complexity iterative algorithm to UE-eNB association. The paper provides an evaluation for the UC model in LTE-A HetNets in terms of energy saving at both eNBs and UEs and the added capacity to the network. Two sets of simulation experiments for different number of UEs (reflecting the network load) were carried out with the same setup; one for NC model and one for UC model. The collected results show the percentage of energy saving in the UC model compared to the NC model and additional gained data rate (data rate in the NC model subtracted from data rate in the UC model). The results show significant energy saving (up to 15% in the downlink and 6% in the uplink) and capacity enhancement in UC model. It is noteworthy that some claims that NC strategies are better than UC ones in terms of achievable throughput but are worse in terms of computational complexity in certain scenarios (i.e., LTE/WiFi coexistence) [7]. This opens research door for further investigation of the UC model as well as hybrid or joint user and network architecture.



The rest of the paper is arranged as follows. Section II provides an overview of the related works. Section III provides a detailed description of the UC-based network model. The problem is formulated in Section IV and proposed method described in Section V. The results are presented in Section VI. Finally Section VII concludes the paper.

## II. RELATED WORKS

Some works have considered the UC architecture for wired networks or Internet for self-organizing, autonomic networks. The architecture is used for sharing network services and resources by installing the device as the owner and controller of its personal data [8]–[11]. Recently, some works have envisioned the UC architecture as one of the core features of the 5G networks [1], [2], [4]–[6], [8], [12]–[14]. Not all the authors considered the full UC paradigm; some of them either considered the separation of uplink and downlink [2] or separating control and data planes [5].

To the best of the authors' knowledge, little work has been done on the evaluation of the UC architecture for wireless communication. The authors of [15] studied the dynamic user association decoupled UL-DL time division duplexing (TDD)-based networks to balance the UL and DL loads in different small cells. The authors of [16] presented a transmission/reception scheme for LTE/LTE-A HetNets that exploits the concept of Coordinated Multipoint (CoMP). The authors of [17] we presented device-centric design, implementation, and testing of optimized data aggregation mechanisms for file downloading and video streaming applications. The uplink and downlink transmissions of a UE are established with different cells assuming the existence of coordination between the cells. In this scheme, the UE is associated to a small cell for uplink transmission and to macrocell for downlink reception. The authors in [18] investigated the potential to enable emergency communications with different radio access technologies such as LTE and WLAN which are the candidates for direct communication in emergency cases [19]. However, the main focus was to enable better emergency communication. It was not in the context of 5G networks, and there was no considerations for using this feature for network efficiency and self-organization. The most significant work done in this regard is in [20] where the authors studied the decoupling of downlink and uplink based on simulation of LTE field trial network in a dense urban HetNet deployment. The authors considered downlink cell association based on the received power and uplink cell association based on the pathloss.

## III. NETWORK MODEL

We consider a a LTE-A HetNet deployed in a given geographical area divided into equal-size cells where an eNB is placed at the center of each cell. The area also includes smaller cells (micro, femto, pico) placed either within the macrocells or to bridge the coverage gaps. ENBs are classified as macro, micro, pico and femto eNBs based on both their transmit power and their antenna heights. In LTE air interface, Orthogonal Frequency Division Multiple Access (OFDMA) is used for the downlink access mechanism and the Single Carrier - Frequency Division Multiple Access (SC-FDMA) is used for the uplink. For OFDM-based access schemes, the available spectrum is divided into subcarriers in the frequency domain. In LTE, the

spectrum is divided into resource blocks (RBs). Each RB is constituted by 12 consecutive subcarriers for a fixed duration of 1 ms. In the UC model, the uplink and downlink are decoupled and are considered two separate networks. The deployment scenario is shown in Fig. 1.

### A. Energy Consumption Model

1) *Power Consumption Model for eNodeBs*: For simplicity, we consider that each eNB in the macro and small cells are equipped with an omni-directional antenna. The  $j$ th active eNB consumed power  $P_j^{\text{eNB}}$  is computed as follows [21]:

$$P_j^{\text{eNB}} = a_j P_j^{\text{tx}} + b_j, \quad (1)$$

where  $P_j^{\text{tx}}$  denotes the radiated power of the  $j$ th eNB. The coefficient  $a_j$  corresponds to the radiated power consumed due to feeder and amplifier losses. The term  $b_j$  is the fixed power offset which is consumed by the site independently of the transmitted power and depends on the eNB type.

2) *Power Consumption Model for UEs*: Each UE in the network is considered to be equipped with a set of omni-directional antennas  $\mathcal{N}_i^{\text{ant}}$  and can communicate with macro and small cells (open access for femto cells). Assuming that the  $i$ th UE is connected to a set of eNBs  $\mathcal{N}_{\text{eNB}}^{\text{DL},i}$  in the downlink and to a set of eNBs  $\mathcal{N}_{\text{eNB}}^{\text{UL},i}$  in the uplink according to the suggested UC model for 5G [1], [2], [4], [6], [16], [22] and given that the  $i$ th user is connected to the  $j$ th eNB through set of antennas, then the consumed power  $P_i^{\text{UE}}$  of the  $i$ th running mobile is computed as follows:

$$P_i^{\text{UE}} = m_i^l \sum_{l \in \mathcal{N}_i^{\text{ant}}} \sum_{j \in \mathcal{N}_{\text{eNB}}^{\text{UL},i}} P_{i,j}^{\text{tx},l} + n_i, \quad (2)$$

where  $P_{i,j}^{\text{tx},l}$  corresponds to the radiated power of the  $l$ th antenna of the  $i$ th UE connected to the  $j$ th eNB. The coefficient  $m_i^l$  corresponds to the radiated power consumed due to system losses which varies from one antenna to another and  $n_i$  is the fixed power consumed to keep the mobile on.

The eNB's energy consumption is segregated into two types namely, the static energy consumption and the dynamic energy consumption. When turned on, each eNB consumes a constant amount of energy (fixed power) depending on its type regardless of the traffic load. This amount of energy is always required just for the equipment to be powered on. Similarly, UE's energy consumption is divided into static and dynamic energy consumption. The second part is the adaptive power consumption which is proportional to the transmission density. For the UEs, they are assumed to be on all the time and there is nothing to optimize regarding their static power consumption. Henceforward, this paper focuses on optimizing both saving the eNBs static and dynamic energy consumption as well as the UEs adaptive energy consumption. The overall power consumed by the HetNet infrastructure and UEs  $E_{\text{Het}}$  for a  $T$  hours of time can be represented by the sum of energy consumed by all active eNBs and UEs as follows:

$$E_{\text{Het}} = \left( \sum_{j=1}^{N_{\text{eNB}}} P_j^{\text{eNB}} + \sum_{i=1}^{N_{\text{UEs}}} P_{i,j}^{\text{UE}} \right) \times \frac{T}{1000} \quad (\text{kWh}), \quad (3)$$

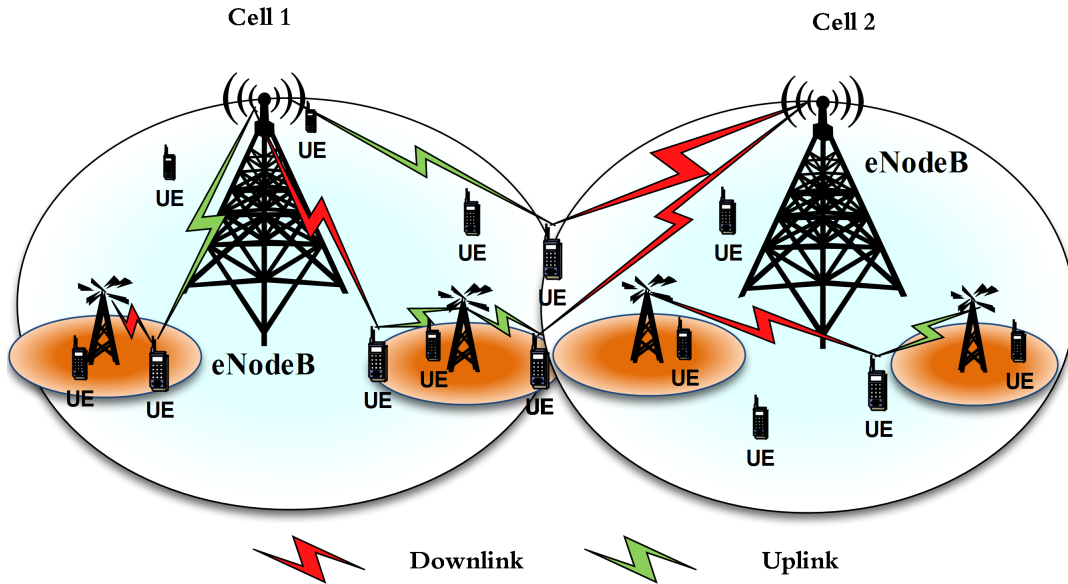


Fig. 1. Deployment scenario with UC model.

### B. Channel Model

The channel gain for both uplink and downlink over subcarrier  $s$  between  $i$ th UE and  $j$ th eNB is given by [23]:

$$H_{i,s,j,\text{dB}} = (-\kappa - v \log_{10} d_{i,j}) - \xi_{i,s,j} + 10 \log_{10} F_{i,s,j}, \quad (4)$$

where the first term represents the propagation loss with  $\kappa$  being the path loss constant,  $d_{i,j}$  being the distance in km from the  $i$ th UE to the  $j$ th eNB and  $v$  being the path loss exponent. The second term,  $\xi_{i,s,j}$ , represents the zero-mean log-normal shadowing with a standard deviation  $\sigma_\xi$ , while  $F_{i,s,j}$  corresponds to Rayleigh fading. The notation  $H_{i,s,j}^{\text{UL}}$  and  $H_{i,s,j}^{\text{DL}}$  will be used to differentiate between uplink and downlink channel gains, respectively. The LTE bandwidth is defined as a function of the number of RBs multiplied by the RB bandwidth,  $B = N_{\text{RB}} \times B_{\text{RB}}$  (kHz). It can be expressed in terms of number of subcarriers and subcarrier bandwidth as  $B = N_{\text{sub}} \times B_{\text{sub}}$  (kHz).

### C. Data Rates Calculation

1) *Data Rates in the Downlink*: Letting  $s_{i,j}$  be a subcarrier assigned by the  $j$ th eNB to the  $i$ th UE,  $\mathcal{I}_{s,i,j}^{\text{DL}}$  be the set of downlink subcarriers allocated to the  $i$ th UE from the  $j$ th eNB and  $R_i^{\text{DL}}$  the achievable downlink rate of the  $i$ th UE. The set of subcarriers given to the  $i$ th UE by the HetNet in the downlink is denoted as  $\mathcal{I}_{s,i}^{\text{DL}}$ . The OFDMA data rate of  $i$ th UE supported by the  $j$ th eNB is given by:

$$R_{i,j}^{\text{DL}}(P_{j,\text{max}}^{\text{tx}}, \mathcal{I}_{s,i,j}^{\text{DL}}) = \sum_{s \in \mathcal{I}_{s,i,j}^{\text{DL}}} B_s \cdot \log_2 \left( 1 + \gamma_{i,s,j}^{\text{DL}} \right) \quad (5)$$

where  $\gamma_{i,s,j}^{\text{DL}}$  is the downlink SINR of the  $i$ th UE over subcarrier  $s$  transmitted from the  $j$ th eNB and is given by:

$$\gamma_{i,s,j}^{\text{DL}} = \frac{P_{s,j}^{\text{tx}} H_{i,s,j}^{\text{DL}}}{I_{s,i,j}^{\text{DL}} + \sigma_{s,i,j}^2}, \quad (6)$$

where  $H_{i,s,j}^{\text{DL}}$  is the channel gain of the  $i$ th UE over subcarrier  $s$ ,  $\sigma_{s,i,j}^2$  is the noise power over subcarrier  $s$  in the receiver of the  $i$ th UE, and  $I_{s,i,j}^{\text{DL}}$  is the interference on subcarrier  $s$  measured at the receiver of the  $i$ th UE. The total data rate provided to the  $i$ th UE by the network is given by:

$$R_i^{\text{DL}} = \sum_{j \in \mathcal{N}_{\text{eNB}}^i} R_{i,j}^{\text{DL}} \quad (7)$$

The total data rate provided to the  $i$ th UE by the network should be equal to or greater than a threshold value,  $R_{i,\text{th}}^{\text{DL}}$ , in order to provide the QoS requested by the user based on the contract. Letting  $\mathcal{N}_{\text{UE}}^j$  set of UEs attached to the  $j$ th eNB, the total data rate that cell  $j$  can support is given by:

$$R_j^{\text{DL}} = \sum_{i \in \mathcal{N}_{\text{UE}}^j} R_{i,j}^{\text{DL}} \quad (8)$$

We assume that bandwidth varies from one eNB to another, so does the total number of subcarriers  $N_{\text{sub}}^{\text{DL}}$  for cell  $j$  in the downlink. As we seek to come out with optimized realistic solutions for power allocation, we consider non-uniform or adaptive power transmission over the subcarriers, i.e.,  $P_{s,j}^{\text{tx}}$  is not constant. This allows eNBs to adjust their transmit power levels according to the distance of the UE, interference and modulation and coding scheme (MCS).

2) *Data Rates in the Uplink*: According to UC model, UE can be associated with one or more eNB in the uplink. Letting  $\mathcal{I}_{s,i,j}^{\text{UL}}$  be the set of uplink subcarriers granted to the  $i$ th UE from  $j$ th eNB,  $P_{i,j}^{\text{UE}}$  the total transmit power of the  $i$ th UE and  $R_{i,j}^{\text{UL}}$  its achievable rate in the uplink, the set of subcarriers guaranteed to the  $i$ th UE by the HetNet in the uplink  $\mathcal{I}_{s,i}^{\text{UL}}$  then, the SC-FDMA data rate of the  $i$ th UE is given by:

$$R_{i,j}^{\text{UL}}(P_{i,j}^{\text{UE}}, \mathcal{I}_{s,i,j}^{\text{UL}}) = B_{\text{sub}} |\mathcal{I}_{s,i,j}^{\text{UL}}| \cdot \log_2 \left( 1 + \gamma_{i,j}^{\text{UL}}(P_{i,j}^{\text{UE}}, \mathcal{I}_{s,i,j}^{\text{UL}}) \right) \quad (9)$$

where  $|\mathcal{I}_{s,i,j}^{\text{UL}}|$  is the cardinality of  $\mathcal{I}_{s,i,j}^{\text{UL}}$  and  $\gamma_{i,j}^{\text{UL}}(P_{i,j}^{\text{UE}}, \mathcal{I}_{s,i,j}^{\text{UL}})$  is the SINR of the  $i$ th UE after frequency domain equalization at the receiver. The uplink SINR of the  $i$ th UE over subcarrier  $s$  served by  $j$ th eNB and is given by [24]:

$$\gamma_{i,s,j}^{\text{UL}} = \frac{P_{i,s,j}^{\text{UE}} H_{i,s,j}^{\text{UL}}}{I_{s,j}^{\text{UL}} + \sigma_{s,j}^2}, \quad (10)$$

where  $H_{i,s,j}^{\text{UL}}$  is the channel gain between the  $i$ th UE and the  $j$ th eNB over subcarrier  $s$ ,  $\sigma_{s,j}^2$  is the noise power over subcarrier  $s$  at the  $j$ th eNB,  $P_{i,s,j}^{\text{UL}}$  is the power transmitted by the  $i$ th UE over subcarrier  $s$  in the  $j$ th cell.

#### IV. PROBLEM FORMULATION

In both uplink and downlink, amount of data rate depends on both number of assigned subcarriers and SINR. SINR is a function of the transmit power and the link quality. However, increasing the power is not necessarily a good choice since it leads, of course, to higher power consumption and increase the interference which degrades the link quality specially in ultra dense deployment of 5G systems. UC approach can reduce the interference and ensure energy savings in designing green wireless cellular networks with higher capacity. With the decoupling of uplink and downlink, UE can be associated with different eNBs in the uplink and downlink so that data rate is maximized with minimum power consumption. Assuming that UE has full knowledge of the channel status which can be sensed or collected from the eNB, the UE will choose the best link for downlink and uplink data transmission. The total power consumption over all subcarriers has to be less or equal to the maximum transmission power of the eNB denoted by  $P_{j,\max}^{\text{tx}}$ . The LTE standard mandates that the RBs allocated to a single user in the uplink be consecutive with equal power allocation over their subcarriers [25], [26]. The model be formulated as follows:

- **Parameters:**

$\mathcal{N}_{\text{eNB}}$  : set of the deployed eNBs within the HetNet  
 $\mathcal{N}_{\text{UE}}$  : set of subscribers in the area

- **Decision Variables:**

$$\delta_{i,j}^{\text{DL}} = \begin{cases} 1 & \text{if the } i\text{th UE is associated to the } j\text{th eNB} \\ & \text{in the downlink,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\eta_{i,j}^{\text{UL}} = \begin{cases} 1 & \text{if the } i\text{th UE is associated to the } j\text{th eNB} \\ & \text{in the uplink,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\vartheta_{x,i,j}^{\text{DL}} = \begin{cases} 1 & \text{if subcarriers } s_x \text{ is allocated to the } i\text{th UE} \\ & \text{from the } j\text{th eNB in the downlink,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\epsilon_{x,i,j}^{\text{UL}} = \begin{cases} 1 & \text{if subcarriers } s_x \text{ is allocated to the } i\text{th UE} \\ & \text{from the } j\text{th eNB in the uplink,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\psi_i^{\text{DL}} = \begin{cases} 1 & \text{if } R_i^{\text{DL}} \geq R_{i,\text{th}}^{\text{DL}} \\ 0 & \text{otherwise.} \end{cases}$$

$$\varrho_i^{\text{DL}} = \begin{cases} 1 & \text{if } R_i^{\text{UL}} \geq R_{i,\text{th}}^{\text{UL}} \\ 0 & \text{otherwise.} \end{cases}$$

- **Mathematical Model:**

$$\text{Minimize: } \sum_j P_j^{\text{eNB}} + \sum_{i=1}^{N_{\text{UEs}}} P_{i,j}^{\text{UE}} \quad (11)$$

$$\text{Subject to: } \sum_{s \in \mathcal{I}_{s,i}^{\text{UL}}} P_{s,j}^{\text{tx}} \leq P_{j,\max}^{\text{tx}}, \quad (12)$$

$$|\mathcal{I}_{s,i,j}^{\text{UL}}| \times P_{i,s,j}^{\text{UE}} \leq P_{i,\max}^{\text{tx}} \quad (13)$$

$$R_i^{\text{DL}} \geq R_{i,\text{th}}^{\text{DL}} \quad (14)$$

$$R_i^{\text{UL}} \geq R_{i,\text{th}}^{\text{UL}} \quad (15)$$

$$1 \leq \sum \delta_{i,j}^{\text{DL}} \leq |\mathcal{N}_{\text{eNB}}| \quad (16)$$

$$1 \leq \sum \eta_{i,j}^{\text{UL}} \leq |\mathcal{N}_{\text{eNB}}| \quad (17)$$

$$\sum_{i \in \{1, |\mathcal{N}_{\text{UE}}|\}} \vartheta_{i,j}^{\text{DL}} \leq 1 \quad (18)$$

$$\sum \epsilon_{i,j}^{\text{UL}} \leq 1 \quad \forall \text{UE}_i \in \mathcal{N}_{\text{UE}} \& \forall \text{eNB}_i \in \mathcal{N}_{\text{eNB}} \quad (19)$$

$$\sum \vartheta_i^{\text{DL}} = 1 \cdot x : x \in \{1, |\mathcal{I}^{\text{DL}}|\} \quad (20)$$

$$\sum \epsilon_i^{\text{UL}} = 2 \cdot y : y \in \{1, |\mathcal{I}^{\text{UL}}|/2\} \quad (21)$$

where (12) and (13) ensure that eNB and UE do not exceed the maximum allowed transmit power while (14) and (15) ensure that the data rates are equal or greater than the required threshold values in order to respect the communication QoS for the downlink and uplink, respectively. Cell association in the uplink and downlink is ensured by (16) and (17), respectively.

#### V. PROPOSED SCHEME

The UE is assumed to have some terminal intelligence and ability to establish connectivity with different cells within the LTE-A HetNet assuming the existence of coordination between the different cells in the network. The UE decides which cells to choose for uplink and downlink transmissions such that the energy efficiency is maximized. Algorithm 1 illustrates the implementation at the UE.

First, the UE searches the available eNBs,  $\mathcal{N}_{\text{eNB}}^i$ , that can establish communication with. Next, the UE calculates the channel gain in the uplink and downlink based on the interference followed by the power consumption required to transmit with the required data rate. If the power required does not exceed a certain limit, the eNB is added to the uplink and/or downlink eNB candidates pool,  $\mathcal{N}_{\text{eNB}}^{\text{DL},i}$  and/or  $\mathcal{N}_{\text{eNB}}^{\text{UL},i}$ . The set of candidate eNBs for the uplink and downlink communications are sorted according to the required power for the uplink and downlink transmission. The UE secures resources from  $\mathcal{N}_{\text{eNB}}^{\text{DL},i}$  and  $\mathcal{N}_{\text{eNB}}^{\text{UL},i}$  for the uplink and downlink communications starting with the eNB requiring less power till satisfying the required data rate. The rest of the eNBs are then neglected. The same approach can be applied for NC with only one difference, which is the eNB of the uplink will be the one of the downlink.

#### VI. NUMERICAL RESULTS

This section presents and analyzes the simulation results and outlines energy-saving and capacity improvement of the

**Algorithm 1:** Decision Algorithm at the UE

```

1 begin
2  $\mathcal{N}_{eNB}^i = \text{searchCandidatCells}(\mathcal{N}_{eNB})$ ;
3 if ( $\mathcal{N}_{eNB}^i \neq \text{null}$ ) then
4    $\mathcal{N}_{eNB}^{DL,i} = \{\}$ ;  $\mathcal{N}_{eNB}^{UL,i} = \{\}$ ;
5   for each  $j \in \mathcal{N}_{eNB}^i$  do
6      $H_{i,s,j}^{DL} = \text{measureGain}(I_{s,i}^{UE})$ ;
7      $P_{s,j}^{DL} = \text{calculateDLPower}(H_{i,s,j}^{DL}, R_{i,th}^{DL})$ ;
8      $\mathcal{N}_{eNB}^{DL,i} = \text{add}(j, P_{s,j}^{DL})$ ;
9      $H_{i,s,j}^{UL} = \text{measureGain}(I_{s,i}^{UE})$ ;
10     $P_{s,j}^{UL} = \text{calculateULPower}(H_{i,s,j}^{UL}, R_{i,th}^{UL})$ ;
11     $\mathcal{N}_{eNB}^{UL,i} = \text{add}(j, P_{s,j}^{UL})$ ;
12  sort( $\mathcal{N}_{eNB}^{DL,i}, H_{i,s,j}^{DL}$ ); sort( $\mathcal{N}_{eNB}^{UL,i}, H_{i,s,j}^{UL}$ );

```

TABLE I. DEFAULT PARAMETERS

Parameter	Settings
Area	2-by-2 km
No. of eNBs	16
Bearer Type	Default
Path loss Model	Free space
Transmission Mode	SISO
Frequency Reuse	1
Cyclic Prefix	Normal
Duplexing Mode	FDD
DL Bandwidth	60 (3×20) MHz
UL Bandwidth	40 (2×20) MHz
BLER	10 <sup>-4</sup>
eNB Antenna Type	Omnidirectional
UE Antenna Type	Omnidirectional
macro eNB $P_{j,max}^{tx}$	40 Watt/46dBm
small eNB $P_{j,max}^{tx}$	20 Watt/43dBm
UE $P_{i,max}^{tx}$	125 mWatt/21dBm
$R_{i,th}^{DL}$	5 Mbps
$R_{i,th}^{UL}$	2 Mbps

LTE-A networks with dense 5G deployments. MATLAB simulation results obtained by comparing the performance of a UC model and NC model. We consider a 2-by-2 km area with four LTE-A macro cells of radius 500 m and 12 small cells of radius 125 m. Each macro eNB is placed at the cell center and surrounded by three small eNBs, all eNBs are equipped with omnidirectional antennas. The number of users are varied between 50 and 400 UEs which indicates the load variation. Table I summarizes the default simulation parameters settings.

We evaluate the performance of the UC architecture in term of energy saving and added capacity to the network which ultimately indicate the impact on the QoS. The ES (%) is percentage of the reduction of consumed energy by the system when deploying the UC model to the energy consumed with NC model deployment and is measured as  $ES(\%) = \frac{E_{NC} - E_{UC}}{E_{NC}}$ , where  $E_{NC}$  is the energy consumed with NC deployment and  $E_{UC}$  is the energy consumed with UC deployment. The added capacity indicates the difference between the data rate of the system with UC and NC models.

In 3GPP LTE, channel quality indication values describe a range of targeted MCSs. The overall size of the Transport Block and the number of allocated RBs are given as the effective spectral efficiency. UEs can be associated with one eNB in the downlink and one eNB in the uplink. The UC model brings many attractive advantages. Fig. 2 shows the number of users that are associated with different eNBs in the uplink and downlink according to the UC model.

With UC, one or more eNBs can be utilized the downlink transmission and the uplink transmission from any other eNBs. This enables the UE to handle the asymmetric traffic. Since UE can exchange data in either downlink or uplink utilizing the best portion of spectrum with best channel gain, it can transmit same amount of data with less energy consumption and/or increase system capacity by using higher-order MCS. Fig. 3 and 4 show the energy saving at the eNBs and UEs,

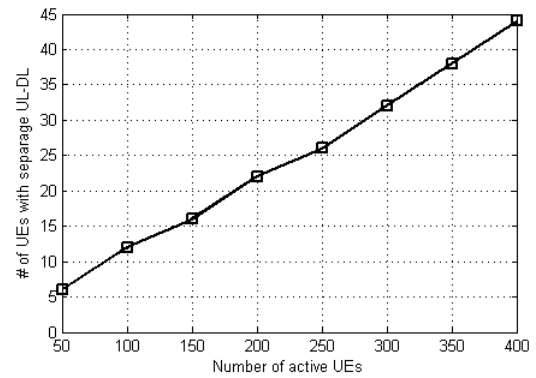


Fig. 2. Number of UEs with separate uplink-downlink connections.

respectively. With 150 active UEs and less, on-off techniques could be implemented to the eNBs which optimize the energy saving, up to 15%. Some eNBs were switched off when active users are 50 and 100 UEs while only one eNB could be switched off with 150 active UEs.

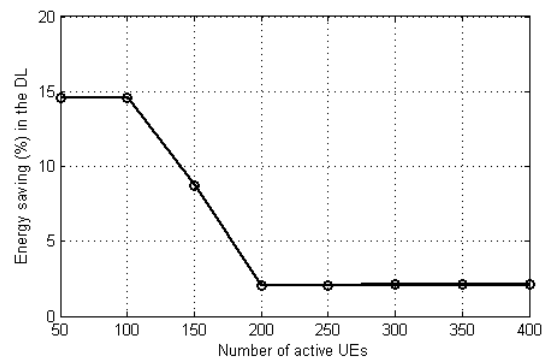


Fig. 3. Energy saving (%) in the downlink (at eNBs).

With implemented UC model, the energy saving at UE's transmit energy is about 5% of the total energy consumed without implementing UC model. Saving energy at the UEs prolong the battery life and has good impact on human health. The data rate in the downlink is generally higher than that of the uplink in LTE, according to [27]. Here, data rate is affected, in both directions, by the MCS order and number of users. Again, better link quality will significantly increase the data rate. The total data rate provided by the network is in factor of Gbps.

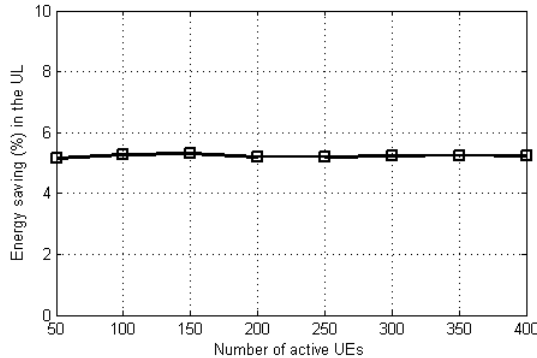


Fig. 4. Energy saving (%) in the Uplink (at UEs).

Fig. 5 shows the added data rate in the uplink and downlink, the added capacity due to the decoupling of the uplink and downlink. The results show that the added data rate is proportional to the number of active UEs and number of UEs with separate uplink/downlink connections. The added capacity is obtained because the number of users served with UC (less UE outage) is higher than that when network-centric model is implemented since UC offers more degree of freedom due to uplink-downlink separation (ability to connect to different multiple eNobeBs) and offer better link quality.

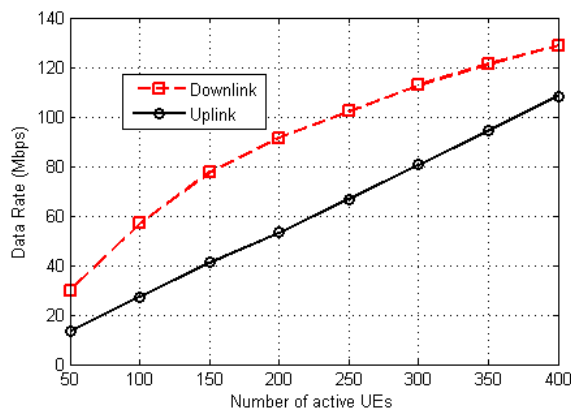


Fig. 5. Added capacity (data rate) to the network.

Table II summarizes the obtained results for the different number of active UEs. With increasing the number of UEs and considering fluctuating radio resources where channel gain is fluctuating, UC model is expected to add more efficiency to the network in terms of added data rates and energy savings.

TABLE II. SUMMARY OF OBTAINED RESULTS

Active UEs	UL/DL UEs	DL ES (%)	UL ES (%)	DL Rate (Mbps)	UL Rate (Mbps)
50	6	14.6	5.15	29.70	13.39
100	12	14.6	5.26	56.94	27.24
150	16	8.73	5.30	77.55	41.09
200	22	2.07	5.18	91.35	53.18
250	26	2.07	5.18	102.29	66.73
300	32	2.09	5.22	112.82	80.59
350	38	2.1	5.25	121.10	94.33
400	44	2.12	5.23	128.84	108.27

### VII. CONCLUSIONS

5G radio access technologies aims to increase the data rates of UEs while reducing the energy consumption per amount of data. User-centric model is foreseen as an interesting feature for minimizing the power consumption at the UEs and eNBs as well. It enables transmission with better link quality and/or, possibly, transmission to the nearest eNB for at least one direction (uplink or downlink) which requires less power for the same amount of data. The results show significant amount of energy savings at the UEs and eNBs. With cooperation between the uplink and downlink, user-centric model adds a degree of freedom to the network planning where a UEs of specific cell can be associated with other cells in uplink and downlink and their cell can be switched off to save energy. Future work include modeling a comprehensive framework for energy efficiency in 5G network including the disruptive 5G features such as massive MIMO. These features can be included to add a degree of freedom to the advanced self organizing 5G network for energy efficiency. The investigation of hybrid/joint user and network centric is also very interesting area of research.

### ACKNOWLEDGMENT

This study is supported by the Fundamental Research Grant Scheme (FRGS), Project: FP007-2016 from Ministry of Higher Education, Malaysia.

### REFERENCES

- [1] B. Bangerter, S. Talwar, R. Arefi, and K. Stewart, "Networks and devices for the 5g era," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 90–96, February 2014.
- [2] J. Andrews, "Seven ways that hetnets are a cellular paradigm shift," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 136–144, March 2013.
- [3] B. Finley and A. Basaure, "Benefits of Mobile End User Network Switching and Multihoming," *CoRR*, vol. 1705.01398, 2017.
- [4] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.
- [5] Y. Kishiyama, A. Benjebbour, T. Nakamura, and H. Ishii, "Future steps of LTE-A: evolution toward integration of local area and wide area systems," *IEEE Wireless Communications*, vol. 20, no. 1, pp. 12–18, February 2013.
- [6] W. H. Chin, Z. Fan, and R. J. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Communications*, vol. 21, no. 12, pp. 106–112, April 2014.
- [7] G. Dandachi, S. E. Elayoubi, T. Chahed and N. Chendeb, "Network-Centric Versus User-Centric Multihoming Strategies in LTE/WiFi Networks," in *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4188-4199, May 2017.

- [8] J. Yelmo, J. del Alamo, R. Trapero, P. Falcarin, J. Yi, B. Cairo, and C. Baladron, "A user-centric service creation approach for next generation networks," in *Innovations in NGN: Future Network and Services, First ITU-T Kaleidoscope Academic Conference*, May 2008, pp. 211–218.
- [9] R. Sofia, P. Mendes, and J. Moreira, Waldir, "User-centric networking: Living-examples and challenges ahead," in *User-Centric Networking*, ser. Lecture Notes in Social Networks, A. Aldini and A. Bogliolo, Eds. Springer International Publishing, 2014, pp. 25–51.
- [10] R. Sofia, "User-centric networking: Bringing the home network to the core," in *User-Centric Networking*, ser. Lecture Notes in Social Networks, A. Aldini and A. Bogliolo, Eds. Springer International Publishing, 2014, pp. 3–23.
- [11] I. Baumgart and F. Hartmann, "Towards secure user-centric networking: Service-oriented and decentralized social networks," in *Fifth IEEE Conference on Self-Adaptive and Self-Organizing Systems Workshops*, Ann Arbor, MI, Oct 2011, pp. 3–8.
- [12] X. Xing, T. Jing, W. Zhou, X. Cheng, Y. Huo, and H. Liu, "Routing in user-centric networks," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 44–51, September 2014.
- [13] M. Katsarakis, G. Fortetsanakis, P. Charonyktakis, A. Kostopoulos, and M. Papadopouli, "On user-centric tools for qoe-based recommendation and real-time analysis of large-scale markets," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 37–43, September 2014.
- [14] S. ping Yeh, S. Talwar, G. Wu, N. Himayat, and K. Johnsson, "Capacity and coverage enhancement in heterogeneous networks," *IEEE Wireless Communications*, vol. 18, no. 3, pp. 32–38, June 2011.
- [15] M. S. ElBamby, M. Bennis and M. Latva-aho, "UL/DL decoupled user association in dynamic TDD small cell networks," in *International Symposium on Wireless Communication Systems*, Brussels, 2015, pp. 456–460.
- [16] A. M. Ghaleb, E. Yaacoub, and D. Chieng, "Physically separated uplink and downlink transmissions in LTE HetNets based on CoMP concepts," in *IET International Conference on Frontiers of Communications, Networks and Applications*, November 2014, pp. 1–7.
- [17] S. Sharafeddine, K. Jahed, and M. Fawaz, "Optimized device centric aggregation mechanisms for mobile devices with multiple wireless interfaces," in *Computer Networks*, Volume 129, Part 1, 2017, Pages 1–16, ISSN 1389–1286.
- [18] Y. Gao, Y. Li, H. Yu, X. Wang, and S. Gao, "Performance analysis of the separation of uplink and downlink under lte-advanced system level simulation: An energy aware point of view," in *2nd International Conference on Computer Science and Network Technology*, Dec 2012, pp. 1289–1293.
- [19] M. Macuha, D. Amgalan, and T. Derham, "Device-centric approach to improve resiliency of emergency communication," in *IEEE Region 10 Humanitarian Technology Conference*, Aug 2013, pp. 124–129.
- [20] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: a disruptive architectural design for 5g networks," *IEEE Global Communications Conference*, Aug 2014, pp. 1798–1803.
- [21] F. Richter, A. Fehske, and G. Fettweis, "Energy efficiency aspects of base station deployment strategies for cellular networks," in *Proc. of the 70th IEEE Vehicular Technology Conference Fall*, Anchorage, Alaska, USA, Sep. 2009.
- [22] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5g wireless communication networks," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122–130, February 2014.
- [23] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.
- [24] J. Lim, H.G. Myung, K. Oh, and D.J. Goodman, "Channel-dependent scheduling of uplink single carrier FDMA systems," in *Proc. of the 64th IEEE Vehicular Technology Conference Fall (VTC 2006-Fall)*, Montreal, Quebec, Canada, Sep. 2006.
- [25] H. G. Myung and D. J. Goodman, *Single Carrier FDMA: A New Air Interface for Long Term Evolution*. Wiley, 2008.
- [26] 3rd Generation Partnership Project (3GPP), "3GPP TS 36.211 3GPP TSG RAN Evolved Universal Terrestrial Radio Access (E-UTRA) Physical Channels and Modulation, version 8.3.0, Release 8," 3GPP, Tech. Rep., 2008.
- [27] A. Ghaleb, D. Chieng, A. Ting, A. Abdulkafi, K.-C. Lim, and H.-S. Lim, "Throughput performance insights of LTE release 8: Malaysia's perspective," in *9th International Wireless Communications and Mobile Computing Conference*, July 2013, pp. 258–263.

# Lifetime Maximization on Scalable Stable Election Protocol for Large Scale Traffic Engineering

Muhammad Asad\*, Arsalan Ali Shaikh, Soomro Pir Dino, Muhammad Aslam and Yao Nianmin  
School of Computer Science and Technology, Dalian University of Technology, Dalian, China

**Abstract**—Recently, Wireless Sensor Networks (WSNs) are getting more fame because of low cost and easy to manage and maintain. WSNs consists of sensor nodes and a Base Station (BS). Sensor nodes are responsible to sense, transmit and receive the data packets from sensing field, and the BS is responsible to collect this data and covert it into readable form. The main issue in this network is lack of power resources. As sensor nodes are restricted to limited energy, so researchers always aims to produce an energy efficient clustered routing protocol. To make the efficient routing protocol, heterogeneity of sensor nodes is a best possible solution. ‘Stable Election Protocol’ was the first heterogeneous network and proposed two level of heterogeneity. SEP Protocol not only improved the network lifetime but also improved the stability of sensor nodes. In order to maximize the network lifetime, we propose the scalability of SEP routing protocol (S-SEP) to check the reliability in large scale networks for traffic engineering. We compare the results of standard SEP routing protocol with fourth level of heterogeneity. Simulation results proves that S-SEP protocol works more better in larger networks.

**Keywords**—Wireless sensor networks (WSN); heterogeneous network; clustered routing protocol; traffic engineering

## I. INTRODUCTION

Wireless Sensor Networks (WSNs) is the unique network consists of small size sensor nodes which contains battery for energy, radio for transmission and reception of data packets, processor for processing on data packets and storage to store the data packets [1]-[3]. These sensor nodes are responsible to sense the sensing field. When the event occurs these sensor nodes sense this event and collect the data. After collection of this data these sensor nodes transmit data packet towards Base Station (BS). As these sensor nodes are limited to energy resources and their embedded batteries are not replaceable once they have deployed in the network, so the design of routing protocol has to be energy efficient in order to maximize the network lifetime [4], [5]. In the past, conventional routing protocols use two featured techniques for data transmission, Direct Transmission (DT) and Minimum Transmission Energy (MTE). In the DT sensor nodes transmit sensed data directly to the BS which results the early death of nodes which are far from BS because of huge amount of energy dissipation due to long range transmission while in MTE, data packets are delivered through routes. Nodes which are supposed to transmit data, forwards the data packets to the nodes which locates near to the BS which results the early death of relay nodes because of huge amount of data transmission [6]-[8]. Both transmission techniques does not suitable for longer network lifetime. LEACH proposed a solution by dividing the network field into clusters and choose a random Cluster Head (CH) in each cluster for a specific round which is responsible

for data transmission of all the nodes locates in its cluster [9]. CH is chosen dynamically through proper CH selection and each node have a equal chance to become CH. This random distribution of energy resources among all the nodes control the network load and produce a better network lifetime. But LEACH was a homogeneous network and all the nodes in the network have the same amount of energy which results the unreliability in the large scale of network. To complete the requirement of some applications of maximum lifetime, SEP was proposed with the heterogeneous network [10]. SEP proposed two level of heterogeneity in which two types of nodes was deployed in the network. First one was normal nodes and the others was advance nodes. Advance nodes have some extra energy resources more than normal nodes and CH selection was purely based on residual energy of sensor nodes. This scheme results in longer lifetime because BS always choose the node as CH which has more energy and rest of the nodes act as member nodes in a specific cluster. Simulation results of SEP routing protocol proved that SEP produce better network lifetime with more stability in the network. But due to the two level heterogeneity SEP does not work for some specific applications. There are some application which require the huge density of sensor nodes in a very large scale network and these application require longer network lifetime due to the cost of sensor nodes. In this regard we proposed Scalability of SEP routing protocol which is enough adaptive to accept new nodes deployed at any time in the network. Our contribution in this paper is to deploy sensor nodes with different level of heterogeneity in different dimensions of network. In order to check the reliability of protocol, we proposes the fourth level of heterogeneity and deployed the SEP routing protocol in four different scenarios. Simulation results proves that scalable SEP extends the stability period and network lifetime in large scale network.

The rest of the paper is organized as follows: Problem definition and detailed explanation of routing protocol is given Section II. Simulation results are discussed in Section III and we conclude this paper in Section IV.

## II. SCALABLE STABLE ELECTION ROUTING PROTOCOL

In this section, we briefly explain the potentials of SEP routing protocol and why it is necessary to be deployed in large scale of networks. This section contains the problem definition and network models which are explained in the subsections.

### A. Problem Definition

As WSNs are emerging from the past decades, the requirements of applications are also getting high with technologies. When SEP routing protocol was proposed, it was easy to

manage due to small scale of applications. But recently, applications like, military surveillance, vehicle monitoring, environmental control, harvest monitoring, etc are getting vast [11]. So the previous protocol is difficult to deploy in such application due to its limited network lifetime. Nowadays, applications requires maximum network lifetime so it work for a longer period. Furthermore, these application are not limited to lifetime of network, they also require the adaptability of network. Because in the past, when the network is completely dead they need to deploy a complete new network. But due to the recent technologies, sensor nodes are enough intelligent to adopt the network at any time [12]. In this regard, we further enhance the scalability of SEP routing protocol to adopt these sensor nodes through CH selection. When the network is about to dead, these intelligent nodes can be added in the network and the whole network will become normal again. Through this addition of sensor nodes, scalability issue is resolved which is proved in simulation results.

### B. Network Topology

In the network model of scalable SEP routing protocol, initially 100 static nodes are deployed in the network of  $100m \times 100m$  sensing field while the energy-free BS is located in the center of the network. All the dimensions of the network is known and BS knows the location of each sensor nodes. As we described earlier that SEP is completely distributed heterogeneous-aware routing protocol, so the CH selection is based on initial energy of sensor nodes for the first round. Heterogeneous nodes are deployed in the network with the initial energy  $1j$  so the probability of being a CH is always between advanced nodes. After finishing of first round, CH selection will based upon remaining energy of sensor nodes. Percentage of advance nodes is 30% and the normal nodes are 70% deployed in the network. Fig. 1 shows the network topology of S-SEP routing protocol, in which normal nodes, advance nodes and BS can be seen clearly while Fig. 2 shows the clustered topology of S-SEP routing protocol.

### C. Heterogeneous Network Model

As we describe earlier, SEP routing protocol is purely heterogeneous network which deploy network in distributed manner. Heterogeneous networks in WSNs are solely proposed to improved the network lifetime. Initially, there were only single type of nodes which has same energy levels. But heterogeneous network provide different levels of energies to sensor nodes which results in better network lifetime [13]. Two level of heterogeneous nodes are deployed in network, one is called as normal nodes and the other is advanced nodes. Normal nodes have the initial of  $0.5j$  while the advanced nodes have initial energy of  $1j$ . This two level of heterogeneity was proposed by SEP routing protocol. In this paper we took heterogeneity to the fourth level with the initial energy of advanced nodes to  $1.5j$  and  $2j$ , respectively.

### D. Radio Model

Sensor nodes depicts the energy in sensing, data collecting, data processing and data transmission but the energy consumed in communication is greater than other dissipation of energies, so the communication energy cannot be negligible. That is why researchers always focus to minimize energy consumed

in transmission process of routing protocols. As we mentioned in the first section that sensor nodes consist of radio and this radio consist of transmitter and receiver shown in Fig. 3. For the fair comparison of lifetime enhancement, we consider first order radio model and free space model which was adopted in previous routing protocols [14]. Proposed scalable SEP routing protocol dissipates  $50n_j/bit$  of energy for transmission and reception of data packet. In order to transmit the data packet ( $dp$ ) at the distance  $d$  the total transmitted energy can be calculated as:

$$T_{TE}(dp, d) = \begin{cases} dp \times E_{elec} + E_{fs}d^2 \times dp, & \text{if } d < d_0. \\ dp \times E_{elec} + E_{amp}d^4 \times dp, & \text{if } d \geq d_0. \end{cases} \quad (1)$$

Where,  $T_{TE}(dp, d)$  is the total transmitted energy and  $E_{fs}d^2$   $E_{amp}d^4$  represents the free space and first order radio models respectively. While,  $dp$  represents the data packet and  $d$  is the distance between source to destination.

### E. Cluster-Head Selection Model

In this section, random CH selection is explained [15]. Algorithm performs the intelligent selection of CH based on residual energy of sensor nodes. First, routing protocol establish the network topology and distribute the nodes equally among clusters to balance the load. After that protocol gives the authority to BS to select the suitable CHs. As described in the network model, two types of nodes are deployed in the network; normal nodes recognized as a  $N_n$  and advance nodes as  $N_a$ . BS calculates the suitability of sensor nodes in order to select the CH. To calculate the suitability of sensor nodes, BS must be familiar with these properties of sensor nodes: initial energy, residual energy, energy consumption ratio ( $ECR$ ) and distance of node to BS. So, in order to calculate the suitability BS needs to calculate the energy consumption ratio of each node in the network. Following equation is used to calculate the  $ECR$ :

$$ECR_N = \frac{E_o}{E_o - E_r} \quad (2)$$

Where,  $N$  denotes the total number of sensor nodes in the network,  $E_o$  is the initial energy and  $E_r$  is the remaining energy of sensor nodes. After getting familiar with  $ECR$  BS will be able to calculate the suitability of each sensor node.

$$Suitability_N = \frac{E_r}{ECR \times d} \quad (3)$$

Where,  $d$  is the distance between sensor nodes and BS. After calculating the suitability, BS selects the desired percentage  $P$  of cluster heads. As SEP routing protocol use the epoch  $N_i = 1/P_{optimum}$  to select CH so after each round new CH will be selected based on residual energy.

When all the CHs are selected, hello packets will be exchanged between all the sensor nodes. At the initialization stage of network, BS transmits the hello packet to all the sensor nodes and the packet contains the node ID, cluster ID, CH and



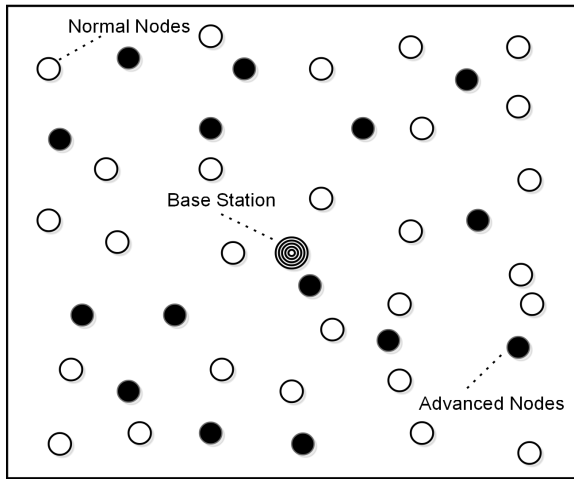


Fig. 1. Network model.

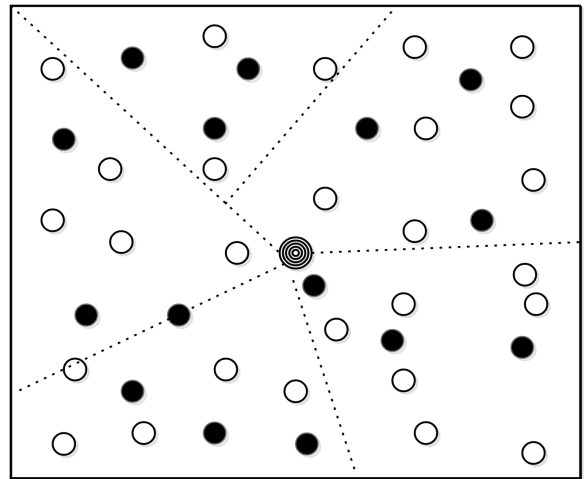


Fig. 2. Clustered.

H

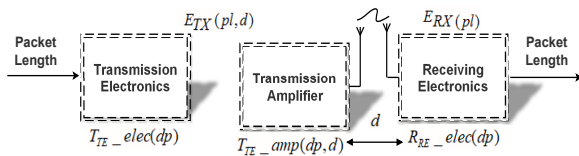


Fig. 3. Radio model.

total network energy. In order to calculate the total network energy following equation is used:

$$Total\ Energy = \sum_{i=1}^{N_i} E_o(1 + \alpha) \quad (4)$$

Average energy of the network will be calculated in the similar fashion:

$$Average\ Energy = \frac{1}{N_i} \sum_{i=1}^{E_i} E_i(A_E) \quad (5)$$

In order to select the appropriate number of CH for the particular nodes, the equation applies:

$$P_{CH} = P_{optimal} \frac{E_i}{E_{i,AE}} \quad (6)$$

$P_{optimal}$  is the percentage to become CH, initially all nodes have same percentage to become CH. As the routing protocol is rotating epoch, so the nodes become CH in first round they won't be able to become CH in the next round because selection of protocol is based on residual energy of sensor nodes, Fig. 4 and 5 shows the selection of CH with respect to rounds. In order to maintain the same probability of CHs in each round, BS choose a random number between [0-1]. If random number is less than value of threshold then the node become CH for the current round otherwise it will be selected as a member node of specific cluster. To calculate the value of threshold, equation applies:

$$Threshold = \begin{cases} \frac{P_{optimal}}{1 - P_{optimal} \times (r \bmod \frac{1}{P_{optimal}})}, & \text{if } N \in G. \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where,  $r$  is the running round,  $N$  are the nodes and  $G$  is the set of CHs and  $d$  is the distance between node to BS. In order to calculate the desired percentage of CH between normal nodes and advance nodes according to their energy levels, the formulation applies:

$$P_{Normal} = \frac{P_{optimal}}{1 + \alpha \times N} \quad (8)$$

$$P_{Advanced} = \frac{P_{optimal}(1 + \alpha \times N)}{1 + \alpha \times N} \quad (9)$$

Equations shows the extra heterogeneous energy in advance nodes. The above two equations shows the two level heterogeneity in standard SEP routing protocol. But in order to add more heterogeneity in the network to make it scalable for large scale, the model applies:

$$P_S = \frac{P_{optimal} \times N(1 + \alpha)}{N + \sum_{i=1}^N \alpha} \quad (10)$$

where  $P_S$  shows the scalability of routing protocol, equation proves that protocol is reliable for large scale network and it is adoptable to add more nodes with different level of heterogeneity in the network.

Above all equation shows the proposed scalability of SEP routing protocol. Models prove that multiple number of heterogeneity can be added in the network to improve the network lifetime and this addition of multiple heterogeneity does not interrupt the CH selection. It is proved that newly added nodes have more chances to become CH due to higher energy and ultimately this feature will enhance the network stability and network lifetime in real-time applications of present day.

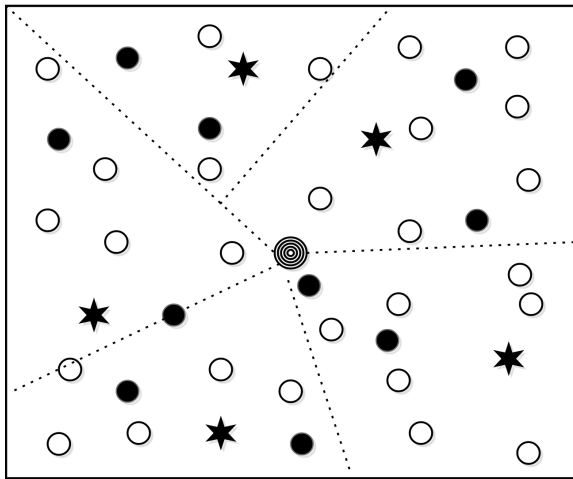


Fig. 4. Selected cluster-heads for running round.

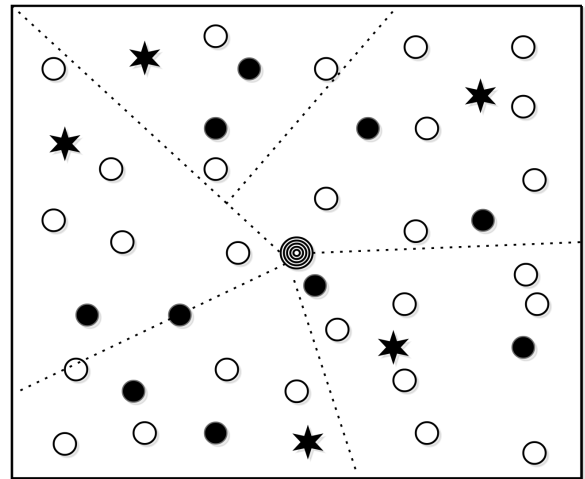


Fig. 5. Selected cluster-heads for running round R++.

### F. Association of Member Nodes to CHs

When all the CHs are chosen, then the association of member nodes with their perspective CH will begin. CH runs the CSMA MAC protocol and sends the hello packet to all neighbour nodes with the flag of CH. All neighbouring nodes will receive this message and response with their location and ID. All the member nodes will remain active during this communication between nodes and CHs. When CHs receive the location and IDs of member nodes it will associate these nodes with itself. In order to calculate the range of CH for associating nodes, the equation applies:

$$Nodes_{association} = \frac{RSSI}{d_{CH}} \quad (11)$$

where,  $RSSI$  denoted the received signal strength indication  $d_{CH}$  shows the distance of CH to member nodes. After the association with CHs, member nodes will receive the TDMA slots from CHs. These slots are allocated to nodes to avoid interruption between communication of member nodes and CHs. Each node will only communicate with CH in its own allocated slot. Only during this time period nodes activate their transmitter to transmit their sensed data to CHs and the rest of the time transmitter will be in sleep mode. This extra property of energy saving will further enhance the network lifetime.

### G. Transmission of Data Packets to BS

After the association of member nodes with their prospective CHs, nodes transmit their sensed data to CH in their allocated time slot. During the whole network time CH will remain active to receive data from sensor nodes and perform the responsibility as a CH. CHs will compress all the data packets using multiple signal processing techniques and aggregate the meaning full data to the BS.

## III. RESULTS DISCUSSION

In this section, we simulate Stable Election Protocol (SEP) in multiple networks with multiple network dimensions. First we simulate SEP routing protocol in MATLAB on different heterogeneity levels and then gather the data in Origin9.1 to

produce the results. In this paper, we check the scalability of SEP routing protocol for large scale network. In order to prove the reliability we took four different scenarios with different parameters and perform the simulations (see Table I). In first scenario, we took 100 nodes in  $100m \times 100m$  network with initial energy  $0.5j$ , which is same as original SEP routing protocol. In second scenario, we took 150 nodes in  $200m \times 200m$  network with initial energy  $1j$ . In third scenario, we took the heterogeneity on next level with  $1.5j$  initial energy while the number of nodes are 200 and the network dimensions are  $250m \times 250m$ .

TABLE I. PARAMETERS USED IN SIMULATION

Parameter	Value
<b>Scenario</b>	<b>1</b>
Number of Nodes	100
Network Dimension	$100m \times 100m$
Initial Energy	$0.5j$
<b>Scenario</b>	<b>2</b>
Number of Nodes	150
Network Dimension	$200m \times 200m$
Initial Energy	$1j$
<b>Scenario</b>	<b>3</b>
Number of Nodes	200
Network Dimension	$250m \times 250m$
Initial Energy	$1.5j$
<b>Scenario</b>	<b>4</b>
Number of Nodes	250
Network Dimension	$300m \times 300m$
Initial Energy	$2j$
Number of rounds	5000
Percentage of being Cluster Head	.1
$E_{DA}$ Energy cost	50pj/bit j
packet size	4000 bit
$E_{TX}$	50nj/bit
$E_{RX}$	50nj/bit
$E_{fs}$	$10pj/bit/m^2$
$E_{amp}$	$100pj/bit/m^2$

Last, but not least we deployed 250 nodes in the field of  $300m \times 300m$  with the initial energy of  $2j$ . In order to further prove the scalability of SEP routing protocol, we took the simulation on higher level of heterogeneity with

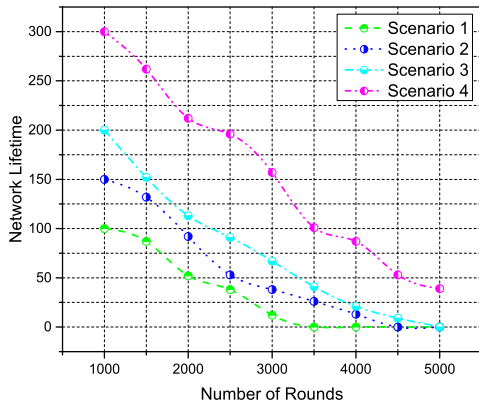


Fig. 6. Network lifetime.

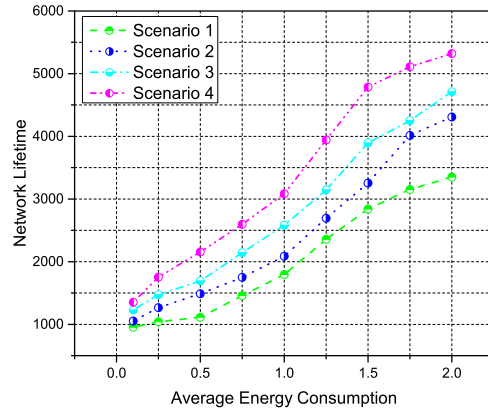


Fig. 8. Average energy consumption.

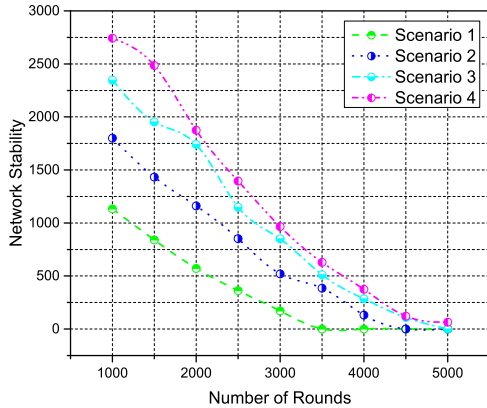


Fig. 7. Network stability.

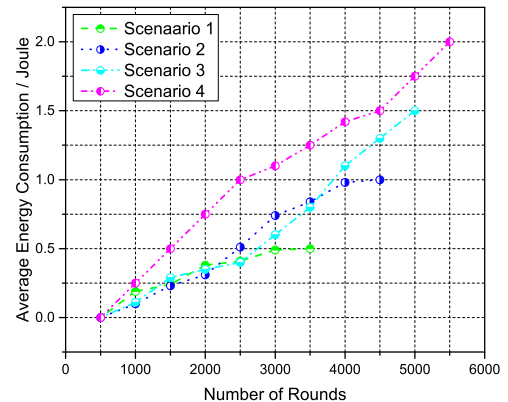


Fig. 9. Average energy consumption per round.

more more sensor nodes in the larger network. Simulation results prove that SEP routing protocol is reliable in large scale network. Detailed simulation parameters are given in Simulation Parameters table.

Fig. 6 shows the network lifetime of SEP routing protocol in four different scenarios. It can be seen that SEP perform more better in large network. Figure shows that in larger network, the network lifetime is 110% increased from the standard SEP routing protocol mentioned in scenario 1.

Network stability is shown in Fig. 7. It can be seen that stability of standard SEP routing protocol is 1110 rounds. Scenario 2 shows the second level of heterogeneity which is 90% higher than original routing protocol with the stability of 1700 rounds. Scenario 3 showing the results in similar fashion, with the stability of 2341 rounds which is 150% improved than standard SEP routing protocol. While in scenario 4, network is stable till 2753 rounds. Figure proves that SEP routing protocol is stable and reliable for large scale network.

As we deploy the network in bigger environment, sensor nodes consumes more energy. Standard SEP routing protocol

consumes minimum energy because of small size of network dimensions. Fig. 8 shows the average energy consumption on different level heterogeneity. Proposed scenarios consumes more energy because of large network size and higher number of nodes.

As we have mentioned in the simulation parameters, we took the four different scenarios to prove the scalability. It can be seen clearly that SEP protocol enhances its network lifetime in large scale. But as the network size increases, sensor nodes consumes more energy because of communication gap between nodes and CHs. Fig. 4 shows the energy consumption with respect to number of rounds. If we focus on scenario 4 in Fig. 9, it can be seen clearly that there is huge difference between standard SEP protocol and scalable SEP protocol. In scalable network, SEP improves network lifetime but consumes more energy which is the only drawback of SEP routing protocol in large scale networks.

Efficient data packet delivery is essential for any routing protocol, Fig. 10 shows the reliability of packet delivery by calculating the delay in network. Figure proves that SEP rout-

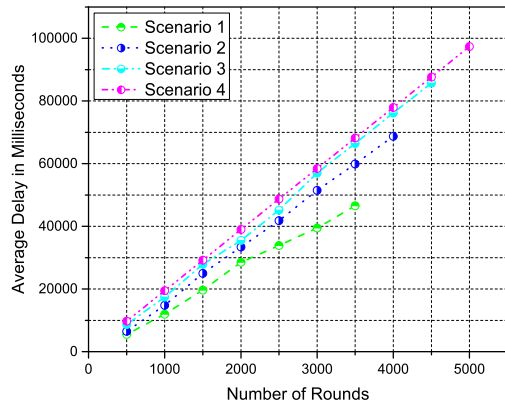


Fig. 10. Network delay.

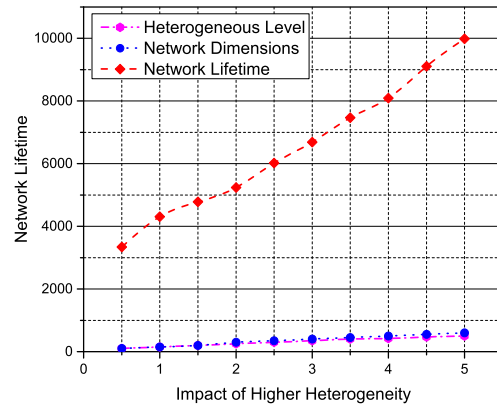


Fig. 12. Impact of higher heterogeneity level.

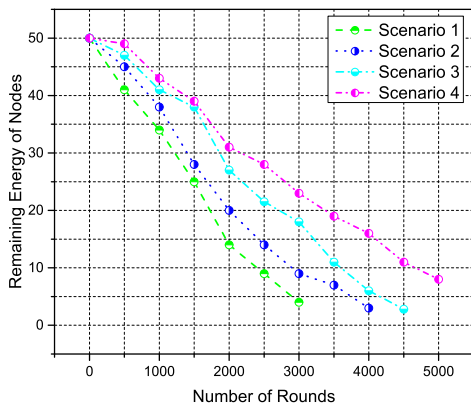


Fig. 11. Remaining energy of sensor nodes.

ing protocol minimize the delay in data packet transmission as well as data packet reception. Figure shows the potential of SEP routing protocol that it enhances the packet delivery ratio by minimizing the average delay in network.

Residual energy of sensor nodes is calculated after every round. As we describe earlier that SEP routing protocol use epoch to select a new CH after each round to balance the load of network. This technique of CH selection is truly based on residual energy of sensor nodes. Nodes with the higher energy level have more probability to become CH. Fig. 11 shows the remaining energy of sensor nodes with respect to nodes. It can be seen clearly that there is no big difference in remaining energy for standard SEP and in large scale. Protocol distribute the network load among all sensor nodes which makes the reliability in large scale network. Small variation in figure is because of larger network and large number of sensor nodes.

After simulating SEP protocol in above mentioned 4 scenarios, we took the heterogeneity from  $0.5 - 5j$  for better comparison. Fig. 12 shows the impact of different heterogeneity levels with respect to different number of sensor nodes

and different level of network dimensions. Figure shows the stability and reliability of SEP routing protocol that it works more stable in large scale network. It can be seen that network lifetime also increases as we increase the heterogeneity level. Figure shows the potentials of routing protocol and it also proves the objective of this paper.

#### IV. CONCLUSION

We propose the scalability of SEP routing protocol with the fourth level of heterogeneity for large scale traffic monitoring. The selection of cluster-heads are independent and elected efficiently through random CHs selection technique. Two types of nodes are deployed in the network with different heterogeneous level as similar in the SEP routing protocol. In order to prove the reliability in large scale network, we deployed four different types of heterogeneous nodes in the simulation experiments. Results shows the behavior of SEP routing routing protocol in large scale network. After proving the reliability in four different network dimensions, we further check the impact of heterogeneity in huge network. Where, the heterogeneity is extended to the 10th level and network dimension is extended to 500% greater than the standard SEP routing protocol. Experimented results shows that SEP routing protocol works more better in large scale network and enhances the lifetime with more stable network.

#### ACKNOWLEDGMENT

We are thankful to the authors of SEP routing protocol for providing code and results online.

#### STATEMENT OF CONFLICT

Authors of this paper: Muhammad Asad, Arsalan Ali Shaikh, Soomro Pir Dino, Muhammad Aslam and Yao Ni-anmin declares that there is no conflict of interest regarding the publication of this research article entitled "Lifetime Maximization on Scalable Stable Election Protocol for Large Scale Traffic Engineering".

## REFERENCES

- [1] Kafi, Mohamed Amine, Jalel Ben Othman, and Nadjib Badache. "A Survey on Reliability Protocols in Wireless Sensor Networks." *ACM Computing Surveys (CSUR)* 50.2 (2017): 31.
- [2] A. Roshini and H. Anandakumar, Hierarchical Cost Effective LEACH for Heterogeneous Wireless Sensor Networks, in *Advanced Computing and Communication Systems*, 2015 International Conference on, Jan 2015, pp. 17.
- [3] M. Pramanick, P. Basak, C. Chowdhury, and S. Neogy, Analysis of energy efficient wireless sensor networks routing schemes, in *Fourth International Conference of Emerging Applications of Information Technology (EAIT)*, 2014, Dec 2014, pp. 379384.
- [4] Rehan, Waqas, et al. "A comprehensive survey on multichannel routing in wireless sensor networks." *Journal of Network and Computer Applications* 95 (2017): 1-25.
- [5] Wang, J.; Zhang, Z.; Xia, F.; Yuan, W.; Lee, S. An energy efficient stable election-based routing algorithm for wireless sensor networks. *Sensors* 2013, 13, 1430114320.
- [6] Wang, Jin, et al. "Analysis of energy consumption in direct transmission and multi-hop transmission for wireless sensor networks." *Signal-Image Technologies and Internet-Based System*, 2007. *SITIS'07*. Third International IEEE Conference on. IEEE, 2007.
- [7] Hammoudeh, M.; Newman, R. Adaptive routing in wireless sensor networks: QoS optimisation for enhanced application performance. *Information Fusion* 2015, 22, 315.
- [8] M Shamsan Saleh, A.; Mohd Ali, B.; A Rasid, M.F.; Ismail, A. A self-optimizing scheme for energy balanced routing in wireless sensor networks using sensorant. *Sensors* 2012, 12, 1130711333.
- [9] Heinzelman, Wendi Rabiner, Anantha Chandrakasan, and Hari Balakrishnan. "Energy-efficient communication protocol for wireless microsensor networks." *System sciences*, 2000. *Proceedings of the 33rd annual Hawaii international conference on*. IEEE, 2000.
- [10] Smaragdakis, Georgios, Ibrahim Matta, and Azer Bestavros. *SEP: A stable election protocol for clustered heterogeneous wireless sensor networks*. Boston University Computer Science Department, 2004.
- [11] Al-Anbagi, Irfan, Melike Erol-Kantarci, and Hussein T. Mouftah. "A survey on cross-layer quality-of-service approaches in WSNs for delay and reliability-aware applications." *IEEE Communications Surveys & Tutorials* 18.1 (2016): 525-552.
- [12] Chang, Feng-Cheng, and Hsiang-Cheh Huang. "A survey on intelligent sensor network and its applications." *Journal of Network Intelligence* 1.1 (2016): 1-15.
- [13] Rostami, Ali Shokouhi, et al. "Survey on clustering in heterogeneous and homogeneous wireless sensor networks." *The Journal of Supercomputing* (2017): 1-47.
- [14] Alwajeeh, Taha, et al. "Efficient method for associating radio propagation models with spatial partitioning for smart city applications." *Proceedings of the International Conference on Internet of things and Cloud Computing*. ACM, 2016.
- [15] Fujii, Shohei, et al. "Optimal cluster head selection and rotation of cognitive wireless sensor networks for simultaneous data gathering." *Information Networking (ICOIN)*, 2017 International Conference on. IEEE, 2017.

## AUTHORS' PROFILE

**Muhammad Asad** received his B.S. degree in Telecommunication and Networks from COMSATS Institute of Information Technology, WAH CANTT, Pakistan in 2014. Now he is pursuing his masters degree in School of Computer Science and Technology, Dalian University of Technology, China. His main research interests include Wireless Sensor Networks, Internet of Things, and information security.

**Arsalan Ali Shaikh** was born in (1990). He received his Bachelor Degree in Software Engineering from the University of Sindh Jamshoro Pakistan in (2012). After that he served as a Software Developer in private software organization in 2013. Currently, he is doing Master Degree in Computer Science and application technology in Dalian University of Technology, P.R. China. His research interest area include Big Data, Cloud Computing and Computer Networks.

**Soomro Pir Dino** was born in (1989); He received his Bachelor Degree in Bachelor of Science Information Technology BSIT (Hons.) from the Sindh Agriculture University (SAU) Tandojam, Pakistan in 2013. Currently he is doing Masters Degree (MS) in Computer Science and Technology in Dalian University of Technology (DUT), Dalian, P. R. China. His research Interest is Data Mining, Machine Learning and Interconnected Networks and Topological Structure.

**Muhammad Aslam** received the B.S degree in Telecommunication System and the M.S degree in Electrical Engineering from BZU Multan and COMSATS Institute of Information Technology Islamabad in 2010, 2012, respectively. He was Lecturer at COMSATS Institute of Information Technology, Wah Cantt. His major research interests are Energy optimization in WSNs, WBANs, and UWSNs. Currently he is PhD scholar at School of Computer Science and Technology, Dalian University of Technology, Dalian, China.

**Yao Nianmin** received the B.E., M.S. and Ph.D degree from Jilin University. He is currently a professor in Dalian University of Technology. He has been a visiting scholar at University of Connecticut. His Primary research interests include Wireless Sensor Networks and Wireless Network Security.

# Comparative Analysis of Raw Images and Meta Feature based Urdu OCR using CNN and LSTM

Asma Naseer, Kashif Zafar  
Computer Science Department  
National University of Computer and Emerging Sciences  
Lahore, Pakistan

**Abstract**—Urdu language uses cursive script which results in connected characters constituting ligatures. For identifying characters within ligatures of different scales (font sizes), Convolution Neural Network (CNN) and Long Short Term Memory (LSTM) Network are used. Both network models are trained on formerly extracted ligature thickness graphs, from which models extract Meta features. These thickness graphs provide consistent information across different font sizes. LSTM and CNN are also trained on raw images to compare performance on both forms of inputs. For this research, two corpora, i.e. Urdu Printed Text Images (UPTI) and Centre for Language Engineering (CLE) Text Images are used. Overall performance of networks ranges between 90% and 99.8%. Average accuracy on Meta features is 98.08% while using raw images, 97.07% average accuracy is achieved.

**Keywords**—Long Short Term Memory (LSTM); Convolution Neural Network (CNN); OCR; scale invariance; deep learning; ligature

## I. INTRODUCTION

The recognition of optical characters of cursive scripts always captures attention of computer scientists and linguists. Urdu is among those languages which use cursive script for writing. It is a widely spoken language which has 60,000,000 to 70,000,000 native speakers all around the world [3]. In Pakistan and in a few states of India, it is official language. Script of Urdu is Arabic, and Nastalique is the most popular font face. In this writing style, characters and ligatures are written at an angle of  $45^\circ$ . Same character differs in shape due to certain reasons such as position of character within a ligature which can be start, middle or end. Adjacent characters also affect shape of same character. When same character joins a different character, it gets a different shape. In isolation, shape of a character is also different from its other shapes. Due to so many shapes of single character, there are so many unique shapes (characters and ligatures) in language. Characters are also overlapping which makes it hard to segment ligatures. Due to such characteristics of Urdu language and its script, Urdu OCR is still an open to research problem.

In this research, Urdu optical characters are recognized by using two different forms of inputs. The first form of input is raw images while the second form of input is graphs of ligature

thickness. Two neural networks, LSTM and CNN, are trained on both these forms of input. Networks extract features from raw images and Meta features from formerly plotted graphs of ligature thickness. At the end, performance of networks using Meta features and raw images is compared and analysed.

LSTM shows considerably better results due to three factors, i.e. equation updating mechanism, memory cell and back propagation dynamics [4]. In this network, LSTM units are used instead of normal nodes. It contains memory cells consisting of self-feedback loops and three adaptive gates i.e. input, output and forget gate as can be seen in Fig. 1. Throughout training, forget gate takes decisions and eventually only important and relevant information is used for next iterations. Irrelevant or less important information is discarded [5], [6].

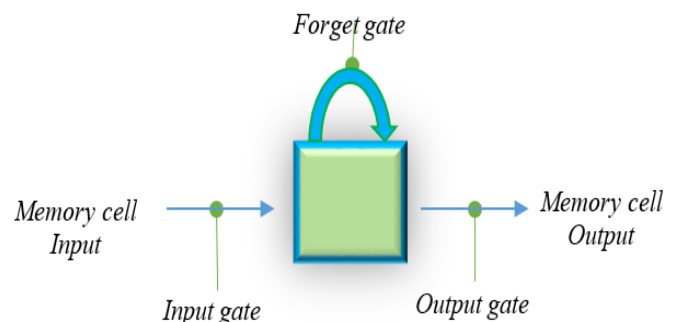


Fig. 1. LSTM, an illustration of memory cell.

CNN consists of different layers such as convolution layer, Rectified Linear Unit (ReLU) layer, SoftMax layer, Pooling layer, fully connected layer, input layer and output layer as can be seen in Fig. 2. These layers extract more relevant features from any type of input either multifaceted or simple. Convolution layers extract features in a progressive manner and pooling layers downscale feature space and memory. Fully connected layers connect two consecutive layers. As multiplication of small numbers arouses vanishing gradient problem so to resolve it ReLU layer is used. Other than these layers there are some more layers such as network in network layer, classification layer and dropout layer [6]-[8].

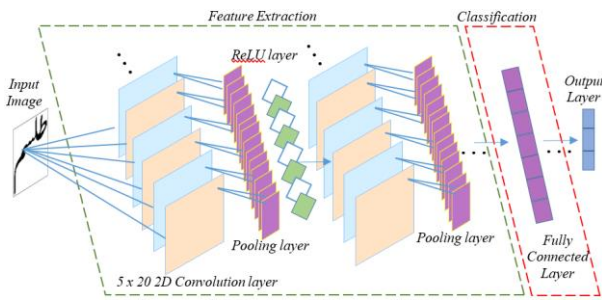


Fig. 2. An Illustration of CNN

Major contributions of this research are:

- Developing Meta Features
- Exploring scale invariance nature of ligature thickness graphs
- Analyzing and comparing performance of networks for Meta features and raw images.
- Comparing performance of CNN and LSTM
- Increasing accuracy of Urdu OCR up till 99.33%

Organization of paper is in such a way that, related work is described in Section II, Section III presents proposed methodology, results are reported in Section IV, Section V describes comparative analysis and finally conclusion and future work is given in Section VI.

## II. RELATED WORK

Although OCR is not a new field for research, and for certain languages like printed English, it is almost a solved problem, still for languages like Urdu; a lot of research is required. In early versions of Urdu OCR systems, Discrete Cosine Transformation (DCT) is calculated using overlapping windows [21]. These features are used to train Hidden Markov Model (HMM). As this system was trained only for a few classes of Urdu characters so it is further extended and all classes of Urdu characters are used [20]. In both these research work, artificially created data set of font size 36 is used. The system shows good performance on 36 font size but variation in font size affects accuracy. To handle font size variation, Tesseract is trained for different font sizes ranging from 12 to 36 [22]. Although performance of previous systems increased but for each font size an independent model is trained.

As there is a paradigm shift due to outstanding performance of deep learning algorithms, usage of deep networks in OCR becomes popular. Multilayer perceptron, CNN, Recurrent Neural Networks (RNN), LSTM and its variations are extensively used for OCR [9]-[13]. Raw images of UPTI [1] data set are used to train Multi-dimensional LSTM [13]. For the same data set, CNN is also trained and 98.1% accuracy is achieved. Text images are divided into different zones and 2DLSTM is trained to identify Urdu text lines [14]. This variation of LSTM is trained on density of pixels and it shows 93.39% accuracy. Other than Urdu, deep networks are widely used for other languages. Segmentation free data set, created synthetically for English and Greek, is used to train LSTM and 98.81% accuracy is achieved [15]. In a similar approach Bi-

directional LSTM is trained for Fraktur (German) and English and 99.4% and 98.36% accuracy is achieved for English and German respectively [16]. LSTM shows 99% average accuracy for French, German and English [17]. This network is trained on normalized raw pixels values.

Besides deep networks, OCRopus, Tesseract and OCRoRACT are also very popular for different languages such as Latin, Urdu and Devanagri [9], [18], [19]. Segmentation based (when ligatures are divided into characters) and segmentation free (ligature based), both approaches are used for OCR systems for these languages. OCRopus, based on LSTM, is also trained on raw pixel values of Devanagri text images and 91% accuracy is achieved [11].

## III. PROPOSED METHODOLOGY

For Urdu character recognition of different font sizes, two deep networks are used. Both networks are trained on raw images as well as on extracted feature i.e. ligature thickness graphs. CLE [2] and UPTI [1] text images are used for training and testing.

### A. Corpora

Two Corpora, CLE Text Images and UPTI are used to train networks. From both corpora, in total 78,714 instances of text images are used. CLE corpus contains images from font size 14 to 40. For developing this corpus, 2912 text documents are scanned. In these text images, text font face is Nastalique. It is full of variety in different dimensions such as different eras, different paper quality, different printing, different ink quality and different domains etc. UPTI Corpus is a collection of 10 thousands text lines. From these lines, 771,339 characters are grouped in 44 classes.

### B. Meta Feature Extraction

To recognize characters of different font sizes, thickness of ligatures and characters is measured. Trend of increase and decrease in thickness value remains same across different font sizes. When ligature thickness graphs are fed to networks they extract Meta features from them.

#### 1) Extracting Thickness:

To extract thickness of ligatures and characters, certain steps are performed one after another. At first images containing text are converted into binary images with value only 1 (white) and 0 (black). From binary images, skeleton (centre line) of ligatures are obtained. At each point of skeleton, tangent line is attained by using current pixel and its next neighbouring pixel. For getting tangent line (1) and (2) are used. After that, perpendicular (normal line) at each tangent line is calculated using equation (3). This normal line is traversed and all back pixels on this line, in binary image, are counted. Number of pixels on this line are stored as thickness of ligature at that point.

$$g_i = \frac{\partial y}{\partial x} = \frac{(y_i - y_{i+1})}{(x_i - x_{i+1})} \quad (1)$$

$$t_i = y_i - y_{i+1} = g_i (x_i - x_{i+1}) \quad (2)$$

$$n_i = y_i - y_{i+1} = -\frac{1}{g_i} (x_i - x_{i+1}) \quad (3)$$

Where,  $g_i$  is gradient,  $t_i$  is tangent and  $n_i$  is normal line at  $i^{\text{th}}$  point.

For each image, a feature vector containing thickness is extracted. Variation in font size and ligature length results in feature vectors which are having different length. To make feature vectors of same size, normalization is performed. Vectors are either scaled up or scaled down while considering average length of all vectors. Once vectors of same length are obtained, graphs are plotted. To get graphs symmetrical for same character or ligature, smoothing is performed by using (4). Smoothing also decreases signal to noise ratio. For smoothing  $\pm 5$  values are considered.

$$S_i = \left\{ \begin{array}{l} \frac{1}{p \times 2 + 1} \sum_{k=i-p}^{i+p} t_k \quad | p \leq i \leq s - p \\ \frac{1}{q \times 2 + 1} \sum_{k=1}^l t_k \quad | i < p, 1 \leq q \leq p - 1 \\ \frac{1}{q \times 2 + 1} \sum_{k=q}^s t_k \quad | i > s - p, s - p + 1 \leq q \leq s \end{array} \right\} \quad (4)$$

Where,  $s$  is length of vector,  $t_i$  and  $S_i$  are thickness values at  $i^{\text{th}}$  point, before and after smoothing, respectively.

Finally, graphs of thickness are plotted and stored as images. The complete process of extracting thickness of ligatures and plotting them is given in Fig 3.

### 2) Raw Images

To explore effects of Meta features (ligature thickness graphs), raw images are also fed to networks. As images are of different sizes due to different font sizes and ligature length so all images are resized as per average height and width.

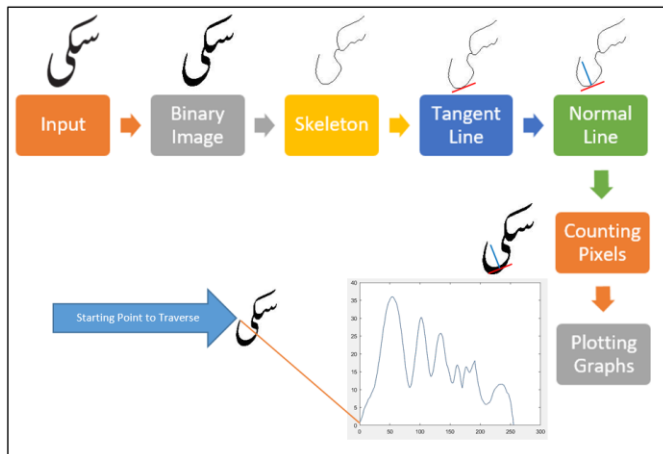


Fig. 3. Process of extracting thickness graphs.

### C. Training

Two deep learning models, i.e. LSTM and CNN, are trained using thickness graphs and raw images so in total four models are trained. Eventually, performance of trained models is evaluated and compared. For training 80% data and for testing 20% data is used.

#### 1) LSTM

At first, LSTM is trained on graphs of ligature thickness. Network is created with fully connected layers. Total number

of fully connected layers are same as total number of classes. Input and output size are set as 16 and 100, respectively. Sigmoid is chosen as activation function of network. Network extracts Meta features from pre extracted features for training the model. Gating mechanism, mainly responsible for maintaining memory related issues, is handled using (5) to (9).

$$M_t^{\sim} = \tanh(W^M x_t + U^M h_{s_{t-1}}) \quad (5)$$

$$Fg_t = \sigma(W^{Fg} x_t + U^{Fg} h_{s_{t-1}}) \quad (6)$$

$$M_t = Fg_t \circ M_{t-1} + i_t \circ M_t^{\sim} \quad (7)$$

$$Og_t = \sigma(W_o x_t + U_o h_{s_{t-1}}) \quad (8)$$

$$h_{s_t} = Og_t \circ \tanh(M_t) \quad (9)$$

Where,  $M_t^{\sim}$  is new memory cell,  $M_t$  is final memory cell,  $Fg_t$  is forget gate,  $Og_t$  is output gate and  $h_{s_t}$  is hidden state.

Once LSTM is trained on thickness graphs, another model of same network is trained using raw images. Both models use same values for all properties and settings so that only thing that varies should be input.

#### 2) CNN

CNN is also trained on both types of input, i.e. ligature thickness graphs and raw images. For both types of input, two independent models are trained while keeping all parameters same. For CNN network 2D convolution layers are created. Size of layer is set as  $5 \times 20$ . Other than convolution layer, input layer, ReLU layer, maximum pooling layers, softmax layer, classification layer and fully connected layers are also created. Size of batch is set as per total number of classes. Stochastic gradient descent with momentum is used. For convolution operation (10) is used.

$$op[i, j] = (w \times g)[i, j] = \sum_{p=-p}^p \sum_{q=-q}^q w[p, q] g[i + p, j + q] \quad (10)$$

Where,  $w$  is weight and  $g$  is input image.

## IV. RESULTS

After training networks, they are tested with 20% of data. Results of CNN and LSTM are described in Table I through Table IV. Total 16 results are reported for each model, considering length of ligature, corpus and ligature and character level accuracy.

Average performance of CNN for thickness input graphs, ranges from 94.91% to 99.22%. On the other hand, average performance of CNN for raw images is between 91.14% and 97.49%. Detail of results for this model, as per character and ligature length, can be seen in Tables I and II..

Table I describes accuracies, achieved by CNN for thickness graphs. CNN reveals same results for character recognition and ligature recognition for UPTI dataset but for CLE dataset, character level accuracy is better than ligature level accuracy by 2.27%. Results, shown by this model for raw images, also have the same trends. Accuracy for character recognition and ligature recognition is almost the same for UPTI dataset. For CLE dataset, character level accuracy is better than ligature level by 2.27%.



TABLE. I. RESULTS OF CNN FOR THICKNESS GRAPHS INPUT

Convolution Neural Network Input: Thickness Graphs				
Ligature Length	UPTI		CLE	
	Character Accuracy	Ligature Accuracy	Character Accuracy	Ligature Accuracy
1 Character	96.88%	96.88%	100%	100%
2 Characters	100%	100%	98.78%	98.44%
3 Characters	100%	100%	92.94%	88.71%
4 Characters	100%	100%	97.00%	92.50%
<b>Average</b>	<b>99.22%</b>	<b>99.22%</b>	<b>97.18%</b>	<b>94.91%</b>

TABLE. II. RESULTS OF CNN FOR RAW IMAGES INPUT

Convolution Neural Network Input: Raw Images				
Ligature Length	UPTI		CLE	
	Character Accuracy	Ligature Accuracy	Character Accuracy	Ligature Accuracy
1 Character	90.63%	90.63%	100%	100%
2 Characters	100%	100%	97.04%	96.24%
3 Characters	100%	100%	95.87%	94.64%
4 Characters	99.32%	99.20%	79.80%	73.68%
<b>Average</b>	<b>97.49%</b>	<b>97.46%</b>	<b>93.18%</b>	<b>91.14%</b>

Average performance of LSTM for thickness graphs is 100% for UPTI data set. For CLE data set average performance on same input is 96.19% for ligatures and 97.93% for characters. For raw images, LSTM reveals 98.16% to 100% accuracy as can be seen in Tables III and IV.

Table III presents accuracy shown by LSTM for thickness graphs. The model shows similar results for character and ligature level recognition for UPTI dataset. For CLE dataset, character level accuracy is 1.74% better than ligature level accuracy. In Table IV, performance of LSTM on raw images is described. LSTM reveals better accuracy for character level recognition by 0.2% for UPTI dataset, while for CLE dataset character level recognition is better than ligature level recognition by 1.22%.

For thickness graphs input, CNN shows 100% accuracy for 8/16 results while for raw images 6/16 results are 100% accurate. LSTM reveals 100% accuracy for all the eight results of UPTI corpus when thickness graphs are used. In total  $\frac{10}{16}$  results are 100% accurate. For raw images LSTM provides  $\frac{16}{16}$  100% results.

TABLE. III. RESULTS OF LSTM FOR THICKNESS GRAPHS INPUT

Long Short Term Memory Network Input: Thickness Graphs				
Ligature Length	UPTI		CLE	
	Character Accuracy	Ligature Accuracy	Character Accuracy	Ligature Accuracy
1 Character	100%	100%	100%	100%
2 Characters	100%	100%	99.35%	98.76%
3 Characters	100%	100%	99.28%	98.76%
4 Characters	100%	100%	93.10%	87.25%
<b>Average</b>	<b>100.00%</b>	<b>100.00%</b>	<b>97.93%</b>	<b>96.19%</b>

TABLE. IV. RESULTS OF LSTM FOR RAW IMAGES INPUT

Long Short Term Memory Network Input: Raw Images				
Ligature Length	UPTI		CLE	
	Character Accuracy	Ligature Accuracy	Character Accuracy	Ligature Accuracy
1 Character	100%	100%	100%	100%
2 Characters	100%	99.20%	99.23%	99.17%
3 Characters	100%	100%	99.63%	96.75%
4 Characters	100%	100%	98.64%	96.71%
<b>Average</b>	<b>100.00%</b>	<b>99.80%</b>	<b>99.38%</b>	<b>98.16%</b>

Overall average performance of both networks is 98.08% for thickness graphs and 97.07% for raw images. Average performance of networks for raw images and thickness graphs for both corpora can be visualized in Fig. 4.

## V. COMPARATIVE ANALYSIS

During this research two networks are trained on different forms of inputs from same data set so that effects of thickness graphs (Meta features) on deep networks can be analyzed. CNN reveals 2.82% better results for thickness graphs as compare to raw images. On the other hand LSTM reveals better performance for raw images by 0.80%. In Fig. 4, comparison of thickness graphs and raw images is given. In six out of eight experiments in which ligatures from length 1 to 4 are tested from both the corpora, accuracy of thickness graphs is better than raw images.

Results also reveal that performance of all the models for UPTI dataset is better in a consistent way. The reason is, UPTI contains cleaner (less noise) corpus as compare to CLE. CLE image dataset is also full of variations in different dimensions which results in a bit lesser accuracy as compare to UPTI dataset.

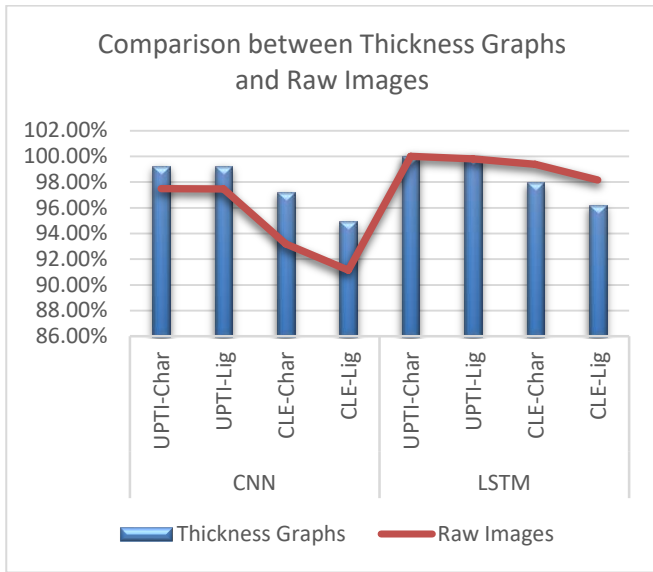


Fig. 4. Comparison between Thickness Graphs and Raw Images

Average performance of both networks is 97.07% for raw images and 98.08% for thickness graphs as can be seen in Fig. 5. Thickness graphs got better performance by 1.01%.

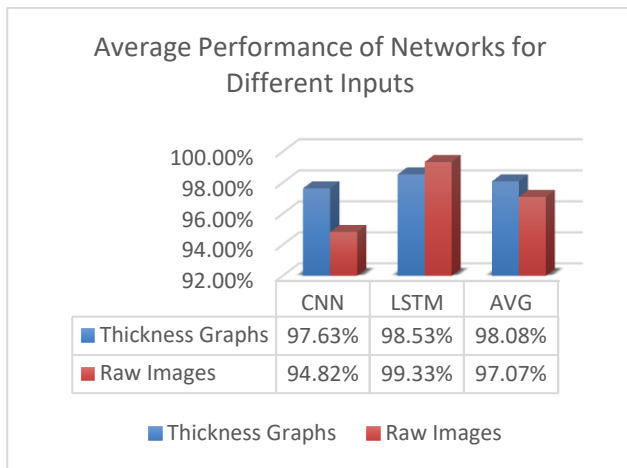


Fig. 5. Average performance of networks.

## VI. CONCLUSION AND FUTURE WORK

In this research thickness graphs are used to extract Meta features to train deep networks. Thickness graphs show trend of thickness for a particular character or ligature. In certain cases, thickness graphs of different characters and ligatures may show same trends as can be seen in Fig. 6. In this figure two graphs are plotted for thickness of character 'alif' ا and 'bay' ب. Both graphs show almost same trend of thickness as low high and then again low. With a little bit more smoothing they may appear more similar. Such scenarios may arouse error. To avoid it, more features can be extracted such as direction of next pixel. In both graphs, illustrated in Fig. 6, there is same trend of thickness but it's in different direction. For letter 'alif' direction is vertical while for letter 'bay' direction is horizontal.

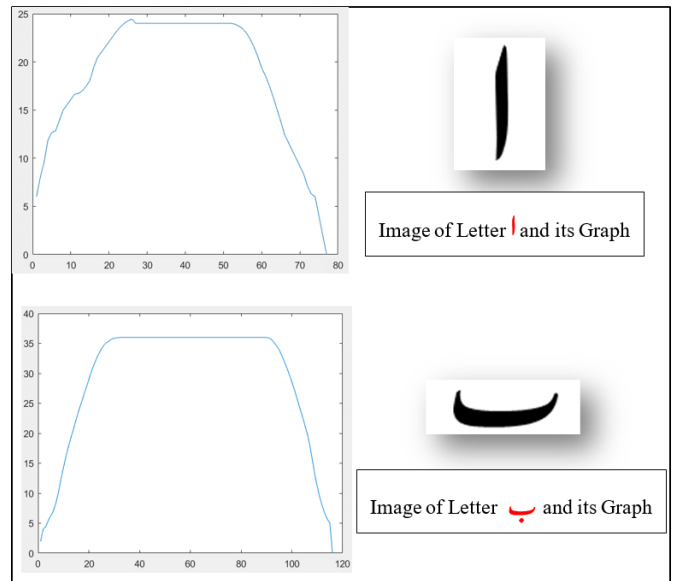


Fig. 6. Same trends in thickness graphs of letter 'alif' and letter 'bay'

Although overall average performance of networks shows that Meta features may bring more accuracy to deep learning frameworks, still more experiment can be carried out to explore its effects. In this research only one feature, thickness graphs, is used. This feature can be combined with more geometric or statistical features and results can be analysed.

## REFERENCES

- [1] Sabbour, Nazly, and Faisal Shafait. "A segmentation-free approach to Arabic and Urdu OCR." In DRR, p. 86580N. 2013.
- [2] Qurrat-ul-Ain, Niazi A., Farrah Adeeba, Urooj S., Sarmad Hussain and Shams S., "A Comprehensive Image Dataset of Urdu Nastalique Document Images", in the Proceedings of Conference on Language and Technology 2016 (CLT 16), Lahore, Pakistan
- [3] Rahman, Tariq. "From Hindi to Urdu: A social and political history." Orientalistische Literaturzeitung 110, no. 6 (2015).
- [4] Senior, Andrew W., and Anthony J. Robinson. "Forward-backward retraining of recurrent neural networks." In Advances in Neural Information Processing Systems, pp. 743-749. 1996.
- [5] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term \*memory." Neural computation 9, no. 8 (1997): 1735-1780.
- [6] Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." Neural networks 61 (2015): 85-117.
- [7] De Vries, Bert, and José Carlos Príncipe. "A theory for neural networks with time delays." In Advances in neural information processing systems, pp. 162-168. 1991.
- [8] LeCun, Yann, Patrick Haffner, Léon Bottou, and Yoshua Bengio. "Object recognition with gradient-based learning." Shape, contour and grouping in computer vision (1999): 823-823.
- [9] Ul-Hasan, Adnan, Syed Saqib Bukhari, and Andreas Dengel. "OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters." In DAS, pp. 174-179. 2016.
- [10] Ul-Hasan, Adnan. "Generic Text Recognition using Long Short-Term Memory Networks." (2016).
- [11] Karayil, Tushar, Adnan Ul-Hasan, and Thomas M. Breuel. "A segmentation-free approach for printed Devanagari script recognition." In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pp. 946-950. IEEE, 2015.
- [12] Ul-Hasan, Adnan, Syed Saqib Bukhari, Faisal Shafait, and Thomas M. Breuel. "OCR-Free Table of Contents Detection in Urdu Books." In Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, pp. 404-408. IEEE, 2012.

- [13] Naz, Saeeda, Arif I. Umar, Riaz Ahmad, Imran Siddiqi, Saad B. Ahmed, Muhammad I. Razzak, and Faisal Shafait. "Urdu Nastaliq recognition using convolutional–recursive deep learning." *Neurocomputing* 243 (2017): 80-87.
- [14] Naz, Saeeda, Saad Bin Ahmed, Riaz Ahmad, and Muhammad Imran Razzak. "Zoning features and 2DLSTM for Urdu text-line recognition." *Procedia Computer Science* 96 (2016): 16-22.
- [15] Ul-Hasan, Adnan, Muhammad Zeshan Afzal, Faisal Shafait, Marcus Liwicki, and Thomas M. Breuel. "A sequence learning approach for multiple script identification." In *Document Analysis and Recognition (ICDAR)*, 2015 13th International Conference on, pp. 1046-1050. IEEE, 2015.
- [16] Breuel, Thomas M., Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. "High-performance OCR for printed English and Fraktur using LSTM networks." In *Document Analysis and Recognition (ICDAR)*, 2013 12th International Conference on, pp. 683-687. IEEE, 2013.
- [17] Ul-Hasan, Adnan, and Thomas M. Breuel. "Can we build language-independent OCR using LSTM networks." In *Proceedings of the 4th International Workshop on Multilingual OCR*, p. 9. ACM, 2013.
- [18] Naz, Saeeda, Arif I. Umar, Riaz Ahmad, Saad B. Ahmed, Syed H. Shirazi, Imran Siddiqi, and Muhammad I. Razzak. "Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks." *Neurocomputing* 177 (2016): 228-241.
- [19] Ul-Hasan, Adnan, Saad Bin Ahmed, Faisal Rashid, Faisal Shafait, and Thomas M. Breuel. "Offline printed Urdu Nastaleeq script recognition with bidirectional LSTM networks." In *Document Analysis and Recognition (ICDAR)*, 2013 12th International Conference on, pp. 1061-1065. IEEE, 2013.
- [20] A. Muaz, and S. Hussain, "Urdu Optical Character Recognition System", MS Thesis, National University of Computer and Emerging Sciences, Lahore, Pakistan, 2010.
- [21] S. T. Javed and S. Hussain, "Improving Nastalique Specific Pre-Recognition Process for Urdu OCR", In *Proceedings of 13th IEEE International Multitopic Conference (INMIC)*, Islamabad, Pakistan, 2009.
- [22] Q. A. Akram, S. Hussain, A. Niazi, U. Anjum and F. Irfan, "Adapting Tesseract for Complex Scripts: An Example for Urdu Nastalique", In *Proceedings of 11th IAPR Workshop on Document Analysis Systems (DAS 14)*, Tours, France, 2014.

# An Empirical Evaluation of Error Correction Methods and Tools for Next Generation Sequencing Data

Atif Mehmood  
Riphah Institute of Computing and  
Applied Sciences (RICAS)  
Riphah International University  
Lahore, Pakistan

Javed Ferzund, Muhammad Usman Ali, Abbas Rehman,  
Shahzad Ahmed  
Department of Computer Science  
COMSATS Institute of Information Technology  
Sahiwal, Pakistan

Imran Ahmad  
Riphah Institute of Computing and Applied Sciences (RICAS)  
Riphah International University  
Lahore, Pakistan

**Abstract**—Next Generation Sequencing (NGS) technologies produce massive amount of low cost data that is very much useful in genomic study and research. However, data produced by NGS is affected by different errors such as substitutions, deletions or insertion. It is essential to differentiate between true biological variants and alterations occurred due to errors for accurate downstream analysis. Many types of methods and tools have been developed for NGS error correction. Some of these methods only correct substitutions errors whereas others correct multi types of data errors. In this article, a comprehensive evaluation of three types of methods (k-spectrum based, Multi-sequencing alignment and Hybrid based) is presented which are implemented and adopted by different tools. Experiments have been conducted to compare the performance based on runtime and error correction rate. Two different computing platforms have been used for the experiments to evaluate effectiveness of runtime and error correction rate. The mission and aim of this comparative evaluation is to provide recommendations for selection of suitable tools to cope with the specific needs of users and practitioners. It has been noticed that k-mer spectrum based methodology generated superior results as compared to other methods. Amongst all the tools being utilized, Racer has shown eminent performance in terms of error correction rate and execution time for both small as well as large data sets. In multisequence alignment based tools, Karect depicts excellent error correction rate whereas Coral shows better execution time for all data sets. In hybrid based tools, Jabba shows better error correction rate and execution time as compared to brownie. Computing platforms mostly affect execution time but have no general effect on error correction rate.

**Keywords**—Next generation sequencing; bioinformatics; errors; error correction; execution time; k-spectrum; suffix tree based; hybrid based

## I. INTRODUCTION

Gigantic amount of data is originated with the help of next generation sequencing technologies at lowest cost and high throughput. As compared to old generation of sequencing data (the first-generation technology) for example Sanger NGS data faces high challenges of error rate. NGS plays a leading

role in the discovery of many applications in bioinformatics research and changed the way of genomic research [1]. NGS demands high-power CPU and various algorithms that can work in parallel mode for bioinformatics studies. It also needs the spacious memory and execution time for total data that may cause issues for data management. NGS takes advantage of big data computing infrastructure that divides the memory in clusters and provides the batch queue system which helps to produce large amount of sequencing reads [2]. Errors in sequencing data mainly occur due to the replacement of correct bases with incorrect bases and indels. NGS technologies produce different tools such as Illumina and Solid to induce the substitution error, whereas the Roche 454 and Ion torrent create the insertion and deletion error. Most of the tools and methods focus on removing the substitution errors [3]. There are three types of biases that cause errors in sequencing data: systematic bias, coverage bias and batch effect bias. The rate of error in data is also different for various NGS technologies. It is key step to remove the data error before any analysis can be made. These errors also disturb the accuracy of algorithm therefore it is beneficial to rectify data before analysis to conclude better results in downstream analysis [4].

Correction of sequencing errors is a critical module for bioinformatics discovery. The basic concept behind correcting the sequencing read errors is to correct the erroneous bases. Many error correction tools subjective of different data structures related to various methods have been developed. The error correction methods are classified into four categories:

- 1) K-spectrum based method such as Quake (2010), Lighter (2014) [5], Reptile (2010) [6], BLESS (2014) [7] hammer (2011), Musket (2013), HECTOR (2014) and RACER (2013). These tools correct the errors on k-mer incidence.
- 2) Multiple sequence alignment based method such as Karect (2015), Coral (2011) and ECHO (2011).
- 3) Suffix tree based method such as SHREC (2009) [8].

4) Hybrid based method such as LoRDEC (2014) [9], Jabba (2016) and Brownie (2015).

Different error correction tools and algorithms have evolved with the passage of time possessing better accuracy and minimum execution time. Evaluation of specific tools is a study matter being provided by various educational sources. In this comparative study, three methods and six tools are selected, each pair of tools belonging to each method. Musket and RACER are selected from the K-spectrum based category, Coral and Karect are selected from the multiple sequence alignment categories, and Jabba and Brownie are selected from the Hybrid based category. These tools run on two different computing platforms. This piece of study aims to answer the following questions:

- Do these tools cope with data scalability?
- Does the computing platform affect the performance of tools?
- Which method of error correction is better?
- Which tool outperforms other tools?
- Which tool is better within the same category?
- Which tool has maximum error correction rate?
- Which tool requires minimum execution time on same dataset?

In addition, performance of different tools will be evaluated for different data sets. The rest of the paper is organized as follows: Section 2 describes the related work and Section 3 presents the experimental details. Results are discussed in Section 4 and paper is concluded in Section 5.

## II. RELATED WORK

Error correction depends on read coverage and error correction rate of different tools. These tools are based on different approaches and data structures. Three main approaches are used to make error correction tools more efficient such as k-spectrum based, suffix array based and hybrid based approach. Li et al. [5] has developed tool that depends on k-mer spectrum based. The authors used 31-55 k-mer length as well as bloom filter and hash table data structure. The authors focused on the removal of substitution errors. They also checked the trusted regions and extracted the optimal solution by using extension mechanism. In this task of material study, the experimental results are targeted on achieving maximum error correction rate. Heo et al. [7] used the hash table data structure. They used k-spectrum approach for error correction. They determined the solid minimum edit path in between solid k-mer. Using the reverse bloom filter, they changed false positive rate. During k-mer counting Bloocoo used 10 bits for storing solid k-mers. They also described the need for 4 GB memory requirement for human genome correction. Song et al. [5] developed memory efficient tool based on k-mer spectrum. The authors used bloom filter and 23 k-mer length, Sequencing reads were processed in three steps and two bloom filters were used for error correction. In this work, k-mer subsample is obtained using first bloom filter and then test is applied on each read on

each position to find solid k-mer. These solid k-mers are stored and second bloom filter is applied. They used greedy approach for error correction which is also used in bless. Lighter corrected substitution errors. They used multiple sequence alignment method and suffix array based data structure. In his paper, two-sided error correction technique was used to correct substitutions errors. Salmela et al. [9] presented hybrid based error correction using de Bruijn graph. They corrected most weak left and right regions by choosing traversal paths in graph. The authors argued that LoRDEC consumes less memory as compared to other tools and error correction rate is 99%. In fiona, used partial suffix array with hierarchical statically method to correct errors in sequencing reads. They used each read  $r$  as reference and corrected first overlap reads. It is also able to corrected substitutions errors produced by illumina platform. They argued that fiona can process the data on inexpensive hardware. The authors used the hash table data structure and confusion matrix error model. Their technique is sufficient to correct short reads without using reference genome.

## III. EXPERIMENTAL DESIGN

For the experiments, six tools are selected based on their reviews. These tools belong to three methods of error correction. Two different computing platforms are used to run these tools. Four datasets of different sizes are used for the experiments. Details are given below:

### A. Tools

A brief description of the selected tools is presented in Table I.

**Coral** is used for multiple alignments of short reads to correct the error. It is the first approach used for the short-read sequencing. Coral can easily understand and run on the data produced by different NGS technologies. It can also read data coming from single molecule sequencing technology. Coral works by first indexing the reads. All the k-mers that are valuable in total data are indexed two times into forward and reverse directions. After this process the list of k-mers are stored in hash table. Next step after indexing is multiple alignments; every alignment depends on base that being generated from neighborhood based read. This alignment helps to correct the overall data and look over the overlapping k-mer read [10]. After the comparison, the new reads are produced to have minimum error rate. Coral is superior approach as compared to SHREC, and Reptile.

**Karect** also belongs to the same category as Coral; however, its working differs from it. Karect uses each read as a reference and stores results in partial order graph (POG). It is also used for multiple-alignment. It is able to correct different type of errors and handles data generated by different NGS technologies. It uses less peak memory during data processing. Its performance is outstanding against low-coverage region and high error rate of data. Karect depends on POG that accumulates partial alignment results. Alignment and normalization are performed based on correction reference reads with respect to alignment of each read [11]. **Musket** uses the k-spectrum based method. It provides more accurate results against the correction reads and has the ability

to execute the large read length of data and provides high coverage level. It mainly comprises of three techniques; one-sided aggressive correction, two-sided conservative method and voting based refinement method. Its time and space complexity are good for large dataset. When compared to other programs like Reptile, SHREC and Musket, it is three times faster than these tools [12].

TABLE I. ERROR CORRECTION TOOLS

Tools	Methods	Overview of Algorithm	Error Correction Type
Karect	Multiple sequence alignment	Partial order graph is used to accumulate partial alignment results. It considers each read r as reference.	Substitution Insertion Deletion
Coral	Multiple sequence alignment	Correction with alignment uses bases from the error in the correction process. Indexing k-mers that occur in reading are connected with a hash table.	Substitution Insertion
Racer	K-mer based	Racer is linked with k-mer counting program. It also uses 2-bit encoding nucleotide and arbitrary replacement of the unknown position and K-mer stored in the hash table.	Substitution
Musket	K-mer based	It is multi-threaded program, uses a master slave model and demonstrates superior parallel scalability. One sided aggressive and voting based refinement.	Substitution
Jabba	Hybrid based	Pseudo alignment approach with seed and extend method using maximal exact matches. This method corrects third generation reads by mapping on de Bruijn graph.	Substitution Insertion Deletion
Brownie	Hybrid based	It depends on de Bruijn graph and works with the help of Jabba and Karect tool. It also needs extra libraries to run the algorithm.	Substitution

**RACER** is another efficient tool that shows maximum error correction accuracy, time and space complexity. RACER ignores installation of extra software for processing the data, whereas other tools have two or three extra software libraries to process data. It uses the hash table for storing the k-mer because it introduces 2-bit encoding of nucleotides for random replacement of unknown position. It has the capability to process different data formats such as fastq and fasta data. **Jabba** uses hybrid method to correct the alignment and error in third generation sequencing to map the reads on de Bruijn graph which is made for second next generation sequencing.

Seudo alignment approach is mostly used by this tool. Jabba processes the data in two phases: in the first phase smaller k-mer size (K=13) are used, in the second phase results are processed on de Bruijn graph that provide the extra accuracy for given data. Jabba also uses larger k-mer size (k=75) for long reads and thus repeating the entire process. It uses less time as compared to Proovread and time consumption is more like RACER.

**Brownie** also uses hybrid method and supports Jabba in its methodology and techniques being adopted. It also creates the de Bruijn graph for Jabba as a result the resultant file is stored in Jabba directory and then different commands are applied to find the result of error correction (Releases. biointec/brownie. GitHub). This tool requires three extra libraries for processing the data. It provides exceptional results on small dataset.

**B. Error Correction Methods**

Tools selected for this study implement different error correction methods. These methods are presented in Table 2..

**K-spectrum based** method decomposes the reads and makes the set of k-mer. Mostly NGS technology introduces substitution error, so k-mer set has small distance to each other if they belong to same genome location. k-spectrum is then constructed using hashing and k-mer frequency is counted to determine the error threshold. During this process, threshold of each type of k-mer (solid k-mers and weak k-mers) is determined. Then both k-mers are compared with the help of bloom filter and results are stored in hash table [1]. These results are converted into the high multiplicity k-mers and algorithm corrects the error in erroneous regions and provides with corrected reads.

TABLE II. ERROR CORRECTION METHODS USED BY THE SELECTED TOOLS

Method Label	Method Type	Tools
M1	K-spectrum based	Musket, Racer
M2	Multi Sequencing Alignment (MSA)	Coral, Karect
M3	Hybrid based	Jabba, Brownie

**Multiple sequence alignment (MSA)** is used for biological sequences such as protein, DNA and RNA. Two approaches are used for MSA; iterative and progressive. MSA starts working on one sequence and then aligns step by step. Working parameters and steps differ for each type of MSA. In the progressive approach, it starts from much similar sequence and aligns the new sequence to each of the previous sequences. After that it creates the distance matrix and phylogenetic guided tree is created from the matrices. Using the guided tree, it defines the next sequence to be added for alignment and preserves the gap. These steps are repeated until the total data is converted into appropriate alignment. In the iterative approach, it starts the alignment in pair wise grouping. Selection of these pairs depends on the sequence relation on the guided tree. Progressive approach is competitively efficient as compared to the iterative approach.

**Hybrid Method** is suitable for third generation sequencing that produces large amount of data with high error

rate. This respective method uses minimum CPU time for data processing [13].

C. Datasets

In this study, four different datasets are used that are generated by the Illumina sequencing machine. A detailed description of the datasets is presented in Table III. These datasets are selected on the basis of varying attributes such as read length, number of reads and size. Some of these datasets were previously used in correction studies. The accession numbers provide the complete details about datasets.

All the datasets are available on National Center for Biotechnology www.ncbi.org.

D. Platforms for Running the Tools

Two different computing platforms are used to evaluate the tools. A brief description of the used platforms is presented in Table IV.

Machine 1 has 2.10GHz CPU (Intel i3) with 8 GB main memory. Operating system is Ubuntu Linux version 14.04 and compiler was g++. Machine 2 has different specification such as 3.10GHz CPU (Intel i7) and 16 GB RAM. Both machines used the same version of Ubuntu Linux and compiler.

TABLE III. DATASETS USED FOR THE EVALUATION OF SELECTED TOOLS

Dataset	Species	Sequencing Platform	Accession Number	No of Bases	Size
D1	S.aureus	Illumina	SRR022868	3100M	2.3Gb
D2	S.aureus	Illumina	SRR022865	821.2M	692.1Mb
D3	Escherichia coli	Illumina	SRR022918	677.2M	386Mb
D4	C.elegans	Illumina	SRR065887	316.5M	207.7Mb

TABLE IV. PLATFORMS USED TO RUN THE TOOLS

Machine Label	Processor	Installed memory	System type	Operating system	Compiler
Machine1	Core(i3) 2.10GHz	8GB	64 bit	Ubuntu Linux version 14.04	g++
Machine2	Core(i7) 3.10GHz	16GB	64 bit	Ubuntu Linux version 14.04	g++

IV. RESULTS

On Machine 1, an experiment was conducted to evaluate the error correction rate. The results are shown in Figure 1. The tools comparison shows that Musket, Racer, Coral, Karect, Jabba and Brownie are best performers on data structures D2, D1, D4, D1, D4 and D3 respectively. If we take into account the data sets, for D1 and D2 Racer, for D3 Racer and Brownie produced best results and for D4 JABBA produced best results. The result from overall perspective depicts that, Racer has shown consistent performance in terms of error correction rate on all data sets. However, on the largest dataset JABBA outperforms other tools for error correction rate. On average basis for error correction rate, Coral and Musket show middle level performance, respectively.

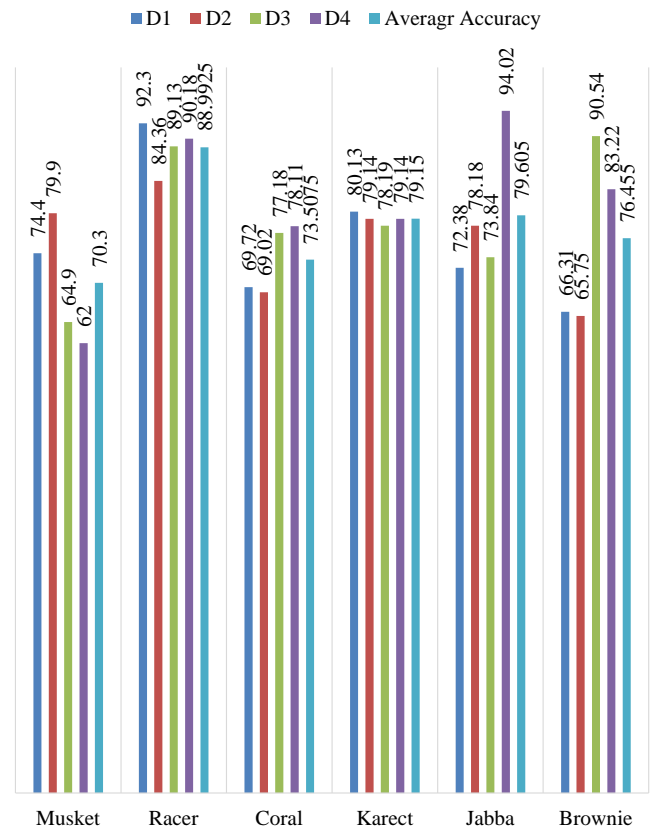


Fig. 1. Error correction rate on Machine 1.

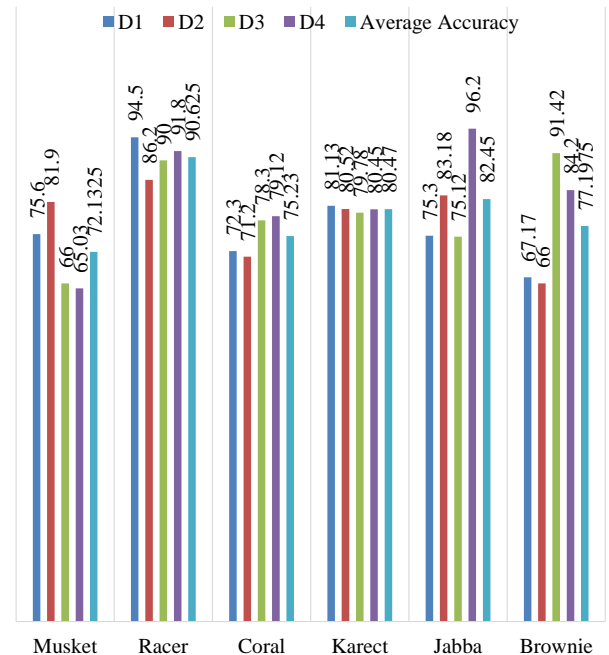


Fig. 2. Error correction rate on Machine 2.

On Machine 2, identical experiment was conducted to evaluate the error correction rate. The results are shown in Fig. 2. Comparison of various tools shows that Musket, Racer, Coral, Karect, Jabba and Brownie performed best on D2, D1, D4, D1, D4 and D3 data sets respectively. Analysis of data structures shows that for D1 and D2 RACER produced best results, for D3 Racer and Brownie produced best results and for D4 JABBA produced best results as compared to other tools. If we look at overall results, Racer has shown consistent performance in terms of error correction rate on all data sets. However, on the largest dataset JABBA is the winner for error correction rate. If we consider the average error correction rate, Coral and Musket are poor performers respectively. So, improvement in processing speed and memory does not affect the error correction rate.

On Machine 1, another experiment was conducted to evaluate the execution time required to process the data for error correction. The results are shown in Fig. 3. The comparative analysis of tools shows that, Musket, Racer, Coral, Karect, Jabba and Brownie best performed on D4, D1, D3, D1, D3, and D3 respectively. Consideration of data structures shows us that for D1 and D2 Racer produced best results, for D3 and D4 Brownie produced best results. From overall perspective, Racer has shown consistent performance in terms of time on all data sets. However, on the largest dataset Racer is at its peak of performance for execution time. On low to average basis Karect and Musket are poor performers respectively. So, enhancement in processing speed and memory reduces the execution time required for error correction. The difference is evident in the case of D4 which is the largest data set used in this study.

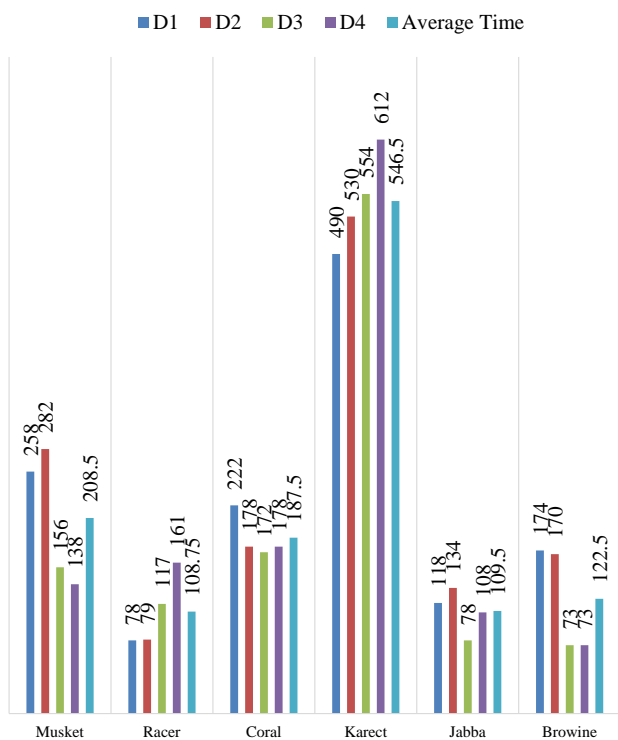


Fig. 3. Execution time on Machine 1.

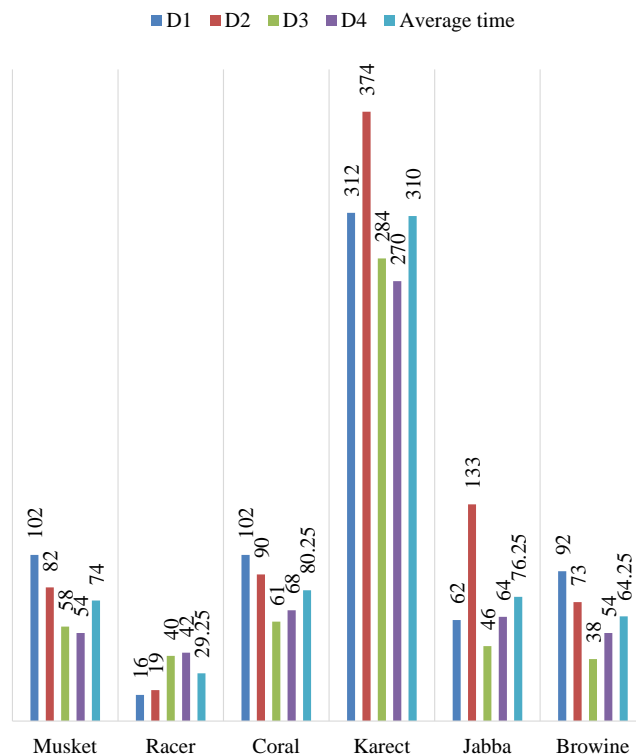


Fig. 4. Execution time on Machine 2.

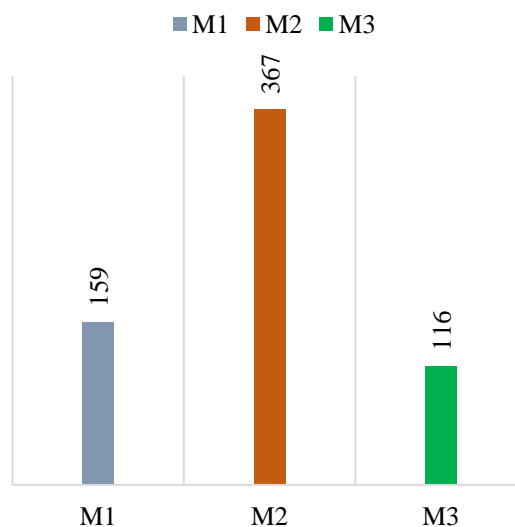


Fig. 5. Execution time on Machine 1 for three methods.

On Machine 2, same experiment was conducted to evaluate the time required to process the data for error correction. The results are shown in Fig. 4. Figure illustrates that, Musket, Racer, Coral, Karect, Jabba and Brownie are best performers on D4, D1, D3, D4, D3 and D3, respectively. Looking on structures of data sets, it is obvious from figure that for D1 and D2 Racer produced best results, for D3 Brownie produced best results and for D4 Racer produced best results with Brownie and Musket at second position. Overall results show that Racer has shown consistent performance in



terms of execution time on all data sets. However, on the largest dataset Racer acts as best performer for execution time. On low to average basis, Coral and Karect are poor performers, respectively. On Machine 1, Browine used the minimum execution time, whereas on Machine 2, Racer used the minimum execution time for D4.

On Machine 1, average execution time was calculated for each method. The results are shown in Fig. 5. If we consider the methods, M3 (Hybrid) based tools produced best results for all data sets as compared to M1 (K-spectrum) based tools. Whereas, M2 (Multi Sequencing Alignment) based tools performed poorly.

On Machine 2, average execution time was also calculated for each method. The results are shown in Fig. 6. If we take into account various methods, M1 (K-spectrum) based tools produced best results for all data sets, as compared to M3 (Hybrid) based tools. Whereas, M2 (Multi Sequencing Alignment) based tools performed poorly. So, improvement in processing speed and memory affects the execution time required by different methods. K-Spectrum based tools perform better on high performance machines, whereas Hybrid based tools can produce better results even on lower specification machines.

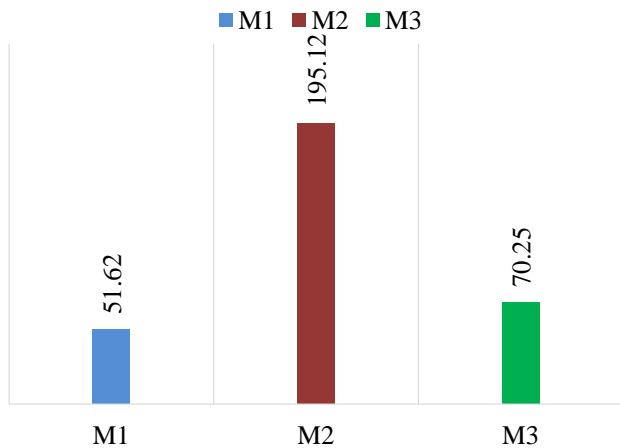


Fig. 6. Execution time on Machine 2 for three methods.

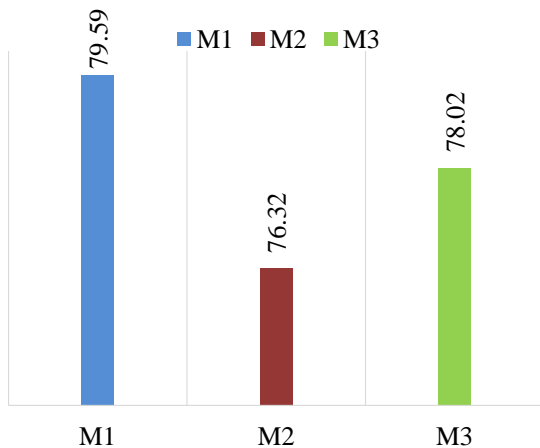


Fig. 7. Error correction rate on Machine 1 for three methods.

On Machine 1, average error correction rate was calculated for each method. The results are shown in Fig. 7. If we consider the methods, M1 (K-spectrum) based tools produced best results for all data sets, with M3 (Hybrid) based tools at second position. Whereas, M2 (Multi Sequencing Alignment) based tools performed poor.

On Machine 2, average error correction rate was also calculated for each method. The results are shown in Fig. 8. If we consider the methods, M1 (K-spectrum) based tools produced best results for all data sets, with M3 (Hybrid) based tools at second position. Whereas, M2 (Multi Sequencing Alignment) based tools performed poorly. So, improvement in processing speed and memory does not affect the error correction rate of different methods.

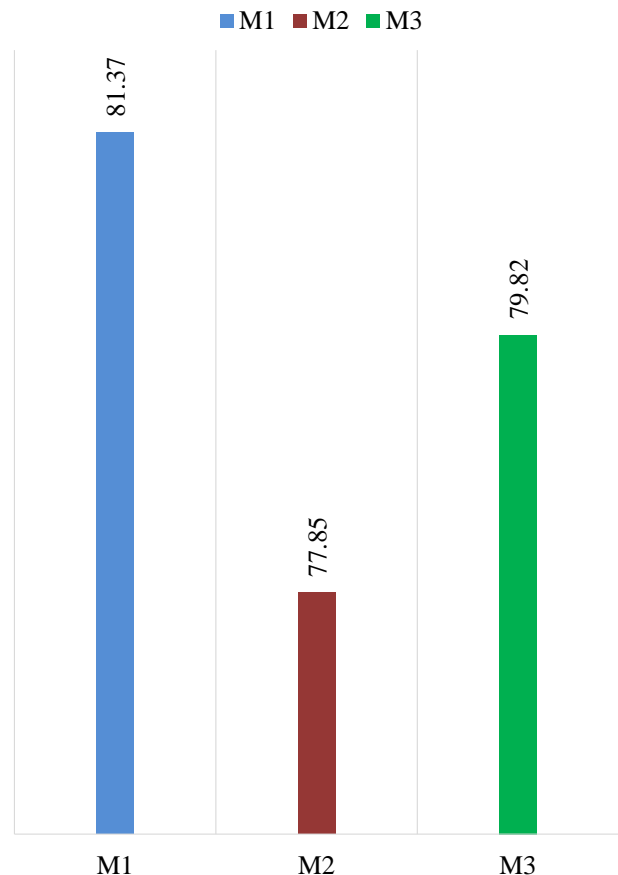


Fig. 8. Error correction rate on Machine 2 for three methods.

On the basis of the above findings, answers to the questions formulated in the Introduction section are presented below:

**Do these tools cope with data scalability?**

No, all tools cannot handle data scalability because the performance of these tools degrades with increase in data size. However, Racer and JABBA perform equally well on small and large datasets.

#### REFERENCES

### Does the computing platform affect the performance of tools?

Yes, the computing platform affects the execution time required to process data. However, it does not affect the error correction rate.

### Which method of error correction is better?

K-spectrum based method produced best results with Hybrid based method at second position.

### Which tool outperforms other tools?

Racer has outperformed other tools both in execution time and error correction rate. JABBA is the second-best performer.

### Which tool is better with in the same category?

Racer is better than Musket in k-spectrum based category. Karect is better than Coral in terms of error correction rate, whereas Coral is better than Karect in terms of execution time in the Multiple Sequence Alignment based category. Jabba is better than Brownie in the hybrid based category.

#### V. CONCLUSIONS AND FUTURE WORK

Among the three methods studied, k-spectrum based method generated good results as compared to other methods. Racer can perform well in error correction rate and time execution on small as well as large data sets. In multi sequence alignment based tools, Karect performed better in error correction rate whereas Coral performed better in execution time for all data sets. Jabba performs well in error correction rate and time execution; however, brownie provided good results in terms of execution time on Machine 2. These tools depend on hybrid based method. Computing platform has effect on execution time but has not significant effect on error correction rate. In future, we want to evaluate tools that can process large datasets in shorter time.

- [1] Isaac Akogwu, Nan Wang, Chaoyang Zhang, and Ping Gong, "A comparative study of k-spectrum-based error correction methods for next-generation sequencing data analysis," *Human Genomics*, pp. 49-59, 2016.
- [2] Xiao Yang, Sriram , Chockalingam , and Srinivas Aluru, "A survey of error-correction methods for next-generation sequencing," *BRIEFINGS IN BIOINFORMATICS*, vol. 14, pp. 56-66, 2012.
- [3] Leena Salmela and Jan Schröder, "Correcting errors in short reads by multiple alignments," *bioinformatics*, vol. 27, pp. 1455-1461, 2011.
- [4] Margaret A Taub, Hector Corrada Bravo, and Rafael A Irizarry, "Overcoming bias and systematic errors in next," *genome medicine*, pp. 1-5, 2010.
- [5] Li Song, Liliana Florea, and Ben Langmead, "Lighter: fast and memory-efficient sequencing error correction without counting," *Genome Biology*, pp. 1-13, 2014.
- [6] Xiao Yang, Karin S Dorman, and Srinivas Aluru, "Reptile: representative tiling for short read error correction," *bioinformatics*, vol. 26, pp. 2526-2533, 2010.
- [7] Yun Heo, Xiao Long Wu, Deming Chen, Jian Ma, and Wen Mei Hwu, "BLESS: Bloom filter-based error correction solution for high-throughput sequencing reads," *bioinformatics*, vol. 30, pp. 1354-1362, 2014.
- [8] Jan Schröder, Heiko Schröde, Simon J Puglisi, and Ranjan Sinha, "SHREC: a short-read error correction method," *bioinformatics*, vol. 25, pp. 217-2163, 2009.
- [9] Leena Salmela and Eric Rivals, "LoRDEC: accurate and efficient long read error correction," *Bioinformatics*, pp. 3506-3514, 2014.
- [10] Leena Salmela and Jan Schröder, "Correcting errors in short reads by multiple alignments," *bioinformatics*, vol. 27, pp. 1455-1461, 2011.
- [11] Amin Allam, Panos Kalnis, and Victor Solovyev, "accurate correction of substitution,insertion and deletion errors for next-generation sequencing data," *bioinformatics*, pp. 3421-3428, 2015.
- [12] Yongchao Liu, Jan Schro der, and Bertil Schmidt, "Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data," *bioinformatics*, vol. 29, pp. 308-315, 2013.
- [13] Giles Miclotte et al., "Jabba: hybrid error correction for long sequencing reads," *Algorithms Mol Biol*, pp. 1-12, 2016.

# Combinatorial Double Auction Winner Determination in Cloud Computing using Hybrid Genetic and Simulated Annealing Algorithm

Ali Sadigh Yengi Kand, Ali Asghar Pourhaji Kazem

Department of Computer Engineering,  
Tabriz Branch, Islamic Azad University,  
Tabriz, Iran

**Abstract**—With the advancement of information technology need to perform computing tasks everywhere and all the time there, in cloud computing environments and heterogeneous users have access to different sources with different characteristics that these resources are geographically in different areas. Due to this, the allocation of resources in cloud computing comes to the main issue is considered a major challenge to achieve high performance. Due to the nature of cloud computing is a distributed system to account, comes to business, economic methods such as auctions are used to allocate resources for decentralization. As an important economic bilateral hybrid auction model is the perfect solution for the allocation of resources in cloud computing, on the other hand, providers of cloud resources similarly, their sources of supply combined addressing. One of the problems auction two-way combination with maximum benefit for the parties to the transaction is the efficient allocation of resources to the problem of determining an auction winner is known. Given that the winning auction is NP-Hard. It results in a problem, several methods have been proposed to solve it. In this dissertation, taking into account the strength simulated annealing algorithm, a modified version of it is proposed for solving the winner determination in combinatorial double auction problem in cloud computing. The proposed approach is simulated along with genetic and simulated annealing algorithms and the results show that the proposed approach finds better solutions than the two mentioned algorithms.

**Keywords**—Cloud computing; double auction; winner determination; genetic algorithm; simulated annealing

## I. INTRODUCTION

Resource management is one of the key challenges in cloud computing and cloud data center management [1]. Most cloud providers use fixed price mechanisms to allocate resources to users. But these mechanisms do not provide an efficient and acceptable allocation of resources, and in fact cannot maximize the profitability of cloud resource providers [2], [3]. In such a situation, cloud-based economic models are suited to tune, deliver, and demand resources. An appropriate option for allocating resources in cloud computing is the use of bidding mechanisms. Among the auctioning mechanisms, the most appropriate mechanism used in cloud computing for

allocation and pricing of resources is a combinatorial auction mechanism.

In this way, prices depend on the conditions of demand and supply rather than the fair exchange between cloud providers and users. Considering the above, the use of combinatorial double auction to allocate resources in cloud computing can be a very appropriate model [4]–[7]. The double auction mechanism consists of two steps. The first step is to determine the winning bidder by solving an optimization model that aims to maximize social welfare by taking into account the payment of users and the profit of the providers. The second step is the allocation and pricing of resources among the winners. However, it has been proved that the problem of winner determination of the auctions is a NP-hard problem and hence the researchers are using heuristic, meta-heuristic and greedy methods to solve it [2], [8]. Considering the above issues regarding allocation of resources in cloud computing, in this paper, a modified simulated annealing algorithm has been used to determine the winner of the auction in the allocation of cloud resources. The modified simulated annealing algorithm in each step, instead of using a neighboring solution, uses several neighboring solutions. This change in the base simulated annealing algorithm leads to early convergence and improves the final solution found. In the rest of this paper, in Section 2, related works about economic models for resource allocation in cloud computing are introduced. In Section 3, formal definition of the problem is provided. The proposed mechanism is outlined in Section 4 and the simulation and experimental results are evaluated in Section 5. Finally, Section 6 concludes the paper.

## II. RELATED WORKS

Economic models provide different policies and tools for allocating resources in cloud systems. In cloud computing, users compete with other users as well as resource owners with other resource owners [9], [10]. Economic models can be based on the transaction or payment of the resource price. In cloud computing, providers and owners of resources with financial incentives provide their users with cloud resources. Taking into account the points mentioned, the use of decentralized methods is a good way to manage resources in cloud computing. Economic solutions are appropriate because

they have a decentralized structure and also motivate the owners of the resources to participate in their resources in the cloud [11], [12]. Another economic model is that both the user's goals and the objectives of the owners of the resource are taken into consideration in the process of resource allocation.

To date, several market-based resource allocation models and algorithms have been proposed for cloud computing environments. In the rest of this section some of them have been discussed. Wang et al. [13] conducted a study for resource allocation using an English combinatorial auction in cloud computing environments. The resource marketing price was resolved by an English combinatorial auction model, which is concentrated principally on maximizing the seller's profit and reducing the execution time for the winner determination. In 2013 [14], a virtualized resource allocation mechanism to assign CPU resources in virtualized machines was proposed. This work tried to overcome the unfairness issue of resource allocation in cloud computing. This work is essentially concentrated on enhancing the system resource utilization, and it is not considered to be of benefit to the user and service provider.

Another deficiency of this model is that it was restricted to virtual machine and different types of resources were not considered.

Xu presented CDA-CCRA, a new cloud computing resource allocation model based on combinatorial double auction mechanism for more effective resource utilization in cloud computing [15]. The CDA-CCRA model can simultaneously satisfy the users and providers requirements and significantly reduce transactions. Sabzevari et. Al have been proposed one double combinatorial auction based resource allocation approach for cloud computing environments [2]. The main goal of their study is to allocate economic resources in a way that lead to increase social welfare. Their proposed approach uses imperialist competitive algorithm for winner determination and genetic algorithm for resource allocation and payment schemes.

### III. PROBLEM DEFINITION

Auctions that bidders can offer a combination of resources has recently been taken into consideration. Compared to a non-trading auction, the combinatorial auction has a high performance. In the cloud computing environment distributed resources, including computing resources, storage resources, network bandwidth, and so on, compete with each other to execute user's work, and as a result, a combinatorial auction is appropriate for allocating resources. In double auction, both buyer and seller can submit their offers. Compared to a one-way auction in which several buyers compete for goods sold by a vendor, a double auction prevents of monopolies. The combinatorial double auction offers not only the benefits of a combinatorial auction, but also the needs of both buyers and sellers and therefore is suitable for cloud resource allocation. The purpose of the combinatorial double auction is to maximize the overall profit by taking into account this limitation that the number of units selected from the resources in the buyer's combined packages does not exceed the number of units provided by the vendors.

Suppose there is a R resource set containing k resources. After both parties offer their offers to the broker, the broker must perform the auction which is known as the winner determination of the auction problem. This problem is described in (1), (2) and (3).

$$\text{Max} \sum_{j=1}^n P_j x_j \quad (1)$$

$$\sum a_{ij} x_j \leq 0, \forall i \in K \quad (2)$$

$$x_j \in \{0,1\}, \forall j \in \{1,2,\dots,n\} \quad (3)$$

The set of proposed packages is  $\bar{B} = \{B_1, B_2, \dots, B_j, \dots, B_n\}$ , where  $n$  is the number of combined packets of resources. Each bid  $B_j$  is  $(a_j, p_j)$ , where  $a_j = (a_{1j}, \dots, a_{ij}, \dots, a_{kj})$  and  $a_{ij}$  represents the number of units requested from resource  $i$ . Also,  $p_j$  is the offered price for package  $j$ . If  $p_j > 0$ , it is a buyer's offer, and if  $p_j < 0$ , then it is considered as the seller's offer. Also, if  $x_j = 1$ , that is, the packet is assigned and if  $x_j = 0$ , that is, not assigned. Finding the best  $x_j$  values to maximize formula (1) with the constraints (2) and (3) is the same as determining the winner of the auction, which is considered as programming 0-1 and is an NP-hard problem.

### IV. PROPOSED APPROACH

In this section, the various stages of the proposed approach, which uses the hybrid genetic and simulated annealing algorithm to determine the winner of the auction in the allocation of cloud computing resources, is presented.

#### A. Hybrid Genetic and Simulated Annealing

In order to increase the efficiency of the simulated annealing algorithm, various researchers have tried to combine this algorithm with genetic algorithm using different methods, and the results of the combination of these two algorithms show an increase in efficiency. To combine these two algorithms, the total number of iteration of the hybrid algorithm is shared between the simulated annealing and genetic algorithms. The genetic algorithm performs the first half of the iteration of the hybrid algorithm and the simulated annealing algorithm uses the best chromosome of the final population of genetic algorithm as the initial solution. In addition, the simulated annealing algorithm performs the other half of the iteration of the hybrid algorithm and finally converges to the optimal solution. Considering that the simulated annealing algorithm does not start with a completely randomized solution, and in fact the best solution obtained from the genetic algorithm is considered as an initial solution for it, it will be seen a significant increase in the quality of the final solution.

#### B. Encoding

An important step in evolutionary algorithms is how to encode and display a solution. Each problem solution in the proposed approach is an array of binary numbers of length  $n$ , in which  $n$  is the number of offers. Array members as mentioned, there are binary numbers in which 1 means

acceptance of the offer and 0 means the denial of the relevant offer. Fig. 1 illustrates an example of a solution for eight offers.

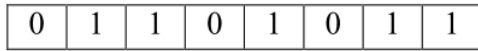


Fig. 1. An example of an offer (n=8) and the corresponding solution.

C. Fitness Function

Fitness function is one of the important concepts in every heuristic and meta-heuristic approach. In determining the winner of the auction, the suitability of a solution is obtained from the sum of the winning bidder's bid. The fitness function is shown in (4).

$$\sum_{j=1}^n \text{Price}_j x_j \tag{4}$$

where  $x_j \in \{0,1\}$

D. Selection Operator

One of the methods for selecting chromosomes in the genetic algorithm is roulette wheel selection. In the proposed approach, this method is used for the selection operator. This approach is because all individuals are mapped to neighborhoods based on their fitness level. The size of the area is determined by each individual according to its fitness size and then a random number is generated and, depending on the size of the number, the individual is selected. This process is repeated so that the desired number of parents (reproductive population) is provided.

E. Crossover Operator

Genetic algorithm of the proposed approach uses uniform crossover. In uniform crossover, for each gene of a chromosome a random number is produced between 0 and 1. If the generated random number is less than 0.5, the gene is inherited from the first chromosome and otherwise it inherited from the second chromosome. The second child's chromosome is also obtained by using reverse mappings. Fig. 2 shows how to perform a uniform crossover operator on two sample chromosomes.

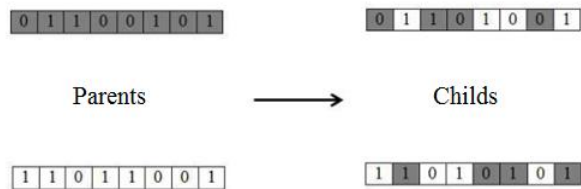


Fig. 2. How to make a uniform crossover on two sample chromosomes.

F. Mutation Operator

In the genetic algorithm of the proposed approach to apply the mutation operator, a random gene is selected from the chromosome, and its value, which is a binary number, is complemented. In fact, if the value of the selected gene is zero, it is converted to one and vice versa.

G. Temperature Initialization and Reduction

Initialization of temperature as well as its reduction in each iteration of simulated annealing algorithm is crucial step. The initial value of the temperature is considered  $T_{in}$ , and (5) is used to reduce it.

$$T_{p+1} = \alpha T_p \tag{5}$$

In (5),  $\alpha$  which is the coefficient of temperature reduction is a real value in the interval (0,1). If the value is close to 1, it causes a slow decrease in temperature, and thus allows the algorithm to search a large space of solutions and accepts many of the displacements to the optimal solution. After several experiments and taking into account the items mentioned in the proposed approach, the value of  $\alpha$  is considered equal to 0.992.

H. Generation of Neighboring Solutions

The generation of neighboring solutions in the simulated annealing algorithm is generally carried out using one or more approaches from known mutation approaches. In the proposed hybrid algorithm, neighboring solutions are generated by random selecting the one component of current solution and completing its value.

V. SIMULATION AND EXPERIMENTAL RESULTS

To evaluate the proposed approach, Matlab software is used for simulation. In addition, to compare the performance of the proposed approach, genetic and simulated annealing algorithms are also simulated. To carry out all tests, a Dell computer with a Core i5, 2 GHz processor and 4 GB of main memory was used.

Taking into account the fact that the proposed method uses a hybrid meta-heuristic algorithm, the results of the convergence as well as stability experiments are presented in the rest of this section. Also, the results of the proposed approach, which is named GASA in the rest of this section, have been compared with the results of genetic algorithm (GA) and simulated annealing (SA). In the convergence experiment, GASA, SA and GA algorithms are executed for two different test scenarios, and Fig. 3 and 4 show the results. In the first scenario, the number of participants is 200, and in the second scenario, the number of participants is 500. In the graphs, the horizontal axis indicates the iteration number of the algorithm and the vertical axis shows the value of the fitness function.

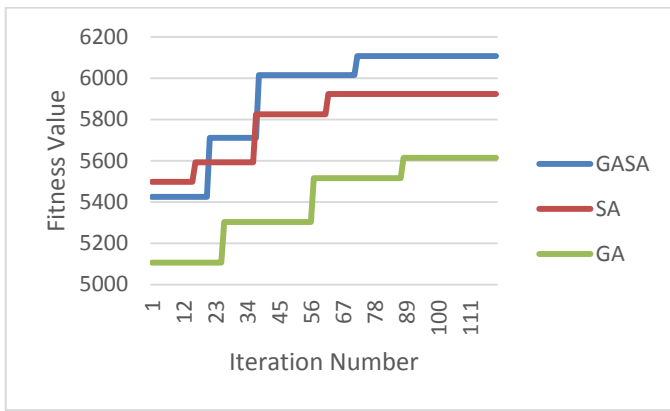


Fig. 3. Convergence test (Scenario 1).

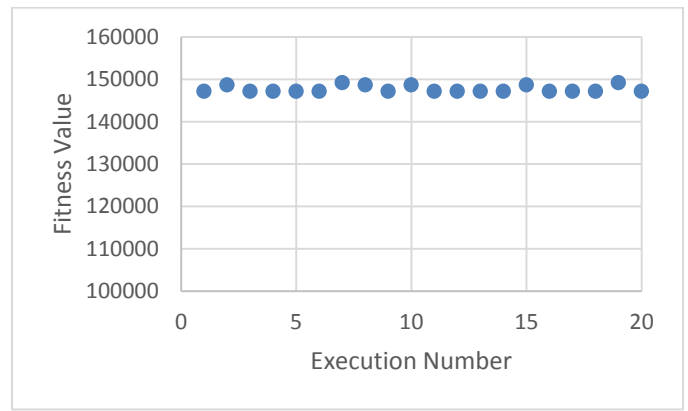


Fig. 6. Stability test (Scenario 2).

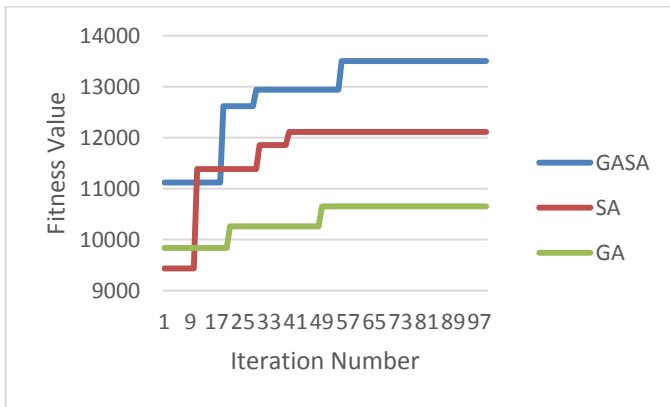


Fig. 4. Convergence test (Scenario 2).

The results of convergence test show that the proposed approach has a good convergence rate and finds better solutions compared to the two SA and GA algorithms.

Meta-heuristic algorithms have a nondeterministic and random nature, so it is necessary to examine the stability of these algorithms. The stability of an algorithm is whether the algorithm generates the same responses for various and different executions. To verify the stability of the GASA algorithm for the two scenarios mentioned, the algorithm is executed 20 times, and the value of the fitness function in each run is shown in Fig. 5 and 6. The horizontal axis in the diagrams shows the execution number of the algorithm and the vertical axis of the fitness value.

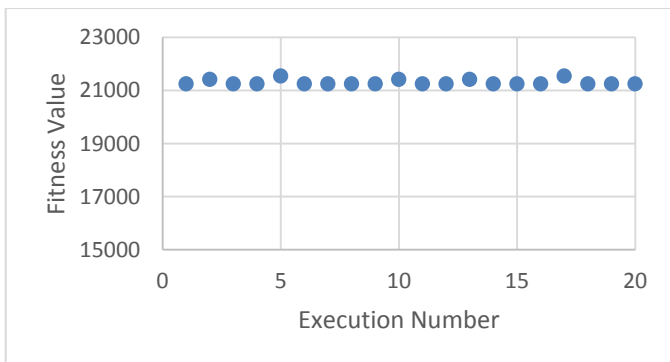


Fig. 5. Stability test (Scenario 1).

Examining the results of the stability test shows proper stability of the proposed approach which converges to optimal solution every time the algorithm is executed.

## VI. CONCLUSION AND FUTURE WORKS

Resource management is one of the key challenges in cloud computing and cloud data center management. Most cloud providers use fixed price mechanisms to allocate resources to users. But these mechanisms do not provide an efficient and acceptable allocation of resources, and in reality they cannot maximize the profitability of cloud resource providers. In such a situation, cloud-based economic models are appropriate for the regulation, presentation and demand of resources. In this paper, combinatorial double auctioning has been used to allocate resources in cloud computing. Given the fact that the issue of determining the winner of the auction is in the category of NP-hard problems, in this paper, a combination of genetic and simulated annealing algorithms are used to solve it. The integration of two genetic and simulated annealing algorithms allows for better solutions to the problem. The results of the experiments performed in the MATLAB environment showed that the proposed approach has a good convergence rate and is also very good in terms of stability. Also, the results of experiments performed on two different scenarios showed that the proposed approach produces more suitable solutions compared to the two genetic simulated annealing algorithms.

In the future, researchers can use new meta-heuristic algorithms such as forest optimization algorithm, krill herd optimization algorithm and etc. to optimize the winner determination problem in combinatorial double auction in cloud computing.

## REFERENCES

- [1] A. M. Sampaio and J. G. Barbosa, "Chapter Three - Energy-Efficient and SLA-Based Resource Management in Cloud Data Centers," in *Energy Efficiency in Data Centers and Clouds*, vol. 100, A. R. Hurson and H. Sarbazi-Azad, Eds. Elsevier, 2016, pp. 103–159.
- [2] R. A. Sabzevari and E. B. Nejad, "Double Combinatorial Auction based Resource Allocation in Cloud Computing by Combinatorial using of ICA and Genetic Algorithms," *International Journal of Computer Applications*, vol. 110, no. 12, 2015.
- [3] H. Wang, H. Tianfield, and Q. Mair, "Auction Based Resource Allocation in Cloud Computing," *Multiagent Grid Syst.*, vol. 10, no. 1, pp. 51–66, 2014.

- [4] X. Deng, P. Goldberg, B. Tang, and J. Zhang, "Revenue maximization in a Bayesian double auction market," *Theoretical Computer Science*, vol. 539, pp. 1–12, 2014.
- [5] I. Fujiwara, K. Aida, and I. Ono, "Applying double-sided combinatorial auctions to resource allocation in cloud computing," in *Proceedings - 2010 10th Annual International Symposium on Applications and the Internet, SAINT 2010*, 2010, pp. 7–14.
- [6] F. Nassiri-Mofakham, M. A. Nematbakhsh, A. Baraani-Dastjerdi, N. Ghasem-Aghaee, and R. Kowalczyk, "Bidding strategy for agents in multi-attribute combinatorial double auction," *Expert Systems with Applications*, vol. 42, no. 6, pp. 3268–3295, 2015.
- [7] Y. LIU, Y. LIU, X. MA, and K. LIU, "Solving {WDP} in combinatorial double auction based on trading strategy," *The Journal of China Universities of Posts and Telecommunications*, vol. 19, Supple, pp. 148–152, 2012.
- [8] F. Gorbanchadeh and A. A. Pourhaji Kazem, "Hybrid Genetic Algorithms for Solving Winner Determination Problem in Combinatorial Double Auction in Grid," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 1, no. 2, 2012.
- [9] M. Mihailescu and Y. M. Teo, "Dynamic resource pricing on federated clouds," in *Proceedings of the 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, 2010, pp. 513–517.
- [10] J. Stöber, D. Neumann, and C. Weinhardt, "Market-based pricing in grids: On strategic manipulation and computational cost," *European Journal of Operational Research*, vol. 203, no. 2, pp. 464–475, 2010.
- [11] A. Haque, S. M. Alhashmi, and R. Parthiban, "A survey of economic models in grid computing," *Future Generation Computer Systems*, vol. 27, no. 8, pp. 1056–1069, 2011.
- [12] L. Rodero-Merino, E. Caron, A. Muresan, and F. Desprez, "Using clouds to scale grid resources: An economic model," *Future Generation Computer Systems*, vol. 28, no. 4, pp. 633–646, 2012.
- [13] X. w. Wang, X. y. Wang, and M. Huang, "A resource allocation method based on the limited English combinatorial auction under cloud computing environment," in *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, 2012, pp. 905–909.
- [14] C. Jiang, L. Duan, C. Liu, J. Wan, and L. Zhou, "VRAA: virtualized resource auction and allocation based on incentive and penalty," *Cluster Computing*, vol. 16, no. 4, pp. 639–650, 2013.
- [15] J. Xu, "A Cloud Computing Resource Allocation Model Based on Combinatorial Double Auction," in *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*, 2016, pp. 5–8.

# QoS-based Cloud Manufacturing Service Composition using Ant Colony Optimization Algorithm

Elsoon Neshati

Department of Mechanical Engineering,  
Tabriz Branch, Islamic Azad University  
Tabriz, Iran

Ali Asghar Pourhaji Kazem

Department of Computer Engineering  
Tabriz Branch, Islamic Azad University,  
Tabriz, Iran

**Abstract**—Cloud manufacturing (CMfg) is a service-oriented platform that enables engineers to use the manufacturing capacity in the form of cloud-based services that aggregated in service pools on demand. In CMfg, the integration of manufacturing resources across different areas and industries is accomplished using cloud services. In recent years, the interest in cloud manufacturing service composition has grown, due to its importance in different manufacturing applications. When no single service is capable of satisfying the need for a manufacturing service requester, the service combination may be useful in order to fulfill the purpose of the manufacturing service requester. Therefore, the problem of how efficient and effective interconnection of cloud manufacturing services has come to fetch many research fields. In this paper, a new algorithm is presented using an ant colony optimization for the problem of cloud manufacturing service composition considering the quality of service.

**Keywords**—Cloud computing; cloud manufacturing; service composition, ant colony optimization

## I. INTRODUCTION

The ‘cloud manufacturing’ term was first introduced by Li et al. [1], which targets in creating an integrated and collaborative platform for distributed manufacturing systems based on cloud computing technology. Cloud manufacturing system enables different users to search the qualified manufacturing services from cloud-based resource repositories and dynamically combine them into a virtual manufacturing environment or solution to finish their tasks [1]. Current research efforts are focusing on the development of appropriate descriptions for manufacturing services. Many of these approaches have chosen for extending web services for the implementation of manufacturing service descriptions [2]. Web services constitute a promising technology perspective for software engineering. The service-oriented system poses several additional challenges in terms of component-based software engineering [3]. Alternatively, several services may be available with the same function (which they call semantic equivalent services). However, they certainly, there are different criteria for service quality. Quality of service features include: cost, response time, accessibility and reliability. In addition, service quality can

have other features like precision and frequency. Choosing between different services, including semantic services, provides a function of the quality of service choices. In addition, a user may have certain limitations on the values of some of the features. For example, the cost should not exceed the amount given, which will affect the selection. On the other hand, the service provider can provide a range for the values of the quality of service as part of the contract with the users Potential (hidden, hidden). Also, the quality assurance of the service for this service can be customer-related, so that each of them will be applied to a different instance of that service. For example, a user who buys a service at a price, does not expect a response time to be less than a certain threshold.

The composition of manufacturing cloud services with a knowledge of the quality of service is a key requirement in service-oriented cloud manufacturing system, since it makes it possible to perform complex user activities by meeting the quality of service constraints [3]. Manufacturing services with the same functionality and quality of service are increasing day by day, and providers of these services always have functional requirements along with a set of service quality limits. Therefore, the choice of manufacturing cloud services with the knowledge of the quality of service plays an important role in the composition of manufacturing cloud services. To solve the problem of choosing manufacturing cloud services with the knowledge of the quality of service, some methods with the help of semantic web and some others based on computations service quality traits have been created, but it is clear that the second approach is a more satisfactory solution to meet the global requirements of service quality, because it is a combination of optimization that combines the best composition of services.

Considering the aforementioned aspect of service composition in cloud manufacturing, in this paper, an approach is proposed for QoS-aware manufacturing cloud service composition using ant colony optimization algorithm. The rest of the paper is organized as follow. In the next section, related work will be reviewed. In Section 3, a formal definition of the problem is presented. In Section 4, simulation and experimental results will be discussed and finally Section 5 concludes the paper.



## II. RELATED WORKS

The problem of choosing a service based on quality of service was first reported by Chang Yu and his colleagues, and was then welcomed by many scholars. In 2007, Hoffman introduced a programmatic formalism to illustrate the composition of Web services, as well as the identification of a particular case of a web service combination called “forward effect” [4]. In 2004, Cardoso, Miller, and Arnold [5], and in 2001, Casati and Sean, as well as in 2004, Greener, discovered dynamic services and service combinations, with the knowledge of service quality, that they were in the service architecture Oriented, the set of constraints is used to describe the functional and non-functional characteristics of the services for search, and the service may be selected according to some desirable criteria of service quality. Michael Jogierer and Grove Mole have used the use of genetic algorithm for optimizing the problem of choosing web services with the knowledge of service quality and implementing this algorithm in their simulation environment in order to compare its efficiency [6]. It has been tested with other methods. In 2009, Zhang and Zhou offered an open cloud computing architecture at an international conference on web services, pointing out that virtualization and service-oriented architecture are two powerful technical key [7].

Cloud manufacturing extends the cloud computing technology with manufacturing infrastructures involved in the entire lifecycle of manufacturing applications. Luo et al. [8] studied the formal description of multidimensional information for manufacturing capability in cloud manufacturing system. Also, Wang et al. [9] discussed standardized data models describing cloud services and relevant features for supporting interoperable cloud manufacturing. Tao et al. [10] presented a modified particle swarm optimization algorithm for manufacturing grid service composition, in which its parameters for particle updating were dynamically tuned. In [11], the variant GA and fruit fly optimization was combined to address the QoS-aware cloud computing service composition.

## III. PROBLEM DEFINITION

The QoS-aware cloud manufacturing service composition is to find a set of cloud manufacturing candidate services with different functionalities that firstly observe user-defined limits and, secondly, optimize a target function. In this section, the above problem is officially stated. An example of the QoS-aware cloud manufacturing service composition is formally expressed as follows:

- A service composition request in the form of workflow that is modeled using a directed acyclic graph  $G=(V,E)$ .
- $V = \{T_1, T_2, \dots, T_n\}$  which  $n$  is the number of tasks in workflow.
- $E$  is a set of edges that shows the priority of the tasks.
- Each task  $T_i$  ( $1 \leq i \leq n$ ) has a set of candidate manufacturing cloud services  $CS_i = \{cs_i^1, cs_i^2, \dots, cs_i^{m_i}\}$  in which  $cs_i^j$  ( $1 \leq j \leq m_i$ ) is candidate cloud service.

- $m_i$  is the total number of available candidate manufacturing cloud services for task  $T_i$ .
- Each candidate manufacturing cloud service  $cs_i^j$  has a set of different quality of service information  $QoS_i^j = \{Q_1, Q_2, \dots, Q_k\}$  in which  $Q_i$  is a quality of service parameter.
- Quality of service related to manufacturing cloud services is stored in the quality of service repository.
- $K$ : The number of quality of service parameters for the manufacturing cloud services which are used in the quality of service model.
- $QC$ : The set of global restrictions defined by the user  $QC = \{C_1, C_2, \dots, C_k\}$ .

With this in mind, the goal of the QoS-aware manufacturing cloud service composition is to find the near optimal composite manufacturing cloud service where:

$$\forall j = 1 \dots K \begin{cases} \sum_{i=1}^n S_i Q_j < C_j & \text{if } Q_j \text{ is additive} \\ \prod_{i=1}^n S_i Q_j > C_j & \text{if } Q_j \text{ is multiplicative} \end{cases}$$

## IV. PROPOSED APPROACHES

ACO is a heuristic algorithm with efficient local search for combinatorial problems. This paper applies a novel ACO algorithm to QoS-aware manufacturing cloud service composition problem. Different parts of the proposed ACO are presented in the rest of this section.

### A. Initialization

The initialization step of the algorithm consists of two processes:

- Creation of initial population
- Initialization of pheromone matrix.

In order to create the initial population, a random solution is created for each  $Ant_i$ ,  $i=1,2,\dots,k$  where  $k$  is the number of ants in the population. A solution in the population is an integer array with size  $n$  that the item of index  $i$  indicates the candidate manufacturing cloud service executing the task  $T_i$  in the workflow. The next step in the initialization phase is the initialization of pheromone matrix. The pheromone matrix is an  $m \times n$  matrix that all of its items are set to an initial value  $\tau_0$ .  $m$  is the maximum number of candidate manufacturing cloud services for a task.

$$\tau_{ij} = \tau_0 \quad 1 \leq i \leq m \text{ and } 1 \leq j \leq n$$

### B. Fitness Function

The main objectives of QoS-aware manufacturing cloud service composition problem are satisfying the user's global constraints while optimizing a fitness function. Therefore, the problem can now be modeled by means of a

fitness function and, finally, some constraints. The fitness function should maximize some QoS parameters of manufacturing cloud services such as reliability and availability, while minimizing others such as cost and response time.

Considering the aforementioned aspects of the fitness function, it can be defined as follows:

$$Fitness(sol) = \frac{w_1 * sol.Resp + w_2 * sol.Cost}{w_3 * sol.Avail + w_4 * sol.Reli}$$

Where  $w_1, w_2, w_3$  and  $w_4$  are positive weights which indicate the importance of QoS parameters identified by the user.

### C. Pheromone Updating

After initialization step that a solution for each ant is created, the pheromone trails are updated. In fact in the proposed ACO algorithm only global pheromone updating is applied. Pheromone updating is first done by decreasing the pheromone value on all paths by a constant factor. This step of pheromone updating is referred to as pheromone evaporation. After evaporation, all ants increase pheromone values in the pheromone matrix according to their solution's feasibility. In the QoS-aware grid service composition problem, a solution is feasible if it satisfies all of the QoS constraints identified by the user.

Considering the aforementioned aspects of the pheromone updating, it can be implemented by:

$$\tau_{ij} = \left[ (1 - \rho)\tau_{ij} + \Delta\tau_{ij}^{best} \right]_{\tau_{min}}^{\tau_{max}}$$

$$1 \leq i \leq m, 1 \leq j \leq n$$

Where  $0 < \rho \leq 1$  is the pheromone evaporation rate;  $\tau_{max}$  and  $\tau_{min}$  are respectively the maximum and minimum bound for the pheromone value and the operator  $[x]_b^a$  is defined as follows:

$$[x]_b^a = \begin{cases} a & \text{if } x > a \\ b & \text{if } x < b \\ x & \text{otherwise} \end{cases}$$

Also,  $\Delta\tau_{ij}^{best}$  is defined as follows:

$$\Delta\tau_{ij}^{best} = \begin{cases} \frac{1}{F_{best}} & \text{if } cs_j^i \text{ is selected for } T_i \text{ in the best ant(Composite Service)} \\ 0 & \text{Otherwise} \end{cases}$$

Where  $F_{best}$  is the fitness value of best ant of the current iteration. The parameter  $\rho$  is used to avoid unlimited accumulation of the pheromone trails.

## V. SIMULATION AND EXPERIMENTAL RESULTS

In this section, the results obtained from the simulation of the proposed approach will be presented in comparison with

the genetic and particle swarm optimization algorithm. The proposed approach is simulated in MATLAB environment. Different experiments are performed and the results of them stated in the rest of this section. To perform experiments, two different test scenario with 20 and 50 tasks are generated randomly. Also different QoS parameters values are generated by random. One of the main features of heuristic algorithms mainly ACO is the convergence of it. Fig. 1 and 2 show the results of the convergence test of proposed ACO compared to genetic and particle swarm optimization algorithms. Convergence results indicate that the proposed ACO converges to the optimal or near optimal solution as quickly as possible and also proposed ACO generates better composite manufacturing cloud services than genetic and particle swarm optimization algorithms.

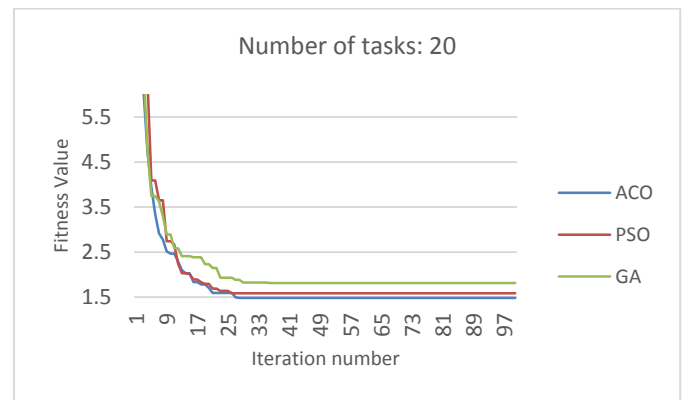


Fig. 1. Convergence test (Number of tasks: 20).

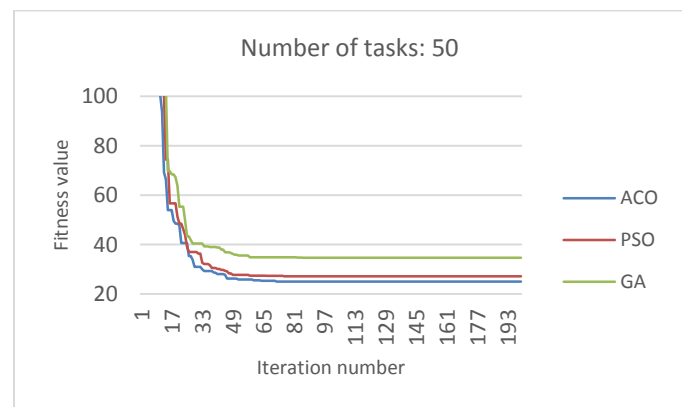


Fig. 2. Convergence test (Number of tasks: 50).

Meta-heuristic algorithms like ACO have an indeterminate and random nature, so it is necessary to examine the stability of these algorithms. The stability of an algorithm is whether the algorithm generates the same results for various executions. To verify the stability of the proposed ACO algorithm for the two scenarios mentioned above, the algorithm is executed 10 times and the fitness value per run is shown in Fig. 3 and 4. Examining the results of the stability test shows good stability of the proposed approach and indicates that the proposed approach converges to optimal solution in every execution of the algorithm.

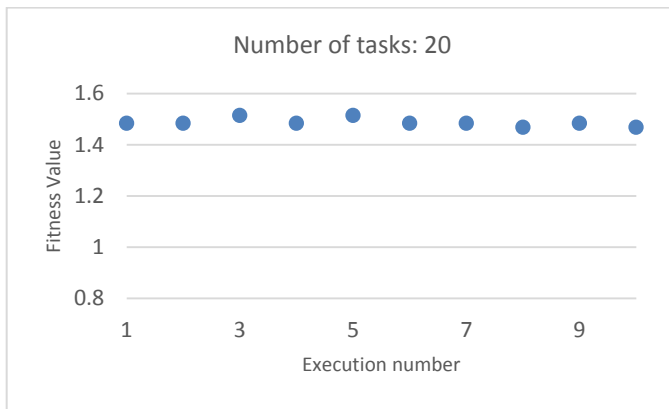


Fig. 3. Stability test (Number of tasks: 20).

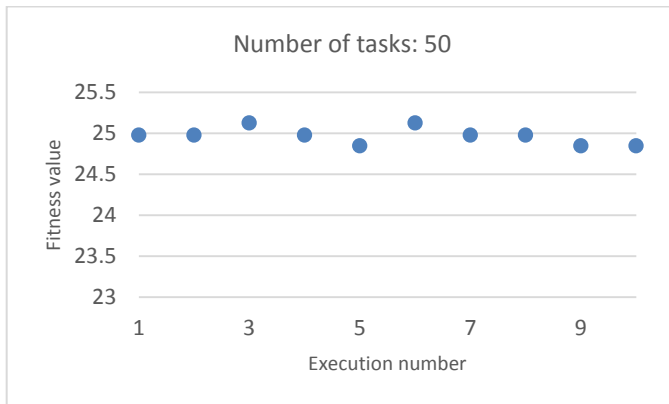


Fig. 4. Stability test (Number of tasks: 50).

## VI. CONCLUSION

The composition of manufacturing cloud services is an important technology for creating value added manufacturing services. The QoS-aware manufacturing cloud service composition with general QoS constraints is a very important problem in manufacturing cloud and service calculations. Designing an excellent algorithm for solving this problem with the capability to find the near optimal solution was the goal of this paper, which has been realized using the ant colony optimization algorithm. Taking into account the uncertainty and dynamic aspect of manufacturing cloud

environments, proposed ACO is an efficient approach for the aforementioned problem. Different experiments are performed to evaluate the proposed approach compared to genetic and particle swarm optimization algorithms and the results of them indicate that it has good convergence speed and stability.

## REFERENCES

- [1] B.-H. Li, L. Zhang, S.-L. Wang, F. Tao, J. W. Cao, X. D. Jiang, X. Song, and X. D. Chai, "Cloud manufacturing: a new service-oriented networked manufacturing model," *Computer integrated manufacturing systems*, vol. 16, no. 1, pp. 1–7, 2010.
- [2] F. Tao, L. Zhang, V. C. Venkatesh, Y. Luo, and Y. Cheng, "Cloud manufacturing: a computing and service-oriented manufacturing model," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 225, no. 10, pp. 1969–1976, 2011.
- [3] V. Diamadopoulou, C. Makris, Y. Panagis, and E. Sakkopoulos, "Techniques to support Web Service selection and consumption with QoS characteristics," *Journal of Network and Computer Applications*, vol. 31, no. 2, pp. 108–130, 2008.
- [4] J. Hoffmann, P. Bertoli, and M. Pistore, "Web Service Composition As Planning, Revisited: In Between Background Theories and Initial State Uncertainty," in *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2*, 2007, pp. 1013–1018.
- [5] J. Cardoso, A. Sheth, J. Miller, J. Arnold, and K. Kochut, "Quality of service for workflows and web service processes," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1, no. 3, pp. 281–308, 2004.
- [6] M. C. Jaeger and G. Muhl, "QoS-based selection of services: The implementation of a genetic algorithm," in *Communication in Distributed Systems (KiVS), 2007 ITG-GI Conference*, 2007.
- [7] L.-J. Zhang and Q. Zhou, "CCOA: Cloud computing open architecture," in *Web Services, 2009. ICWS 2009. IEEE International Conference on*, 2009, pp. 607–616.
- [8] Y. Luo, L. Zhang, F. Tao, L. Ren, Y. Liu, and Z. Zhang, "A modeling and description method of multidimensional information for manufacturing capability in cloud manufacturing system," *The International Journal of Advanced Manufacturing Technology*, vol. 69, no. 5–8, pp. 961–975, 2013.
- [9] X. V. Wang and X. W. Xu, "An interoperable solution for Cloud manufacturing," *Robotics and Computer-Integrated Manufacturing*, vol. 29, no. 4, pp. 232–247, 2013.
- [10] F. Tao, D. Zhao, H. Yefa, and Z. Zhou, "Correlation-aware resource service composition and optimal-selection in manufacturing grid," *European Journal of Operational Research*, vol. 201, no. 1, pp. 129–143, 2010.
- [11] F. Seghir and A. Khababa, "A hybrid approach using genetic and fruit fly optimization algorithms for QoS-aware cloud service composition," *Journal of Intelligent Manufacturing*, pp. 1–20, 2016.

# Envisioning Internet of Things using Fog Computing

Urooj Yousuf Khan

Department of Computer Science  
SZABIST Dubai Campus  
Dubai, United Arab Emirates

Tariq Rahim Soomro

College of Computer Science & Information Systems  
Institute of Business Management (IoBM)  
Karachi, Pakistan

**Abstract**—Internet of Things is the future of the Internet. It encircles a wide scope. There are currently billions of devices connected to the Internet and this trend is expecting to grow exponentially. Cisco predicts there are at present 20 billion connected devices. These devices, along with their varied data types, transmission rates and communication protocols connect to the Internet seamlessly. The futuristic implementation of Internet of Things across various scenarios demands the real time performance delivery. These range from RFID connected devices to huge data centers. Until date, there is no single communication protocol available for envisioning IoT. There is still no common, agreed upon architecture. Hence, huge challenges lie ahead. One of the ways to envision Internet of Things is to make use of Fog Networks. Fog is essentially a cloudlet, located nearer to the ground. It offers lower latency and better bandwidth conservation. The Fog or Fog computing is a recent concept. The OpenFog Consortium is a joint effort of many vendors. Its latest work is the background study for realizing Fog as a possible platform for activating Internet of Things. This paper revolves around Envisioning Internet of Things using Fog computing. It begins with a detailed background study of Internet of Things and Fog Architecture. It covers applications and scenarios where such knowledge is highly applicable. The paper concludes by proposing Fog Computing as a possible platform for Internet of Things.

**Keywords**—IoT; fog computing; cloud computing

## I. INTRODUCTION

Internet of Things (IoT) is the future of the Internet. It enables inter connectivity among devices and platforms. It is expected that by 2020, IoT will expand up to 26 billion devices, a huge leap since 2009. One of the main challenges to enabling IoT is the ability to identify each device uniquely. Many techniques exist in this regard. A popular approach is to assign each device, an IPv6 address. This enables the devices to be identified exclusively. On a more advanced level, Uniform Resource Identifiers (URI) can be deployed. These include both IPv6 addresses and Uniform Resource Locators (URLs). URI is used in parallel with Domain Name Service (DNS). Another important domain is that of location sharing and identification. Participating gadgets: *Things* need to find out one another, approve each other and then share information over a link. Sharing information using protocols and data formats must be common to both participating devices [1]. Uninterrupted Internet connection is ensured by Cloud platforms. The “Pay-As-You-Go” model is an efficient way of managing data centers, for processing client applications and batch processing. This shift in the paradigm towards centralized cloud computing is primarily due to the

ease in management. It further includes scalability, expansion of data centers, automatic backup and elevated physical data security. However, uninterrupted Internet connection might be an expensive requirement for many small systems. They could be embedded devices, pervasive systems or simple battery operated circuits [2].

Such devices require mobility, location awareness and low latency. In reality, it might be ideal to have one local device to provide required Internet access to all devices. If the majority of device requests can be entertained locally, IoT vision can be materialized. To meet these requirements, a new platform, namely Fog Computing or simply Fog is coming into perspective. Fog Computing or Fog is a framework where large number of heterogeneous devices collaborate with each other to perform various networking tasks within the network, without the intervention of third party. The purpose of the study is to explore the potential implementation of IoT. It discusses Fog as a possible platform for enabling IoT [3].

Fog is essentially a cloudlet, placed near to host device [2]. It enables devices to identify and interact with each other locally. It thus prevents the redundant cloud access. This, in turn, leads to efficient data processing and better security. Fog is still a newer concept and is in its primary development phase. The paper is organized as follows: Section 2 covers the Literature Review; Section 3 discusses Architecture for Internet of Things, while Section 4 discusses Cloud Computing and the Fog. Findings are covered in Section 5. Discussion and Future work is covered in Section 6.

## II. LITRATURE REVIEW

The term Internet of Things was first coined by Kevin Ashton in 1999 [4]. It is an effortless interweaving of sensors, actuators and drivers. It is a concept that has taken firm grounding in the recent research years. IoT incorporates universal communication using existing, valid protocols. It does so by uniquely addressing each of the participating devices by using smart, interactive interfaces. Future Web i.e. the next generation of the Internet combines many aspects such as Internet of Things (IoT), Intelligent Networks (IN), Internet of Service (IoS), Internet of Content (IoC) and Internet of Media (IoM) [5]. Many technologies support this concept. These include Radio Frequency Infrared Detector (RFID), Near Field Communication (NFC), Optical tagging and Quick Response (QR) codes and assigning an IPv6 addresses to each device. As the web progresses from static, HTML based websites to more dynamic, AJAX-enabled (Asynchronous JavaScript and XML) Web sites and social networking towards Ubiquitous Computing, the need for smart

representation and all-time availability of data has increased by many folds [4].

RFID was one of the oldest in this regard. To be exact, a new Ultra High Frequency (UHF) RFID tag standard was created by EPC (Electronic Product Code) Global. The aim was to replace bar codes with machine-readable ones that could be read from a distance. The idea, however could not take firm ground due to limitations such as poor product design. Another visualization for Internet of Things came from Near Field Communication (NFC). Although currently limited, the scope for this field is bright. It is a step forward in RFID. The purpose for this implementation was to enable smart phones to read passive NFC tags. These tags could store Universal Resource Identifier (URI). They were portable, thin and small chips that could be attached to any device. Apple's share in this regard was also significant. In September 2014, Apple announced that iPhone6 would embed NFC support, promoting NFC to be a key player in IoT. A varied implementation for IoT comes from Optical tagging or Quick Response (QR) tags. These tags can serve as identifiers for devices. QR is particularly popular as every smart phone is equipped with a high-resolution camera. A QR code is extracted and read from the scene using the image-processing techniques. These methods yield a hidden text, number or URI. These QR codes are found in numerous products [1].

None of these technologies, however, has the scalability to fully enable Internet of Things. The scope and application of IoT are enormous. The major areas include agriculture, aerospace industries, environmental technology, and intelligent, embedded systems to name a few [5]. The participating devices range from very small Wi-Fi enabled devices to huge data centers. The key cost effective factors include size and battery requirements of the embedded devices. In order to reduce battery consumption and increase device portability, chip designers aim at producing smaller and more independent devices at a minimum investment [2].

The term "Cloud Computing" became popular after Google's CEO Eric Schmidt used it. It is another paradigm in the field of research. According to National Institute of Standard and Technologies (NIST), Cloud Computing is defined as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction". Cloud services have gained popularity over last few years. It is due to the fact that Cloud services provide ease of management and lesser hassle on the client's end. It is a platform in which system assets namely processor, memory, bus and bandwidth are provided as utilities that can be claimed and released per the user need via Internet. This low cost model with seemingly infinite storage and processing power has gained huge market acceptance over very short period. It is evident from the fact the tech-giants like Google and Amazon have shifted their services to Cloud. Cloud computing contains certain relatable features such as resource provisioning. A layered model of Cloud computing reveals 4 layers. These are the hardware layer, the infrastructure layer,

the platform layer and the application layer [6], [7]. Cloud services fall into three layers of the stack: Infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS). IaaS is the most basic category. It permits the purchase of IT infrastructure from a cloud-provider. PaaS implies the deployment of cloud computing services for software development. It enhances development, testing and managing software. The developers need not worry about underlying details of servers, networks and databases. SaaS is a scenario where delivered over the internet on subscription basis [8].

It is imperative here to understand the various Cloud deployment scenarios as well. They also fall into three major classes: Public, Private and Hybrid. Hybrid clouds combine public and private clouds, permitting greater flexibility and data sharing. In spite of the fact that Cloud services provide reduced cost and better development, there are certain downsides of it. Latency and intermittent connectivity are major contributors. Poor bandwidth results in slower performance and affected quality [9]. Security and non-negotiable terms of services also become bottleneck in production environment. Lack of support, minimal flexibility and limited knowledge about the system hardware can also become a retarding factor in visualizing Cloud services to its full potential. In certain applications such as health monitoring and emergency response, delay and downtime cannot be tolerated. Fog answers these downfalls using nearer to ground connection and larger coverage area. Fog computing brings the Cloud closer to the ground. It is a paradigm of managing a largely available, varied data requirements and data sets [10].

The purpose of Fog computing or fogging is to enhance efficiency and lessen the amount of data travelling to the cloud. It does so by entertaining the majority of device requests itself without forwarding them to the cloud. It results in better security and efficient utilization of bandwidth. In a Fog scenario, data processing takes place in a data hub or local gateway router. It is worth mentioning here that Fog Network complements the Cloud computing. It provides quick, short-term response and analysis at the edge. It provides deeper insight to system as multiple data points provide it with data [11]. Fog computing can be an ideal platform for enabling Internet of Things as it is the cloud nearer to ground [3]. Fog devices can be termed as nano Data Centers (nDCs). These are tiny servers located at deployment site. These nDCs can distribute data at the client site and entertain client requests locally [12]. If Internet of Things is implemented using Fog computing, the vision of IoT can be realized. It will enhance larger number of applications and practical scenarios [6]. In this study, we discuss a possibility of implementing IoT using Fog. It proposes a possible platform for IoT using Fog computing.

### III. ARCHITECTURE OF INTERNET OF THINGS

The research approach adopted in this paper is qualitative nature backed by literature review and analysis. Based on this pattern we derive a nearly predictable, temporal conclusion of implementing IoT using Fog. Majority of the papers reviewed are published during 2008-2017 [6].

A. Internet of Things Framework

IoT is still in its infancy. It is based on a multi-layered architecture. This layered approach divides the functionality such that varied requirements of different industries and businesses can be met [5]. The precise and accurate layered structure of IoT is still debatable and there is no common agreed-upon model for its implementation. Largely, IoT comprises of at least three layers: Perception, Network and Application Layer [13]. This model was developed in the early stages of its development. It defines the main theme of IoT but does not cover its implementation details very precisely.

- **The Perception Layer:** It is the physical layer. It comprises of sensors and actuators in the physical environment. The purpose of this layer is to gather the data from the environment. The main task of this layer is to gather data from various sources. These sources are heterogeneous in nature, ranging from 6LoWPAN to embedded systems deploying RFID or Optical tags.
- **The Network Layer:** This layer is all about connectivity. It works as a linking layer among various devices. It serves to connect smart devices, network hubs and servers. It is also responsible for transmitting and processing data received from various layers.
- **The Application Layer:** This layer is responsible for entertaining the user requests. It defines various scenarios in which IoT can be deployed.

These layers are further elaborated in the five-layer model as depicted in Fig. 1.

A more detailed layered approach is a five-layer model. It additionally includes two more layers of processing and business. The remaining three layers are discussed as follows:

- **Transport Layer:** It moves data from lower perception layer to higher layer for processing. It does so by deploying various network techniques such as Wireless, 3G, LAN, Bluetooth, RFID, etc.
- **The Processing Layer:** This layer is also known as the middleware layer. Its task is to store, analyze and process huge amount of data that comes from the lower layer i.e. the transport layer.
- **The Business Layer:** It manages the whole IoT system, including applications, deployed business models, etc.

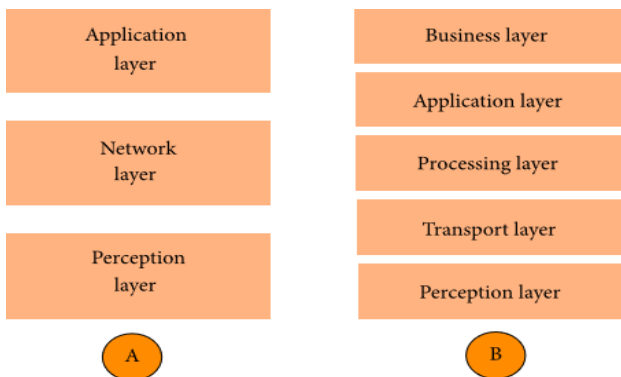


Fig. 1. Layered architecture of IoT [13].

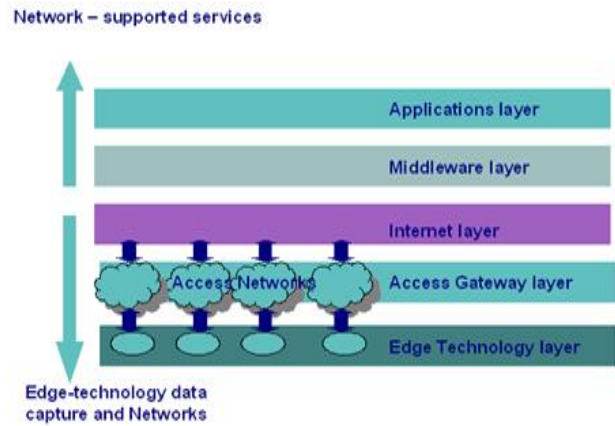


Fig. 2. Alternate architecture for IoT [5].

Another way to describe the layered architecture of IoT is by using Top-down approach. The layers essentially have the same functionality. This particular architecture focuses on standardization and interoperability [14]. This model has following layers: Application Service layer, Utility layer, IoT Service layer, and the Environment layer. Another literature describes this architecture as Technology Edge layer, Access Gateway layer, Middleware layer, Application layer and Internet layer [5]. These layers essentially perform the same functions.

Technology Edge layer is the physical layer comprising of sensors and actuators and are depicted in Fig. 2.

- **Access Gateway layer** is the network layer.
- **Middleware layer** is transport and processing layer together.
- **Internet layer and Application layer** are domain specific and application dependent.

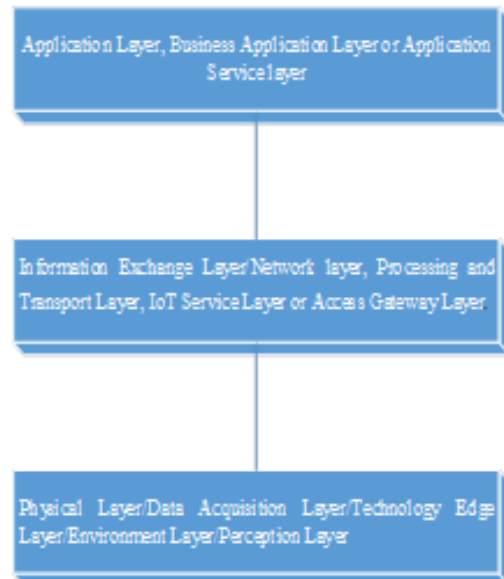


Fig. 3. Consolidated layered approach for IoT.

Another taxonomy describing the layered architecture of IoT breaks down the layers into Data Acquisition Layer, Information Exchange Layer and Application Layer [15] and as shown in Fig. 3. Let's explore the functionality of each layer in detail.

### B. Data Acquisition Layer

This layer is referred as the Physical Layer, Technology Edge Layer, Environment Layer or the Perception Layer. This is the layer closest to input devices. It is the layer that captures the user data. It comprises of multiple identification schemes and methods. Input devices include Bar Code Readers, RFID tags, Global Positioning System (GPS), sensors and actuators, embedded devices, Wireless Sensor Networks (WSN), Low-powered Wireless Personal Area Network (6LoWPAN) or simple Wi-Fi enabled devices. The main purpose of this layer is to capture data from the user in real time. It also must have object identification system. It comprises of multiple data acquisition methods and modules. It serves to complete the information perception of the IoT. It does so to provide a solid ground for the above layers in the model. Here the main challenge is to manage device heterogeneity and ability to capture large amount of data in real time.

### C. Information Exchange Layer

This layer is termed as Network layer, Processing and Transport Layer, IoT Service Layer or Access Gateway Layer. This layer is the heart of enabling IoT. It comprises of various communication technologies that form the backbone for envisioning IoT. Defining technologies here include network routing protocols, mobile communication technologies and the Internet. It is at this layer that rich information exchange and routing takes place. It includes information processing, information management and data convergence. This layer handles huge, intelligent and varied data. It is at this infrastructure layer that the vision of IoT can become a universal reality. A key player in this layer of information processing and management is Cloud. Cloud computing has already revolutionized the way information is stored, processed and delivered to business clients.

### D. Application Layer

This layer is termed as Application Layer, Business Application Layer or Application Service layer. It refers to the merging of various IoT solutions and technologies with industries. This layer addresses wide range of industrial problems. It is through this layer that IoT realizes its deep connection with real world. It includes a variety of servers. Its main function is the analysis of gathered data, adaptation to user needs, socialization and security of the data [15].

## IV. CLOUD COMPUTING AND THE FOG

The term Cloud computing has taken the industry by storm. It essentially means storage and access of data and instructions over the Internet [16]. It is a general term used for services that are controlled, managed and delivered through the Internet. It equips the companies with resource sharing and virtually infinite memory. Companies pay to the services provider by assuming Internet as a utility. It frees

organizations to focus on Solution design rather than worrying about maintaining the computing infrastructure [17].

### A. Cloud Computing Architecture

This section discusses in detail the underlying architecture and various deployment models for Cloud. The model for Cloud computing follows a modular or layered approach. The top and the bottom layer lightly bind each of the participating. It simplifies understanding and provides service-based understanding of the Cloud. The layered architecture also enhances modularity of resources. It means that the owner of each layer has its own milestones. Broadly speaking, the Cloud is divided into four layers: The Hardware/Datacenter Layer, the Infrastructure Layer, the Platform Layer and the Application Layer [18], [19]. These layers are depicted in the Fig. 4.

Here we describe each of these in detail:

- **The Hardware Layer:** This layer is implemented at the Data Centre and is responsible for all the physical devices, nodes or computers connected to the Cloud. These resources include servers, routers, switches, power and cooling systems, etc.
- **The Infrastructure Layer:** This layer is also termed as Virtualization Layer. It is a core layer in Cloud as it provides dynamic resource management. It creates a huge, virtual pool for hardware resources.
- **The Platform Layer:** Next in the stack is the Platform Layer. It consists of Operating Systems and Application frameworks. Its main purpose is to reduce the workload of the Infrastructure layer.
- **The Application Layer:** This layer is the top most layers. It contains the actual Cloud applications. These are different from traditional ones as they provide scalability and Bandwidth conservation.

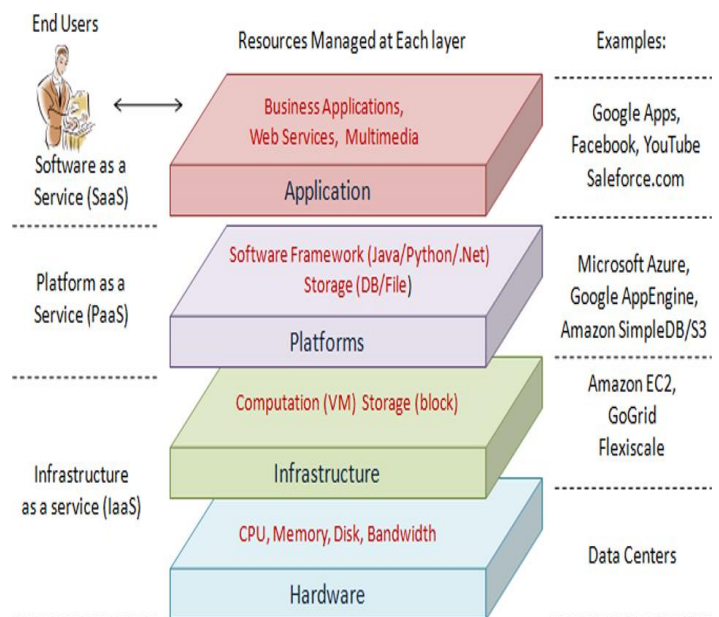


Fig. 4. Layers in a cloud [18].

### B. Disadvantages of Cloud Computing

Following are the main disadvantages of Cloud Computing [20] :

- **Downtime:** It is in fact the biggest downside for Cloud computing. Since Cloud services are dependent on the internet connection, it implies a loss of Internet connection can bring the entire service down. One of the most notable incidents took place in 2014, where DropBox faced an outage of around two days.
- **Privacy and Security:** By outsourcing the company data to an external provider, the company faces security and risk issues. One of the most notable examples was Code Space which was forced to close down after its console was hacked.
- **Vulnerability to Attack:** Nothing connected to the internet is perfectly secure. Even the best of teams and the most secure of the systems face cybercrimes and Vulnerability attacks.
- **Limited Control and Flexibility:** Cloud users have coarse grained control over the services provided by the Vendor. They often lack flexibility and agility.
- **Platform Dependency:** It is also termed as “Vendor Lock-in”. Finer differences between different vendor systems sometimes make migration to a newer platform very difficult.
- **Cloud Computing Costs:** For smaller scales and short term projects, Cloud can be hefty.

With the above mentioned disadvantages, it can be seen that there are certain downsides of the Cloud that need to be improved. But even with those downsides, Cloud has huge potential and scope for enhancing businesses and enterprise.

### C. The Fog

Fog computing is a term originally coined by Cisco [21]. Fog network are defined by their nearness to the Things and “Computational Density at the Edge of the Network”. An alternate definition is “system-level horizontal architecture that distributes resources and services anywhere along the continuum from Cloud to Things [22].” In contrast to Edge Computing, a fog may perform analysis on anything in the network from the core to the edge. Fog is essentially an extension of the Cloud, nearer to ground. The connected devices are called Fog nodes that can be placed anywhere in a Control System. Ideally, any smart device i.e. device with computational power, memory and network connectivity can be a Fog device. These typically include switches, routers, embedded sensors and surveillance cameras [23]. The applications can be developed utilizing specific, tiny pieces of code with minimum or no interaction to the Cloud [2]. It has following distinguishing features [22]:

- **Horizontal Architecture:** It distributes the services and applications horizontally along the industries and enterprises.
- **Cloud-to-Thing Continuum:** Real-time response for the Connected Vehicles and Smart grids. Bringing

Internet closer to Things. It is in fact, the connection between the Things: sensors, actuators, and Internet in real-time.

- **Location Awareness:** The participating Nodes are location sensitive and fully aware of their surroundings.
- **Low Latency:** Fog provides lower latency as it entertains the devices requests locally and in Real-time.
- **Wireless Access:** The participating devices are equipped with wireless sensors and receivers.

### D. How does Fog Work?

It follows a layered structure. It is divided into three layers: User Device Layer, Edge Node or the Fog Layer and finally the Cloud Layer [24]. They are shown in Fig. 5 and described below:

- **User-Device Layer:** It is the layer closest to the Things or the participating nodes. Here the data generated by the node is collected. It can be called as the Physical layer as well. It deals with various types of sensors, grids and real-time processing applications.
- **Edge-Node Layer or the Fog Layer:** It is the heart of the entire architecture. Here the data is stored, interpreted and analyzed. The most latency sensitive applications are processed immediately. These may include grid sensors or connected vehicles input. Here the data can be checked for errors and redundancy. Data that can tolerate delay is sent for analysis and action to aggregation node. Data collected for long term studying and prediction is sent to Cloud. It is also possible that plenty of Fog nodes collect similar data and send periodic updates for analysis and design study.
- **The Cloud Layer:** This layer is essentially the Cloud, described in the previous section.

A more detailed approach to Fog is shown in Fig. 6 and described below:

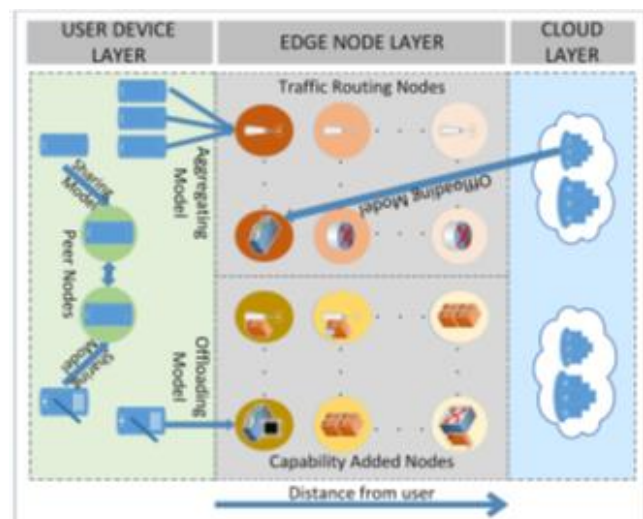


Fig. 5. Fog devices and the cloud [24].



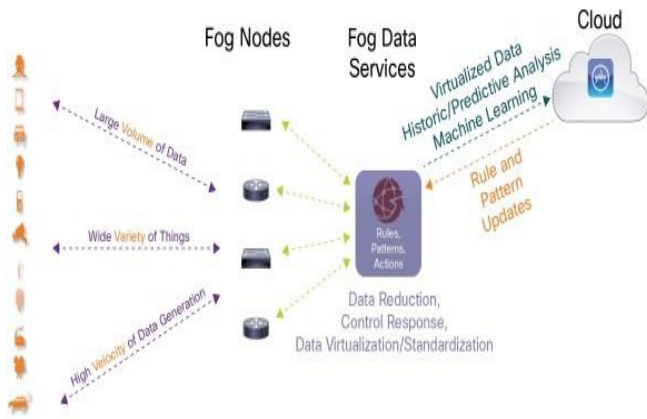


Fig. 6. Fog architecture.

TABLE I. COMPARATIVE ANALYSIS BETWEEN CLOUD AND FOG

Feature	Cloud	Fog
Downtime	Greater Downtime	Lowest possibility
Connectivity	Fewer devices	Multiple heterogenous devices
Platform Dependability	Platform dependence	Platform independence
Flexibility and Control	Coarse-grained	Fine-grained

The layers are essentially the same.

- The first layer or first tier, closer to the node is designed for Human To Machine (H2M) interaction. It enables the data collection, processing and control of the data.
- The second and third layer or tier is designed for Machine to Machine (M2M) interaction. It essentially performs data analysis, error detection [3].

Table I sums up the advantages of Fog over Cloud.

## V. RESULTS AND FINDINGS

As discussed above, the various Things, or devices that are connected to the IoT are heterogeneous in nature. They range in size, data type and communication speed. Most of them embody real-life applications and connecting them directly to the Cloud would be practically impossible. To handle the volume, variety and velocity of the IoT devices, we need a novel computational and analytical edge namely Fog. Traditional Cloud architecture does not meet all of these requirements completely. Moving the complete data to Cloud adds latency, consumes bandwidth and can seriously cripple Real-time applications. Moreover, Cloud communicates only based on IP and not on any other protocol used in industries. Thus, the realistic approach signifies that the IoT data must be processed and analyzed near the nodes. We call this scenario Fog Computing or simply the Fog. There are many valid reasons for it as already described.

### A. Conceptual Model of IoT using Fog Computing

Devices can be equipped with IoT-based applications, specifically for Fog nodes as shown in Fig. 7 and 8. The devices closest to these can handle this data and based on the required response, decide the ideal location. The time-critical

applications are processed first, nearest to the nodes, and then the data that can bear delayed processing or requires multiple inputs or multiple samples of input stream is sent to the level above. Data that requires analysis and long-time storage is sent to the Cloud. This layered approach reduces the latency time and speeds up the processing of the information. Response time can be reduced from minutes to seconds, days and weeks can be mapped in the similar analogy. The time for which the data is stored on the Cloud can also be adjusted accordingly.

Once at the Cloud, the received data can be aggregated, analyzed and summarized; from many Fog nodes to gain better insight into business applications, Big Data analysis and predicting the expected device/system response. These details can be further used to control the device modelling behavior and system response in general. There are great benefits associated with this approach. It promotes better business agility, enhanced security, much finer privacy controls and security and a lower operating expense.

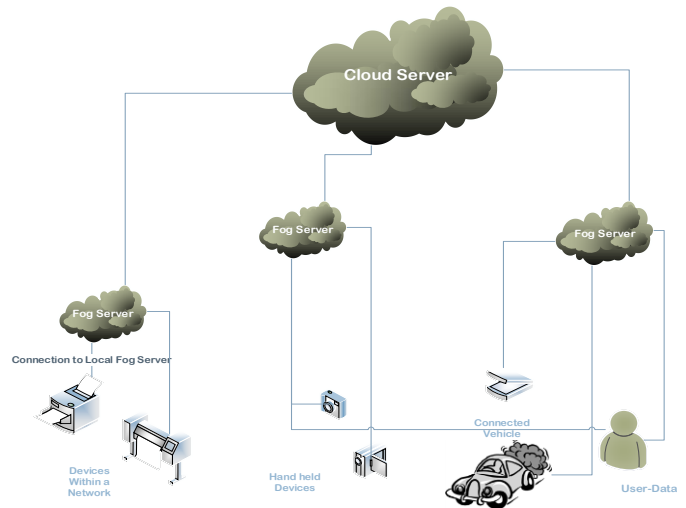


Fig. 7. Conceptual model of IoT using fog computing.

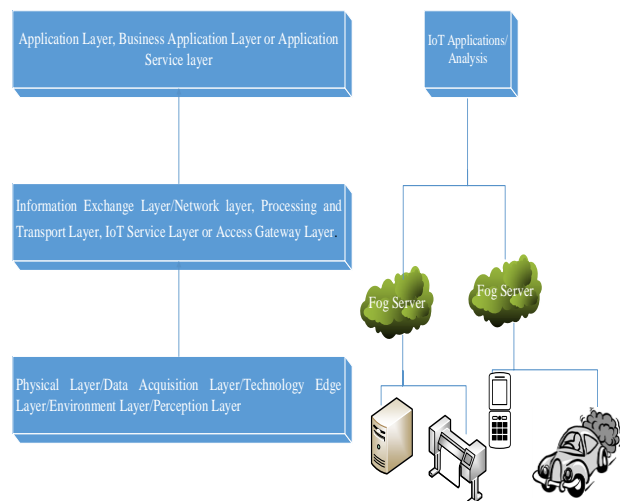


Fig. 8. Mapping IoT layers to devices.

### B. Internet of Things Architecture and the Fog

The proposed architecture for the Fog computing in IoT is depicted in the Fig. 9 below:

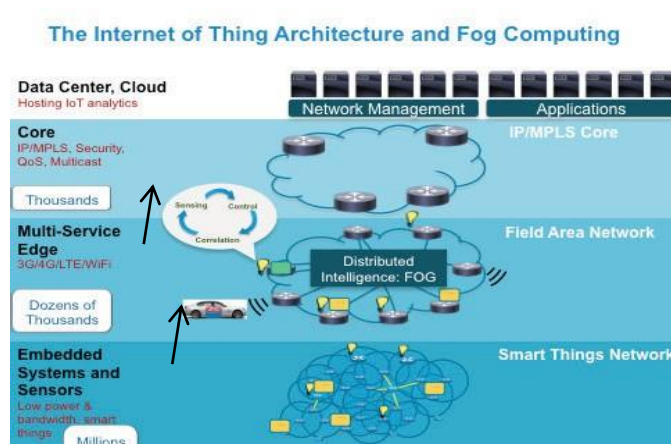


Fig. 9. The Internet of Things architecture and fog computing [3].

Beginning with the bottom most layer, the first layer is of “Things”. This layer defines the Human-to-Machine interaction. The first layer of this architecture collects the data and sends control instructions to the actuators. Here is the rich interconnected network of internet-connected devices. The second layer deals with Machine-to-Machine interaction, system processing, reporting and visualization. Here the time scale ranges from microseconds to minutes. This is the core Fog layer. Consequently, Fog must support a number of storage types and multiple inputs at the same time. It must also have wide geographical coverage and broader time scale. The Cloud provides the top most level coverage and data transmission. It is also used as a central repository for data storage; the storage capacity of which is virtually infinite and unbounded by time [3].

### C. Benefits of Fog Computing

There are numerous benefits of Fog. Some of the advantages are discussed below [23].

- **Greater Flexibility** for business development. Fog applications can be developed quickly and are more tailored for user needs.
- **Better Security:** Since majority of the data remains localized, the chances of identity theft and misuse are reduced.
- **Finer Control:** A finer control can be implemented to access, analysis and control of data.
- **Reduced Operational Cost:** Network Bandwidth can be conserved, as there is less data transfer to the Cloud.

## VI. DISCUSSION AND FUTURE WORK

The implementation of IoT using Fog computing is a still a new field and is in its infancy. Huge challenges lie ahead. These include definite layered architecture for IoT, communication technologies, Data and Signal processing technologies, Hardware modifications, Network technologies to name a few [5], [22].

- **Definite Layered Architecture:** The definite architecture and layered approach for IoT is still debatable. There is no set or standard that clearly defines the IoT layers.
- **Communication Technologies:** Since the participating devices are heterogeneous in nature, there is a lack of common communication technologies and protocols.
- **Network Technologies:** Lack of networking topologies, technologies and infrastructure in general is a major field of study, still unexplored.

Besides, Fog and its layered structure is still debatable. Hence, implementing IoT using Fog is a major field of research for future. If implemented successfully, IoT can revolutionize the way Internet operates today. Few application areas where IoT can be significant are discussed below:

### A. Transportation

In 2016, it was predicted that an average person produces 650MB of data every day and it is expected to double soon in future. IoT is important in transportation as it enables the creation of connected vehicle. It enhances low latency, user privacy and resource sharing at different layers. Information collection during peak hours, downloading content while travelling by pooling resources, tracking the position of the vehicle, surveillance cameras on roads and ticketing system in public transport can be visualized. These sub-applications can be entertained locally using Fog computing instead of sending updates to the Cloud. Similarly, safety systems can be activated. The tiebreaker here, the Fog preserves Bandwidth and provides response in Real-Time. Cloud can provide analysis and results such as which routes to choose. A Fog computing scenario for Connected Vehicle is described here [25]. Smart, self-directed vehicles will generate huge amount of data from various sources such as Light Detection and Ranging (LIDAR), Global Positioning System (GPS), etc. Here, the Things cover a broad spectrum: various types of sensors such as roadside sensors, on-vehicle sensors, numerous systems, associated data and functions. These Fog devices also manage actuators. In-vehicle Fog nodes can provide other certain services such as infotainment, advanced driver assistance systems (ADAS), possible collision detection and avoidance. The participating technologies include Dedicated Short Range Communications (DSRC), 3G, LTE etc. These technologies ensure connectivity and network availability [25].

### B. Smart Cities

The infrastructure and design of new cities can be greatly enhanced if IoT is visualized. Envisioning it using IoT can enhance basic city operations such as security, broadband connectivity and safety. Most of the cities provide internet access that leaves little room for the high-ended maintenance services, disaster management advanced municipal services [22]. Smart city is a mega application of IoT as it includes multi-dimensional data and various processing speeds. This majorly includes camera surveillance and camera deployments in places and remote areas that do not have Internet connectivity or Network coverage. This type of coverage is needed for uploading the collected videos, Real-time

monitoring and anomaly detection. Security and anomaly detection pose a significant challenge in this application. Intrusion detection, elderly health care and fire alarm are all low latency applications that require timeliness and quick response. Moreover, image identification is another addressable context. It implies that while capturing image data, confidential contextual data should not be made public. Fog enables such scenarios to be envisioned accurately and efficiently due to its close proximity with the participating nodes and horizontal distribution. Vector Algorithms, Video analysis and concurrent application processing can hasten the current applications. This is a brief application of Fog in Smart cities. The concept is further applied as smart parking detectors, shopping infrastructure, interlinking hospitals for greater and better services, intelligent highways, and factories that are all interconnected [25].

### C. Smart Buildings

Building automation and control are also model cases that demonstrate utilization of IoT. There may be thousands of sensors for recording various parameters such as temperature, humidity and parking space [22]. Smart buildings contain a rich interwoven system of sensors and actuators. They can be used to store different parameters such as temperature, humidity, number of people currently in the building, fire detectors, security and alarm. Some of the data is latency-sensitive and cannot be delayed. This data is to be analyzed and calculated locally. Typical examples include Security Breach and Fire Alarm [25].

#### REFERENCES

- [1] B. N. S. a. S. J. Roy Want, "Enabling the Internet of Things," THE IEEE COMPUTER SOCIETY, 2015.
- [2] L. R.-M. Luis M. Vaquero, "Finding your Way in the Fog: Towards a Comprehensive Definition of Fog Computing," ACM SIGCOMM Computer Communication Review, October 2014.
- [3] R. M. J. Z. S. A. Flavio Bonomi, "Fog Computing and Its Role in the Internet of Things," CISCO, Helsinki, Finland, August 17, 2012.
- [4] R. B. S. M. M. P. Jayavardhana Gubbia, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. Volume 29, no. Issue 7, p. Pages 1645–1660, September 2013, .
- [5] M. U. R. S. Gul Ahmad, "INTERNET OF THINGS (IOT): AN OVERVIEW," *Journal of Information & Communication Technology*, Vols. Vol. 10, No. 1, no. Vol. 10, No. 1, pp. 122-130, Spring 2016.
- [6] W. d. D. ., V. P. ., A. P. Alessio Botta, "Integration of Cloud computing and Internet of Things: A survey," *Future Generation Computer Systems*, p. 684–700, 3 October 2015.
- [7] L. C. ., R. B. Qi Zhang, "Cloud computing: state-of-the-art and research challenges," *J Internet Serv Appl (2010)*, vol. 1, pp. 7-18, 2010.
- [8] "https://azure.microsoft.com/en-in/overview/what-is-cloud-computing/," MICROSOFT, Saturday June 2017. [Online]. Available: https://azure.microsoft.com/en-in/overview/what-is-cloud-computing/. [Accessed 10 June 2017].
- [9] U. Singh, "https://www.linkedin.com/pulse/11-pros-cons-cloud-computing-everyone-should-know-umesh-singh," LinkedIn, April 27, 2015 April 27, 2015 April 27, 2015. [Online]. Available: https://www.linkedin.com/pulse/11-pros-cons-cloud-computing-everyone-should-know-umesh-singh. [Accessed 10 June 2017].
- [10] H. G. R. N. C. S. K. G. R. B. Amir Vahid Dastjerdi, "Fog Computing: Principles, Architectures, and Applications," in *Internet of Things: Principles and Paradigms*, ELSEVIER, 28 Jan 2016, pp. 61-75.
- [11] M. Rouse, "http://internetofthingsagenda.techtarget.com/definition/fog-computing-fogging," TECHTARGET NETWORK, December 2016 December 2016. [Online]. [Accessed 10 June 2017].
- [12] K. H. ., R. A. ., T. A. ., a. R. S. T. Fatemeh Jalali, "Fog Computing May Help to Save Energy in Cloud Computing," *IEEE Journal on Selected Areas in Communications*, 2015.
- [13] P. S. a. S. R. Sarangi, "Internet of Things: Architectures, Protocols, and Applications," *Journal of Electrical and Computer Engineering*, Vols. Volume 2017, , p. 25, 26 January 2017.
- [14] M. S. S. V. Bhagyashri Katole, "Principle Elements and Framework of Internet of Things," *International Journal Of Engineering And Science*, Vols. Vol.3, Issue 5, no. Vol.3, Issue 5, pp. 24-29, July 2013.
- [15] B. L. P. D. Hong ZHOU, "The Technology System Framework of the Internet of Things and its Application Research in Agriculture," Funding Project for Academic Human Resources Development, Beijing, 2010-2011.
- [16] B. P. M. M. TEAM, "http://me.pcmag.com/networking-communications-software-products/1758/feature/what-is-cloud-computing," Altus Inc, APRIL 18, 2015, 3:20 P.M. APRIL 18, 2015, 3:20 P.M. APRIL 18, 2015, 3:20 P.M.. [Online]. Available: http://me.pcmag.com/networking-communications-software-products/1758/feature/what-is-cloud-computing. [Accessed 11 June 2017].
- [17] S. J. B. Margaret Rouse, "http://searchcloudcomputing.techtarget.com/definition/cloud-computing," TECHTARGET NETWORK, October 2016 October 2016. [Online]. Available: http://searchcloudcomputing.techtarget.com/definition/cloud-computing. [Accessed 11 June 2017].
- [18] Q. Z. ., L. C. ., R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *Springer:J Internet Serv Appl (2010)*, vol. 1, pp. 7-18, 2010.
- [19] K. M. Yashpalsinh Jadeja, "Cloud Computing - Concepts, Architecture and Challenges," in *2012 International Conference on Computing, Electronics and Electrical Technologies [ICCEET]*, Gujarat, India, 2012.
- [20] S. Seshachala, "https://cloudacademy.com/blog/disadvantages-of-cloud-computing/," Cloud Academy Blog, 17 Mar, 2015 17 Mar, 2015. [Online]. Available: https://cloudacademy.com/blog/?s=Disadvantages+of+Cloud+Computin g. [Accessed Tuesday,13 June 2017].
- [21] J. McKendrick, "Fog Computing: a New IoT Architecture?," RTInsights, 2017. [Online]. Available: https://www.rtinsights.com/what-is-fog-computing-open-consortium/. [Accessed 13 June 2017].
- [22] OpenFog, "https://www.openfogconsortium.org/resources/," OpenFog, [Online]. Available: https://www.openfogconsortium.org/resources/#definition-of-fog-computing. [Accessed 14 June 2017].
- [23] "www.cisco.com/c/dam/en\_us/solutions/trends/iot/docs/computing-overview.pdf," CISCO, 2015. [Online]. Available: www.cisco.com/c/dam/en\_us/solutions/trends/iot/docs/computing-overview.pdf. [Accessed 13 June 2017].
- [24] N. W. D. S. N. Blesson Varghese, "Feasibility of Fog Computing," 19 Jan 2017 19 Jan 2017 19 Jan 2017. [Online]. Available: https://www.openfogconsortium.org/wp-content/uploads/Varghese-FogComputing.pdf. [Accessed 13 June 2017].
- [25] O. Consortium, "https://www.openfogconsortium.org/wp-content/uploads/OpenFog\_Reference\_Architecture\_2\_09\_17-FINAL-1.pdf," OpenFog Consortium, February 2017. [Online]. Available: https://www.openfogconsortium.org/ra/technical-document-download/. [Accessed 19 July 2017].

# A Group Decision-Making Method for Selecting Cloud Computing Service Model

Ibrahim M. Al-Jabri

Department of Accounting and  
Management Information Systems  
College of Industrial Management  
King Fahd University of Petroleum  
and Minerals  
Dhahran, Saudi Arabia

Mustafa I. Eid

Dammam Community College  
King Fahd University of Petroleum  
and Minerals  
Dhahran, Saudi Arabia

M. Sadiq Sohail

Department of Management and  
Marketing  
College of Industrial Management  
King Fahd University of Petroleum  
and Minerals  
Dhahran, Saudi Arabia

**Abstract**—Cloud computing is a new technology that has great potential for the business world. Many business firms have implemented, are implementing, or planning to implement cloud computing technology. The cloud computing resources are delivered in various forms of service models which make it challenging for business customers to select the model that suits their business needs. This paper proposes a novel group-based decision-making method where a group of decision makers is involved in the decision process. Each decision maker provides weights for the cloud selection criteria. Based on weight aggregations and deviations, decision makers would select the alternative which has the highest ratio of deviation to mean is selected. The method is illustrated with an example on the selection of cloud service models. This method is useful for IT managers in selecting the appropriate cloud service model for their organizations.

**Keywords**—Cloud computing; cloud service models; multi-criteria decision-making; group decision-making

## I. INTRODUCTION

According to Forrester [1], the projected public cloud market will generate a revenue of US\$191 billion by 2020. This includes US\$133 billion for cloud applications, US\$44 billion for cloud platforms, and US\$14 billion for cloud business services. Etro [2] reported that cloud computing tends to increase in new business formation in European economies as it reduces cost of entry into a market by saving in capital expenditure on IT. The European Business Research Center estimates cloud computing would generate, between 2010 and 2015, a cumulative increase in output of €763 billion in five European countries (France, Germany, Italy, Spain and the UK), and an increase in employment of 2.3 million [3]. During these five years, the CEBR predicted that the annual economic benefits would be more than €177 billion and an annual increase of 446,000 jobs.

In an era of information and globalization, an immense computing power is required to empower business intelligence and competitive gains. Nonetheless, operating a private data centre and managing software licensing to meet a growing computing processing demands is complex and costly. Cloud computing represents a shift in computing paradigm which comprises outsourcing of computing resources with characteristics like on-demand self-service, resources scalability, zero up-front investment, and measured services; it

also promises to provide a solution in the form of on-demand computing, swift deployment, little required maintenance, fewer IT staff and low cost [4]. Such captivating promises has made this technology a primary academic research and business media topic over the last few years. However, serious security and privacy concerns have made businesses reluctant to deploy cloud computing [5]-[13].

Due to the immense benefits, opportunities and serious concerns in adopting cloud computing [14]-[16], it is important to select the right cloud service model that satisfies the business requirements. Businesses face a number of challenging decisions with respect to the selection of the appropriate cloud service models, like SaaS, PaaS, or IaaS. The decision involves considering organizational and technological factors, business information needs, and budget requirements. The decision is a challenging and complex because it requires due consideration of several conflicting factors that need to be dealt with simultaneously. This study proposes a novel method that is based on MCDM and incorporates a group of decision makers in the selection of cloud service model. The method aggregates the weights of the selection criteria for each decision maker and rank the cloud service models based on the ratio of deviation among to aggregate mean of the decision makers.

The rest of the paper is organised as follows. Section 2 provides an overview of cloud computing and the selection factors of cloud computing models. Section 3 presents different multi-criteria decision-making methods in cloud computing. Section 4 explains in details the proposed method in selecting cloud computing service model. Section 5 demonstrates the proposed method with a numerical example. Section 6 concludes the papers and offers future research direction.

## II. OVERVIEW OF CLOUD COMPUTING

The idea of delivering software application and computing processing power from a computer network herein labelled “cloud computing” is not entirely new. Cloud computing has its roots within grid computing, service-oriented architecture, distributing computing, and virtualization [17]-[20]. John McCarthy, in his speech at the MIT Centennial in 1961, predicted that computing would become a public utility [21]. Carr [22] predicted that IT resources are going to be supplied as services in a manner similar to the supply of electricity by

power companies. Power companies based on demand and charged based on use, and no need for households or factories to run dedicated power generators to supply electricity. He also argued that since IT services would be available to everyone, companies would not anymore consider IT as a competitive weapon. Carr's viewpoint of IT being supplied based on demand and charged based on use like electricity is very much supported by the emergence of cloud computing technology.

Cloud computing has become the topic of almost every IT forum today. Over the years, IT has made remarkable advancements. Most notable advances are in the areas of virtualization, hardware and software infrastructure and web technologies. Cloud computing enables user to gain access to information and to lower the barriers to computing. With cloud computing, the need to maintain technology infrastructure fades out as the burden of system management and data protection is shifted to cloud service providers [19]. The adoption of cloud computing is changing IT service delivery models, enabling changes in IT agility, re-engineering business processes, revolutionizing the use of applications, and interacting with consumers and other companies.

### A. Defining Cloud Computing

Both academics and industry have proposed various definitions for cloud computing but no one definition has gained mutual consensus so far. Buyya et al. [18] defined cloud computing as "a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resource(s) based on service-level agreements established through negotiation between the service provider and consumers." According to Armbrust et al. [23], "Cloud computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those service." Kramer [24] defined it as "a new computing paradigm, which changes the purchasing, maintenance and disposal process of IT by providing on-demand procurement of a dynamic basket of IT resources, these resources are hosted in specialized data centers and can be purchased and scaled over the Internet, on-demand and location independently." The most acknowledged definition of cloud computing in literature is the one presented by the National Institute of Standards and Technology (NIST). NIST defines cloud computing as "A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction", [25, p. 11]. Furthermore, NIST described the cloud computing as consisting of five essential characteristics, three service delivery models, and four deployment models. The essential characteristics are broad network access, measured service, rapid elasticity and on-demand self-service. The service models are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS); and the deployment models are private cloud, public cloud, hybrid cloud, and community cloud. The five essential characteristics, three cloud service models, and four cloud deployment models are depicted in Fig. 1.

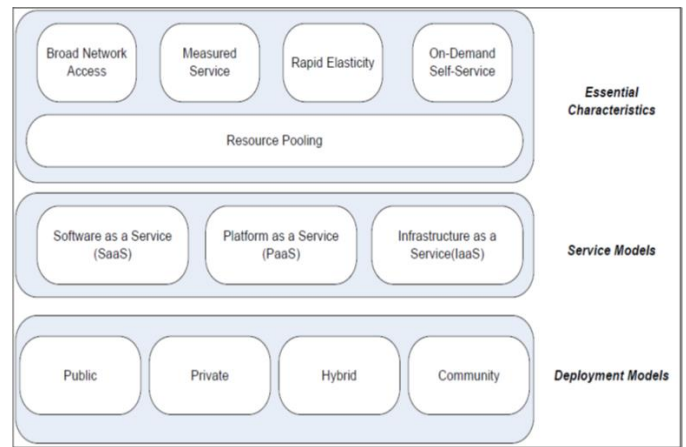


Fig. 1. NIST visual model of cloud computing definition [26].

### B. The Service Delivery Models

- Software as a Service (SaaS) provides applications and software to the consumer. The applications are accessible from various client devices through either a thin client interface, such as a web browser, or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems or storage devices.
- Platform as a Service (PaaS) is the deployment of operating systems, programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, or operating systems, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.
- Infrastructure as a Service (IaaS) is the provision of basic hardware and software needed for processing power, storage space, communication networks, and other necessary computing resources where the consumer is able to deploy and run the needed system and application software.

### C. The Selection Factors of Cloud Computing Service Models

Eid et al. [27] conducted an extensive review of literature where they identified the factors that affect the adoption of cloud computing in organizations. We will select the most important factors we believe very relevant to the selection of a service delivery model. The following seven factors along with their brief descriptions represent organizational and technical conditions relevant to the selection of service delivery models:

- Cost: the cost of leasing service delivery model (SaaS, PaaS, or IaaS). It includes total amount charged by the cloud provider as well as maintenance and support cost [28].
- Adaptability: the level of service delivery model flexibility with respect to changing user requirements, and needs of the organization adopting cloud computing [29].

- Available IT skills: degree of IT skills availability in the organization adopting cloud computing [30].
- Urgency: degree of urgency of needed cloud service which allows for faster deployment and immediate cloud service provisioning [31].
- Security of data: security level of the used service and client data maintained by the cloud deployment model. This includes: 1) data integrity (data accuracy and recovery), 2) level of audibility and 3) access control [28], [32], [33].
- Privacy of data: degree of confidentiality of data maintained by the cloud provider [32].
- Service reliability: the extent to which the service is available without interruption or with minimum downtime.

### III. MULTI-CRITERIA DECISION-MAKING IN CLOUD COMPUTING

Multi-criteria decision-making (MCDM) is a branch of operations research/management science that is concerned with the methods and techniques to solve the multi-criteria decision problems. Gavade [34] classified multi-criteria decision-making problems into two categories:

- Multiple attribute decision-making (MADM): MADM involves the selection of the “best” alternative from pre-specified alternatives described in terms of multiple attributes; and
- Multiple objective decision-making (MODM): MODM involves the design of alternatives which optimize the multiple objectives of Decision Maker (DM).

Multi-Criteria Decision-Making (MCDM) provides effective approach in many economical, manufacturing, material, service selection problems [35]. It specifically plays an important role in areas of investment decision, product evaluation, staff appraisal and others [34]. Despite a long history, researchers constantly develop methods based on the MCDM approach. These methods differ in both implementation details and scope of application. Each method has its own strengths and weaknesses. There are several methods of multi-criteria decision-making. In their research, Whaiduzzaman et al. [35] provided a taxonomy of MCDM-

based methods along with their objectives, criteria/approach, strengths, and limitations. Examples of the reviewed MCDM methods are:

- Analytic Hierarchy Process (AHP).
- Analytic Network Process (ANP).
- Technique for Order of Preferences by Similarity to Ideal Solution (TOPSIS).
- Elimination and Et Choice Translating Reality (ELECTRE).
- Preference Ranking Organization METHod of Enrichment Evaluations (PROMETHEE).
- Decision-Making Trial and Evaluation Laboratory (DEMATEL).
- Grey Relational Analysis (GRA).
- Simple Additive Weighting (SAW).
- Fuzzy MCDM.
- Data Envelopment Analysis (DEA).

As cloud computing technology adoption has become more popular during the last few decades, researchers have paid more attention to address the managerial decision-making issues faced by organizations interested to adopt cloud computing. Conway and Curry [35] addressed the management of cloud computing adoption from a lifecycle approach perspective. They developed a lifecycle model for managing cloud-computing adoption. Whaiduzzaman et al. [36] focused on addressing the service selection for cloud computing using the multi-criteria decision-making approach. They described the multi-criteria decision analysis (MCDA) types and characteristics and compared several methods by synthesizing and reviewing the present literature. The selection of cloud service models by organizations necessitates the consideration of a number of related conflicting factors that are relevant to the cloud service models and organizational requirements. In such multiple criteria decision situations, a compromise or tradeoff has to be made because in most real-world situations, no single alternative satisfies all criteria but one alternative may be better in terms of some of the criteria while other alternatives may outperform it, if judged based on the remaining criteria [37].

TABLE I. SELECTION PROBLEMS IN CLOUD COMPUTING

SN	Selection Problems	MCDM Method	Reference
1	Cloud service selection	BSC, FDM, FAHP	[38]
2	Cloud computing vendor selection	TOPSIS, SAW, AHP	[39]
3	Adoption of cloud computing services	ANP	[16]
4	Selecting cloud computing deployment model	AHP	[40]
5	Ranking of cloud computing services	AHP	[28]
6	Public cloud service selection	SAW	[41]
7	SaaS vendor selection	AHP	[42]
8	Selection of IaaS cloud service	ANP	[43]
9	Selection of public cloud service	TOPSIS	[29]

Recent technological developments in cloud computing and its adoption by organizations presents a new set of problems to managers. One of the major problems is the selection of the right MCDM method. Several examples are found in literature where MCDM approach are applied to the decision-making process of cloud computing technology adoption by top managers. Table I provide examples of such problems and applied MCDM methods.

IV. PROPOSED METHOD

The origin of the proposed method is based on the Multi-Criteria Decision-Making (MCDM). The novelty of this method is the use of MCDM in a group decision-making setting where alternatives are ranked with the least amount of variability relative to the mean. Theoretically, the best alternative is the one with the high aggregate mean and low variation values among decision makers. The ratio of deviation to the mean is the coefficient of variation. The proposed method involves the following steps:

Step 1: List the alternatives (A1, A2 A3, Am) where j ∈ m and m is the number of alternatives

Step 2: Identify the selection criteria (C1, C2, C3, Cn) where i ∈ n and n is the number of criteria

Step 3: Invite participating decision makers (D1, D2, D3, Dd) where k ∈ d and d is the number of decision makers

Step 4: Assign importance score (s<sub>i</sub><sup>k</sup>) of i<sup>th</sup> criterion given by k<sup>th</sup> decision maker. The importance score (s<sub>i</sub><sup>k</sup>) is assigned using a Likert-type scale ranging from 1 to 7, where:

- 1= not at all important
- 2= not important
- 3= Somewhat not important
- 4= neutral
- 5= somewhat important
- 6= important
- 7= very important

The importance score matrix is shown in Table II.

TABLE II. IMPORTANCE SCORE OF ATTRIBUTES (s<sub>i</sub><sup>k</sup>)

D. Maker (k) \ Criteria (i)	1	2	.	.	.	d
1	s <sub>1</sub> <sup>1</sup>	s <sub>1</sub> <sup>2</sup>	.	.	.	s <sub>1</sub> <sup>d</sup>
2	s <sub>2</sub> <sup>1</sup>	s <sub>2</sub> <sup>2</sup>	.	.	.	s <sub>2</sub> <sup>d</sup>
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	s <sub>n</sub> <sup>1</sup>	s <sub>n</sub> <sup>2</sup>	.	.	.	s <sub>n</sub> <sup>d</sup>

TABLE III. RELATIVE IMPORTANCE OF ATTRIBUTES (w<sub>i</sub><sup>k</sup>)

D. Maker (k) \ Criteria (i)	1	2	.	.	.	d
1	w <sub>1</sub> <sup>1</sup>	w <sub>1</sub> <sup>2</sup>	.	.	.	w <sub>1</sub> <sup>d</sup>
2	w <sub>2</sub> <sup>1</sup>	w <sub>2</sub> <sup>2</sup>	.	.	.	w <sub>2</sub> <sup>d</sup>
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
N	w <sub>n</sub> <sup>1</sup>	w <sub>n</sub> <sup>2</sup>	.	.	.	w <sub>n</sub> <sup>d</sup>

Step 5: Calculate the relative importance (w<sub>i</sub><sup>k</sup>) of i<sup>th</sup> criterion given by k<sup>th</sup> decision maker, where:

$$w_i^k = s_i^k / (\sum_{i=1}^n s_i^k) \text{ for } k = 1, 2, \dots, d \tag{1}$$

The relative importance score matrix is shown in Table III.

Step 6: Assign importance score (v<sub>ij</sub><sup>k</sup>) of i<sup>th</sup> criterion with respect to the j<sup>th</sup> alternative given by the k<sup>th</sup> decision maker. The importance score (v<sub>ij</sub><sup>k</sup>) is assigned using a Likert-type scale ranging from 1 to 7, where:

- 1= not at all important
- 2= not important
- 3= somewhat not important
- 4= neutral
- 5= somewhat important
- 6= important
- 7= very important

The importance score matrix is shown in Table IV.

Step 7: Calculate the decision values (u<sub>ij</sub><sup>k</sup>) of i<sup>th</sup> criterion with respect to j<sup>th</sup> alternative given by the k<sup>th</sup> decision maker, where:

$$u_{ij}^k = w_i^k * v_{ij}^k \text{ for } k = 1, 2, \dots, d \text{ and } \begin{cases} i = 1, 2, \dots, n \\ j = 1, 2, \dots, m \end{cases} \tag{2}$$

Step 8: Aggregate all decision values (z<sub>j</sub><sup>k</sup>) of j<sup>th</sup> alternative by k<sup>th</sup> decision maker, where:

$$z_j^k = \sum_{i=1}^n u_{ij}^k \text{ for } k = 1, 2, \dots, d \text{ and } j = 1, 2, \dots, m \tag{3}$$

Step 9: Calculate the average decision value (z̄<sub>j</sub>) of j<sup>th</sup> alternative by all decision makers, where:

$$\bar{z}_j = (\sum_{k=1}^d z_j^k) / d \text{ for } j = 1, 2, \dots, m \tag{4}$$

Step 10: Calculate the standard deviation (σ<sub>z<sub>j</sub></sub>) of aggregate decision values of j<sup>th</sup> alternative, where:

$$\sigma_{z_j} = \sqrt{\frac{\sum_{k=1}^d (z_j^k - \bar{z}_j)^2}{d-1}} \text{ for } j = 1, 2, \dots, m \tag{5}$$

Step 11: Calculate the coefficient of variation ( $cv_{z_j}$ ) of aggregate decision value of  $j^{th}$  alternative, where:

$$cv_{z_j} = \frac{\sigma_{z_j}}{\bar{z}_j} \quad \text{for } j = 1, 2, \dots, m \quad (6)$$

The decision value ( $u_{ij}^k$ ), aggregate decision values ( $z_j^k$ ), average aggregate decision value ( $\bar{z}_j$ ), and standard deviation

( $\sigma_{z_j}$ ) and coefficient of variation ( $cv_{z_j}$ ) of aggregate decision values are presented in Table V.

Step 12: Rank all alternatives and select the best alternative. The best alternative is the one which has the highest mean and lowest standard deviation. That is, the lowest coefficient of variation.

TABLE IV. IMPORTANCE SCORE OF ATTRIBUTES WITH RESPECT TO ALTERNATIVES ( $v_{ij}^k$ )

Alternative (j)	1				2				.				m			
D. Maker (k)	1	2	.	d	1	2	.	d	1	2	.	d	1	2	.	d
Criteria (i)	1	2	.	d	1	2	.	d	1	2	.	d	1	2	.	d
1	$v_{11}^1$	$v_{11}^2$	.	$v_{11}^d$	$v_{12}^1$	$v_{12}^2$	.	$v_{12}^d$	.	.	.	.	$v_{1m}^1$	$v_{1m}^2$	.	$v_{1m}^d$
2	$v_{21}^1$	$v_{21}^2$	.	$v_{21}^d$	$v_{22}^1$	$v_{22}^2$	.	$v_{22}^d$	.	.	.	.	$v_{2m}^1$	$v_{2m}^2$	.	$v_{2m}^d$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
N	$v_{n1}^1$	$v_{n1}^2$	.	$v_{n1}^d$	$v_{n2}^1$	$v_{n2}^2$	.	$v_{n2}^d$	.	.	.	.	$v_{nm}^1$	$v_{nm}^2$	.	$v_{nm}^d$

TABLE V. DECISION VALUE STATISTICS

Alternative (j)	1				2				.				m			
D. Maker (k)	1	2	.	d	1	2	.	d	1	2	.	d	1	2	.	d
Criteria (i)	1	2	.	d	1	2	.	d	1	2	.	d	1	2	.	d
1	$u_{11}^1$	$u_{11}^2$	.	$u_{11}^d$	$u_{12}^1$	$u_{12}^2$	.	$u_{12}^d$	.	.	.	.	$u_{1m}^1$	$u_{1m}^2$	.	$u_{1m}^d$
2	$u_{21}^1$	$u_{21}^2$	.	$u_{21}^d$	$u_{22}^1$	$u_{22}^2$	.	$u_{22}^d$	.	.	.	.	$u_{2m}^1$	$u_{2m}^2$	.	$u_{2m}^d$
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
n	$u_{n1}^1$	$u_{n1}^2$	.	$u_{n1}^d$	$u_{n2}^1$	$u_{n2}^2$	.	$u_{n2}^d$	.	.	.	.	$u_{nm}^1$	$u_{nm}^2$	.	$u_{nm}^d$
$z_j^k$	$z_1^1$	$z_1^2$	.	$z_1^d$	$z_2^1$	$z_2^2$	.	$z_2^d$	.	.	.	.	$z_m^1$	$z_m^2$	.	$z_m^d$
$\bar{z}_j$	$\bar{z}_1$				$\bar{z}_2$				.	$\bar{z}_m$						
$\sigma_{z_j}$	$\sigma_{z_1}$				$\sigma_{z_2}$				.	$\sigma_{z_m}$						
$cv_{z_j}$	$cv_{z_1}$				$cv_{z_2}$				.	$cv_{z_m}$						

V. AN EXAMPLE

In this section, we present an example to illustrate the process and application of the proposed approach. The example is about a firm wants to identify which cloud service model is suitable for its business requirements and technical environment. There are three options/alternatives (A1, A2 and A3) for service cloud models. There are five decision makers (D1, D2, D3, D4 and D5) who will collaborate to make the suitable decision for the firm. The decision makers will evaluate the alternatives according to seven criteria (C1, C2, C3, C4, C5, C6 and C7). The decision goal, alternatives, evaluation criteria and decision makers are listed in Table VI.

The criteria are assessed in linguistic terms, using Likert-scale, as follows:

- 1= not at all important
- 2= not important
- 3= Somewhat not important
- 4= neutral
- 5= somewhat important
- 6= important
- 7= very important



TABLE VI. GOAL, CRITERIA AND ALTERNATIVES

Decision goal	Select the cloud computing service model that will be the most suitable for the firm
Alternatives	A <sub>1</sub> : Software as a Service (SaaS) A <sub>2</sub> : Platform as a Service (PaaS) A <sub>3</sub> : Infrastructure as a Service (IaaS)
Criteria	C <sub>1</sub> : Cost of cloud service C <sub>2</sub> : Service adaptability C <sub>3</sub> : IT skills availability C <sub>4</sub> : Urgency of needed cloud service C <sub>5</sub> : Data security C <sub>6</sub> : Data privacy C <sub>7</sub> : Performance
Decision makers	D <sub>1</sub> : Chief executive officer (CEO) D <sub>2</sub> : Chief information officer (CIO) D <sub>3</sub> : Chief technology officer (CTO) D <sub>4</sub> : Chief technology officer (CTO) D <sub>5</sub> : Consultant

TABLE VII. IMPORTANCE SCORE OF ATTRIBUTES (s<sub>i</sub><sup>k</sup>)

Attribute (i)	Decision Maker (k)				
	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>
C <sub>1</sub>	7	7	7	6	7
C <sub>2</sub>	6	6	7	5	5
C <sub>3</sub>	4	4	6	3	6
C <sub>4</sub>	6	6	6	6	6
C <sub>5</sub>	5	6	4	4	4
C <sub>6</sub>	7	7	5	6	4
C <sub>7</sub>	7	6	6	7	7
( $\sum_{i=1}^n s_i^k$ )	42	29	41	24	39

TABLE IX. IMPORTANCE SCORE OF ATTRIBUTES WITH RESPECT TO ALTERNATIVES (v<sub>ij</sub><sup>k</sup>)

Criterion (i)	A <sub>1</sub> =SaaS					A <sub>2</sub> =PaaS					A <sub>3</sub> =IaaS				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C <sub>1</sub>	7	5	3	7	7	1	5	5	6	5	7	6	6	5	6
C <sub>2</sub>	7	6	2	5	5	3	7	5	6	6	7	6	4	7	6
C <sub>3</sub>	5	4	2	4	7	2	6	7	5	7	5	7	6	7	7
C <sub>4</sub>	5	6	7	7	7	3	5	6	4	6	3	6	5	3	5
C <sub>5</sub>	4	5	6	6	6	7	6	5	5	6	7	7	6	5	6
C <sub>6</sub>	6	5	5	5	5	7	6	6	5	5	6	6	7	7	5
C <sub>7</sub>	7	6	6	6	4	6	7	6	7	5	5	6	6	7	7

The decision makers uses the linguistic terms above to rate the general importance of the selection criteria (s<sub>i</sub><sup>k</sup>), as shown in Table VII.

The relative importance of criteria (w<sub>i</sub><sup>k</sup>) is calculated using (1), as shown in Table VIII.

TABLE VIII. RELATIVE IMPORTANCE OF ATTRIBUTES (w<sub>i</sub><sup>k</sup>)

Attribute (i)	D.M (k)				
	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>4</sub>	D <sub>5</sub>
C <sub>1</sub>	0.304	0.304	0.269	0.300	0.292
C <sub>2</sub>	0.261	0.261	0.269	0.250	0.208
C <sub>3</sub>	0.174	0.174	0.231	0.150	0.250
C <sub>4</sub>	0.261	0.261	0.231	0.300	0.250
C <sub>5</sub>	0.119	0.207	0.098	0.167	0.103
C <sub>6</sub>	0.167	0.241	0.122	0.250	0.103
C <sub>7</sub>	0.167	0.207	0.146	0.292	0.179

The decision makers rate the importance of criteria with respect to alternatives (v<sub>ij</sub><sup>k</sup>) using the linguistic terms mentioned above, to rate the importance of the selection criteria, as shown in Table IX.

Using (2), (3), (4), (5) and (6), we compute the decision values (u<sub>j</sub><sup>k</sup>), aggregate decision values (z<sub>j</sub><sup>k</sup>), average aggregate decision value (z̄<sub>j</sub>), and standard deviation (σ<sub>z<sub>j</sub></sub>) and coefficient of variation (cv<sub>z<sub>j</sub></sub>) of aggregate decision values, as shown in Table X.

TABLE X. DECISION VALUE STATISTICS

Alternative (j) Attribute (i)	A <sub>1</sub> =SaaS					A <sub>2</sub> =PaaS					A <sub>3</sub> =IaaS				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
C <sub>1</sub>	2.130	1.522	0.808	2.100	2.042	0.304	1.522	1.346	1.800	1.458	2.130	1.826	1.615	1.500	1.750
C <sub>2</sub>	1.826	1.565	0.538	1.250	1.042	0.783	1.826	1.346	1.500	1.250	1.826	1.565	1.077	1.750	1.250
C <sub>3</sub>	0.870	0.696	0.462	0.600	1.750	0.348	1.043	1.615	0.750	1.750	0.870	1.217	1.385	1.050	1.750
C <sub>4</sub>	1.304	1.565	1.615	2.100	1.750	0.783	1.304	1.385	1.200	1.500	0.783	1.565	1.154	0.900	1.250
C <sub>5</sub>	0.476	1.034	0.585	1.000	0.615	0.833	1.241	0.488	0.833	0.615	0.833	1.448	0.585	0.833	0.615
C <sub>6</sub>	1.000	1.207	0.610	1.250	0.513	1.167	1.448	0.732	1.250	0.513	1.000	1.448	0.854	1.750	0.513
C <sub>7</sub>	1.167	1.241	0.878	1.750	0.718	1.000	1.448	0.878	2.042	0.897	0.833	1.241	0.878	2.042	1.256
$z_j^k$	6.000	7.724	4.244	9.042	5.897	5.000	8.655	5.707	8.500	5.692	5.738	9.034	5.634	8.958	6.077
$\bar{z}_j$	6.581					6.711					7.088				
$\sigma_{z_j}$	1.846					1.729					1.750				
$cv_{z_j}$	0.280					0.258					0.247				
Ranking	3					2					1				

Based on the least ratio of variability  $cv_{z_j}$  among the decision makers, the alternatives are ranked as IaaS, PaaS and SaaS. Therefore, IaaS is the most preferred cloud service model for the firm, followed by PaaS and SaaS.

### VI. CONCLUSION AND FUTURE WORK

In this paper, we have identified a set of important factors for selecting cloud computing service models. These factors are available budget, adaptability to changing user requirements, available IT skills, urgency of needed service, data security, data privacy and reliability of service. We also proposed a novel multi-criteria approach that takes into consideration the aggregate mean of decision values and the deviations values among the decision makers. Despite the novelty of this approach, it can be improved by adding more selection criteria, like interoperability, performance, scalability, compatibility, complexity and vendor credibility and support.

### ACKNOWLEDGMENT

This project was funded by the National Plan for Science, Technology, and Innovation (MAARIFAH) – King Abdulaziz City for Science & Technology through the Science & Technology Unit at King Fahd University of Petroleum & Minerals – Kingdom of Saudi Arabia, award number (14-INF83-04).

### REFERENCES

[1] A. Bartels, J.R. Rymer, J. Staten, K. Kark, J. Clark, and D. Whittaker, "The Public Cloud Market Is Now in Hypergrowth: Sizing The Public Cloud Market, 2014 To 2020", Forrester, available at <https://www.forrester.com/report/The+Public+Cloud+Market+Is+Now+In+Hypergrowth/-/E-RES113365> (accessed 10 January 2017).

[2] F. Etro. "The Economics of Cloud Computing", The IUP Journal of Managerial Economics, Vol. 9, No. 2, 2012, pp. 7-22.

[3] CEBR. Economic impact of cloud computing. Centre for European Business Research, 2011.

[4] A. Azadegan and J. Teich, "Effective benchmarking of innovation adoptions: A theoretical framework for e-procurement technologies", Benchmarking: An International Journal, Vol. 17, No. 4, 2010, pp. 472–490.

[5] I.M. Al-Jabri and M.H. Alabdulhadi, "Factors affecting cloud computing adoption: perspectives of IT professionals", International Journal of Business Information Systems, Vol. 23, No. 4, 2016, pp.389-405.

[6] I.M. Khalil, A. Khreishah, and M. Azeem, "Cloud Computing Security: A Survey", Computers, Vol. 3, No. 1, 2014, pp. 1-35.

[7] K.M. Khan, A. Erradi, S. Alhazbi, and J. Han, "Addressing security compatibility for multi-tenant cloud services" International Journal of Computer Applications in Technology, Vol. 47, No. 4, 2013, pp. 370-378.

[8] T. Radwan, M.A. Azer, and N. Abdelbaki, "Cloud computing security: challenges and future trends", International Journal of Computer Applications in Technology, Vol. 55, No. 2, 2017, pp.158-172.

[9] D. Servos, S. Mohammed, J. Faiidhi, and T. Kim, "Extensions to ciphertext-policy attribute-based encryption to support distributed environments", International Journal of Computer Applications in Technology, Vol. 47, No.2/3, 2013, pp. 215-226.

[10] S. Singha, Y. Jeong, and J.H. Park, "A survey on cloud computing security: Issues, threats, and solutions", Journal of Network and Computer Applications, Vol. 75, 2016, pp. 200-222.

[11] A. Tashkandi and I. Al-Jabri, "Cloud Computing Adoption by Higher Education Institutions in Saudi Arabia: An Exploratory Study", Cluster Computing, Vol. 18, No. 4, 2015, pp. 1527-1537.

[12] M.F. Mushtaq, U. Akram, I. Khan, S.N. Khan, A. Shahzad, and A. Ullah, "Cloud Computing Environment and Security Challenges: A Review", International Journal of Advanced Computer Science and Application, Vol. 8, No. 10, 2017, pp. 183-195.

[13] M. Kazim and S.Y. Zhu, "A survey on top security threats in cloud computing", International Journal of Advanced Computer Science and Application, Vol. 6, No 3, 2015, pp.109-113.

[14] R. El-Gazzar, E. Hustad, and D.H. Olsen, "Understanding cloud computing adoption issues: A Delphi study approach", The Journal of Systems and Software, Vol. 118, 2016, pp. 64–84.

- [15] A. Basahel, M. Yamin, and A. Drijan, "Barriers to Cloud Computing Adoption for SMEs in Saudi Arabia", *BVICAM's International Journal of Information Technology*, Vol. 8, No. 2, 2016, pp. 1044-1048.
- [16] S. Khan, M.S.A. Khan, and C.S. Kumar, "Multi-criteria Decision in the Adoption of Cloud Computing Services for SME's based on BOCR Analysis", *Asian Journal of Management Research*, Vol. 5, No. 4, 2015, pp. 606-619.
- [17] F.M. Aymerich, G. Fenu, and S. Surcis, "An Approach to a Cloud Computing Network", 1st International Conference on the Applications of Digital Information and Web Technologies, Ostrava, Czech Republic, August 4-6, 2008, pp. 120-125.
- [18] R. Buyya, C.S. Yeoa, and S. Venugopala, J. Broberg and J. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility", *Future Generation Computer Systems*, Vol. 25, 2009, No. 6, pp. 599-616.
- [19] Jaeger, P.T., J. Lin, and J.M. Grimes. Cloud computing and information policy: computing in the policy cloud?. *Journal of Information Technology & Politics*, Vol. 6, No. 3, 2008, pp. 269-283.
- [20] M.A. Vouk, "Cloud Computing - Issues, research and implementations", *Journal of computing and information technology*, Vol. 16, No. 4, 2008, pp. 235-246.
- [21] S. Biswas, "Cloud Computing vs Utility Computing vs Grid Computing. Cloud Tweaks", available at <http://cloudtweaks.com/2011/02/cloud-computing-vs-utility-computing-vs-grid-computing-sorting-the-differences/> (accessed 10 January 2017)
- [22] N.G. Carr, "IT Doesn't Matter", *Harvard Business Review*, available at <http://hbr.org/2003/05/it-doesnt-matter> (accessed 22 March 2017)
- [23] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A View of Cloud Computing", *Communications of the ACM*, Vol. 53, No. 4, 2010, pp. 50-58.
- [24] F. Kramer, "Musings on the cloud - A customer oriented concept formation on cloud computing with respect to SME", *European, Mediterranean & Middle Eastern Conference on Information Systems (EMCIS2012)*, Munich, Germany, June 7-8, 2012.
- [25] P. Mell and T. Grance, *The NIST Definition of Cloud Computing*. National Institute of Standards and Technology (NIST). US Department of Commerce. Special Publication 800-145. 2011.
- [26] Computer Security Alliance, "Security Guidance for Critical areas focus in Cloud Computing v3.0", 2011, available at <https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf> (accessed 22 March 2017).
- [27] M.I. Eid, I.M. Al-Jabri, M.S. Sohail, and K.J. Syed "Cloud computing adoption: a mapping of service delivery and deployment models", 15th International Conference on Electronic Business, Hong Kong, December 6-10, 2015, pp. 160-165.
- [28] S.K. Garg, S. Versteeg, and R. Buyya, "A framework for ranking of cloud computing services", *Future Generations Computer Systems Journal*, Vol. 29, No. 4, 2013, pp. 1012-1023.
- [29] P. Saripalli and G. Pingali, "MADMAC: Multiple Attribute Decision Methodology for Adoption of Clouds", 4th International Conference on Cloud Computing, Washington, DC, USA, July 4-9, 2011, pp. 316-323. *IEEE Xplore*, DOI: 10.1109/CLOUD.2011.61
- [30] F. De Borja, "Cloud Computing Skills Required for IT Employees. Cloud Times", available at <http://cloudtimes.org/2013/02/06/cloud-computing-skills-required-for-it-employees/> (accessed 23 February 2017).
- [31] E.O. Yeboah-Boateng and K.A. Essandoh, "Factors Influencing the Adoption of Cloud Computing by Small and Medium Enterprises in Developing Economies", *International Journal of Emerging Science and Engineering*, Vol. 2, No. 4, 2014, pp. 13-20.
- [32] E.O. Güner and E. Sneider, "Cloud Computing Adoption Factors in Turkish Large Scale Enterprises", *PACIS 2014 Proceedings*, Chengdu, China, June 24-28, 2014. <http://aisel.aisnet.org/pacis2014/353> (accessed 17 April 2017).
- [33] L. Sun, H. Dong, F.K. Hussain, O.K. Hussain, and E. Chang, "Cloud service selection: State-of-the-art and future research directions", *Journal of Network and Computer Applications*, Vol. 45, No. 1, 2014, pp. 134-150.
- [34] R.K. Gavade, "Multi-Criteria Decision Making: An overview of different selection problems and methods", *International Journal of Computer Science and Information Technologies*, Vol. 5, No. 4, 2014, pp. 5643-5646.
- [35] G. Conway, and E. Curry, "The IVI Cloud Computing Life Cycle", in: Ivanov, I.I., M. van Sinderen, F. Leymann, T. Shan (Eds) *Cloud Computing and Services Science, Communications in Computer and Information Sciences*, Vol. 367, Springer, Cham, 2013, pp. 183-199. DOI: 10.1007/978-3-319-04519-1\_12
- [36] M. Whaiduzzaman, A. Gani, N.B. Anuar, M. Shiraz, M.N. Haque, and I.T. Haque, "Cloud Service Selection Using Multi-criteria Decision Analysis", *The Scientific World Journal*, Volume 2014, Article ID 459375, pp 1-10.
- [37] Z. Rehman, O.K. Hussain, and F.K. Hussain, "IaaS Cloud Selection using MCDM Methods", 9th International Conference on e-Business Engineering, Hangzhou, China, September 9-11, 2012, pp. 246-251. *IEEE Xplore*, DOI: 10.1109/ICEBE.2012.47
- [38] S. Lee and K. Seo, "A Hybrid Multi-Criteria Decision-Making Model for a Cloud Service Selection Problem Using BSC, Fuzzy Delphi Method and Fuzzy AHP", *Wireless Personal Communications*, Vol. 86, No. 1, 2016, pp. 57-75.
- [39] S. Liu, F.S. Chan, and W. Ran, "Decision making for the selection of cloud vendor: An improved approach under group decision-making with integrated weights and objective/subjective attributes", *Expert Systems with Applications*, Vol. 55, No. 1, 2016, pp 37-47.
- [40] N. Ramachandran, P. Sivaprakasam, G. Thangamani, and G. Anand. "Selecting a suitable Cloud Computing technology deployment model for an academic institute: A case study", *Campus-Wide Information Systems*, Vol. 31, No. 5, 2014, pp. 319-345.
- [41] C.P. Muir, "Decision Making Model for the Adoption of Cloud Computing in Jamaican Organizations", 19th Americas Conference on Information Systems, August 15-17, 2013, Chicago, Illinois.
- [42] C. Yiming and Z. Yiwei. "SaaS vendor selection basing on Analytical Hierarchy Process", 14th International Joint Conference on Computational Sciences and Optimization, Yunnan, China, April 15-19, 2011 pp. 511-515, *IEEE Xplore*, DOI: 10.1109/CSO.2011.232.
- [43] M. Menzel, M. Schonherr, J. Nimis, and S. Tai, "(MC2): A Generic Decision-Making Framework and its Application to Cloud Computing", available online: <https://arxiv.org/abs/1112.1851> (accessed on 2 December 2017).

# Prediction of Stroke using Data Mining Classification Techniques

Ohoud Almadani, Master of Health Informatics  
(MHI), and Registered Pharmacist (R.Ph)  
Pharmaceutical care department at King Abdulaziz  
Medical City  
Riyadh, KSA

Riyad Alshammari  
King Saud bin Abdulaziz University for Health Sciences  
King Abdullah International Medical Research Center  
(KAIMRC)  
Ministry of National Guard Health Affairs  
Riyadh, KSA

**Abstract**—Stroke is a neurological disease that occurs when a brain cells die as a result of oxygen and nutrient deficiency. Stroke detection within the first few hours improves the chances to prevent complications and improve health care and management of patients. In addition, significant effect of medications that were used as treatment for stroke would appear only if they were given within the first three hours since the beginning of stroke. A framework has been designed based on data mining techniques on Stroke data set that is obtained from Ministry of National Guards Health Affairs hospitals, Kingdom of Saudi Arabia. A data mining model was built with 95% accuracy. Furthermore, this study showed that patient with the following medical conditions, such as heart diseases (hypertension mainly), immunity diseases, diabetes militias, kidney diseases, hyperlipidemia, epilepsy, or blood (platelets) disorders has a higher probability to develop stroke.

**Keywords**—Stroke; data mining; classification

## I. INTRODUCTION

Knowledge Discovery from Data (KDD) is a growing field of computer science that deals with information gain and decision support through large data analysis and automated extraction of patterns.

Information gain from health data may lead to innovative solution or better treatment plan for patients. In order to gain knowledge intelligently from stroke data, a data mining technique is utilized to semi-automatically process data and generate data mining model that can be used by health care professionals [1].

A stroke is a neurological disease that occurs when a brain cells die because of oxygen and nutrient deficiency. Occlusion of brain blood vessel by a clot or blood vessel rupturing are the major causes of oxygen and nutrient supply deficiency [2]. Cerebro-Vascular Accident (CVA) is the previous name of stroke, which divided nowadays into three types known as Hemorrhagic stroke, Acute Ischemic stroke, or Transient Ischemic Attack [2], [3].

Stroke detection within the first few hours improves the chances to prevent complications and improve health care and management of patients [4]. In addition, significant effect of medications that used as treatment for stroke will appear only

if they were given within the first three hours since the beginning of stroke [4].

According to heart disease and stroke statistics update of 2015 [5], 11.13% of deaths globally were accounting for stroke. With 33 million affected persons, stroke is the second leading cause of death worldwide. Furthermore, it is the first leading cause of adult disability with 16.9 million affected persons [5], [6].

According to the Global Burden of Diseases (GBD), Disability-Adjusted Life Years (DALYs) measure showed that the rate of DALYs in Saudi Arabia increased by 50% from 1990 to 2010 because of stroke [6], [7]. In addition, the number of Years of Life Lost (YLLs) statistics of 2010 showed that stroke was the fourth reason of death in Saudi Arabia with an increased rate of 52% since 1990 [8]. Therefore, early prediction of stroke will facilitate effective therapy administration within an appropriate period [9].

The main objectives of this research are twofold: i) Use data mining techniques to predict patient at risk of developing stroke; and ii) Find the patient with who has higher chances to develop stroke. Therefore, three classification algorithms, namely C4.5, Jrip, and multi layers perceptron (MLP), are used on stroke patient data set collected from National Guard hospitals in three different cities in Kingdom of Saudi Arabia. The three classifiers are compared with each to find the best the performance with the goal of finding the best predication model. Hence, framework has been developed to identify stroke patients using proper decision support tool that would help in achieving the following goals: i) Decrease the impact of stroke on patient life; ii) Improve country's population life expectancy and health; and iii) Reduce health care budget.

The remaining segments of this research article are arranged as following: Section 2 presents the literature review on using data mining to predict stroke. Section 3 explains the methodology while Section 4 presents the results and discussion. Finally, Section 5 includes the conclusion and future work.

## II. LITERATURE REVIEW

Stroke has a high impact on public health and countries' economies that lead to build several stroke associations with the aim of improving lives quality by providing public health

education, lifestyle modification, evidence-based treatment guidelines, and CardioPulmonary resuscitation (CPR) training [5]-[9]. In addition, it leads to conduct multiple researchers, which focus on finding preventive and educational materials for stroke [5]-[9]. Defining risk factors were the main goal for several researches about stroke [10]-[21]. A research was conducted in Taiwan, that showed age had a significant risk factor for stroke with patients older than 65-year-old with hypertension, and diabetes mellitus (DM), while, gender and cerebral ischemic events were non-significant factors [10]. Another study took its place in United States revealed that the main risk factors were hypertension, DM, hyperlipidemia, smoking, obesity, and congestive heart failure [11].

With the advance development of technology and the high performance of data analysis tools, health care researchers seek a suitable tool to prevent or detect acute stroke in its early stages. Data mining technique provides researchers with a helpful tool to analyze a large amount of data, such as in the case of health care organization, and facilitate the detection of common patterns for such conditions. Therefore, it could provide a prediction model to identify possible individuals to develop such disease [12]-[18].

Decision support tools were the main outcome for many health-related data mining articles. Sheng-Feng Sunga, et al. [19] analyzed data of acute ischemic stroke patients to develop a prediction model for the severity of the disease. In their study, they used K-nearest neighbor model, multiple linear regression, and regression tree model, that resulted an accuracy of 0.743, 0.742, and 0.737, with 95% confidential interval [19].

Ahmet K. Arslan et al. [20] used three data mining algorithms, namely: Support Vector Machine (SVM), Stochastic Gradient Boosting (SGB) and penalized logistic regression (PLR) to predict stroke. SVM achieved an accuracy of 98% [20]. In addition, by using K-nearest neighbor and C4.5 decision tree, Leila Amini et al. [18] achieved an accuracy of stroke prediction equal to 94.2% and 95.4% respectively. Artificial Neural Network (ANN) prediction model achieved a predictive accuracy of thrombotic stroke equal to 89% as shown in Shanthi et al. study [21]. Stroke is being observed as a rapidly growing health issue in Saudi Arabia. It is the second cause of death by killing 14.4 thousand people in 2012. Therefore, it becomes one of the health care issues in Saudi Arabia. The lack of researches that focus on the role of technology, mainly KDD, in predicting of stroke in the Saudi Arabia, leads to this research.

### III. METHODOLOGY

In this section, the methodology is explained including on how the data sets are obtained, attributes, the data mining algorithms and the evaluation criteria.

#### A. Data Collection

Data received from the data governance department at King Abdulaziz Medical City (KAMC). KAMC opens on 2001 at Riyadh city. KAMC grows up to become one of the top hospitals in the Middle East with a bed capacity of more than 1500 by 2016. KAMC is serving 2.5 million outpatients and around 60,000 in-patients annually. The data set was

extracted from KAMC contained all patients who were diagnosed as stroke case or stroke mimic case on 2016 from the 2<sup>nd</sup> of January to 31<sup>st</sup> of September. The reason behind using this time frame is due to the installation of new Health Information System at KAMC. There are two classes in the data set. The first class includes the medical records for patient's known to have stroke while the second class includes records of stroke mimic patients, who usually misdiagnosed as stroke patient due to the similarity of the symptoms. This data set consists of 969 instances, 69 of them classified as stroke mimics while 899 classified as stroke patients. The data set contains 360 females (37.15 %) 33 of them diagnosed as stroke mimic and 327 as stroke cases. As well, 607 males (62.6%) 36 of them are stroke mimic while 571 are stroke cases (Fig. 1).

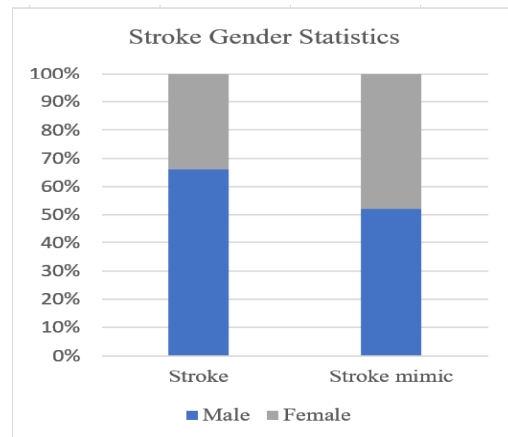


Fig. 1. Stroke gender statistics.

#### B. The Attributes

The obtained data set contained 1004 attributes. Its main attributes are the class (stroke, stroke-mimic), age (ordinal), gender (female, male), number of medication (numeric), medication name (taken, not-taken), and lab test name (normal, abnormal).

Attribute selection was applied then to reduce data dimensionality. Attribute selection is a data mining technique that used to select the most relevant attributes [1], [22]. Principle Component Analysis (PCA) is utilized [22]. It is a data mining technique that works on dimensionality reduction. The idea of this technique is to create a new alternative attribute that combines several previous attributes essence. This algorithm has the ability of reducing attribute noise by transforming it to the PCA space, eliminating the worst eigenvector then transform it back to the original space [1], [22].

The final attribute sets are related to patient age, gender, lipid disorder, lab test abnormalities, hypertension medications, diabetes medications, and other medications are included, which resulted of data set that contains 147 attributes. Moreover, data are divided to two separate data sets: training data set to build the model, and test data set to evaluate the model.

### C. Data Mining Algorithms

For their known high accuracy rate, J48 (C4.5), JRip, and Neural Network (multilayer perceptron [MLP]) algorithms were applied on the stroke training data set to build a model. All Data mining algorithms have been applied using Weka Software (Version 3.8, Machine Learning Group, University of Waikato, Hamilton, New Zealand). C4.5- J48 in WEKA, is an algorithm that works first by choosing the root attribute through attribute selection (gain ratio) [22], [24], [25]. It then works to build decision tree branches from that attribute values and distribute instances into its corresponding branch [22], [24], [25]. This process will be repeated until all instances are assigned to their correct class [22], [24], [25]. On the other hand, RIPPER algorithm, Jrip in WEKA, is a rule based algorithm [22], [26]. The algorithm takes certain steps to build its model. First, a set of rules will be constructed using incremental reduced error [22], [26]. Then each class will be examined against those rules repeatedly until all instances of each class are covered [22], [26]. At the end, rules that cover all classes will be used to build the model [26].

The third algorithm is a Neural Network called Multilayer perceptron (MLP). MLP is a forward feed neural network [22], [27], [28]. It uses one direction feed of input through one or more layers to produce output layer [22], [27], [28]. To train this algorithm a back-propagation learning algorithm usually used, and it helps to solve non-linearity problem [22], [27], [28].

### IV. RESULTS

Results are shown for data with all attributes (row data) and data after attributes section (data after using PCA).

#### A. Results of Prediction Model with All Attributes (Raw Data)

The comparison of the data mining algorithms used with 10-fold cross validation method, were data set first performed on training data set before any attribute reduction methods. It is shown that Jrip has the highest accuracy rate with 86.96% followed by MLP with 85.7% and C4.5 with 84.67%. As well, when test data set supplied to the model a better accuracy is achieved as the following: Jrip has the highest accuracy rate with 92.6%, followed by 89.4% for both MLP and 85.53% for C4.5.

The comparison of the data mining algorithms performed after applying PCA on Stroke data, showed that C4.5 has the highest accuracy on the test data set (95.25%).

TABLE I. CLASSIFIERS PERFORMANCE USING ACCURACY

Algorithms	Training set (10 fold-cross validation)	Test set
Performance on raw data		
MLP	85.70%	89.40%
Jrip	86.96%	92.60%
C4.5	84.77%	85.50%
Performance on after principle component analysis		
MLP	89.85%	94.42%
Jrip	88.81%	93.18%
C4.5	88.81%	95.25%

Generally, it can be seen that C4.5 and Jrip are the highest classifiers in name of accuracy after PCA on the unseen data set (test data set), Table I.

### V. DISCUSSION

In an attempt to use stroke data for the predication of stroke patients, the difference between the process on raw data, and after principle component analysis were examined. The result obtained in this research confirmed the benefit of PCA.

The technique can be used in collaboration with C4.5, Jrip, and MLP as a new framework in identifying new stroke patient. This framework works by reducing number of attributes (variables) to the optimal number using Csf subset evaluation, followed by PCA and then supplies the new data set to the three chosen algorithms. The research found that the accuracy of this approach is approximately 95% (for C4.5 algorithm), compared to un-processed data. The proposed approach showed an improvement of classification accuracy on the test data by 9.72%, 0.58%, 5.02% for J48, Jrip, and MLP respectively. Finally, based on this experiment the highest achieved accuracy was for C4.5 by 95.25%.

The results of this study showed that among the important lab test abnormalities to diagnose stroke, creatine kinase-MB (CKMB) came first, followed by lymph auto, eGFR, and HbA1C. CKMB is a test used to determine if the elevation of creatine kinase is due to heart muscle damage or skeletal muscle damage [23]. Lymph auto, is a lab test that measure white blood cells to exclude any immune system diseases [23]. The estimated glomerular filtration rate (eGFR), is a lab test used to screen renal function and evaluate it [23]. Finally, Hemoglobin A1c (HbA1C) used mainly to diagnose diabetes militias by measuring the average blood glucose level through three months periods [23]. In addition, the result of attribute ranking using Information gain attribute evolution algorithm showed that patient receiving following medication has high risk to develop stroke: Atorvastatin (lipid-lowering medication) [3], Amlodipine (hypertension medication) [3], Levetiracetam (anticonvulsant medication) [3], Metformin (diabetes militias medication) [3], Aspirin (antiplatelet medication) [3], Clopidogrel (antiplatelet medication) [3].

This means that patient who develop heart diseases (hypertension mainly), immunity diseases, diabetes militias, kidney diseases, hyperlipidemia, epilepsy, or blood (platelets) disorders, has a higher probability to develop stroke.

### VI. CONCLUSION

Health care organizations have gained big benefit from data mining in name of big data analysis and decision support system. In this research, stroke patient data has been collected from kanc-ngha that ended up with 17 attributes rather than class attribute.

### ACKNOWLEDGMENT

This work was in part supported by MITACS and, NSERC granting agencies as well as the cfi new opportunities program. The King Abdulaziz City for Science and Technology (KACST) grant supported this research paper

(grant# KACST 37-1895 ط ١). In addition, we would like to thank King Abdulaziz medical city and King Abdullah International Medical Research Center for their support and for providing us with the required data.

#### DISCLOSURES

None of the authors have any competing interests.

#### REFERENCES

- [1] Jiawei Han, et al. Data Mining Concepts and Techniques (2011). Third edition; P84-88.
- [2] The American Heart and Stroke Association.
- [3] Edward C Jauch, et al. Ischemic Stroke. Medscape.
- [4] Guidelines for the early management of patients with acute ischemic stroke. A guideline for healthcare professionals from the American Heart Association/American Stroke Association. March 2013.
- [5] Mozaffarian D, et al. on behalf of the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics (2015 update) a report from the American Heart Association [published online ahead of print December 17, 2014]. Circulation. doi: 10.1161/CIR.000000000000152
- [6] World health organization. The top 10 causes of death. <http://www.who.int/mediacentre/factsheets/fs310/en/>. Mayo 2014.
- [7] Ziad A. Memish, et al. Burden of Disease, Injuries, and Risk Factors in the Kingdom of Saudi Arabia, 1990–2010. Preventing Chronic Disease public health research, practice, and policy volume 11, e169 October 2014.
- [8] Institute for Health Metrics and Evaluation. GBD Profile: Saudi Arabia. [www.healthmetricsandevaluation.org](http://www.healthmetricsandevaluation.org). 2010.
- [9] UPMC Presbyterian. A Designated Comprehensive Stroke Center what to expect: recovering from stroke. [www.UPMC.com/services/stroke-institute](http://www.UPMC.com/services/stroke-institute). 2015.
- [10] Lian-Yu Lin, et al. Risk factors and incidence of ischemic stroke in Taiwanese with nonvalvular atrial fibrillation—A nationwide database analysis. *Atherosclerosis* 217 (2011) 292–295. doi: 10.1016/j.atherosclerosis.2011.03.033. Epub 2011 Apr 5.
- [11] José Rafael Romero, et al. Stroke prevention: modifying risk factors. *Therapeutic Advances in Cardiovascular Disease*. (2008) August; 2(4): 287–303. doi:10.1177/1753944708093847.
- [12] Antonio Coca, et al. Predicting Stroke Risk in Hypertensive Patients With Coronary Artery Disease. *Stroke*. AHA Journals Home (2008) Feb;39(2):343-8.
- [13] Tanika N. Kelly, et al. Cigarette Smoking and Risk of Stroke in the Chinese Adult Population. *AHA journals*. June 2008. DOI: 10.1161/STROKEAHA.107.50530.
- [14] Wenbin Liang, et al. Tea Consumption and Ischemic Stroke Risk Case–Control Study in Southern China. *AHA journals*. July 2009. DOI: 10.1161/STROKEAHA.109.548586
- [15] Fuk-hay Tang, et al. An image feature approach for computer-aided detection of ischemic stroke. *Computers in Biology and Medicine* 41 (2011) 529–536. doi: 10.1016/j.combiomed.2011.05.001.
- [16] A. Przelaskowskia, et al. Improved early stroke detection: Wavelet-based perception enhancement of computerized tomography exams. *Computers in Biology and Medicine* 37 (2007) 524 – 533. DOI: <http://dx.doi.org/10.1016/j.combiomed.2006.08.004>
- [17] Kartheeban Nagenthiraja, et al. Automated decision-support system for prediction of treatment responders in acute ischemic stroke. *Frontiers in Neurology* (2013) Volume4:Article140. doi:10.3389/fneur.2013.00140
- [18] Leila Amini, et al. Prediction and Control of Stroke by Data Mining. *International journal of preventive Medicine*. 2013 May; 4(Suppl 2): S245–S249.
- [19] Sheng-Feng Sunga, et al. Developing a stroke severity index based on administrative data was feasible using data mining techniques. *Journal of Clinical Epidemiology*. Volume 68, Issue 11, November 2015, Pages 1292–1300
- [20] Ahmet K. Arslana, Cemil Colaka, and Ediz Sarihanb. Different Medical Data Mining Approaches Based Prediction of Ischemic Stroke. *Computer Methods and Programs in Biomedicine*. March 2016.
- [21] D.Shanthi.,et al. Designing an Artificial Neural Network Model for the Prediction of Thromboembolic Stroke (IJB), 2008, Volume 3. pp.10-18.
- [22] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). *The WEKA Workbench*. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
- [23] Lab test online. ©2001 - 2017 by American Association for Clinical Chemistry. <https://labtestsonline.org/>.
- [24] Tjortjis C, et al. Using T3, an improved decision tree classifier, for mining stroke-related medical data. *Methods of Information in Medicine* (2007) ;46(5):523-9. Doi: <http://dx.doi.org/10.1160/ME0317>.
- [25] A. Sudha. et al. Effective analysis and predictive model of stroke disease using classification methods. *international journal for computer application* (0975-8887). April 2012. Voulume 43-No.14.
- [26] Poonam Gupta, Rohit Miri, S.R.Tandan, Decision Tree Applied For Detecting Intrusion, *International Journal of Engineering Research & Technology* (IJERT) Vol. 2 Issue 5, May – 2013 ISSN: 2278-0181.
- [27] Anil Rajput , Ramesh Prasad Aharwal, et al. J48 and JRIP Rules for E-Governance Data. *International Journal of Computer Science and Security* (IJCSS), Volume (5) : Issue (2) : 2011
- [28] Multi layer Perceptron. <http://neuroph.sourceforge.net/tutorials/MultiLayerPerceptron.html>

# Hardware Implementation for the Echo Canceller System based Subband Technique using TMS320C6713 DSP Kit

Mahmod. A. Al Zubaidy  
Ninevah University  
Mosul, Iraq

Sura Z. Thanoon  
(MSE student) School of Electronics Engineering  
Mosul University  
Mosul, Iraq

**Abstract**—The acoustic echo cancellation system is very important in the communication applications that are used these days; in view of this importance we have implemented this system practically by using DSP TMS320C6713 Starter Kit (DSK). The acoustic echo cancellation system was implemented based on 8 subbands techniques using Least Mean Square (LMS) algorithm and Normalized Least Mean Square (NLMS) algorithm. The system was evaluated by measuring the performance according to Echo Return Loss Enhancement (ERLE) factor and Mean Square Error (MSE) factor.

**Keywords**—Acoustic echo canceller; Least Mean Square (LMS); Normalized Least Mean Square (NLMS); TMS320C6713; 8 subbands adaptive filter

## I. INTRODUCTION

The acoustic echo problem that appears in the communications systems was impeding the performance of systems efficiently and to solve this problem, it must define the echo and knowledge of its characteristics. Echo signal is defined as the delayed and attenuated version of the original signal produced by some device, such as a loud speaker. If we consider the transmitted signal  $x(t)$ , then the attenuated and delayed version of it is the echo signal and it is given by the following (1) [1]:

$$x_d(t) = \alpha x(t - t_d) \quad (1)$$

Where  $\alpha$  is the attenuation factor and  $t_d$  is time delay of the echo replica.

The echo signal starts to be sensible about the listener after 35 msec and the amount of its delay and attenuation depend on the surface that the signal reflected from such as walls, floors and furniture, and the path that will travel through [2].

To solve this problem the adaptive filter was used to design acoustic echo cancellation system (AEC), where we will review the techniques used in the design of the adaptive filter in Section II, as well as the algorithms used in each technique in Sections III and IV. Then we present the circuits designed to delete the acoustic echo in Section V, performance evaluation in Section VI and the results obtained from each technique and algorithm used and the comparisons between the systems in Section VII followed with conclusions and future works in Sections VIII and IX, respectively.

## II. ADAPTIVE FILTER TECHNIQUES

### A. Fullband Technique

Fullband adaptive filter is the first solution to acoustic echo where in this way the whole sound would process by using one adaptive filter. Here the adaptive filter deals with the input that contains wide spectral dynamic range as one band so the speed of convergence would be slow. Fig. 1 shows Echo Canceller System using fullband adaptive filter method.

Fullband adaptive filter method suffers from two main problems which are [3]:

- First, when the input signal has wide spectral dynamic range such as that found in speech then the input correlation matrix is ill-conditioned or equivalently so that lead to be the convergence and tracking of a gradient-based adaptive filter can be very slow.
- Second, this method need very high-order adaptive filters and so it is computationally expensive. Subband adaptive filter technique used to overcome these problems, by decomposing the signal into subbands and each subband signal used separated adaptive filter.

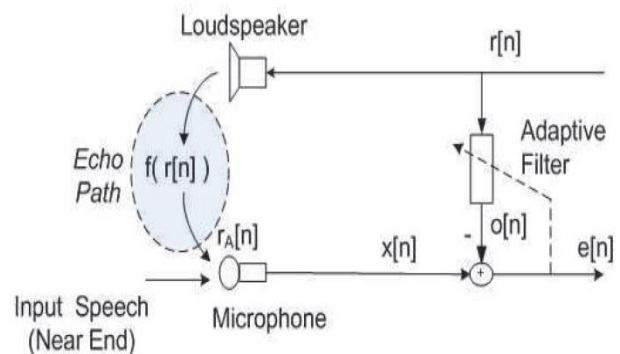


Fig. 1. Acoustic echo cancellation system.

### B. Subband Technique

The subbands echo canceller technique consists of the subband analyzer which splits the input signal into N sub-



bands, and N subband adaptive filter, each subband adaptive filter deals with one separated frequency band [4]. The main advantages of a subbands echo canceller are reduction in filter length, reduction in computational complexity and increasing the speed of convergence [2]. Fig. 2 shows the subband adaptive filter.

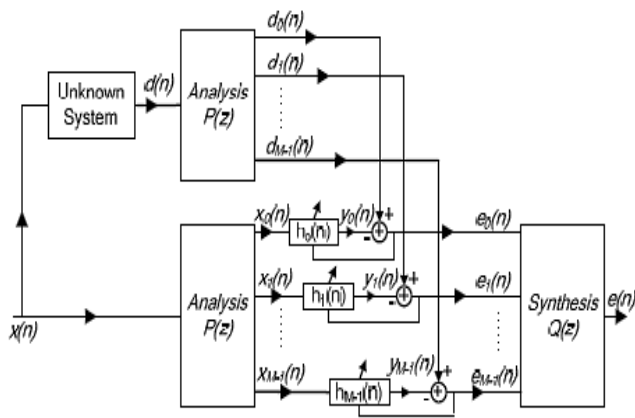


Fig. 2. Subband system identification.

However the adaptive filter can deal with any changing in the environments because it self-adjusting his weights continuously to get the highest performance for the system, the adaptive filter adjusts his weights according to algorithms such as LMS, NLMS, VLMS and RLMS. By choosing appropriate algorithm for the application, suitable length and step size for the filter, the performance of the designed system it becomes better.

### III. LEAST MEAN SQUARE ALGORITHM (LMS)

The LMS adaptive filter is the first, most popular and widely used in adaptive system, appearing in numerous commercial and scientific applications the reasons behind this fame to it's easily in the implementation and there is no complexity in the calculation. The following equations described the LMS adaptive filter operation [5]:

$$W(n+1) = W(n) + \mu(n) e(n) X(n) \quad (2)$$

$$e(n) = d(n) - W^T(n)X(n) \quad (3)$$

Where

$W(n) = [w_0(n) \ w_1(n) \ \dots \ w_{L-1}(n)]^T$  is the coefficient vector,  $X(n) = [x(n) \ x(n-1) \ \dots \ x(n-L+1)]^T$  is the input signal vector,  $d(n)$  is the desired signal,  $e(n)$  is the error signal, and  $\mu(n)$  is the step size.

### IV. NORMALIZED LEAST MEAN SQUARE ALGORITHM (NLMS)

The NLMS adaptive filter is derivative form LMS adaptive filter where it is also easy to implement but with a change in the calculation of the step size where the step size is varying continuously to get a better performance where the convergence speed is increased and this property has been

used in the design of acoustic echo cancellation system. The equation for calculating the step size is [6]:

$$\mu = \frac{\beta}{c + \|x(n)\|^2} \quad (4)$$

Where

$\mu(n)$  = step size,  $\beta$  = Normalized step size ( $0 < \beta < 2$ ),  $c$  = small positive constant

The following equation to clarify how the filter calculates his weights [7]:

$$w(n+1) = w(n) + \mu(n)e(n)x(n) \quad (5)$$

### V. HARDWEAR IMPLEMENTATION FOR THE AEC SYSEM USING TMS320C6713 DSK

To implement AEC system in real-time required: one DSP TMS320C6713 Starter Kit (DSK), Two Personal Computers (PC) and Display Unit (Oscilloscope), first the system must design using MATLAB beside the Embedded Target for C6000 DSP, the procedure of the implementation in general was shown in Fig. 3.

The TMS320C6713 DSK starter kit is a low cost stand-alone DSP development platform that can be used to develop applications for the TMS320C67xx DSP family. It includes the C6713 floating-point digital signal processor (DSP) and a 32 bit stereo codec (AIC23) for input and output. The AIC23 codec uses a sigma delta technology that provides Analogue to Digital conversion (ADC) and Digital to Analogue conversion (DAC) and has got variable sampling rates from 8 kHz to 96 kHz. It includes 16 MB synchronous dynamic random access memory (SDRAM) and 256 KB of flash memory. Furthermore it includes two inputs (LINE IN, MIC IN) and two output ports (LINE OUT, HEADPHONE). The DSK operates at a frequency of 225 MHz and has got a single power supply of 5 V [8]. Fig. 4 and 5 show the photograph and the block diagram of the TMS320C6713 DSK KIT.

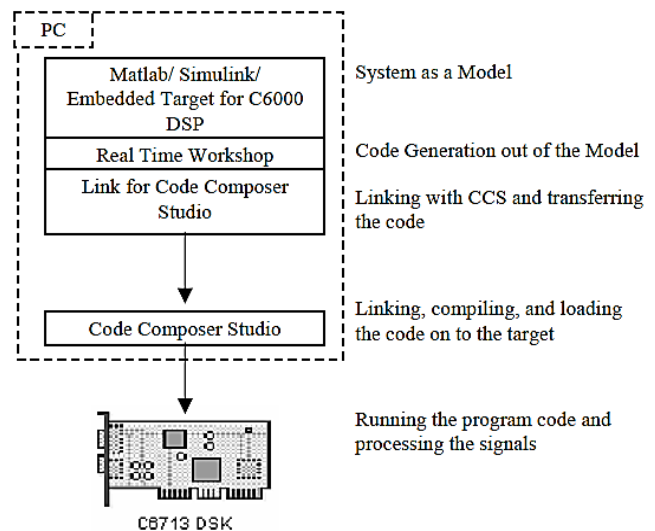


Fig. 3. Flow diagram of the procedures of implementation the model on C6713 DSK.



Fig. 4. The photograph of the C6713 DSK.

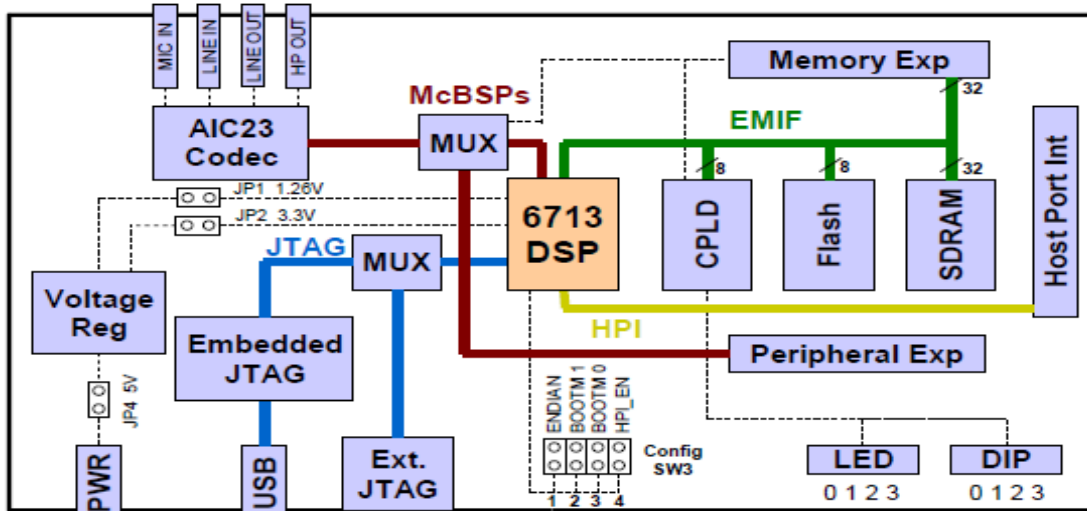


Fig. 5. C6713 DSK block diagram.

The real time hardware implementation of the AEC system needs two PCs, the first PC was employed to run the MATLAB model on DSP kit to install the definitions of the DSP kit. The second PC was used for running the models of sending the recorded speech to the DSP kit and saving the results, the two PCs must contain the MATLAB program.

By using MATLAB\_ SIMULINK program the fullband and 8 subband AEC system based LMS and NLMS algorithm for adaptive filter was designed and implemented on C6713 DSK, Fig. 6 shows the fullband AEC system based LMS and NLMS algorithm for the adaptive filter.

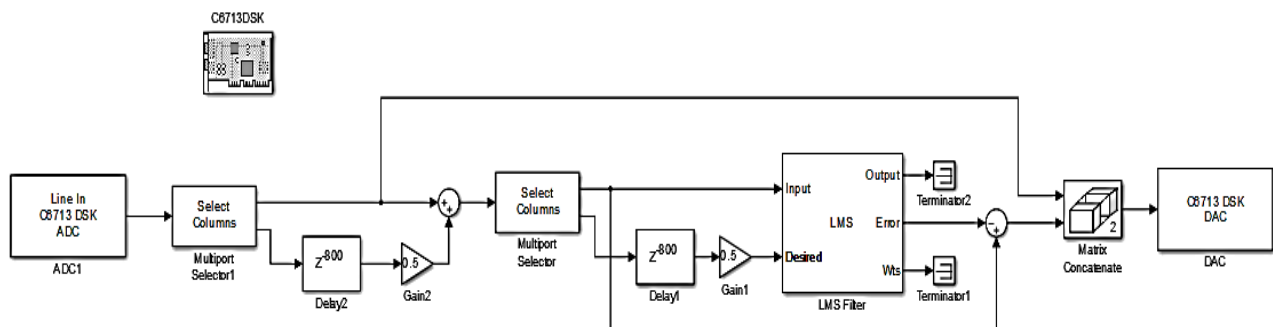


Fig. 6. Fullband AEC system using LMS algorithm and recorded speech signal as input on C6713 DSK.

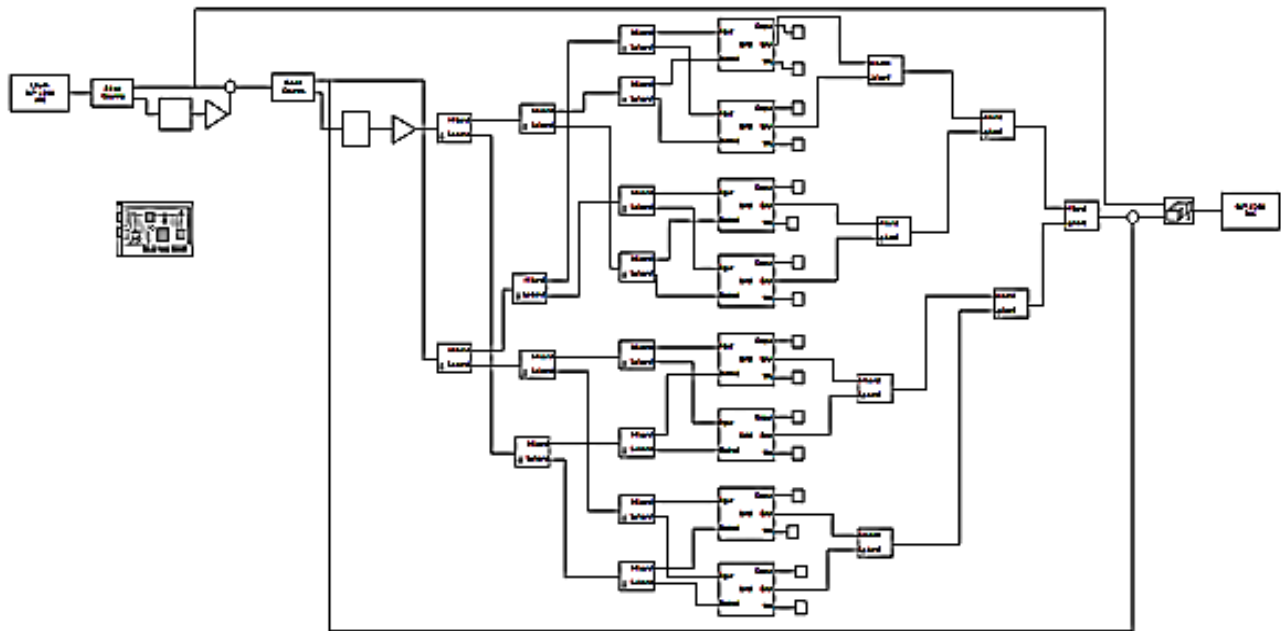


Fig. 7. 8 Subbands echo canceller system using LMS algorithm and recorded speech signal as input on C6713 DSK.



Fig. 8. The photograph of the hardware system implemented.

To display the results of the DSP kit a digital storage (500Msample/sec) oscilloscope was used to achieve this purpose. The complete hardware implemented system is shown in Fig. 8.

#### VI. PERFORMANCE EVALUATION

The AEC system can be evaluated according to two factors, the echo return loss enhancement (ERLE) and mean square error (MSE). ERLE defined as the ratio of the is defined as the ratio of the microphone power signal,  $d(n)$ , to

the power of the residual error signal after cancellation,  $e(n)$ . ERLE measured in dB and it can be expressed as [9]:

$$ERLE = 10 \log \frac{p_d(n)}{p_e(n)} = 10 \log \frac{E[d^2(n)]}{E[e^2(n)]} \quad (6)$$

The mean square error is measure give us the quantity of error level between two signals, or the similarity between them as shown in the following equation [10].

$$MSE = \frac{\sum e^2}{n} \quad (7)$$

### VII. HARDWEAR IMPLEMENTATION RESULTS

The AEC system was implemented based TMS320C6713 DSK by using LMS and NLMS algorithm for the adaptive filter and choosing the parameters of the filter where the length of the filter was 1024 and the step size was 0.002 for both algorithms. The speech used to test the implemented system was recorded with sampling rate 8000 per second then the echo generated from this recorded speech with attenuation 0.5 and delay 100 msec and summed with original speech to get the input signal to the system as shown in Fig. 7.

The output signals of the 8 subbands AEC system using LMS algorithm for the adaptive filter shown in Fig. 9.

The output signals of the 8 subbands AEC system using NLMS algorithm for the adaptive filter shown in Fig. 10.

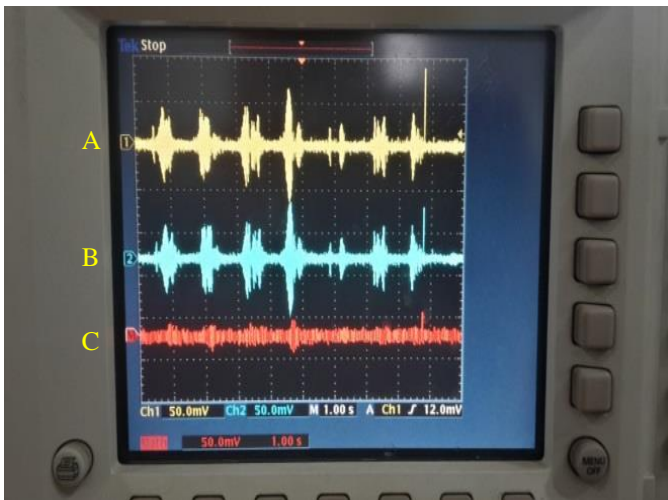


Fig. 9. 8 subbandsAEC system results using LMS algorithm and speech signal as input A- The original signal, B- The output signal, C- The error signal.

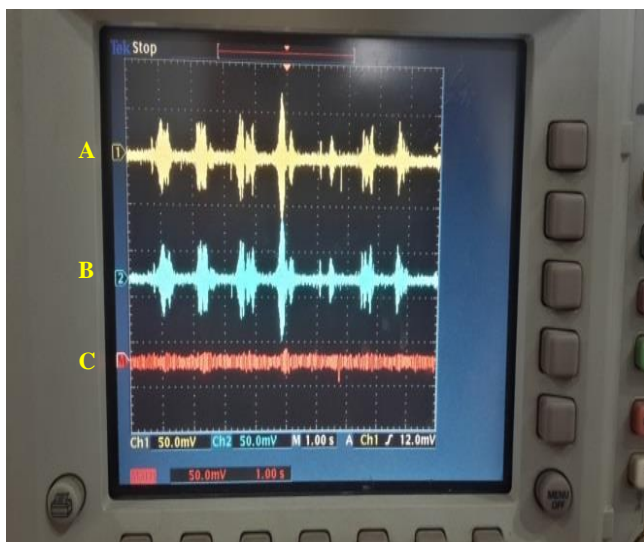


Fig. 10. 8 subbands AEC system results using NLMS algorithm and speech signal as input A- The original signal, B- the output signal, C- the error signal.

By comparing the output signals of the both systems it can be notice that the error signal in Fig. 10 is less than the error signal in Fig. 9, and by calculating ERLE and MSE factors for these two systems, For LMS algorithm ERLE was 11.23 dB and MSE was 0.0000408189 and for NLMS algorithm ERLE was 13.56 dB and MSE was 0.0000279061.

By comparing between the above results for LMS algorithm and NLMS algorithm the different in the performance was clearly noticed and proof that NLMS algorithm is better than LMS algorithm as shown in Fig. 11 and 12.

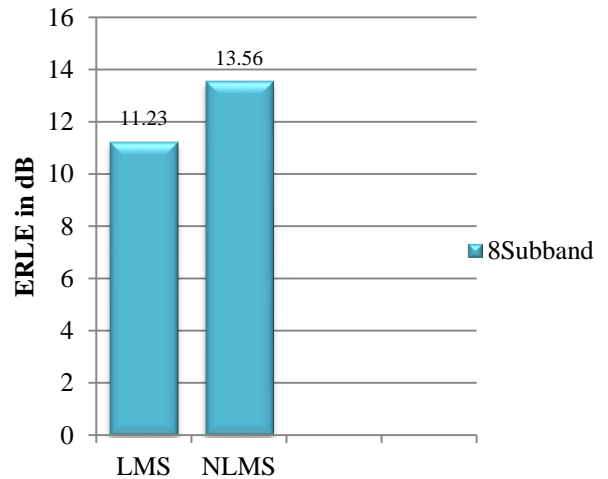


Fig. 11. ERLE values for the 8subbandsAEC system on C6713 DSK based on LMS and NLMS tested by speech input signal.

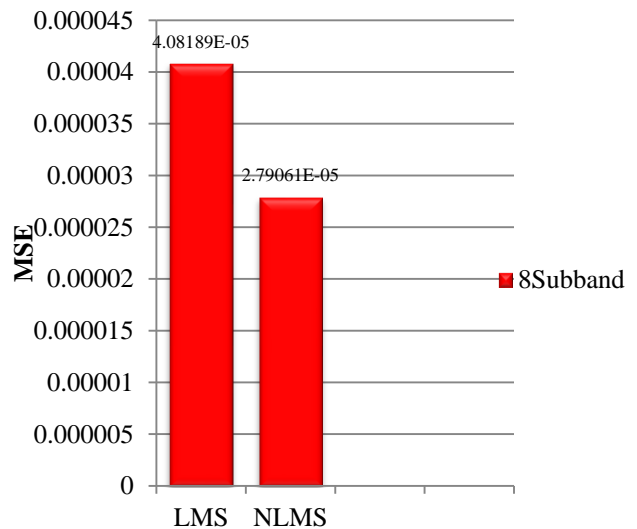


Fig. 12. MSE values for the 8subbandsAEC system on C6713 DSK based on LMS and NLMS tested by speech input signal.

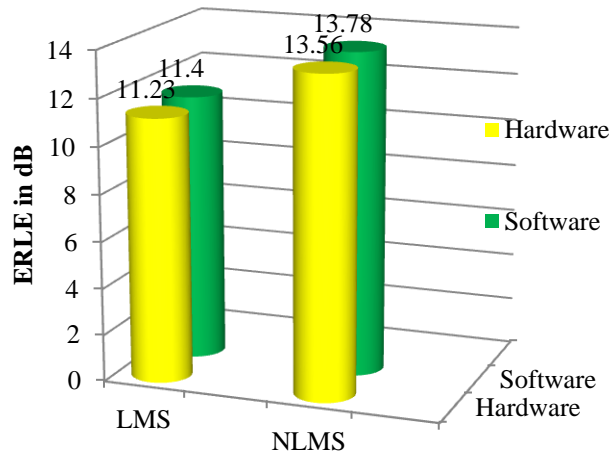


Fig. 13. Comparison between software and hardware results in ERLE values of the AEC system using LMS and NLMS algorithm and recorded speech as input signal.

The AEC system was designed primary by using MATLAB SIMULINK and tested also by using the same recorded speech, by looking at the results of the simulation model and compare it with the hardware results it can be notice that the results using TMS320C6713 DSK converge from the simulation results where the different is very small as shown in Fig. 13 [11].

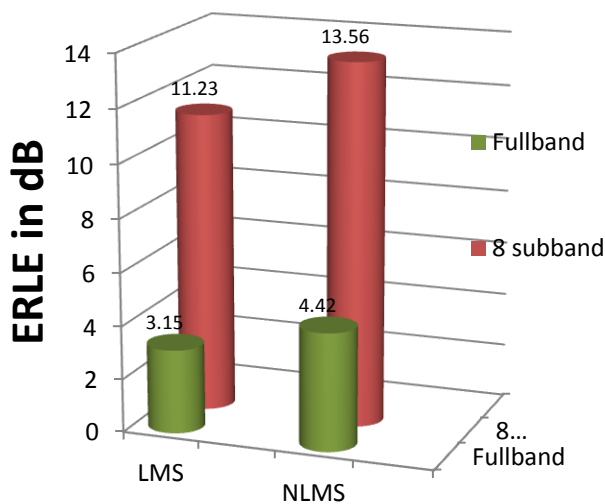


Fig. 14. Comparison between fullband and 8subband techniques in ERLE values of the AEC system using LMS and NLMS algorithm and recorded speech as input signal.

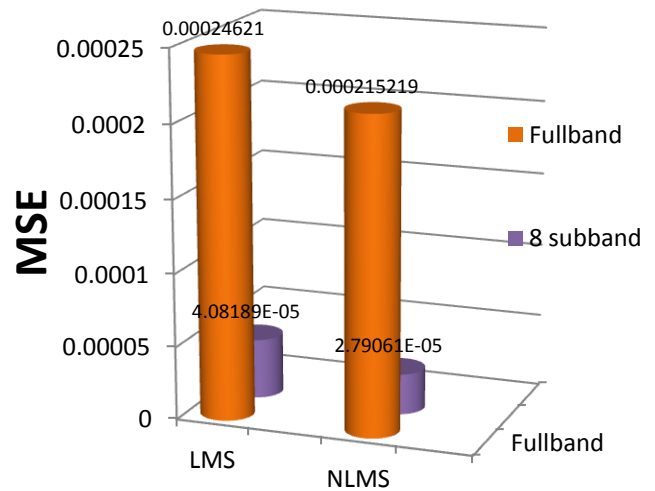


Fig. 15. Comparison between fullband and 8subband techniques in MSE values of the AEC system using LMS and NLMS algorithm and recorded speech as input signal.

The AEC system was implemented using two technique fullband and 8 subband, to determine which system cancel the echo signal more efficiently we compare the results between these two systems according to ERLE and MSE factors. In AEC system using fullband adaptive filter, for LMS algorithm, ERLE was 3.15 dB, and MSE was 0.00024621. For NLMS algorithm, ERLE was 4.42 dB, and MSE was 0.000215219. Comparing the results using fullband technique with the results when using 8 subband technique either using LMS or NLMS algorithm it can be notice that 8 subband technique is better than fullband technique as Fig. 14 and 15 shown.

### VIII. CONCLUSION

- 1) The Echo Cancellation system is an effective system for deleting acoustic echo by generating an echo similar to the echo comes with the original signal using an adaptive filter and then subtracting it from the received signal.
- 2) The appropriate option for implementing the systems in the real-timer is to use TMS320C6713 DSP kit.
- 3) Returning to the results obtained from the hardware and using the factors of evaluation ERLE and MSE we find that the NLMS algorithm that used for the adaptive filter is better than the LMS algorithm.

### IX. FUTURE WORK

- 1) Using 16 subbands technique to implement AEC system.
- 2) Implementation of AEC system in mobile applications.

REFERENCES

- [1] Artur Ferreira and Paulo Marques, (2011), "Echo Cancellation for Hands-Free Systems", Adaptive Filtering, DrLino Garcia (Ed.), ISBN: 978-953-307-158-9, InTech, Available from: <http://www.intechopen.com/books/adaptive-filtering/echo-cancellation-for-hands-free-systems>.
- [2] Saeed V. Vaseghi, "Advanced Digital Signal Processing and Noise Reduction", Second Edition, ISBNs: 0-471-62692-9 (Hardback): 0-470-84162-1 (Electronic), 2000.
- [3] Chaitanya M. Patil and R.A.Deshpande, "Comparative Analysis Based on Statistical Parameters of LMS and NLMS Algorithms on ECG Signal by Using Matlab Simulink", International Journal of Advances in Electronics and Computer Science, Vol. 2, Issue 8, August 2015.
- [4] Phillip L. De Le\_on II and Delores M. Etter, "Acoustic echo cancellation using subband adaptive filtering ", ch 11 wavelet book.pdf
- [5] Widrow, B. and Hoff, M.E., "Adaptive switching circuits", IRE WESCON Conv. Rec.,4, 96–104, Aug. 1960. Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.
- [6] Monsoon H. Hayes, "Statistical Digital Signal Processing and Modeling", John Wiley & Sons, Inc., 1992.
- [7] J. Dhiman, S. Ahmad and K. Gulia, "Comparison between adaptivefilter algorithms (lms, nlms and rls)", International Journal of Science,Engineering and Technology Research, vol. 2, no. 5, pp. pp-1100, 2013.
- [8] E Kaymak , M A Atherton, K R G Rotter and B Millar, " Real-time adaptive filtering of dental drill noise using a digital signal processor", Conference Paper, June, 2006.
- [9] SrinivasaprasathRaghavendran, "Implementation of an acoustic echo canceller using MATLAB", University of South Florida, 2003.
- [10] Zhou Wang and Alan C. Bovik, "Mean Squared Error: Love It or Leave It?",IEEE,pp 98-117,JANUARY 2009.
- [11] Mahmod. A. Al Zubaidy and Sura. Z. Thanoon, "Evaluation and the Performance of LMS and NLMS Algorithm in Acoustic Echo Cancellation using the 8 sub-bands Techniques", International Journal of Engineering and Innovative Technology (IJEIT), Volume 7, Issue 3, PP. 1-3, September 2017.

# SME Cloud Adoption in Botswana: Its Challenges and Successes

Malebogo Khanda  
Lecturer, Faculty of Computing,  
Botho University,  
Gaborone, Botswana

Srinath Doss  
HOD, Faculty of Computing  
Botho University  
Gaborone, Botswana

**Abstract**—The standard office or business in Botswana hosts their resources in-house. This means that a company will have their hardware, software and support staff as part of their daily work operations. Technology has brought a shift to the office environment with Cloud Computing. Botswana has seen the growth of the Cloud Technologies, within its own boundaries where companies have embraced the new technology to mobilize and push their operational agenda with the same tenacity as the rest of the world using the technology. Cloud computing has taken root in Botswana and it shows that a lot of SMEs are using cloud computing, whilst some are non-adopters to the technology. Edgar Tsimane reported the take up on cloud computing in Botswana. Botswana uses the National ICT policy to guide on technological advances and development, i.e the Maitlamo policy. This paper is considering aspects influencing the company's decisions on utilizing the cloud as a service, both opportunistic and challenges. Some of the questions to address for the study are: how effective is cloud computing for businesses in Botswana; what challenges and successes these companies have had, is there any particular framework they had to follow to guide them in adopting the services? Finally, the paper was to take consideration in recommending a framework that can be adopted within the Botswana.

**Keywords**—Cloud computing; SME's; cloud computing services; cloud business processes; cloud computing framework; cloud deployment models; cloud computing services model

## I. INTRODUCTION

Botswana Government and various Information and Communications Technology (ICT-based) companies have investments in improving local Internet connectivity. The Internet acts as the major platform for cloud usage by almost everyone who uses cloud computing, as resources are online. As indicated by Dr Seleka, Botswana is still working on a mandate to utilize the cloud; therefore identifying the challenges and successes of the service, combined with efforts made by the government, will help equip SME's with the necessary tools to guide them to adopt cloud services. The study first explores what the cloud means to the general public; then, it considers the usage of the Cloud by Small to Medium Enterprises (SME's) in their operational activities. The study will also explore the challenges and successes involved in cloud computing contrasting companies that are using the cloud and those that are not using it. This will help build up or identify a framework that can be adopted in Botswana to help SME's in their transition to supporting cloud based services.

Botswana has seen a number of companies adopting Cloud Computing. For example, Botswana Post launched Poso Cloud for its clients in 2013. Some of the companies also using cloud computing are IT-IQ and Dimension Data. Edgar Tsimane reported that BIH "seeks to utilise cloud computing in the near future to provide services to its start-up companies and willing tenants". Dr Geoffrey Seleka, Senior Director BIH, added that "the intention is to see how BIH can also replicate cloud usage in a national setting".

Cloud Computing involves service and product under one umbrella, thus the "Cloud". The usage and implementation of cloud services are also influenced by the legal systems and policies in institutions or guided by international standards from the Institute of Electrical and Electronics Engineers (IEEE). With new technologies, or advancing technologies, concepts are adopted to influence standardization of resources to allow for easy communication and integration. This then denotes, that the Cloud should have some basic concepts that one could analyze before selecting the cloud services.

## II. LITERATURE REVIEW

Different countries in the world have shown that cloud computing adoption is necessary. For examples, countries like Kenya, Nigeria, South Africa, Ghana, India, England and Australia are amongst a number of countries across the globe that have embraced the use of Cloud Computing Technology.

### A. Small-To-Medium Enterprises Growth and Technology Adoption

SME's are recorded to be more than 90% of all businesses in the Sub-Saharan Africa, consequentially thus considered to contribute a lot to a country's GDP. Albeit the controversy on which size business contributes more to a country's GDP, the fact that SME's are greater in number than larger enterprises, indicates a vital role SME's play in the economy. The SMEs' growth and sustenance hence becomes detrimental, and Robert indicated that there is need to focus on "technology advisory" for SMEs' [9], [10]. In his own words, his resolve was:

"These firms, then, need new and innovative types of technologies, such as Software-as-a-Service (SaaS), virtualization, and cloud computing; these firms also need (and deserve) a better kind of technology advisory delivery model to provide them with research, analysis and insight."

Botswana is a developing country, and is recognized as one of the fast growing economies in the world [18], [20].

Botswana has seen growth in businesses in general, and technology seeing a big growth whereas internet access, telecommunications (including mobile), transport, banking and education have also improved. The Maitlamo policy and the implementation of the E-government portal in 2012 are acclaimed to boost the use of technology in the country, Botswana.

### B. Botswana E-Readiness

The Botswana government rolled out the Maitlamo policy since 2004, final report, which represented Botswana's aim to have an internet environment for e-services in government department. These included areas on e-health, e-education (termed as e-Learning), e-legislation and connectivity to homes and communities, the Information and Technology sector, and making sure that infrastructure resources for availing these other services is also available. To continue the many efforts to implementing these pillars for Maitlamo ICT Policy, in 2012 the Ministry of Transport and Communications, Honorable Minister Nonfo Molefi, launched one of the most awaited service, the Botswana National e-Government Strategy, to tag along government services to be availed online [6]. This brought about the creation of Botswana Fiber Networks enterprise (BOFINET), whose mandate is to provide network infrastructure (use of FIBER connections), and connectivity to rural and urban areas (BOFINET). BOFINET runs as a sprectra over 9000km across the country, and has seen implementation in rural and urban areas. Areas like Selebi Phikwe, Tsabong, Kachikau, and Mohembo projects on the fiber deployment are ongoing.

Many companies and government entities have taken advantage of the introduction EASSy and WACS. This was a big investment by Botswana to provide a spread on connectivity to the country [5].

#### **Some of the efforts in Botswana on internet access are:**

Botswana Post, a private organization reached out to 81 communities, mostly villages in Botswana through e-centers, called kitsong centers [13], [16]. ICT training, e-services like faxing, emailing, research and telecommunication and postal services are provided in the Kitsong centers.

1) Botswana is hosting the TRASA, to regulate and improve Postal and ICT business environment in the SADC region (IST-Africa). This put, a subtle trust on the infrastructure of Botswana Telecommunication organization, allowing exploitation of available resources and known available services to support the mandate of TRASA.

2) The introduction of internet services, like the BOFINET, the growth of internet service providers across the country to allow home and office access, indicates development in internet access. BOCRA reported having registered 64 facilities, including satellite hospitality facilities, which are being used in urban areas where a form of on ground physical infrastructure is unavailable in BOCRA.

3) Mobile services grew with the country boasting with 3 Private Telecommunication Organization, i.e. Orange Botswana, Mascom Wireless and BeMobile, being the sub-

entity of Botswana Telecommunications Corporations. Allowing for mobile growth, inter-relations between international and local PTO's has given benefit to an increased access, as all these network provide affiliate and local mobile internet access.

Many can be said on the Botswana development on technology access. Education, health, agriculture, tourism, the libraries and many other departments, have initiated projects to merge their sectors with the ICT sector.

### C. Cloud Computing and SME Growth

1) *In Botswana:* There is neither record nor study of the number of companies that are using cloud services for business operations in Botswana. Companies like Dimension Data, Botswana Post, Acutec and IT-IQ are well known in Botswana as using cloud services, either as a platform, infrastructure, or software service [5]. These being ICT companies, we also learn of universities like Botho University, with the University of Botswana, widely using cloud as a service, for storage, applications, software platforms and development for engines and assisting their clients, student or staff, to have access to more cloud resources.

This identifies "The Cloud" as no foreign tool for businesses in Botswana. The use of Dropbox, Facebook, Flickr, and Google Apps and Drive has emerged with cloud usage, amongst many internet users. Some of these services are widely used by the population, as indicates:

a) It has been recorded in the internet world statistics that Botswana has 31.2% penetration rate for Facebook subscribers in 2016 Statistics, indicating a highest subscriber in Africa. University of Botswana.

b) The University of Botswana, School of Graduate Studies uses Google Apps for a collaboration of web-based programmes and storage, to allow for communication, collaboration on student learning, repository for students and institution, institutions application availability online with the benefit of any-time access, reduced risk implications, without requiring to purchase software nor hardware, with an easy interface and use by UB.

There are more services on the cloud that are being used by many companies. The sales force and ERP online have also become widely used by projects and client-base companies in Botswana.

2) *In Africa:* Countries like Nigeria, Kenya, Ghana, and South Africa has done a number of studies on Cloud Computing in Africa. A study in the sub-Saharan Africa identifies Nigeria as an "early adopter" on issues of cloud computing in Africa[2]. South Africa in Southern Africa Development Community (SADC) is known as one of the fastest growing in technology, alongside Mauritius and Zambia, and countries like Madagascar and Mozambique being in the top 10 internet access countries in Africa. Some examples of these countries adoption of Cloud Computing by SME's are:



a) Nigeria has proven adaptable to cloud computing services. A number of researches show that a number of institutions, and SME's in the country have adopted to the use of cloud services.

b) In one research, about 10 universities were used to investigate the adoption of cloud computing in local institutions [3]. The investigation showed that 90% of the universities was using cloud computing services. And the services common between the institutions were SaaS, PaaS and IaaS, which each rated at 70%, 20% and 10% usage (respectively) on average by the universities utilizing the services

c) In another research on SME's adoption issues of cloud computing, showed that though there are issues on the ground to cloud adoption by SME's, that opportunities offered by Cloud Computing are real [2].

d) Kenya has shown great growth in technology development. With the inception of MPESA project and Safaricom enterprise, amongst other technology dynamic stakeholders Kenya is widely and worldly known for its establishments in advancing technology innovations. Considering the SME business at 98% in Kenya, inevitably it is being considered to provide significant impact on employment (4,6 million people - 30% of the population), and brings about 18.4% of the annual country revenue [17]. Safaricom cloud has now taken root in Kenya, and the results of the study indicated that Kenya SME's have adopted to cloud technology in areas as payroll, call conferencing, accounting services, and even online domain hosting for websites. MPESA, a big financial project and widely adopted by many in Kenya, is also part of Safaricom initial projects.

e) South Africa is one of the fastest growing economies in cloud resources. The Deloitte report (2012), shows that SME's are majority players in cloud computing [7]. For example, South Africa hosts about 21 data centers across the country, and IBM launched its first Cloud Data Center in Johannesburg in March 2016. IBM will be using the already available Vodacom data center infrastructure. Development on cloud technology and growth in this case, can piggy back on already existing resources with South Africa, thus it can advance the use of cloud computing.

A telecommunications report indicated that African countries too are adept in developing Cloud services for its clients. MTN, one of the largest telecommunications providers in Africa, is providing a number of cloud services for Nigerian and Ghanaian SMEs [3]. MTN is known to provide telephony services to countries even in the south of Africa. It could then only be expected that such services could be adopted by the

rest of the world, if they are effectively utilized and scalable as other cloud services.

3) *In the World:* The Cloud resource is now a world commodity, a necessity from household to business. This is true for most countries, where internet access is readily available, cheaper, and infrastructure is available to house cloud computing services with limited challenges. The developed world, with countries like Australia, England, Japan, New Zealand, and Ireland have shown studies identifying the adoption of Cloud Computing in within their local business, or for regions in such countries.

a) *Ireland* – a study was conducted where 250 SME companies were surveyed, and the results were concluded that, indeed Irish SME's are well aware of the significance of the adoption of cloud services. Albeit knowing the benefits of cloud computing, Irish SME's were reluctant to adopt this resource, and acknowledged fears concerning “security, poor internet infrastructure, and trust”. These fears influenced their lack of adopting the cloud services [7].

b) *North east England* – is one of the fastest growing technology countries in the world, as one of the developed countries. Hence, it is provided by scalable technologies that deliver services and deliverables any market industry would require. A research conducted in the north east of England showed that cloud computing in the region was matured, even with the complexity and context sensitivity issues [4].

c) *Australia* – showed very impressive results for 2012 employment by SME's to be at 70%, with input of AUS\$480 billion (Australian Dollars), showing a critical role played by SME's for the Australian economy [8]. The survey was conducted to find out the success of cloud computing by SME's in Australia, but limitations on data collection delayed the findings [11].

#### D. Cloud Adoption Challenges

Like many technologies, and its advancement and new developments, Cloud Computing too has a number of challenges like deployment issues, data migration, security, privacy, regulation and cyber-attacks. In developing countries, other issues as water, communication resources, power, and gas can contribute to such challenges. Abubakar et al. added in his study that internet access, online security (trust and privacy) and economic development are common issues for Sub-Saharan Africa [1].

A research in Nigeria, considered about 10 universities in a survey to measure how they have adopted cloud computing for university operations. The study indicated that there were challenges that the institutions faced. Table I below shows the

challenges these universities faced [3]. The results depicted very high percentages for challenges the institutions faced in cloud computing adoption.

TABLE I. NIGERIA UNIVERSITIES (2014) - CHALLENGES OF CLOUD COMPUTING ADOPTION. SOURCE: AKIN, ET AL. 2014

S/N	Challenges of using Cloud Computing	% of Respondents
1	Data insecurity	89.3
2	Unsolicited Advertising	64.6
3	Lock-in	77.6
4	Reluctance to eliminate staff positions	64.6
5	Privacy Concerns	68.9
6	Reliability Challenge	64.2
7	Regulatory compliance concerns/user control	80.0
8	Institutional culture / Resistance to change in technology	59.2

1) *Cloud Adoption Challenges in Universities*: A number of the issues indicated in the Nigerian universities are shared by another study considering cloud adoption in the Sub-Saharan Africa [2]. These challenges seem very common amongst many countries, and they can be categorized into three: managerial, relational and technical [11].

#### E. Technology Frameworks

1) *TOE Framework*: The TOE framework is widely adopted for cloud computing. This model was developed in 1990 by Rocco DePietro, Edith Wiarda and Mitchell Fleicher. It is based on Technical, Organizational and Environmental (TOE) contexts/aspects of an enterprise that influence organizational operations. It is denoted as a “multi-perspective framework” that can help an organization identify barriers and benefits a technology innovation can bring to an organization, as denoted by the diagram below. A decision on adopting the cloud will be based on the consolidated positive feedback of these aspects.

2) *TMR Framework*: Companies and customers utilize the TMR frameworks which consider aspects as: Technical, Managerial and Relational, to inform on implementing cloud technologies.

The framework structure first considers the available and current cloud services; then secondly analyzes the organizations needs or requirements looking at the organizations Relations, Managerial and Technical (both internal and external) aspects; then thirdly maps the organizational results to the cloud variables; which finally should influence the adoption decision.

3) *ITIL Framework*: Information Technology Infrastructure Library (ITIL) is one of the popular IT

management courses, platforms for technology mobilization and management. The framework is also commonly used for IT management. The diagram depicts the variables involved in ITIL framework layout. The ITIL framework is reported to fill existing needs for ITIL evaluations and projects. The report also suggests that the framework can be modified for use in dynamic projects, on a wider scale [15].

4) *COBIT Framework*: This was developed in 1996 by ISACA and the IT Governance Institute. The CobiT framework is set of best practices for IT management solutions and services. The latest of these frameworks is COBIT 5, which was founded in 2012 after 4 other frameworks evolved.

This would allow for a holistic governance and management of the information and related technology in an enterprise.

COBIT 5 “principles and enablers are generic and useful for enterprises of all sizes, whether commercial, not-for-profit or in the public sector” [12]. Therefore, an ordinary user, without much skill in technology can benefit from using the model for adoption processes of any technology.

### III. METHODOLOGY

The research is intended to study issues that hinder and those allowing for cloud adoption in Botswana. Hence, this research will involve both qualitative and quantitative methods for analyzing these factors. With the findings, the study will help identify and influence the need for a local cloud computing framework for SME’s in Botswana. A questionnaire tool was used to collect data from a number of selected companies around Gaborone and its surroundings.

#### Qualitative and Quantitative research are defined as:

1) *Quantitative approach* – uses numerical values to either infer, experiment or simulate[14]. In the case of this research, data collected will be used to infer findings from the population common characteristics and relationships from the sample data collected.

2) *Qualitative approach* – consider opinions, descriptions or feelings, rather than numerical data [19]. The data collected from the survey will bring out opinions of participants on factors in cloud computing adoption, and these will be used along research data to validate issues around cloud computing.

About 51 companies were contacted for sampling. These companies were selected from various industries, as follows: About 41 of these companies are registered as Information Technology (IT) companies in Botswana. These include IT sales, maintenance, support services, managed services, software developers, and infrastructure support services; 5 companies are academic institutions; 1 Private Telecommunications Organizations (PTO’s); 1 Postal service company; and 3 Internet Service Providers SP’s.

A convenience sampling method is a sampling method which depends on collecting data from a population available to participate in the research [19]. The researcher has to conveniently make available the questionnaire for the

participants, and collect the questionnaires after a given and agreed time with the participants. Due to time factors and costs on the other methods of sampling, Convenience Sampling was favorable considering the limits to the research.

IV. FINDINGS

This is based on the key findings from the collected data. These findings are matched with the research objectives. At the end, these results are intended to guide on whether a framework is required nor not, and if there is need to developing one for the local SMEs. These major findings are:

A. Section 1: Demographics

The section’s intention is to establish that the company has an IT department and the number of employees per company. The findings show that:

Statistics of position of the various personnel participants

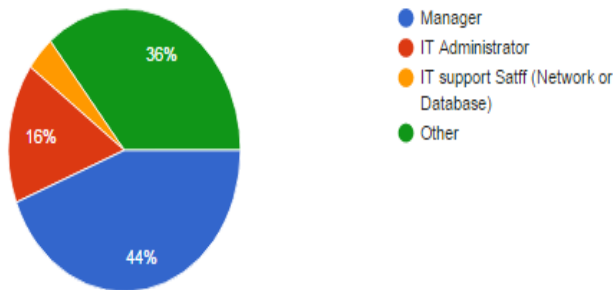


Fig. 1. Participants’ position in the company.

1) **Participant Position:** The questionnaire was targeted for people or staff in companies who had knowledge or were part of the IT department. The results show that, most of the participants were company managers. Then other greater percentage was “others” option, and these mostly included technicians in the company, and in one case, a finance and accounting manager, who was involved with the IT department operations. Fig. 1 gives the statistics of position of the various personnel participants.

2) **Company Employment Number:** Considering the research is focusing on SMEs’ in Botswana, one of the questions was intended to identify the differences between participating companies by the number of employees they have. The chart shows that only 28% of companies had employment of more than 250 employees. The remaining statistics indicate that the rest are SME companies, as they are employing 250 or less employees. What is significant yet, is that more than half of the indicated SMEs in this chart, are those employing at least 50 or less people (48%), are indicated as small or micro companies as reported by Fjose. Fig. 2 gives the organizational population for participants employers.

Organizational Population for participants’ employers

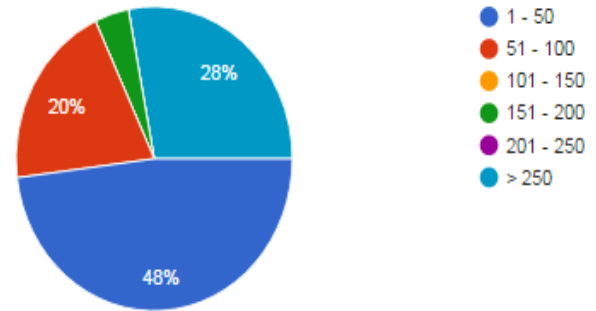


Fig. 2. Number of employees in the company.

B. Section 2: Cloud adopters

The section is intended to identify if the company uses cloud services in their business operations. One of the questions in this section was used to give examples of cloud services to assist the participant to understand what cloud services are commonly used by consumers on a general use basis. The examples used were common to Botswana context for most users (for example: Facebook, Google apps, MS Azure and Google drive).

1) **Companies using Cloud Computing:** There is no uncertainty on the participant whether they are using cloud or not. The results show that 60% of companies are using cloud services, whilst 40% said “no” to the use of cloud computing services. Fig. 3 shows the companies using cloud services in Botswana.

Companies using Cloud services in Botswana

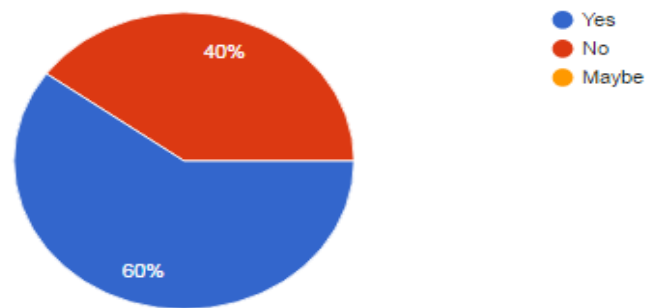


Fig. 3. Number of companies using Cloud Computing Services.

2) **Cloud Computing Services models used:** Cloud computing service models are categorized into SaaS, IaaS and PaaS. Participants needed to identify which models they were using for their organizations. The results show that they 66.7% of these companies that are cloud adopters are

using IaaS (66.7%), more than SaaS (53.3%) and PaaS (13.3%). The analysis of this results show that some of the companies are using more than one service model for their operations. Fig. 4 gives the service models used by cloud adopter companies.

**IaaS, PaaS, SaaS services for participating Companies**

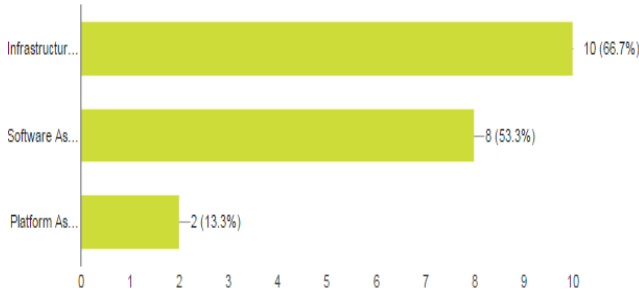


Fig. 4. Service models used by cloud adopter companies.

3) **Examples of cloud services:** examples used were included in the questionnaire to assist the participants have an idea of what common services the cloud has, which are be-known to Botswana cloud users. The results showed that companies could relate to cloud usage, and even the most unique software Microsoft Azure by Microsoft products, had an 8.7% acceptance of use. Facebook was only at 60.9%, whilst google apps and DropBox were both at 69.6%, being the highest in use. Only about 26.1% indicated to be using other cloud services. Fig. 5 shows the examples of cloud services that are used by companies.

**Examples of Cloud services that companies are using**

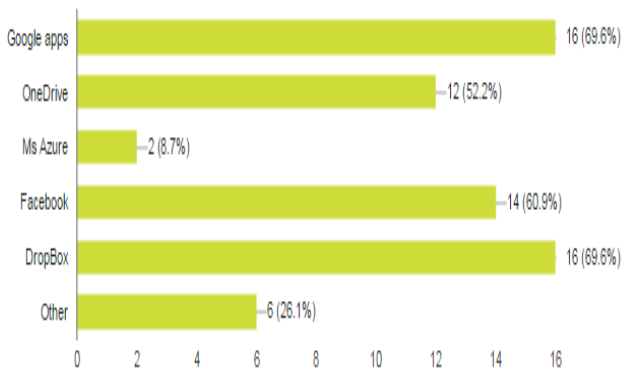


Fig. 5. Examples of cloud services common for local use.

**C. Section 3: Cloud Providers**

This section considers cloud services the participating companies provide to their clients, and to identify the extent of these companies' provisions, and what services they are providing to the clients. The results show that:

1) **Participant Providing Client with Cloud Business/Service:** At least 28% of companies mentioned providing cloud services to their client, whilst 12% were unsure if there were doing so. Most of the companies, of 60%, reported not to be providing these services for their clients or service consumers. Fig. 6 shows the companies providing cloud services to other companies.

**Companies providing cloud services to other companies**

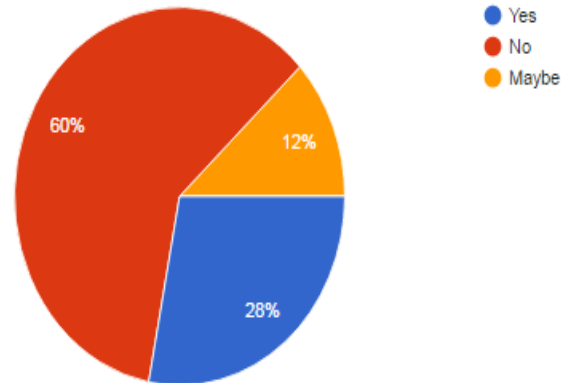


Fig. 6. Participant providing cloud services to their clients.

**Cloud services provided by other companies in Botswana**

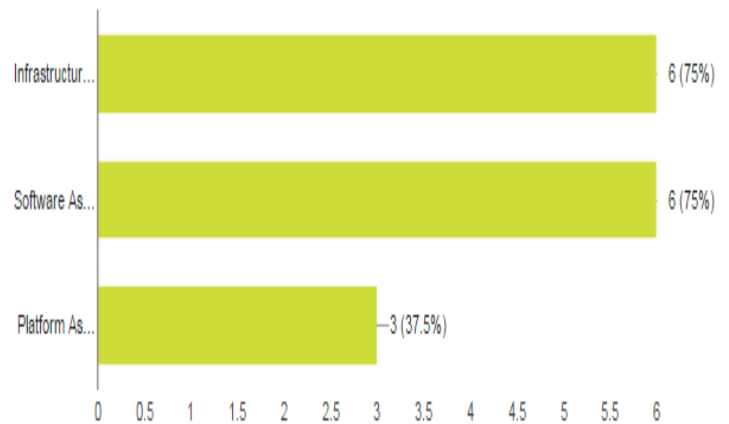


Fig. 7. Service models used by participants' clients.

2) **Service Models Provided by Participants to Clients:** The fact that some companies are providing cloud services to their clients as business, it is interesting to learn results that most of the service models provided are IaaS and SaaS, which are both rated at 75% from the graph below. PaaS does not seem to be common for business at first hand or third party, as it is reported only at 37.5% for client usage or consumer service. Fig. 7 shows the cloud services provided by other companies in Botswana.

3) **Number of Companies Receiving Service from Local Companies:** The results show that 85.7% of companies receiving business for cloud services from a local company, they are at least 10 or less. In comparison, only 14.3% of companies are reported to provide cloud services to more than 40 companies. There is no information on the range between 10 and 40. Fig. 8 shows the number of client companies for cloud services by the participants company.

**SME's Cloud Providers in Botswana**

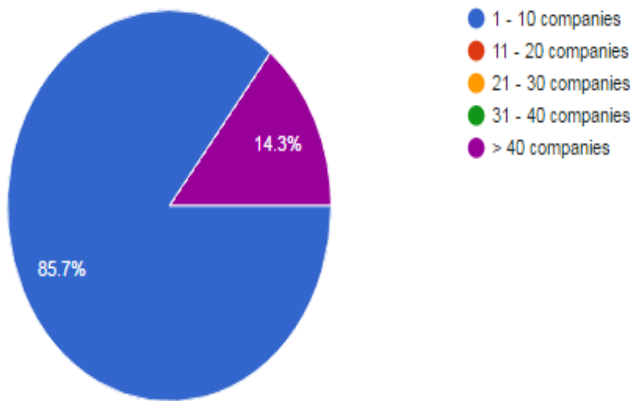


Fig. 8. Number of client companies for cloud services by the participants' company.

**D. Section 4: Cloud Challenges**

This section shows results on challenges that participants have reported to face in their organizations.

1) **Location Access Challenges:** Most of the participants showed they struggle to access cloud services from their home networks (56%), than on their mobile devices (48%), and least at work (20%). Fig. 9 gives the challenges accessing cloud at home, work or in mobile device.

**Challenges of using the cloud: Home or work?**

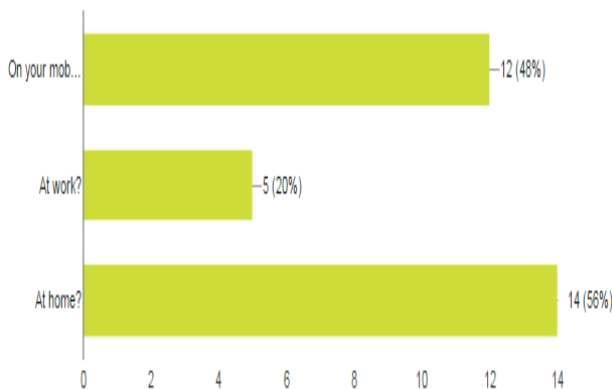


Fig. 9. Any Challenges accessing Cloud at home, work or in your mobile devices.

2) **Access For Work or Personal:** It needs to be established if the services being accessed were for personal or work activities from the locations noted above in Section 4(1). It shows that most of these failing accesses were for personal use at 81.8%, than for work purposes at 31.8%. Fig. 10 shows the results for access at home or work for personal or work purposes.

**Use of cloud services for participants**

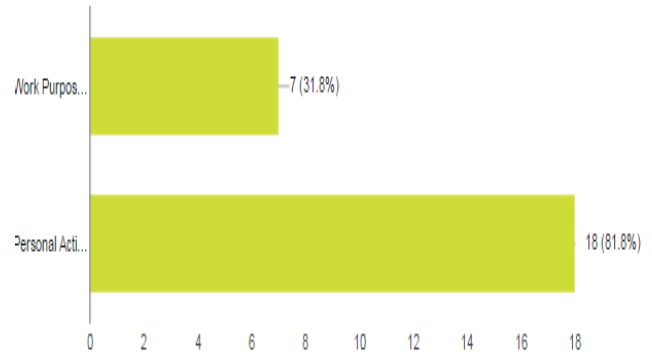


Fig. 10. Results for access at home or work for personal or work purposes.

3) **Local Challenges with Cloud Services:** Since the literature review indicated that a number of surveys have indicated that there are issues in cloud computing, the results from the survey indicate the same. The table below shows the results from the survey. Fig. 11 shows the challenges of cloud services for companies.

**Challenges of cloud services for companies**

**Do these issues affect your company negatively?**

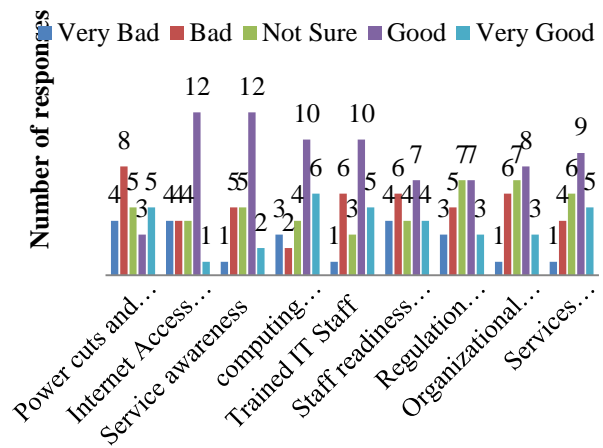


Fig. 11. Issues affecting Cloud services in local companies.

4) **Participants Opinions on their cloud use challenges:**

Some of the results in this section indicate participants opinions on challenges they face in their companies. These include points such as:

“My company is not aware of these services, they don’t want to use them”

“Abuse by clients and/or employee”

“We also sell cloud services to our clients”

“Resistance to change I guess”

“Unreliability of ISP”

“Security”

“The need for skilled personnel both in-house and customer. BYOD devices serve as a vulnerability, so there is need for security awareness regardless of security measures put in place”

“Security issues e.g. DDOS Attacks as well as Faulty Interoperability of Cloud service APIs”

These indicate the concerns of the participants they face in the business as they use cloud services. These are mostly not included in the previous parts of the section.

E. Section 5: Cloud Successes

This section considers participants results on benefits of cloud services to their organization, and to some point personal benefit. The section includes the participant opinions on their successes and benefits from the services.

1) A question in this section considered thoughts and opinions of the participants on what impact and benefits the cloud has provided for them, and some of these responses were:

“e save memory” translated as “It saves memory”

“There is greater market, ease of acquisition of services offered by my company”

“efficiencies, accessibility and lowering costs”

“Reliable backup of important information like student exam codes for modules like java programming”

“boosted productivity and service delivery”

“storage and scalability”

“our company specializes in multimedia production, which requires us to share big video and audio data files with clients across the globe. In the past we had to spend a lot of money sending these files through couriers like DHL. This has saved us time and money”

2) The statistical results of these improvements are shown below in the diagram. It shows that these benefits were across different levels of the participants daily activities, be it personally (83.3%), at work (58.3), and for the delivery of the company objectives (75%) and has seen other staff members

benefiting from its use (45.8%). Fig. 12 gives the benefits to client for using services.

Cloud services resourcefulness for participants

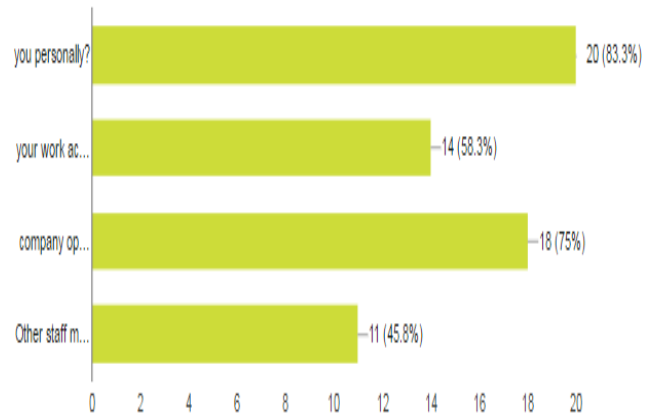


Fig. 12. Benefits to client for using cloud services.

3) Using other positive responses from other surveys, gave a list of issues that the participants may also be identifying as benefits for their organizations. The results are shown in the diagram below. The results show that many of these benefits, the participants have also benefited from them as we see rates of “good” and “very good” showing a greater number of responses in the diagram. Fig. 13 shows the positive impacts to cloud computing users.

Cloud access challenges impact in Companies

Do these issues affect your company positively?

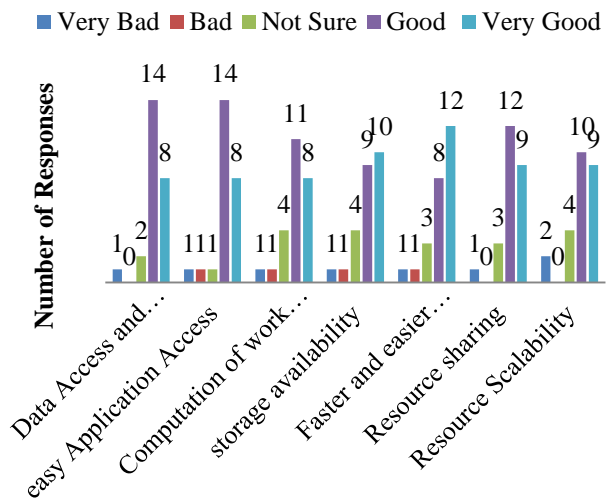


Fig. 13. Positive impacts to cloud computing users.

4) The final part of the section was to enquire from the participants if they have any successes to share on how cloud services benefited their organizations. Their responses were:

*“Re baya sengwe le sengwe mo lerung” Translated as “we keep everything on the cloud”*

*“The use of our private cloud has increased ICT awareness and use in the organization.. This has also brought in product innovation and better efficiency.”*

*“We have implemented more than 7 clients on cloud in the last 4 months”*

*“we have been able to store our previous work on the cloud making it easy for us to share these with prospective clients and organizational use only at a touch of a button.”*

*“Advertising”*

*“Reliable backing up and retrieval of important information anywhere within the institution”.*

## V. SUMMARY OF FINDINGS ON OBJECTIVES

### A. Objective 1: SMEs and Cloud Computing Services Effectiveness

#### 1) Cloud Use

- There are a lot of SME's, more than larger organizations that are using cloud services. The results showed that 60% of companies are using cloud services, and it also showed that at least 72% of these companies are SMEs. This indicates likelihood that since the majority of companies are SMEs, some of them are included in providing cloud services or business for their clients.
- What is more interesting is that, the majority of the companies providing cloud services to other companies, shows to be providing to a maximum of 10 companies per survey responses by participant. This could either be because it is a small company which is providing these services, and is still establishing itself to reach out for more business; or that there are challenges in cloud adoption by companies. The results does not show much challenges in Botswana adopting to cloud services,

#### 2) Access

- Access of cloud services did not seem a problem at work for the participants. They only showed a 20% challenge at work, but more challenges for their home access.
- The challenge in access of cloud services was mainly on personal usage or activities, hence showing a reduced percentage (31.6%) for challenges on work activities, than personal activities which are at 81.8%.

### B. Objective 2: Challenges and Successes for SMEs

#### 1) Challenges

Some of the challenges seen from other surveys were used for the survey, and the scores showed positive results from the

respondents. Most challenges, the participants indicated that they did not have them as challenges for their organizations. These include Internet access and availability, services awareness, computing resources availability, trained staff, staff readiness and awareness, organizational trust and services availability and access.

Though some of these challenges didn't show as challenges, most participants were not sure if these were challenges or not. For instance, “organizational trust” and, “standards, regulation and policy” showed about 30% uncertainty, whilst others (32%) felt it was bad and others (40%) felt it was good. These results are not showing sufficient disparities to understand if these are challenges or are good variables for cloud adoption. The biggest rated challenges were power outages and power cuts. Concerns mentioned and were common between participants were:

- Unreliability of Internet Service Provision
- Security issues
- Lack of skilled personnel
- Awareness issues
- Integration issues and interoperability

#### 2) Benefits

The feedback from participants was very positive as all the variables in the tables indicated for a positive impact in an organization, and were rated above 76%. Very few respondents indicated that they were unsure on a range of 4 – 16%, whilst a maximum of 8% showed having these variables indicating challenges in their organizations.

Many respondents are happy with the cloud as it brought out benefits such as:

- Backup of data in the cloud, and reliable
- ICT awareness
- Increased productivity and efficient service delivery
- Low costs
- Online readily available storage
- Easy acquisition of services

### C. Objective 3: Frameworks used for Technology Adoption

Participants indicated slim disparities on whether policy, regulation and standards could be a problem for their organizational business. Since there is no policy in Botswana or regulation for cloud computing, a question on whether regulation or policies are considered on cloud adoption was never included in the questionnaire.

## VI. IMPLICATION RESEARCH

Finally, conclusions on the results collected from participants are revealing positive cloud adoption in Botswana. Though, the adoption rate is very low considering that issues of awareness are impending cloud adoption decisions. Objectives for research are being answered as follows from the survey:

#### A. Objective 1: SMEs and Cloud Computing Services Effectiveness

A number of companies, especially SME's show that they are using cloud services in Botswana. A report about Sub-Saharan Africa countries agrees with the fact that cloud computing is slowly taking root in these countries and that SMEs are also stakeholders in this growth[2]. The report showed that SMEs make the most of the firms per country, and the results indicated too that a lot of companies reflected being either micro, small or medium businesses, and only very few were large enterprises.

#### B. Objective 2: Challenges and Successes for SMEs

The statistics are showing a very positive uptake of cloud adoption, as the results under challenges companies may be facing for cloud usage, show that the indicators used, many seemed happy with them. Hence, the challenges were very few, that were indicated to be a challenge for cloud uptake. These were power outages and cuts, and staff readiness and awareness, standards and policy, and organizational trust. Albeit these challenges, showing a marginal difference between the bad and good side of these indicators, other indicators were showing high positive scaling. The indicators showing greater rates as challenges are confirmed by the Nigerian research for universities, for example, 80% of universities showed that a lack of regulation on cloud adoption is a security concern [3]. Though the survey results showed a low matrix result on challenges companies may face, in the participants views, many showed concerns on security issues.

The results showed that cloud adopters were happy with the indicators used. Most importantly, some indicated successes on usage of using cloud services. One company revealed registering about 7 clients, for cloud business in a period of 4 months. Some companies recorded "increased productivity and efficiency" in their services.

#### C. Objective 3: Frameworks used for Technology Adoption

Various frameworks were discussed in the literature review. These are TMR, TOE, ITIL and COBIT frameworks. Botswana only has the Maitlamo (2004) ICT policy for all ICT aspects in Botswana. The Botswana ICT Master Plan in 2012 was intended to assist the Botswana government to come up with ways to compliment the initiative on Maitlamo policy, where all resources in ICT are to be harnessed to improve ICT adoption in the country.

#### D. Objective 4: Need to Develop a Cloud Computing Framework for SMEs in Botswana

Considering that cloud adoption does not seem to be having so many challenges in Botswana, there is no need for a framework to be developed yet. Further research is needed on cloud adoption in Botswana to see if a framework is absolutely necessary. With the research survey done, it still shows that there is further survey needed to find out the frameworks companies use or some may need, and also consideration on the understanding of how a framework works, despite have regulatory policies and standards.

## VII. RECOMMENDATIONS FROM THE STUDY

The recommendations are made with relation key findings of this research. A number of issues and questions were also raised to be included in this research, but a need for further research on this topic is needed. These recommendations include:

1) As Botswana has a good number of SMES adopting cloud computing, there is not clear understanding of how many of these companies are technology driven and using such services. There is no record either of the SMEs who are using cloud computing, which could be recommended as future work or research. This would help the statistics department of Botswana, BOCRA, ICT companies listing to know needs to be addressed.

2) A lack of awareness was indicated as a challenge. An establishment of what kind of awareness is needed can be undertaken as a research to survey the need for information dissemination or training to empower SME's on benefits of cloud computing.

3) Given enough time and resources for the research would have given more grounded results, and probably reached other areas around the country. Therefore, further research on the topic could reach out to the rest of the country's SME's.

4) Frameworks present various models on implementing technologies. Another research, could target these various technology frameworks, and with the assistance of this research, map aspects or principles that could be suitable for developing countries, on cloud adoption or technology adoption.

## VIII. CONCLUSION

The research indicates that indeed Botswana SME's are adopting cloud computing services. An ICT policy for Botswana is the only tool that guides ICT adoption and related issues in Botswana. The most challenging to conclude is the recommendation or suggestion of a framework to use by local SMEs in Botswana. The results show very positive adoption by cloud adopters, and showing they are not facing many challenges on cloud services usage. One of the main issues recurrent in a number of participants was lack awareness by users and companies. This could answer for the 40% of results showing non-adopters, that it could be a lack of awareness or disinterest to the services.

## REFERENCES

- [1] A.D Abubakar, J. M. B. I. A., ". Cloud Computing: Adoption Issues for Sub-Saharan African SMEs", EJISDC, 62(1), pp. 1-17,2014
- [2] Abubakar, A., Bass, J. M. & Allison, I., " Cloud Computing: Adoption For Sub-Saharan African SMEs". Electronic Journal of Information System in Developing Countries, 62(1), pp. 1-17,2014.
- [3] Akin, O. C., Matthew, F. T. & Comfort, D. Y., "The Impact and Challenges of Cloud Computing Adoption on Public Universities in South western Nigeria". IJACSA, 5(8), pp. 13-19.,2014.
- [4] Alshamaila, Y. & Papagiannidis, S., "Cloud Computing adoption by SMEs in the north east of England: A multi-perspective framework". Emerald Insight, 26(3), pp. 250-275,2012.



- [5] BOCRA, "Survey on Internet Connectivity in Key Strategic Areas in Botswana (Hospitality Facilities)", Gaborone: BOCRA,2014.
- [6] Botswana, U. o., 2009. [www.ubotho.net/MDP/it-support](http://www.ubotho.net/MDP/it-support). [Online] Available at: [www.ubotho.net](http://www.ubotho.net) [Accessed 10 11 2016].
- [7] Carroll, M. & Ramsingh, K., "Cloud Computing 2012 Survey Results", Johannesburg: Deloitte,2012.
- [8] Connolly, E., Norman, D. & West, T., "Small Business: An Economic Overview in Small Business Finance Roundtable", Sidney: Reserve Bank of Australia,2012.
- [9] Doherty, E., Carcary, M. & Conway, G., "Migration to the Cloud: Examining the drivers and barriers to adoption of cloud computing by SMEs in Ireland: an exploratory study. Emerald Insight", 22(3), pp. 512 – 527,2012.
- [10] Eastman, R., 2010. SMB Research - Sizing up Small-to-Medium Business (SMB). [Online] [Accessed 01 12 2016].
- [11] Fakieh, B., Blount, D. Y. & Busch, D. P., "Success in the Digital Economy: Cloud Computing", SMEs and the Impact to National Productivity. Auckland, New Zealand, 25th Australasian Conference on Information Systems,2014.
- [12] Garsoux, M., "COBIT 5 ISACA's new framework for IT Governance", Risk, Security and Auditing, s.l.: ISACA,2012.
- [13] .IST-Africa, 2015. [www.ist-africa.org/home/default.asp](http://www.ist-africa.org/home/default.asp). [Online] Available at: [www.ist-africa.org](http://www.ist-africa.org),[Accessed 10 11 2016].
- [14] Kothari, C., "Research Methodology: Methods and Techniques". 2nd ed. New Delhi: New Age Internationa (PTY) LTD,2014.
- [15] .McNaughton, B., Ray, P. & Lewis, L., "Designing an evaluation frameowkr for IT service management",. ELSEVIER, 47(003), pp. 219-225,2010.
- [16] Sebina, P. M. I. M., Moahi, K. H. & Bwalya, K. J., "Digital Access and E-Government: Perspectives from Developing ad Emerging Countries". Hershey, United States: Information Science Reference,2014.
- [17] .Simba, J. K., "Adoption of Cloud CComputing among Small to Medium Enterprises in Kenya". Erepository,2014.
- [18] Tsimane, E., "Botswana Looks to Cloud Computing", Gaborone: Sunday Standard,2013.
- [19] Walliman, N., "Research Methods: The Bacis". 1 ed. London and New York: Routledge: Taylor and Francis Group,2012.
- [20] Yevgeniy, S., "IBM Launches Cloud Data Center in South Africa", Johannesburg: IBM,2016.

# Survey Paper for Software Project Team, Staffing, Scheduling and Budgeting Problem

Rizwan Akram, Salman Ihsan, Shaista Zafar, Babar Hayat

Department of Software Engineering  
The University of Lahore, Chenab Campus  
Gujrat, Pakistan

**Abstract**—Software project scheduling is a standout amongst the most imperative scheduling zones looked by Software project management team. Software development companies are under substantial strain to finish projects on time, with budget, quality and with the suitable level of values and qualities. Inexperienced development team or potentially poor management can cause deferrals and costs that given scheduling and spending limitations are regularly unsuitable, prompting business basic disappointments. Software development companies frequently battle to convey extends on time, inside spending plan and with the required quality. For a fruitful project, both software building and software management are exceptionally vital. One conceivable reason for this issue is poor Software project management and, specifically, insufficient project scheduling and inadequate team staffing. Software project schedule issue is one of the essential and testing issues come across by the product project directors in the much focused software companies. Since matter is winding up hard with the expanding quantities of workers and tasks, just a couple of calculations exist and the execution is as yet not fulfilling, to build up an adaptable and powerful model for Software project arranging. In this paper we have attempted to expand a few systems and strategies and results yielded are explained.

**Keywords**—Software engineering; project management; software project resources; project scheduling; budgeting; team

## I. INTRODUCTION

Software development companies get by in a focused market by benefitting from the change of designer' push to helpful and effective software products. To fabricate such items, the organization as a rule takes after a procedure that partitions the development exertion into a few exercises. Each of these exercises requires particular attributes (for example, abilities, capacities, and experience) [2].

Software development companies frequently battle to convey extends on time, inside spending plan and with the user requirements or required quality [1]. One conceivable reason for this issue is low Software project management level, insufficient project scheduling, budgeting and unpracticed team staffing. Staffing a software project is exhausting movement [4], [14].

To build up a software project, the task chief needs to appraise the project amount of work and budget and choose the project calendar and asset distribution. Software project errands need representatives with various abilities, and expertise capability of workers altogether impacts the

effectiveness of project execution. Appointing workers to the well-suited tasks is trying for Software project supervisors, and human asset designation has turned into a significant part in Software project arranging in light of the abilities and encounters of the representatives [3].

Project management systems for the most part respect task planning and human asset portion as two isolated exercises and leave the activity of human asset designation to be finished by project directors physically, bringing about wasteful asset allotment and poor management execution. Principle assets in Software development are people rather than enormous machines, assets in Software projects can more often than not be distributed in a more adaptable manner than those in development or assembling projects [3].

The principle objective of our approach is to assist the project administrator with staffing his tasks, recommending teams that fulfill all requirements associated with the issue naturally. We likewise propose teams that fulfill the requirements, as well as streamline some factor of the issue. In our approach, project staffing is tended to as a requirement fulfillment issue [18], in light of utility capacities that ought to be boosted or limited by the chose development team, with a specific end goal to give more prominent incentive to the organization. A few utility capacities are displayed, to be chosen by the director as per authoritative requirements or imperatives [2].

The survey is planned about important issues counting engineering, planning, scheduling, cost estimation, and monitoring and control of the project operations, with the objectives to optimize cost and time throughout the capable uses of constrained/unconstrained resources [13].



Fig. 1. Project management.

## II. LITERATURE REVIEW

### A. Software Project Scheduling

Software Project scheduling is a system to communicate what tasks need to be done and which organizational resources should be designated to do those tasks in what time span. A project schedule is an archive assembling all the tasks estimated to deliver the task on time.

Nonetheless, concerning influencing a task to plan, well, that is something few have significant organization with.

What and who is being planned, and for what purposes, and where is this schedule occurring, in any case [12], [24]?

A project is contained many tasks, and each task is given a start and end (or due date), so it can be done on time. So also, people have particular timetables, and their availability and escape or leave dates ought to be record keeping in mind the end goal to effectively design those tasks [4], [23].

One of the fundamental obligations of a software project administration to make sense of what work will be done, when and how it will be done. This duty comprises of distinguishing the different items to be conveyed, assessing the exertion for each task to be attempted, and additionally building the project's timetable. Because of the significance of this movement, it ought to have need over all others, and besides, project's calendar should be refreshed routinely to agree with the project's present status.

### B. Software Team Staffing

Staffing a software project isn't a basic action. There are a few engineer-to-action blends to assess, since the manager is generally required to pick a group or team from a bigger arrangement of accessible developers. In addition, team choice is generally compelled by project and hierarchical needs, for example, most extreme team month to month cost, assessed improvement time, and engineer's capacity confuse to exercises, necessities [9].

In these duties a software project administrator is to decide people and their part in teams for project work. Which project contains what level of work team, much the same as experienced or unpracticed representatives [5].

As expressed in [6], the most well-known staffing techniques accessible to software extend chiefs depend vigorously on the task manager's close to home encounters and learning. Be that as it may, these are exceedingly one-sided systems and objectivity does not generally produce the right or finest outcomes. Additional matter is the way that in light of the fact that each project is one of a kind, the utilization of a particular enrolling and staffing technique on a project may not yield the normal outcomes as it was connected on another task as a result of the distinctions in project qualities [7]. And this connects to the way that expertise based and encounter based techniques are not sufficiently reasonable for project chiefs to manage relational connections and social perspectives which emphatically exist in software development organizations [8].

## III. PROJECT MANAGEMENT PLAN – MAJOR COMPONENTS

Projects don't oversee themselves. Proficient project management requires the development of an arrangement that blueprints how it will be overseen. These are the segments that give the center paying little heed to industry or sort of project.

PMP Major Components:

**Scope Management** – defines the natural surroundings for the deliverables of the software project produced to fulfill the project needs and organization desires. It helps to define:

- Scope – what is out and what is in the scope
- Specification – of each of the main products

**Resources and Resource Management** – e.g. machinery and apparatuses, human and their abilities, crude equipment's and semi-completed items, common assets (vitality, water, arrive, and so on.), data, cash and so forth,

At the very least each project must to have a project sustenance and project chief!

This section includes:

- Duty of project network – which describes who is in charge of the fulfillment of every item.
- Organization failure structure, which exhibits the organization progression of the project.
- Delegation schedules, which describes the position of specialist inside the project for the confirmation of records, intakes and response.
- Role descriptions which defines the basic requirements.

**Scheduling** – high level schedule, which features the strategic hopes as a topic of reference schedule. This is must to similarly integrate reviews of cost and quality needs. Particular substance includes:

- Preference diagrams
- Gantt charts
- Resource histograms
- Project lifespan

**Cost Management and Budgeting** – is the surveying of costs and the setting of an agreed spending plan, and the organization of genuine and gage costs against that money related calendar. Having the ability to anticipate with some affirmation the rate at which the task is spending its advantages is basic to knowing whether the endeavor is on track.

**Stake Holders** – Each project has stakeholders. Stakeholders are individuals who have an enthusiasm for the effective fulfillment of the project. There are a wide range of sorts of stakeholders, and the stakeholders change by project. Yet, the essential thing to recollect is that the stakeholders ought to have some part in characterizing the project targets, since they are the general population will's identity influenced by the result. When characterizing project stakeholders, the

project manager and individuals from her or his team ought to deliberately thoroughly consider will identify the end clients of the item, regardless of whether it be managements or merchandise, and whether the item will have a beneficial outcome, and how it is probably going to be gotten. Some of the stakeholders are Customers/customers, Sponsors, Company, Team individuals and the Project Manager [4].

#### IV. SOFTWARE PROJECT SCHEDULING PROBLEMS (SPSP)

SPSP is one of the communal issues in organization software projects. It comprises in choosing who ensures what amid the software product lifespan. SPSP ought to consider pay rates and employees abilities which should be appointed to extend tasks as indicated by the necessities of these tasks [11], [20]. We show the model in Table I.

Scheduling is setting a succession of time-subordinate capacities to play out an arrangement of ward tasks that make up a project [11]. Reliance of tasks is imperative as far as need and priority. So it is conceivable that doing a task identified with doing a few tasks which for this situation, it is said that project contains need confinements.

Deciding a scheduling program has been finished with thinking about the reason or determined purposes. Nearly, there are needs constraints between projects in the majority of the activities; however, notwithstanding this impediments might be there is another sort of confinements between errands in light of asset restrictions. So in project scheduling for expansion to thinking about need impediments, scheduling ought to be done so as to be reliable with asset imperatives. In SPSP, considering an arrangement of uses, for example, assets et cetera are required for tasks [18].

TABLE I. SPSP MODEL

Item	Description
$S = \{s_1, \dots, s_{sk}\}$	Set of skills associated with software projects
$T = \{t_1, \dots, t_T\}$	Set of task necessary for the project
$G(V,A)$	Precedence graph defined in the project's Gantt
$V = \{t_1, t_2, \dots, t_T\}$	Is vertex set consisted of all tasks
$A = \{(t_i, t_j), \dots, (t_n, t_T)\}$	Is an arc set, the task $t_i$ must be done before $t_j$
$t_j^{skills}$	Is a set of skills for the task $j$ . It is a subset of $S$
$t_j^{efforts}$	Is an effort person-months to complete the task $t_j$
$EM = \{e_1, \dots, e_E\}$	Is a set of employees
$e_i^{skills}$	Is the set of skills of $e_i$ . It is a subset of $S$
$e_i^{mazed}$	Is the maximum degree of dedication of $e_i, e_i \in (0, 1)$
$e_i^{salary}$	Is the monthly salary of $e_i$

Additionally identifies with the choice of who does what amid a product project lifetime, therefore including basically the two human serious exercises and HR. Two noteworthy clashing objectives emerge when scheduling a product project: lessening the two its budget and span. A multi-target methodology is accordingly the normal method for confronting the SPS issue. As organizations are getting associated with bigger and bigger software projects, there is a real need of calculations that can manage the gigantic pursuit spaces forced.

#### V. DYNAMIC TRAVELLING SALESMAN PROBLEM

Dynamic travelling salesman problem (DTSP) is one of the optimization problems which it isn't solvable with classical methods. To tackle this issue, different solutions in the literature can be seen that each one have advantages and disadvantages. Genetic Algorithm (GA) and Ant Colony Optimization (ACO) have regarded settle the DTSP [19].

#### VI. METHODOLOGY

##### A. Genetic Algorithm Overview

In 1975 John Holland introduced Genetic algorithms [5], work repetitive with populaces of contestant results challenging as an age band, in order to accomplish the set of individuals reflected as abets result to an issue.

Team staffing and Project scheduling may be reflected optimization issues, and as such will need specialized techniques to be solved. Hereditary calculations are one such enhancement strategy, with which it is conceivable to satisfactorily display the numerical idea of project scheduling and team staffing.

GAs keeps up a populace on a specific size. Every person, which speaks to a speculative answer for issue, is aggressively controlled by applying some variety managers to locate a worldwide ideal [10].

To accomplish the objective of finding a worldwide ideal the issue factors are programmed into what are famous as the chromosome. Along these lines, one individual is related with one programmed arrangement (chromosome) as well as its related wellness comparing to the arrangement. GAs enhance the person wellness, which implies the streamlining level of arrangement, by utilizing sorts of focused operations. Also, with the expansion of age quality, wellness of genetic material is ending up well [4].

In view of the procedure of regular advancement, their point is for acceptable single answers for 173 C. A.S. Andreou and Stylianou beat those that are fewer hard at every age. To accomplish this, the wellness of each single arrangement is assessed utilizing a few criteria in respect to the issue, and in this manner those assessed exceedingly are more plausible to frame the number of inhabitants in the people to come. Advancing sound, happier people and disposing of less reasonable, weaker people in a given age is helped by the utilization of varieties of the choice, hybrid, and change managers, which are in charge of picking the people of the following populace and modifying them to expand wellness as

ages advance – making in this manner the entire procedure take after the idea of ‘survival of the fittest’.

### B. Encoding and Representation

For the issue of team staffing and software project scheduling, the contestant results for optimization need to signify two sections of material. From one perspective, schedule imperative data, in regards to when and in which arrange projects are executed and, then again, expertise limitation data, concerning the task of workers to errands in view of ranges of abilities and experience necessary for a task. Fig. 1 beneath gives a case of the portrayal of a product project schedule holding four tasks and five possible workers. As appeared, the hereditary calculation utilizes a blended sort encoding: schedule data is spoken to by a positive, non-zero number symbolizing the begin day of the errand, though representative task data is spoken to by a twofold code, wherein each piece implies whether a worker is (an estimation of 1) or isn't (an estimation of 0) doled out to complete the task [5].

TABLE II. SOFTWARE PROJECT SCHEDULE REPRESENTATION EXAMPLE

1	10100	11	00010	16	01001	31	00110
---	-------	----	-------	----	-------	----	-------

In Table II, the

- 1<sup>st</sup> task starts at day one
- worker 1 and 3 will fulfill it
- Task 2<sup>nd</sup> starts at day 11 with only worker 4 allotted to it, and so on.

### C. Ant Colony Optimization

A heuristic optimization technique for most limited way and other advancement issues which obtains thoughts from natural ants. In view of the way that ants can discover most limited path between their nest and source of food.

Ant Colony Optimization (ACO) Overview “Ant Colony Optimization (ACO) thinks about simulated frameworks that take motivation from the conduct of real ant colonies and which are utilized to take care of discrete optimization issues.”

Ant Colony optimization procedure is an arrangement of guidelines in view of look artificial intelligence algorithms for ideal results; here is the notable part is ANT System, as suggested by Maniezzo Colorni and Dorigo [20]-[22]. Ants are outwardly disabled and little in dimension and still can locate the straight path to their sustenance source. They all make the utilization of projections and pheromone fluid to be in contact with one another. ACO encouraged from the performance of living ants, are equipped for management with scanning answers for nearby issue by keeping up exhibit rundown to keeping up past data accumulated by every ant.

Also, ACO manages two vital procedures, specifically: Pheromone statement and trail pheromone dissipation. Pheromone testimony is the marvel of ants including the pheromone all ways they take after. Pheromone trail vanishing implies diminishing the measure of pheromone kept on each way regarding time. Refreshing the trail is executed when ants either entire their hunt or catch the most limited way to

achieve the sustenance source. Every combinatorial issue characterizes its specific particular refreshing criteria relying upon its specific nearby hunt and worldwide inquiry individually [18].

Artificial ants left an essential trail gathered on the route sector they follow. The route for every ant is nominated on the source of quantity of “pheromone trail” extant on the thinkable route begin from the present ant node. In situation of equivalent on nearby routes, ants arbitrarily select the route.

On a route Pheromone trail growths the possibility of the route being taken after. Then Ant achieves the next node and another time organizes the route selective progression as defined above. This progression carries on till the ant achieves the present node. This completed visit gives the result for limited or best route which would then be assessed for optimality.

It may be useful for different combinatorial implementing problems. ACO algorithms are utilized by easy mediator called “ants” that is iterative developed contender result to combinatorial executing problems [15].

The “ants” result development has been conducted by artificial pheromone track & issue-dependent experimental information. By law, ACO algorithms may be useful to any combinatorial executing problems by significant “solution components” that the ants utilized to iterative developed contender result and on which they can put down pheromone [11].

Fig. 2 Flow chart of ACO Algorithm is among the best swarm based algorithm propounded by Dorigo and Di Caro in 1999 [20], [21]. It is a meta heuristic motivated by the scrounging activities of ants in the wild, and in addition, the marvels known as stigmergy, term utilized by Grasse in 1959. Stigmergy alludes to the backhanded correspondence among a self-organizing emanant framework through people adjusting their nearby condition.

The most intriguing part of the communitarian conduct of a few subterranean ant animal types is their capacity to discover briefest ways between the ants' nest and the food sources by following pheromone trails Then, ants pick the way to take after by a probabilistic choice one-sided by the measure of pheromone: the more grounded the pheromone trail, the higher its attractive quality. Since ants thus store pheromone on the way they are following, this conduct brings about a self-fortifying procedure prompting the development of ways set apart by high pheromone fixation. By displaying and reenacting subterranean insect searching conduct, brood arranging, settle building and self-collecting, and so on calculations can be created that could be utilized for unpredictable, combinatorial advancement issues [25].

A satisfactory model for software project planning needs to manage the issue of undertaking assignment planning as well as the issue of human asset portion. In any case, as both of these two issues are troublesome, existing models either experience from the effects of an extensive inquiry space or need to confine the adaptability of human asset assignment to improve the model. To build up an adaptable and powerful model for software project planning [16].

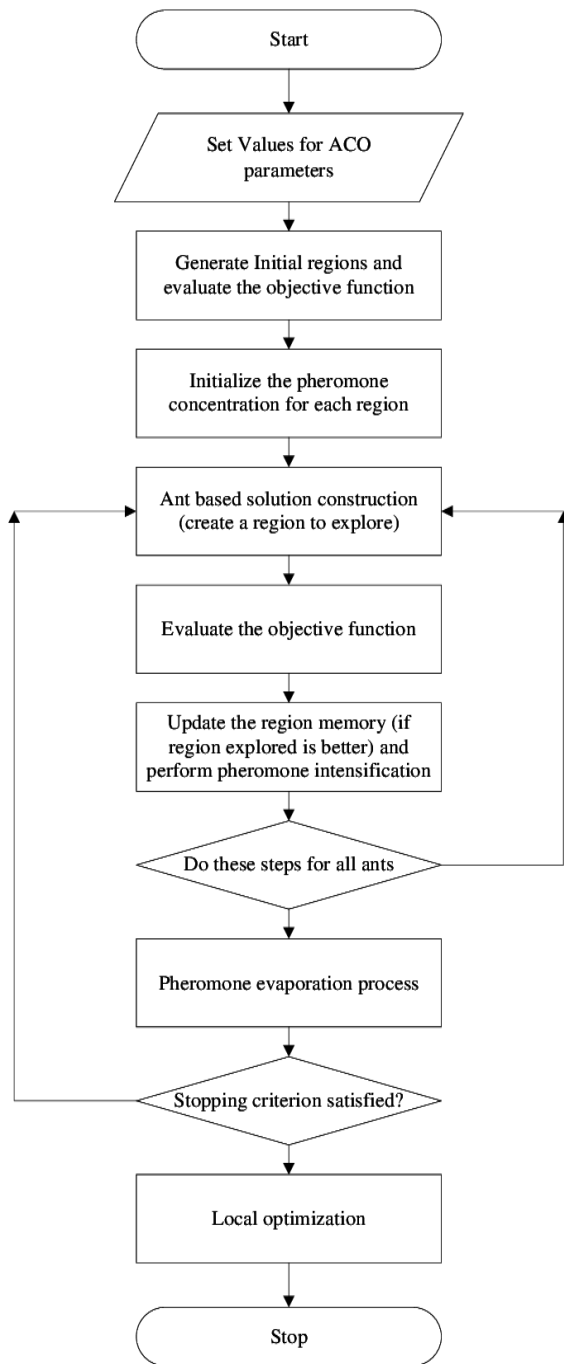


Fig. 2. ACO Flow Chart.

Network analysis provides an effective practical system for planning and controlling large projects in construction and many other fields. Ant Colony System is a recent approach used for solving path minimization problems [17].

#### D. Differential Evolution (DE) Algorithm

The algorithm DE is of meta-heuristic algorithm which was created in 1995 [18]. DE algorithm is a populace-based probabilistic search algorithm which solves optimization issues. This algorithm using the distance and direction information from the existing populace carries out the search processes. The benefits of this algorithm are speediness,

setting the constraints, its effectiveness in finding optimal results, being parallel, high accuracy and absence of need to sorting or matrix duplication. DE Algorithm in instruction to search the optimal results, has the ability to proficiently search the process in the route of coordinate axes of optimistic variables and also changes in the route of the coordinate axes in the right direction. DE Algorithm starts the evolutionary search process from a random initial population. DE Algorithm begins the developmental hunt process from an arbitrary introductory populace. Three managers of change and choice, and the incorporation and three control parameters, including the quantity of populace, scale factor and the likelihood of coordination are essential in the DE algorithm. DE algorithm forms are as per the following (shown in Fig. 3):

- Initial Population Generation
- Mutation Operator
- Crossover Operator
- Selection Operator
- Stopping Criteria

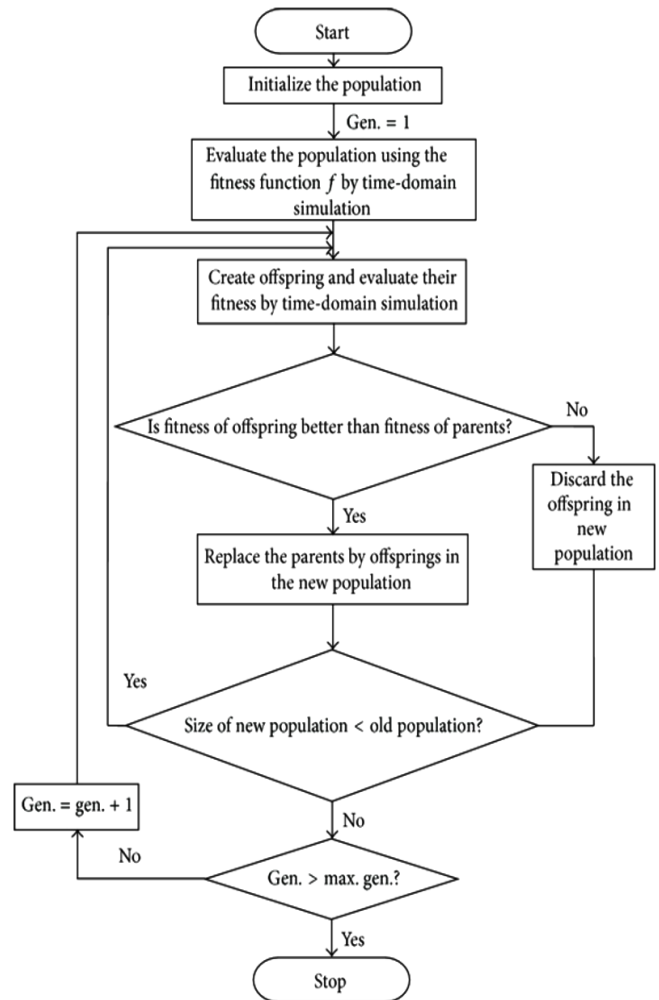


Fig. 3. Differential Evolution (DE) Algorithm Flow Chart.

## VII. PREEMPTABILITY

In many prototypes of project scheduling issues it's accepted that exercises are non-preemptible, however in a few projects this presumption is casual and it is permitted to seize exercises. All in all, for each the project exercises may be either preemptible or (non) preemptible. In any case, normally it is accepted that (non) preemptibility concerns all exercises on the double. Under this supposition, we discuss an arrangement of non-preemptible exercises if none of them might be seized, while we discuss an arrangement of preemptible exercises if every movement can be appropriated whenever and revived later with zero amount. Appropriation might be either distinct, if movement acquisition is permitted toward the finish of eras just, or ceaseless, if appropriation may happen at a self-assertive time moment.

## VIII. CONCLUSION

This paper exhibited a way to deal with tackling the issue of team staffing and software project schedule by receiving a hereditary calculation as a development system to build project's ideal calendar and to dole out the most experienced representatives to tasks.

Software development includes time, ability, and cash. In an aggressive market, a product development organization's real objective is to boost esteem creation for a given project. Subsequently, an appropriate use of each accessible asset in a product project is imperative.

Among a few strategies ACO noises well as it fabricates arrangements in a well ordered and iterative way empowering the utilization of issue based heuristics to control the inquiry bearing of ants, it is conceivable to outline valuable heuristics to guide the ants to plan the basic errands as right on time as could be allowed and to relegate the project tasks to appropriate representatives with required abilities.

This survey of different methods will be useful for better investigation and developing new thoughts for far and away superior schedule systems.

## REFERENCES

- [1] Optimizing Software Project Management Staffing and Work-Force Deployment Processes using Swarm Intelligence; Mazhar Hameed, Usman Qamar, Hiba Khalid, Syed Khizer Abass, Computing Conference 2017 18-20 July 2017 | London, UK.
- [2] Staffing a software project: A constraint satisfaction and optimization-based approach Ahilton Barreto, Márcio de O. Barros. Computers & Operations Research 35 (2008) 3073 – 3089.
- [3] A Hybrid Approach for Software Project Scheduling V.Karthiga and K.Sumangala, International Journal of Computer Applications (0975 – 8887) Volume 59– No.16, December 2012.
- [4] Survey paper for Software Project Scheduling And Staffing Problem Nandkishor Patil, Kedar Sawant, Pratik Warade and Yogesh Shinde, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2014.
- [5] Intelligent Software Project Scheduling and Team Staffing with Genetic Algorithms. Stylianou C., Andreou A.S. In: Iliadis L, 2011.
- [6] , S.T., Juristo, N., Moreno, A. M.: Emphasizing Human Capabilities in Software Development. IEEE Softw. 23(2), 94–101 (2006).
- [7] Wi, H., Oh, S., Mun, J., Jung, M.: A Team Formation Model Based on Knowledge and Collaboration. Expert Sys. Appl. 36(5), 9121–9134 (2009).

- [8] Amrit, C.: Coordination in Software Development: The Problem of Task Allocation. In: 27th International Conference on Software Engineering, pp. 1–7. ACM, New York (2005).
- [9] Barreto, A., Barros, MdO, Werner, C.M.L.: Staffing a software project: a constraint satisfaction and optimization-based approach. Comput. Oper. Res. 35(10), 3073–3089 (2008).
- [10] Review of various Software Project Scheduling techniques, Ramandeep Kaur, Sukhpreet Singh, International Journal of Computer Science & Engineering Technology (IJCSET).
- [11] Solving software project scheduling problems with ant colony optimization Jing Xiao a,n, Xian-TingA, Computers & Operations Research 40 (2013) 33–46.
- [12] A New Approach to Solve the Software Project Scheduling Problem Based on Max–Min Ant System, Broderick Crawford, Ricardo Soto, Franklin Johnson, May 2014.
- [13] Liao, T.W., Egbelu, P., Sarker, B., Leu, S.: Metaheuristics for project and construction management a state-of-the-art review. Autom. Constr. 20(5), 491–505 (2011).
- [14] N. Nan, D.E. Harter, "Impact of Budget and Schedule Pressure on Software Development Cycle Time and Effort", *IEEE Trans. Software Eng.*, vol. 35, no. 5, pp. 624-637, Sept./Oct. 2009.
- [15] Crawford, B., Soto, R., Johnson, F., Monfroy, E.: Ants can schedule software projects. In: Stephanidis, C. (ed.) HCI International 2013—Posters Extended Abstracts, volume 373 of Communications in Computer and Information Science, pp. 635–639. Springer, Berlin (2013).
- [16] Chen, W., Zhang, J.: Ant colony optimization for software project scheduling and staffing with an event-based scheduler. *Softw. Eng. IEEE Trans.* 39(1), 1–17 (2013).
- [17] Abdallah, H., Emara, H.M., Dorrah, H.T., Bahgat, A.: Using ant colony optimization algorithm for solving project management problems. *Expert Syst. Appl.* 36(6), 10004–10015 (2009).
- [18] Maghsoud Amiri, Javad Pashaei Barbi. : New Approach For Solving Software Project Scheduling Problem with Differential Evolution Algorithm. International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.5, No.1, (2015).
- [19] F.S. Gharehghogh, I. Maleki, M. Farahmandian, "New Approach for Solving Dynamic Traveling Salesman Problem with Hybrid Genetic Algorithms and Ant Colony Optimization", International Journal of Computer Applications (IJCA), Vol.53, No.1, pp.39-44, 2012.
- [20] Crawford, B., Soto, R., Johnson, F., Monfroy, E.: Ants can schedule software projects. In: Stephanidis, C. (ed.) HCI International 2013—Posters Extended Abstracts, volume 373 of Communications in Computer and Information Science, pp. 635–639. Springer, Berlin (2013).
- [21] Chang, C.K., Yi Jiang, H., Di, Y., Zhu, D., Ge, Y.: Time-line based model for software project scheduling with genetic algorithms. *Inf. Softw. Technol.* 50(11), 1142–1154 (2008).
- [22] Dorigo, M., Stutzle, T.: Ant Colony Optimization. MIT Press, USA (2004).
- [23] M. di Penta, M. Harman, G. Antoniol, The use of search-based optimization techniques to schedule and staff software projects: an approach and an empirical study, *Software – Practice and Experience* 41 (5) (2011) 495 – 519.
- [24] M. Harman, S. A. Mansouri, Y. Zhang, Search-based software engineering: Trends, techniques and applications, *ACM Computing Surveys* 45 (2012) 11.
- [25] Anukaran Khanna, Akhilesh Mishra, Vineet Kumar Tiwari, P N Gupta; A literature based survey on swarm intelligence inspired optimization technique: International Journal of Advanced Technology in Engineering and Science Volume No 03, Special Issue No. 01, March 2015.
- [26] Subramaniam Sumithra and T. Aruldoss Albert Victoire; Differential Evolution Algorithm with Diversified Vicinity Operator for Optimal Routing and Clustering of Energy Efficient Wireless Sensor Networks: Hindawi Publishing Corporation, The Scientific World Journal Volume 2015, Article ID 729634, 7 pages (2015).

# Real-Time Experimentation and Analysis of Wifi Spectrum Utilization in Microwave Oven Noisy Environment

Yakubu S. Baguda

Information System Department, Faculty of Computing and Information Technology,  
King Abdulaziz University, Rabigh, Saudi Arabia.

**Abstract**—The demand for broadband wireless communication in home and office has been increasing exponentially; thus, the need for reliable and effective communication is very crucial. Both theoretical and experimental investigations have clearly shown that electromagnetic radiation from external sources such as microwave oven (MWO) has detrimental impact on the wireless medium and the media content. Therefore, this drastically degrade the signal strength in wireless link and consequently affects the overall throughput due to noise and interference. This experimental study is primarily aimed at critically analyzing and evaluating the impact of electromagnetic radiation on spectrum utilization under different experimental scenarios. The experimental results clearly show that electromagnetic noise radiation from microwave oven can seriously affect the performance of other devices operating in 2.4GHz frequency band, especially, delay sensitive applications and services.

**Keywords**—*Electromagnetic radiation; microwave oven; spectrum utilization; bandwidth; ISM band; signal strength; throughput; wireless channel*

## I. INTRODUCTION

Wireless communication is increasingly becoming popular and widely used in recent years due its flexibility, scalability and low cost of deployment. It can be potentially applicable in health, public, and commercial sectors. More importantly, it has been used in home environment and hence its potential applications cannot be over emphasized. The 2.4 GHz frequency band is primarily dedicated for industrial, scientific and medical (ISM) usage [1], [2]. Hence the electromagnetic radiation from other equipment operating in the same frequency band could cause interference and subsequently degrades the performance of the wireless network. Cellular phones and microwave ovens are the most common household appliances operating in the frequency band which can electromagnetically radiates noise, and it interferes with other devices. The fact that IEEE802.11b uses the 2.4 GHz frequency band [10], there is every tendency that ISM equipment can reduce the signal strength of the transmitted radio signal over wireless LAN. In [3], [11]-[14], the effective dynamic spectrum utilization for the future wireless networks has been described. The electromagnetic radio frequency (RF) power radiated from microwave oven can cause loss data or connectivity in wireless networks [4], [5]. The impact of some devices on WiFi networks has been studied in [6].

Several theoretical and experimental works related to microwave ovens interference in wireless network have been conducted in order to evaluate and analyze the overall performance of the network [7]. Investigating such phenomenon will assist tremendously in designing efficient strategies on how to mitigate the impact of the interference caused by microwave oven radiation on signal quality. Indeed, this is extremely important with rapid growth in demand for different application over wireless network. Most of the devices and gadget has the capability to seamlessly stream multimedia wirelessly through WLAN at home or office. Microwave oven radiation can reduce the network performance and wireless devices attached to it. By critically analyzing the impact of noise generated by microwave oven on the network and content, adequate provision and network planning can be made by network administrators in order to achieve optimal performance.

Both stochastic and empirical models have been developed using amplitude probability distribution (APD) to model the microwave oven EM radiated noise for simulation purpose. In [1], bit error rate (BER) has been used to evaluate the noise generated by the microwave oven. Higher bit error rate can eventually lead to increase in packet error rate (PER) and it can be computed from the BER. Some researchers used stochastic model to mimic the noise in the wireless link. All these approaches are primarily developed for theoretical analysis of microwave oven noise. In this work, practical noise pulse generated by microwave oven is considered in the experimentation in order to exactly quantify and evaluate its detrimental impact. The fact that large amount of power is released by microwave oven when compared to wireless LAN devices, the performance of wireless devices need to be investigated while microwave oven is actively ON. In [8], [9], it has been clearly indicated that the performance of wireless network can be greatly affected by microwave oven noise. In this paper, real-time experimentation has been conducted to critically evaluate and analyze the interference from external sources which affect the channel quality, throughput and the spectrum utilization based on the real world electromagnetic interference source. There is need to clearly understand and characterize the dynamic nature of interference in WiFi networks due to their significant role in today's communication and networking industry. Undoubtedly, this will eventually leads to the development of highly efficient



and effective strategies to mitigate the interference in WiFi networks.

Section 2 mainly focuses on the effect of microwave ovens on wireless LAN devices from both theoretical and numerical perspectives. Section 3 discusses the experimental procedure and measurement setup used to conduct the experiments. Section 4 discusses the results from the experiments. Finally, conclusions are enumerated in Section 5.

## II. ELECTROMAGNETIC RADIATION EFFECTS

The electromagnetic energy release by microwave oven can be disastrous to other devices and application using the same ISM band – it can eventually increase the bit error rate and consequently can lead to incessant packet lost. For example, multimedia applications are sensitive to noise and delay which can cause errors while decoding the media content at destination. The degradation of the multimedia content due to EM radiation has not been fully exploited and requires more understanding of the impact of electromagnetic radiation on wireless networks.

### A. Theoretical and Experimental Background

It has experimentally been proved that the power density of microwave oven does not exceed up to  $1 \text{ mW cm}^{-2}$  from the surface where it radiates the noise energy. Based on statistics, it can be assumed that only 50% of microwave ovens used today emit less than  $0.062 \text{ mW per cm}^2$ .

The interference between microwave oven noise and signal transmitted by wireless device can be model as shown in Fig. 1. The amplitude of the pulse generated as noise could be model theoretically. It is very obvious that the amplitude and frequency varies with the source voltage. The mathematical expression representing the noise is described in (1). The model is mainly based on the assumption that both the frequency and amplitude of the noise pulse directly varies with the source voltage variation. Fig. 1 shows a typically model illustrating the effect of noise generated A(t) by microwave oven on transmitted signal. Therefore, the noise pulse model can be represented diagrammatically as follows.

The electromagnetic noise A(t) generated by the microwave oven can be modeled and represented mathematically as shown in (1):

$$A(t) = I_0 U[\vartheta(t)] * e^{\{2\pi j(f_0 t + f_{\max} \int_{-\infty}^t \vartheta(q) dq)\}} \quad (1)$$

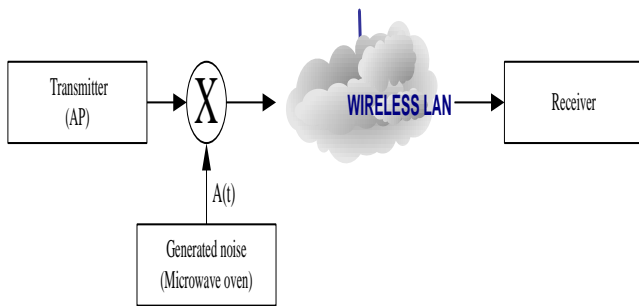


Fig. 1. Model of interference between microwave oven noise and wireless LAN.

Where,  $f_0$  is the carrier frequency and  $f_{\max}$  represents the maximum frequency of deviation.

The amplitude of the pulse radiated is determined by  $I_0$  and its phase is assumed to be distributed uniformly in order to simplify the problem. As seen from (2), the value of U is a function of  $\vartheta$  and it depends on the threshold voltage  $\vartheta_0$  value.

$$U(\vartheta) = \begin{cases} \vartheta, & \text{for } \vartheta \geq \vartheta_0 \\ 0, & \text{for } \vartheta < \vartheta_0 \end{cases} \quad (2)$$

$$\vartheta(t) = \begin{cases} \cos(2\pi f_v t), & \text{for transformer type} \\ \cos(2\pi f_v t) \cos(2\pi f_s t), & \text{for inverter type} \end{cases} \quad (3)$$

Where,  $f_v$  and  $f_c$  represents the A.C. supply and inverter switching frequency respectively as it has been described in (3).

$$f(t) = f_0 + f_m(\vartheta(t)), \quad \vartheta(t) \geq \vartheta_0 \quad (4)$$

The noise generated by microwave oven can be categories as band-limited noise and it can be represented mathematically as shown in (5) below:

$$N(t) = \int_{-\infty}^{\infty} h(\tau) n(1 - \tau) d\tau \quad (5)$$

Where,  $h(\tau)$  represents the complex impulse response for the filter.

## III. MATERIALS AND METHODS

### A. EM Microwave Radiation Source

The experimental work has been conducted using PANASONIC® (NN-S215MF) microwave oven with the specifications described in Table I. The microwave oven output power is around 800 watts and it operates at the frequency of 2.45 GHz. It is used in the experiment as a source for generating the electromagnetic noise which interferes with other devices operating in the ISM frequency band. Basically, the magnetron in the microwave oven generates RF energy when the voltage exceeded the threshold value. The driving voltage is produced by AC supply source is fed to step-up transformer or inverter depending on the microwave oven type. Consequently, it generates RF pulse within the 2.4 GHz band at the frequency of the supply main or switching frequency of the inverter. The description about the microwave oven used for the experimental work is shown in Table I.

TABLE I. MICROWAVE OVEN DESCRIPTION

Microwave Oven Type	Description	
PANASONIC NN-S215MF	Microwave oven type	Transformer / Inverter
	Input power	5.4A, 240 volts AC supply
	Output	800w
	Frequency	2.45 GHz
	AC mains frequency	50 Hz

The impact of electromagnetic noise generated at the frequency of 2.45 GHz was constantly monitored and analyzed. A spectrum analyzer and wire shark were used to keep track of variation in signal strength as the distance between the access point (AP) and microwave oven is adjusted. The AP has been maintained at the same position throughout the experiment.

The AP is used in experimentation to serve as the medium through which devices can connect to the network. External radiation from the microwave oven was used to study the variation in signal transmitted over the medium. When the microwave oven is close to the AP, the signal strength reduces due to the interference signal generated by the oven. At every distance, the signal strength is measured using the spectrum analyzer, and PC equipped with wireshark and Hand Held Software Tool (HHS). The signal strength of the channel 6 to 11 within the ISM band has been evaluated to determine the effect of the radiation source. It is very obvious that channel 11 will be affected more when compared to the other channels. This is because of the fact that the microwave frequency of operation falls within its frequency range.

Table II shows the description of the AP used to conduct the experimentation. The AP operates within the frequency range of 2.4GHz to 2.487GHz which eventually covers the frequency of operation for the microwave oven. The electromagnetic interference (EMI) and susceptibility of the AP are according to FCC Part 15.107 and 15.109 Class B.

The spectrum analyzer has been set to monitor the frequency range of 100 KHz to 3GHz and can be able measure the power between the ranges -110dBm to 20dBm. More detail about the spectrum analyzer has been presented in Table III.

TABLE II. ACCESS POINT DESCRIPTION

Access Point Type	Description	
<b>CISCO AERONET 1200</b>	Antenna	2.2 dBi
	Input power	4.75w
	Typical range	121.9 m
	Frequency	2.400 to 2.497 GHz
	Temperature	-4 to 131°F (-20 to 55°C)
	EMI and Susceptibility	FCC Part 15.107 and 15.109 Class B

TABLE III. SPECTRUM ANALYZER DESCRIPTION

Analyzer	Description	
<b>Spectrum Master (MS2711D)</b>	Frequency range	100 kHz to 3.0 GHz
	Frequency Accuracy	± 10 Hz, 99% confidence level
	Measurement Range	+20 dBm to -110 dBm
	Channel Power	±1 dB typical (±1.5 dB max)
	Temperature	-10 to +50°C
	Adjacent channel power accuracy	±0.75 dBc

**B. Measurement Setup**

In order to experimentally evaluate and analyse the impact of microwave oven electromagnetic radiation on the transmission medium, there is need to determine the characteristic of microwave oven interference and ultimately determine its detrimental impact on spectrum utilization and wireless network performance. The experiment was conducted in an indoor environment which is typical to normal home setup. A single microwave oven was used in experimentation. Only the personal computer is connected to the network and all other devices have been disconnected from the network in order to effectively monitor the network condition. Fig. 3 shows the setup used for the experiment. The equipment used for the experiment includes spectrum analyzer, Cisco Aeronet 1200 AP, microwave oven and personal computer equipped with wireshark software. The wireshark software plays an integral role in networks by capturing and analyzing packets within an environment it has been set to monitor. More importantly, it can be used in monitoring and troubleshooting networks from bits up to the packets level. This tool provided all the necessary information about the network and it has been used to critically examine the network performance under different network scenarios.

The spectrum master has been designed mainly to monitor, measure and analyze signals. It has the capability to measure in-band interference, transmit spectrum analysis, antenna isolation and cell area interference. More importantly, the spectrum analyzer has been used to determine amount of radio energy within the ISM band. In order to track the impact of the microwave oven radiation, channel 6-11 of the IEEE802.11b was monitored using spectrum analyzer and readings were taken at different distances. This is mainly due to the fact the MWO operates within the frequency range of 2.437 GHz to 2.462 GHz as shown in Fig. 2. The bold line in between 2.437 GHz and 2.462 GHz frequencies shows the frequency at which the microwave operates.

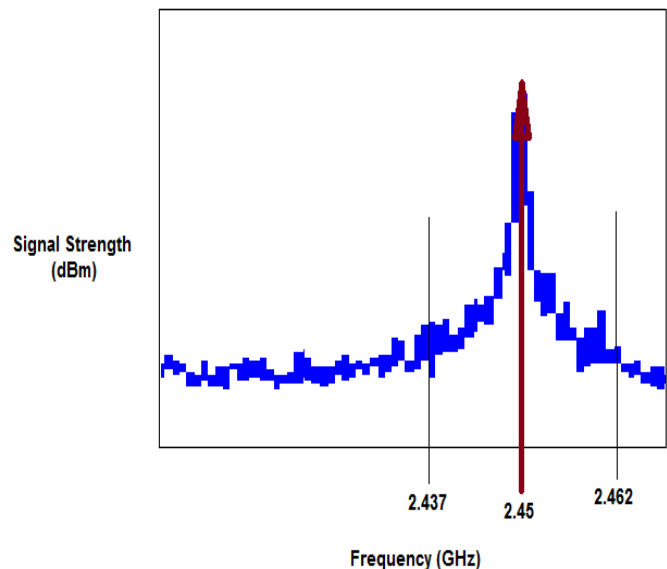


Fig. 2. Microwave oven operating frequency and study range.

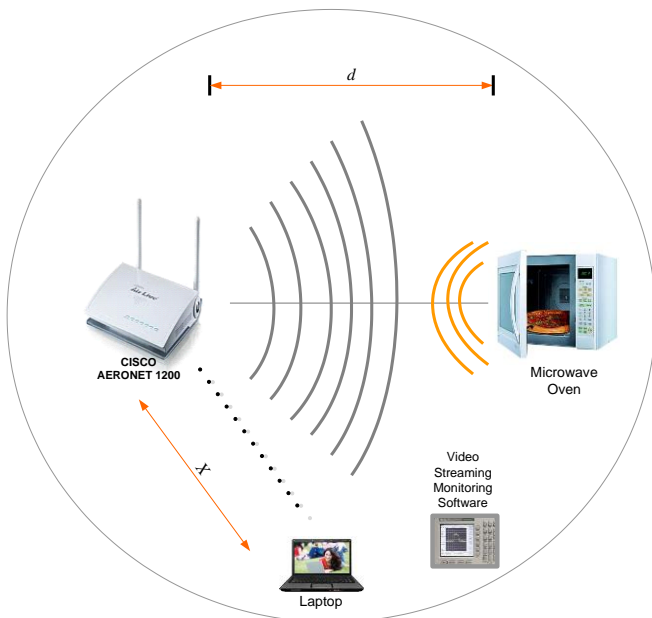


Fig. 3. Experimental setup.

As it has been mentioned earlier, microwave oven is a major source of electromagnetic radiation noise in ISM band. Their fundamental principle of operation is based on generation of microwave energy through the use of magnetron tubes. Interestingly, the microwave oven operates at the center frequency of the ISM band (i.e. 2.45 GHz). Fig. 3 illustrated how the experimentation was conducted in order to observe how the microwave operation affects the spectrum under different scenarios. The proximity of the devices was a key fundamental issue to monitor in the experimentation.

The available channels were constantly monitored as the distance  $d$  and  $x$  are varied in order to detect the variation in signal strength and throughput as well. The fact that the position of the AP is fixed throughout the experiment while the microwave oven and spectrum analyzer position are changed intermittently based on the experimental scenario needed. The distance between the AP and microwave oven is varied from 2m to 10m at an interval of 2m. This is mainly to verify the detrimental impact of the microwave oven radiation on the wireless channel which eventually reduces the spectrum utilization, and consequently it leads to an increased in bit error rate and packet loss within the frequency band.

#### IV. MEASUREMENT RESULTS AND ANALYSIS

In the previous section, the procedure on how the experiment was conducted has been discussed. The analysis of the results obtained from the measurements will be covered in this section. The experimental data was collected using spectrum analyser and the wire shark program. Fig. 4 shows the experimental arrangement used for the experimentation. The wireshark software is used to thoroughly check the wireless channels under scrutiny based on the channel signal strength and interference. The network information captured by the spectrum analyzer was downloaded, and processed by the HHS tool as shown in Fig. 4.



Fig. 4. Measurement equipment and Tools.

Initially, the experiment was conducted in an environment where there is no external interference from any source of interference. Measurements were taken at different distance (2, 4, 6, 8, 10m) between the AP and PC, and the procedure is repeated 5 times in order to determine the average of the readings taken. Again, the same procedure was repeated when the microwave oven has been switched ON. Subsequently, comparison was made between when there is no interference and the scenario when there is interference from external source (MWO). In both cases, the signal strength, throughput, noise traffic and utilization of the network has been recorded.

Based on the experimental setup in Fig. 3, the distance  $d$  between microwave oven and AP was set to a fixed value. The spectrum analyzer monitors the frequency band as the distance between the AP and the microwave oven is adjusted. The signal strength decreases as the proximity between the microwave oven and AP decreases. The network throughput has been used as a metric to measure the performance of the network when subjected to external interference. It is primarily achieved by using the wire shark software to critically analyze the network performance in terms of parameters such as network traffic under different experimental scenarios. Interestingly, it has enabled us to analyze packets individually as the traffic flows through the network.

##### A. Bandwidth Utilization

The bandwidth utilization has been considered as a metric to measure the performance of the network when subjected to external interference. In order to effectively utilize the network resources, the interference from external sources should be mitigated. The spectrum utilization is closely monitored using the wire shark software since it has been installed in the PC which was wirelessly connected to the AP. The bandwidth utilization has been observed when the microwave oven is switched ON and OFF as well. The PC in which the wire shark software is installed has been placed at different distance (2, 4, 6, 8 & 10m) from the WiFi transmitter (AP). The entire spectrum is monitored in order to analyze the performance of the network based on the aforementioned parameter.

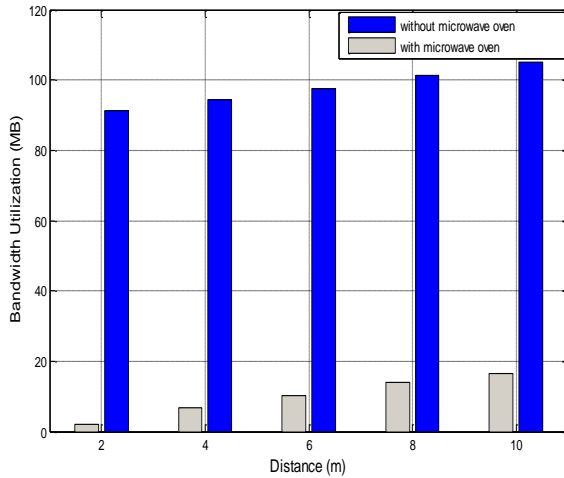


Fig. 5. Bandwidth utilization.

As can be noticed from Fig. 5 that the total traffic level is low when the microwave oven is ON compared to a situation when it is OFF. This is mainly due to the fact that the connectivity between the PC and AP is poor when the radiation source is at close proximity. As the monitoring device moves far away from the external interference source, the wireless spectrum could be assessed better and it leads to an increment in the network traffic and vice versa. In a nutshell, the bandwidth utilization is very much high when the PC has been placed far away from the microwave oven. It is obvious that the bandwidth utilization is less when the AP is within close proximity to the microwave. It is primarily due to the fact the external radiation from the microwave oven will destabilize the spectrum and eventually leads to low bandwidth utilization.

**B. Signal Strength**

Fig. 6 illustrated the performance of the WiFi network in terms of received signal strength. The received signal strength has been used as a metric to measure the network performance when the microwave oven is ON and OFF as well. Due to detrimental impact of the external radiation from the microwave which reduces the strength of the signal radiated from the AP as it travels along the medium. This adverse effect can be noticed from Fig. 6 which clearly shows that the received signal strength has reduced tremendously with the introduction of microwave oven onto the network. This is primarily due to the fact the MWO operates within the same frequency with the other device connected to the WiFi network. When the MWO is switched ON, the received signal strength reduces with increase in distance from the AP.

The ability to receive signal with high strength depend on many factors, but it is very obvious that the noise generated in the medium which the signal traverse play significant on the received signal strength. Interference from the external sources will cause detrimental impact to the signal quality and channel as well. It can be seen from Fig. 6 that the signal strength decreases with increase in distance. The signal strength varies from -87dBm to -96dBm as the distance was adjusted from 2m to 10m in a noisy microwave oven

environment. Also, it can be noticed that the signal strength varies from -18dBm to -38dBm for the same distance, but the microwave oven has been switched off completely. Through the experimentation, it can be noticed that the rate at which the signal strength decreases is much higher in a noisy microwave oven environment when compared to an environment which is not subjected to external interference. In a nutshell, the microwave oven operation within an environment which utilizes wireless connection to communicate and transmit data has detrimental impact on the received signal strength. Non WiFi devices hinder the performance of the WiFi network and ultimately cause incessant packets loss and interference in the network. Developing new techniques on how to handle the problem above will assist greatly in ensuring effective and reliable communication over wireless networks.

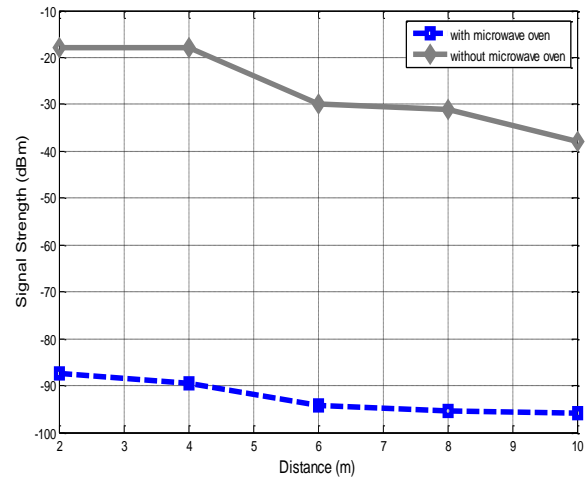


Fig. 6. Signal strength at different proximities.

**C. Network Throughput**

The network throughput has been examined in order to evaluate the impact of microwave oven operation on WiFi network throughput. The experimental approach used to collect the result is based on the scenario shown in Fig. 3. The impact of microwave oven on the throughput in terms of bits per second (bps) is measured to determine the network the performance under different scenarios. Table IV shows the average value of the throughput captured at different distance when the microwave oven is activated ( $T_{ON}$ ) and deactivated ( $T_{OFF}$ ) as well. It very clear that the throughput reduces as the PC is closed to the microwave oven due its negative influence on the number of bits transmitted per second.

TABLE IV. ACHIEVABLE NETWORK THROUGHPUT WITH AND WITHOUT MICROWAVE OVEN

Distance (m)	Throughput (Kbps)		$\zeta$ (%)
	$T_{ON}$	$T_{OFF}$	
2	395.7	566.5	30
4	562.8	695.1	18
6	615.5	737.3	16.5
8	728.05	819.6	11.2
10	1000.1	1122.1	10.8

The percentage deviation ( $\zeta$ ) in throughput when the MWO is ON and OFF can be computed using (6) as shown below:

$$\zeta = \left( \frac{T_{OFF} - T_{ON}}{T_{OFF}} \right) \times 100 \quad (6)$$

$\zeta$  is the percentage change in throughput when the microwave is active and deactivated. It shows the percentage at which the overall network throughput drops by juxtaposing  $T_{ON}$  and  $T_{OFF}$  at different distance. In order to determine by how much the throughput reduce as a result of microwave oven presence within the wireless environment,  $\zeta$  is employed to quantify the rate at which the throughput as the distance increases. (6) has been used in computing the value of  $\zeta$  under different scenario. The key fundamental issue is to compare the throughput when the microwave oven is activated and deactivated at the same position starting from 2m to 10m. For instance, at 2m distance, it is clear that the microwave oven operation greatly affects the throughput of the network. Even though the PC has close proximity to the AP but due to intense radiation from the external source, it immensely reduces the achievable throughput. Indeed, it can be notice that the achievable throughput increases with increase in distance away from the activated microwave oven and reverse is the case when the microwave oven is inactive. The detrimental impact caused as a result of noisy microwave oven environment and distance on throughput can be observed from the graph shown in Fig. 7.

As it has been mentioned already that the need to study the impact of this phenomenon on network performance will lead way toward developing scheme to effectively deals with the adverse impact of the devices on parameters such as throughput, packet loss and bandwidth utilization. This experimental work has examined the impact of radiation on wireless environment, and later much effort and resources would be dedicated to mitigation strategies to reduce the adverse effect of these home devices on the performance of WiFi. Using cognitive radio scheme, it will intelligently detect and determine the best frequency band to operate. This will reduce the incessant packet loss and increase the throughput. More importantly, the network performance will be enhanced.

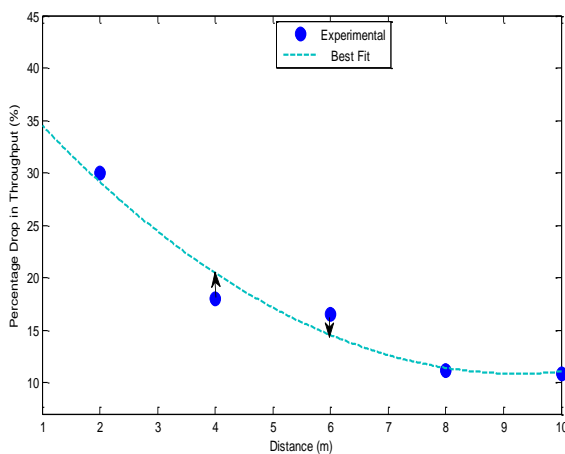


Fig. 7. Percentage in throughput drop with increase in distance.

In this work, only one source of radiation and user has been considered, but there is need to consider multi-users environment which mimic typical real world scenario. There is need to consider different non WiFi devices operating within the frequency and evaluate their detrimental impact on the network performance. More importantly, it would help in designing schemes to address the external radiation and interference problem in wireless network. In our future works, much emphasis will be place on how to significantly improve the network capability by detecting the radiation from non WiFi devices and switching to other frequency band can improves the network performance and ensures more seamless network connectivity. The work will be based on cognitive radio approach in conjunction with other optimization algorithms to enhanced searching ability.

## V. DISCUSSION

Having examined the problem of external interference within the WiFi networks, the need for highly sophisticated scheme to intelligently mitigate the problem is very crucial and important. Understanding the nature of the interference could dramatically improve the performance of the network since can led to the development effective and efficient solutions to mitigate the interference from the external sources.

It is very obvious that there is dramatic need for efficient utilization of spectrum in order to meet with increasing demand for bandwidth and to efficiently mitigate the bottlenecks which reduces the performance of the network. It is challenging to efficiently mitigate the interference form external sources but there is need to dynamically adapt with environmental changes in a noisy microwave environment in order to achieve efficient spectrum utilization and management. Cognitive radio is a promising technology to ensure effective spectral utilization and management in wireless system. The cognitive radio transmission should be properly controlled such that the disturbance of the primary network is within acceptable limit. More research effort is needed for distributed power control in multi-user cognitive radio network to ensure that the power radiated from different sources does not significantly affect the network performance. An intelligent network management scheme will ensure efficient spectrum utilization, coverage and fairness amongst multiple secondary users. Most of the interference aware approaches focused on wireless devices interference [15]-[18], but more emphasis need to be place on non WiFi devices which eventually affect the network performance and it subsequently effects the applications and services as well.

## VI. CONCLUSION

In this paper, an experimental investigation of effects microwave oven noise has been presented in order to evaluate the electromagnetic radiation impact on WiFi spectrum band. An experimental investigation is conducted to analyze and quantify the performance degradation in link quality and spectrum utilization of wireless channels in a noisy microwave oven environment. This is extremely important as the usage of microwave oven and wireless technology become prevalence at home and offices. The electromagnetic radiation noise impact is verified experimentally based on the signal strength

and the proximity of the microwave oven to the access point. The proximity of the microwave oven to the device plays an integral role because it determines the amount of distortion to be anticipated by the wireless networks. Our future work will primarily focus in developing an intelligent scheme to cooperatively manage the WiFi spectrum with low packet loss and high throughput while at the same time ensuring co-existence among devices connected to the network.

#### ACKNOWLEDGMENT

The author would like to thank all those who contributed toward making this research successful. Also, I would like to thanks to all the reviewers for their insightful and valuable comment. This work was supported by the Deanship of Scientific Research (DSR), King Abdulaziz University, Saudi Arabia, under grant No. 830-129-D1437. Therefore, the author gratefully acknowledges the DSR technical and financial support.

#### REFERENCES

- [1] T. Murakami, Y. Matsumoto, K. Fujii, and Y. Yamanaka, "Effects of multi-path propagation on microwave oven interference in wireless systems", IEEE International Symposium on Electromagnetic Compatibility, Vol. 2, pp 749-752, 2003
- [2] T.W. Rondeau, M.F. D'Souza, and D.G. Sweeney, "Residential microwave oven interference on Bluetooth data performance", IEEE Transactions on Consumer Electronics, Vol. 50, pp 856-863, 2004.
- [3] I. F. Akyildiz, W. Lee, M. C. Vuran, and S. Mohanty. "Next generation/dynamic spectrum access/ cognitive radio wireless networks: A survey." Computer Networks, vol 50, pp. 2127-2159, 2006.
- [4] S. Srikanteswara and C. Maciocco, "Interference mitigation using spectrum sensing," in Proc. Int. Conf. Comp. Comm. & Net., pp. 39-44, 2007.
- [5] S. Miyamoto, S. Harada, and N. Morinaga, "Performance of 2.4 GHz band wireless LAN system using orthogonal frequency division multiplexing scheme under microwave oven noise environment," in Proc. IEEE Int. Symp. on Electromagn. Compat., vol. 1, pp. 157-162, 2005.
- [6] A. Baid, S. Mathur, I. Seskar, T. Singh, S. Jain, D. Raychaudhuri, S. Paul, and A. Das, "Spectrum MRI: Towards diagnosis of multi-radio interference in the unlicensed band," IEEE WCNC, 2011.
- [7] Y. Matsumoto, M. Takeuchi, K. Fujii, A. Sugiura, and Y. Yamanaka, "Performance analysis of interference problems involving DS-SS WLAN systems and microwave ovens" IEEE Transactions on Electromagnetic Compatibility, Vol. 47, pp. 45-53, 2005
- [8] P.S. Neelakanta, and J. Sivaraks, "A Novel Method to Mitigate Microwave Oven Dictaded EMI on Bluetooth Communications," Microwave Journal, pp. 70-88, 2001
- [9] C.R. Buffler, and P.O. Risman, "Compatibility Issues between Bluetooth and High Power Systems in the ISM Band," Microwave Journal, pp. 126-131, 2000
- [10] IEEE, "IEEE Std.802. 11b-1999 Wireless LAN medium access control and physical layer specifications: Higher-speed physical layer extension in the 2.4 GHz band," IEEE Std. 802.11b, 1999.
- [11] S. Yin, Z. Qu, and S. Li, "Achievable Throughput Optimization in Energy Harvesting Cognitive Radio Systems" IEEE Journal on Selected Areas in Communications, vol 33, 2015.
- [12] C. S. Karthikeyan, and M. Suganthi, "Optimized Spectrum Sensing Algorithm for Cognitive Radio," Wireless Personal Communications, Vol 94, 2017.
- [13] S. Shantanu, and S. A. Kumar, "On detecting termination in cognitive radio networks." International Journal of Network Management, vol 24, 2014.
- [14] K. A. Ali, R. M. Husain, and R. Martin, "Cognitive Radio for Smart Grids: Survey of Architectures, Spectrum Sensing Mechanisms, and Networking Protocols" IEEE Communications Surveys & Tutorials, vol 18, 2016.
- [15] Y. Chungang, L. Jiandong, N. Qiang, A. Alagan, and G. Mohsen, "Interference-Aware Energy Efficiency Maximization in 5G Ultra-Dense Networks" IEEE Transactions on Communications, vol 65, 2017.
- [16] D. V. Son, N. V. Dinh, and S. O. Soon, "Interference-Aware Transmission for D2D Communications in a Cellular Network" Wireless Personal Communications, vol 98, 2018.
- [17] G. Verma and O. P. Sahu, "Interference Aware Sensing Scheme in Cognitive Radio System" Wireless Personal Communications, vol 94, 2017.
- [18] K. Hyunsoon, and K.Hwangnam, "Designing interference-aware network selection protocol for WLAN mobile devices" Transactions on Emerging Telecommunications Technologies, vol 28, 2017.

# Deep Learning Technology for Predicting Solar Flares from (Geostationary Operational Environmental Satellite) Data

Tarek A M Hamad Nagem, Rami Qahwaji, Stan Ipson  
School of Electrical Engineering and Computer Science  
University of Bradford  
Bradford, United Kingdom

Zhiguang Wang  
GE Global Research  
San Ramon, CA, United States of America

Alaa S. Al-Waisy  
School of Electrical Engineering and Computer Science  
University of Bradford  
Bradford, United Kingdom

**Abstract**—Solar activity, particularly solar flares can have significant detrimental effects on both space-borne and ground based systems and industries leading to subsequent impacts on our lives. As a consequence, there is much current interest in creating systems which can make accurate solar flare predictions. This paper aims to develop a novel framework to predict solar flares by making use of the Geostationary Operational Environmental Satellite (GOES) X-ray flux 1-minute time series data. This data is fed to three integrated neural networks to deliver these predictions. The first neural network (NN) is used to convert GOES X-ray flux 1-minute data to Markov Transition Field (MTF) images. The second neural network uses an unsupervised feature learning algorithm to learn the MTF image features. The third neural network uses both the learned features and the MTF images, which are then processed using a Deep Convolutional Neural Network to generate the flares predictions. To the best of our knowledge, this work is the first flare prediction system that is based entirely on the analysis of pre-flare GOES X-ray flux data. The results are evaluated using several performance measurement criteria that are presented in this paper.

**Keywords**—Convolutional; neural; network; deep; learning; solar; flare; prediction; space; weather insert

## I. INTRODUCTION

The concept of space weather has been defined by the US National Space Weather Program as “Conditions on the Sun and in the solar wind, magnetosphere, ionosphere and thermosphere that can influence the performance and reliability of space-borne and ground-based technological systems and can endanger human life or health” [1]. There are several influences, originating from space weather phenomena that detrimentally affect important industries relying on avionics, satellites, mobile communication networks, and electricity distribution [2]. All these industries touch our daily lives and this means that space weather can impact our lives dramatically.

Painstaking efforts are currently being made in a number of international centres to create accurate solar flare prediction systems. This is because many infrastructures could be affected by significant flares and the cost of building an accurate solar

flare prediction system would be much cheaper than the cost of repairing damage caused by such a flare. In this work, the proposed prediction system generates two probabilities for Event and No-event. Event predictions cover significant X and M class flares that might be harmful, while No-event predictions cover no-flares and the non-harmful A, B and C class flares.

Although scientific progress has increased enormously the rate of generation of data monitoring solar activity, scientists are not yet able to fully understand all the detailed causes of solar flares. Consequently, efforts are being made to develop methods to predict solar storms, making direct use of the data using advances in data analysis.

Since 1987, there have been many approaches that attempted to predict solar flares. The first solar flare prediction system (called THEOPHRASTUS) was launched by the Space Environment Services Centre at NOAA, and it predicts X-ray flares with a time window of 24 hours [3]. More recently, three solar flare prediction systems, ASSA (Automatic Solar Synoptic Analyser), MAG4 (Magnetic Forecast system) [7] and ASAP (Automated Solar Activity Prediction), have become a part of the NASA Integrated Space Weather Analysis (ISWA) system [5] and these three systems are briefly described below.

The first system, ASSA, is based on an artificial neural network technique and the ASSA coronal hole data archive, from the period 1997 till 2013, including SDO solar images, to predict solar flares, solar radiation storms and geomagnetic storms. ASSA predicts C, M and X flares. ASSA predictions are based on statistical analysis of the ASSA sunspot catalogue [6]. The second system, MAG4 was developed at the University of Alabama in Huntsville, to assist NASA Space Radiation Analysis Group (SRAG) at the Johnson Space Flight Centre. MAG4 is using Magnetogram data for the Sun. MAG4 forecasts X and M class flares, CMEs, and Solar Proton Events (SPE) using McIntosh active-region (AR) classes as the basis of their forecasts [7]. The University of Bradford developed a forecasting model, the Automated Solar Activity Prediction (ASAP) system in 2009. ASAP uses McIntosh classes and other sunspots features which it generates from the solar data.

ASAP uses SDO/HMI Continuum and Magnetogram images as an input to the system, also it uses two neural networks to predict solar flares [3].

Recently, the new field of deep learning neural network research has achieved remarkable successes compared with previous artificial intelligence methods [5]. These include complex tasks like medical diagnoses, dealing with huge amounts of data, pattern recognition and numerous others, such as the virtualization frameworks for big data reported in [8]. Using the deep learning technology for space weather prediction is still a novel area of research, which needs to be investigated to help analyse the huge amount of solar activity data that are publically available.

UFCORIN (Universal Forecast Constructor by Optimized Regression of Inputs) is open-source software available online which has been used to predict general time series and solar flares. This system uses HMI image data and GOSE X-ray data as input to predict X, M, and C solar flare class. In 2016, UFCORIN was extended to use deep learning, and provides 24-hour-ahead predictions of solar flares, every 12 minutes by using a deep learning approach.

In this paper, we introduce a solar flare prediction system, summarised in the following subsection, working solely with GOES X-ray flux data that integrates three neural networks to deliver these predictions and provides an automated prediction of solar flares by utilising deep learning techniques.

GOES data are available in real-time (available every minute) and they provide a general indication of flaring across the solar disk. These data come in soft and hard x-ray and are available from 2002. However, GOES data provide an indication of flaring without much info about the exact location of flaring on the solar disk. This could be one of the reasons why it is not used heavily for space weather prediction. The format of GOES Data is also challenging as it is represented as a time-series signal, which makes it challenging for machine-learning based prediction (Deep learning in particular).

### A. Overview of the System

Fig. 1 shows the system model which consists of three units. Starting from the input (GOES X-ray flux time series data) to the output (solar flare prediction) and including the evaluation of the predictive performance.

Unit 1 in Fig. 1 converts a sequence of GOES X-ray flux 1-minute data time series data to a  $64 \times 64$  MTF image in two stages. Firstly, it converts the original text data to a Markov Transition Matrix. Then it encodes the Markov Transition Matrix as a  $64 \times 64$  Markov Transition Field (MTF) image as illustrated in Fig. 6. Unit 2 in Fig. 1 learns the features within the MTF images. Unit 2 pre-processes and normalizes the images and then divides the  $64 \times 64$  images into  $64 \times 8 \times 8$  patches. These patches are encoded using a Back-propagation Auto-encoder to obtain learned feature mappings as indicated in Fig. 1. Unit 3 in Fig. 1 provides predictions for solar flares using a CNN. This unit starts by utilising the historical knowledge and linking the MFT images with the Flare or No-Flare labels. Subsequently, datasets are created for training and testing the neural networks. After training on the associated dataset is carried out, the trained CNN is run on the test dataset to generate prediction results, which are evaluated using space weather verification metrics.

The rest of this paper is organized as follows. Section 2 describes the operation of Unit 1 which converts GOES X-ray flux time series data to  $64 \times 64$  MTF images. Section 3 describes Unit 2, which learns features within MTF images using an unsupervised learning algorithm by applying back-propagation. Section 4 describes Unit 3, which makes solar flare predictions using a Deep Convolutional Neural Network. Section 5 discusses the evaluation and performance of the whole system and Section 6 presents concluding remarks and suggestions for future work.

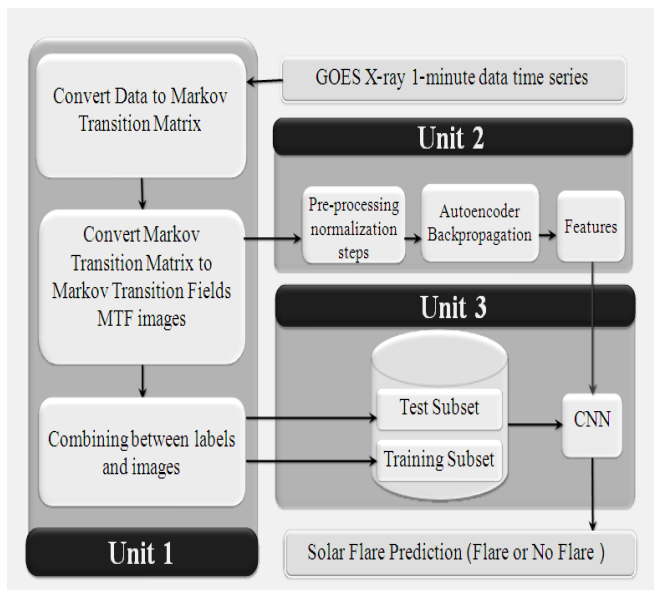


Fig. 1. The diagram showing the internal procedures of the system.

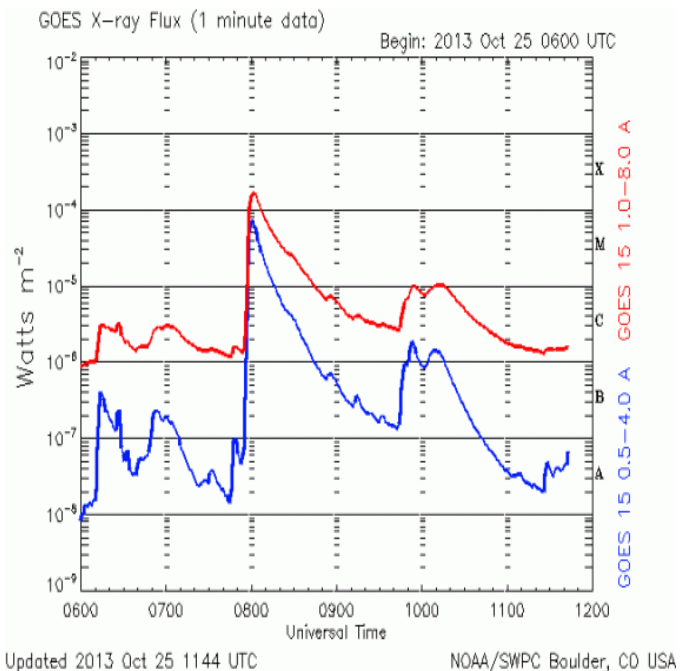


Fig. 2. A sample 6 hour plot of GOES X-ray flux 1-minute data.



## II. PREPARATION OF THE DATA

### A. The Source X-Ray Data

In this work, 1-minute X-ray flux data from the American Geostationary Operational Environmental Satellites (GOES) are used. The data used are provided from four GOES satellites, GOES-10, GOES-11, GOES-14, and GOES-15. All the data produced are archived and available, and it can be found online at [9]. Two X-ray channels are available as shown in Fig. 2; a harder X-ray channel (0.05-0.4 nm), and a softer X-ray channel (0.1-0.8 nm) [10]. For this work, the soft channel is used because provides information about the intensity of solar flares and is used in this work to investigate its suitability for investigating the temporal evolution of flares [10].

### B. Extraction of Relevant X-Ray Flux Data

The temporal evolution of solar flares generally occurs in three phases [4].

- Pre-flare phase: This is the region shown in Fig. 3 which consists of fluctuations and a slow increase of X-ray flux before the start of the flare event.
- Impulsive phase: Here the X-ray flux increase quickly and the main flare energy release occurs during this phase.
- Gradual phase: In this phase, the X-ray flux gradually decreases to the background level.

Fig. 4 shows the cropped AIA images of a flaring region corresponding to the GOES X-ray data regions in Fig. 3. The left image in Fig. 4, captured in the pre-flare phase, shows two sets of nested loops. The middle image in Fig. 4, captured during the main phase, shows inner loops becoming significantly brighter. In the right-hand image, the flare launches a CME. There are many relationships which have been recognized between the pre-flare activities and flaring, and these appear as loop brightening activities [15]. However, the method introduced here bases its prediction solely on changes in the overall X-ray flux during the pre-flare phase.

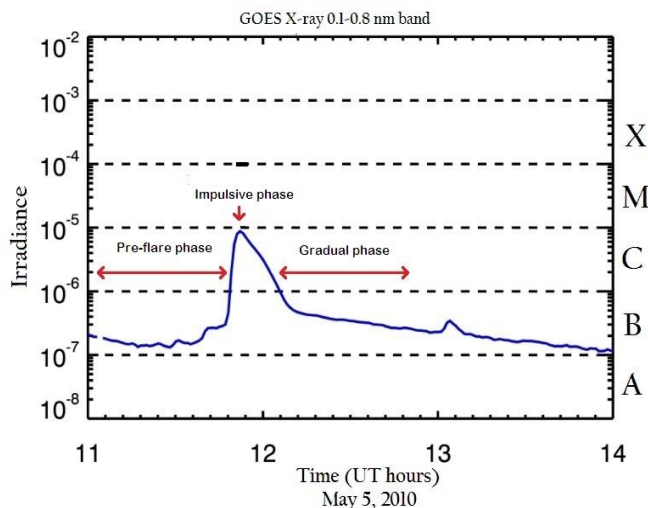


Fig. 3. The solar flare phases on C8.8 flare that occurred on 5<sup>th</sup> May 2010 – From NASA [11].

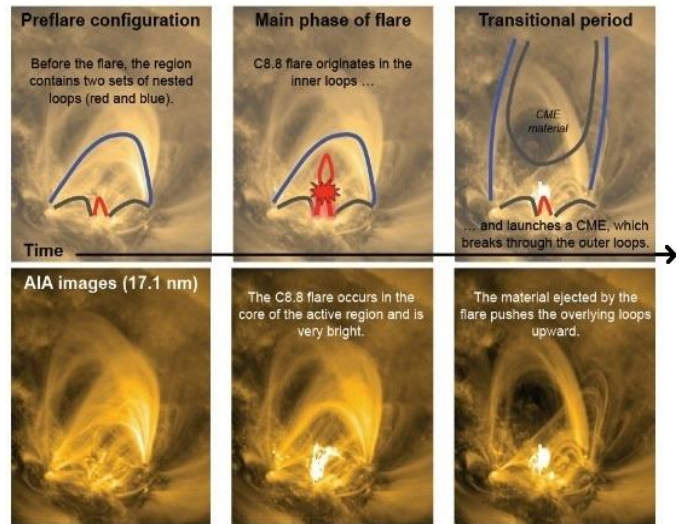


Fig. 4. Cropped AIA images showing three phases of the solar flare which contributes to the GOES data shown in Fig. 3– From NASA [11].

### C. Prediction Optimization for Different Time Windows

The Time windows of 20, 30, 60 and 120 minutes between the end of a data sample and the start of a flare/no-flare are investigated, using the Quadratic score QR, to determine the time window with the best prediction performance. QR is widely used as a verification measure to evaluate the accuracy of prediction. The prediction accuracy is calculated by finding the mean square error between the predictions and the observations as given by [2].

$$QR = \frac{1}{N} \sum_{t=1}^N (f_t - o_t)^2 \quad (1)$$

where  $o_t$  are the binary observation outcomes where 1 means that flare occurred and 0 means that a flare did not happen,  $N$  is the sample size, and  $f_t$  is the prediction probability. QR ranges from 0 (perfect prediction) to 1 (worst possible prediction) [18].

The result for each time window is shown in Table I. It is clearly seen that the best QR is when the time window equals 20 minutes. To find the prediction window duration that would provide the best QR value, we followed the method presented in [2] and applied QR to determine the best prediction window duration.

TABLE I. THE QUADRATIC SCORE (QR) RESULTS FOR 20, 30, 60 AND 120 MINUTES

Sample Size	20 minutes	30 minutes	60 minutes	120 minutes
2124	0.136	0.153	0.249	0.590

### D. Data Presentation

Fig. 5 shows a sub-system that has been created to generate datasets by selecting specific data from GOES X-ray flux 1-minute data using three steps. The first step identifies a flare. Then selects 120 minutes of data, starting 140 minutes before the beginning of the flaring event. Finally, the selected data is saved in a matrix as described in the next subsection.

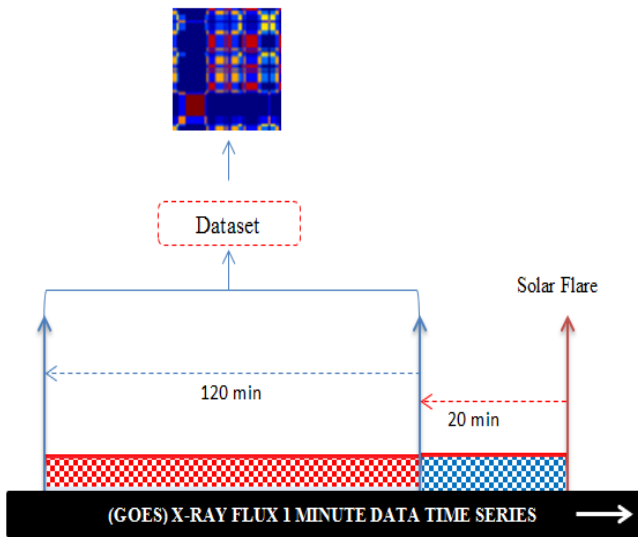


Fig. 5. Creating dataset of a time series of X-ray flux data with a 20-minute data window before the flare occurs.

E. Conversion of Time Series Data to MTF Images

Temporal and frequency correlations are major dependencies embedded in time series data. To build a comprehensive but intuitive visualization, the extracted features of the designed data transformation framework should be able to represent the dynamics in both time and frequency while there should exist a reverse operation to map the information back to the raw GOES time series. The following sub-sections describe how to encode the dynamical frequency information in the temporal ordering, illustrated in Fig. 6, step by step.

The main idea of this stage is to use GOES time series data to generate Markov transition field while maintaining the time-series properties. The method applied in this research is taken from [14]. MTF images were generated by applying the code used in [14] to GOES data.

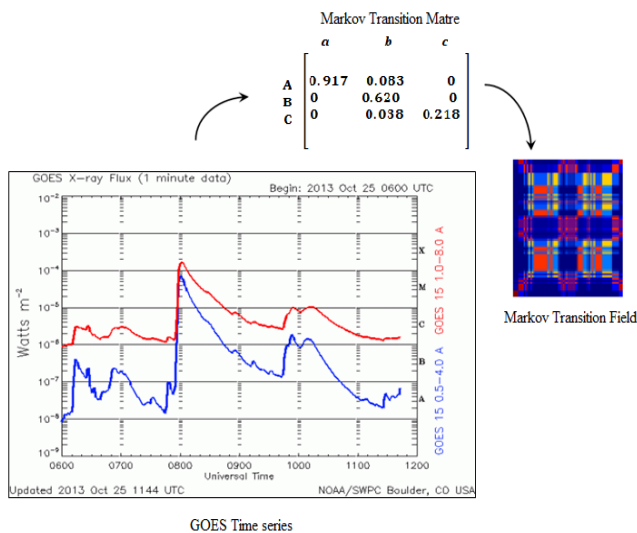


Fig. 6. Conversion of GOES X-ray data time series data to MTF images.

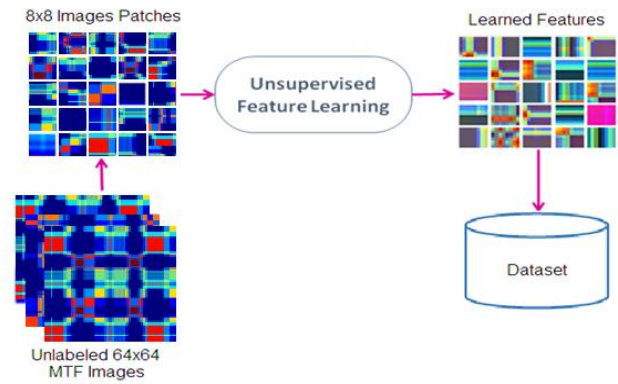


Fig. 7. Learning the features within MTF images.

III. LEARNING THE FEATURES WITHIN MTF IMAGES

The Auto-encoder is an unsupervised back-propagation neural network which tries to learn a function  $hW_b(x) \approx x$ , and is adjusted so that the input values correspond to the target  $y^{(i)} = \hat{x}^{(i)}$  [12]. In this work, we assume  $x$  is the input corresponding to the pixel intensity values for an  $8 \times 8$  MTF image patch with 64 pixels so  $x = 64$ , and there are  $s_2 = 32$  hidden units in layer  $L_2$ . The network is required to learn a compressed representation of the input, because there exist only 32 hidden units. Therefore the auto-encoder should attempt to reconstruct the input to  $8 \times 8$  images (64 pixels) [16] as illustrated in Fig. 7.

IV. PREDICTION OF SOLAR FLARES USING A DEEP CONVOLUTIONAL NEURAL NETWORK

As you can see in Fig. 8 the Convolutional Neural Network (CNN) consists of convolutional layers and sub-sampling layers followed by fully connected layers.

A. The Convolutional Layer

The input to this layer is a  $d \times d \times ch$  MTF image where  $d$  is the height and the width of the image ( $d = 64$  in this case) and  $ch$  is the number of channels. Since the MTF images are RGB images,  $ch = 3$ . As illustrated in Fig. 9 the convolutional layer uses  $K_f$  filters (also called Kernels) of size  $n \times n \times ch$  where  $n$  is the dimension of the filter and  $n = 8$  to produce feature maps. The  $K_f$  filters are convolved over the MTF image to create  $K_f$  feature maps of size  $d - n + 1$  [16].

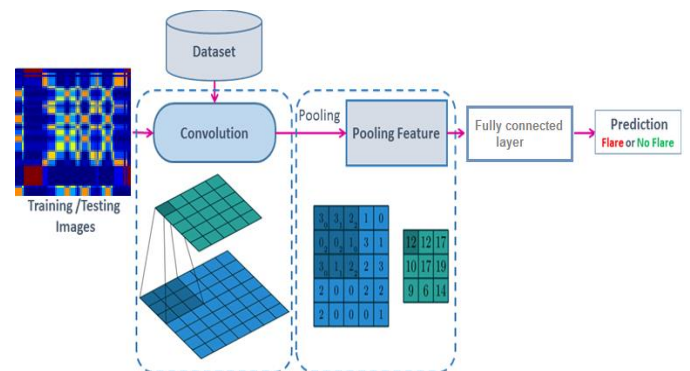


Fig. 8. Convolutional neural network designed to predict solar flares.

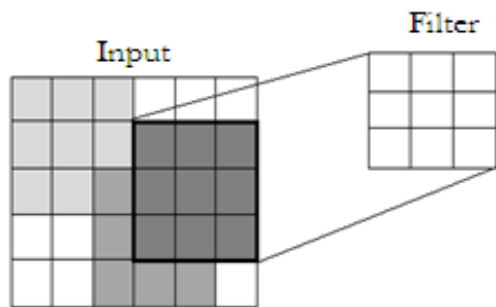


Fig. 9. Convoluting filter over an input image in convolutional layer.

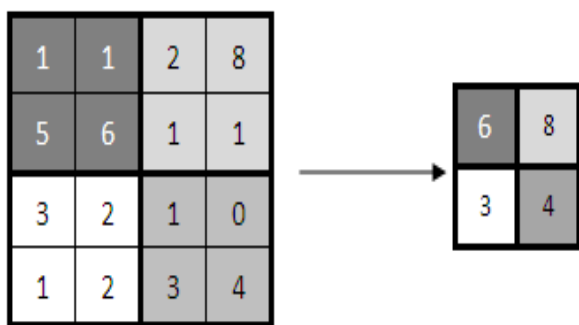


Fig. 10. An example of Max pooling.

### B. The Pooling Layer

After the generation of the feature maps by the convolutional layer, the features are then used for classification. Fig. 10 shows each feature map is down-sampled by max-pooling to size  $p \times p$ . Typically,  $p$  ranges from 2 to 5, for small to big images respectively, and in this work  $p=4$  [16].

### C. The Fully Connected Layer

This layer takes the outputs from the previous layers which were reduced to a one-dimensional feature vector. This layer is fully connected and there is just one output for each class label. The high-level inference in the CNN is performed by this fully connected layer.

## V. IMPLEMENTATION AND EVALUATION OF THE SYSTEM

Three neural networks are integrated into the system to predict solar flares. Fig. 4 shows the integrated system starting from the input (GOSE data) to the output of the system (Flare/No-Flare prediction).

The first part of the system, which encodes the GOES data to MTF images, is implemented in Python and the rest of the system is implemented in Matlab [17]. The system makes flares predictions based on embedded learning rules. The system was trained using training sets covering data from 3rd Dec 2002 till 30th Jan 2017, to ensure this covered a range of activity including both solar Maximum and solar Minimum of the solar cycle.

### A. System Evaluation

The performance evaluation was done by comparing the generated predictions with the actual flare occurrences as reported by 1-minute GOES data. The data were taken from four satellites, GOES-10 data covering (03 Dec 2002 -22 Jun 2006) and (11 Apr 2007-30 Dec 2009); GOSE-11 data covering (23 Jun 2006-10 Apr 2007); GOSE-14 data covering (01 Nov 2009 -26 Oct 2010); and finally GOSE-15 data covering (27 Oct 2010 -30 Jan 2017). The number of flaring and No-flaring events for each satellite is detailed in Table II. All GOES X-ray data were taken from [9].

As noted earlier in this paper, the data is classified as flaring if they produced at least one M or X class flare in the following 20 min period and No-flare if they did not cause any M or X class flares during that period. To determine the flare prediction capability we carried out experiments with 1-minute GOES data covering (Dec 2002-Dec 2005, Jun 2009- Dec 2012) to train the deep learning algorithm. The data covering (Jun 2006 - Dec 2008, Jun 2013 - Jan 2017) are used to test the system as shown in Table III. Table IV details the number of flare and no-flare data that were used in these experiments. The time coverage of the training set was chosen so that the remaining testing set would contain flare activity from periods around the maximum and minimum levels of solar activity.

TABLE II. THE NUMBER OF FLARING AND NO-FLARING FOR GOES-10 DATA COVERING (03 DEC 2002 -22 JUN 2006), (11 APR 2007-30 DEC 2009); GOSE-11 DATA COVERING (23 JUN 2006-10 APR 2007); GOSE-14 DATA COVERING (01 NOV 2009 -26 OCT 2010); GOSE-15 DATA COVERING (27 OCT 2010 -30 JAN 2017) USED IN THIS EXPERIMENT

GOES-10 From 03 Dec 2002 To 22 Jun 2006 and From 11 Apr 2007 To 30 Dec 2009		GOES-11 from 23 Jun 2006 To 10 Apr 2007		GOES-14 From 01 Nov 2009 To 26 Oct 2010		GOES-15 From 27 Oct 2010 To 30 Jan 2017	
Flare events	No Flare events	Flare events	No Flare events	Flare events	No Flare events	Flare events	No Flare events
518	1592	22	265	24	213	763	2070

TABLE III. NUMBER OF FLARE AND NO-FLARE DATA COVERING (03 DEC 2002-30 JAN 2017)

03 Dec 2002-30 Jan 2017		
Flare	No-flare	Total
1327	3981	5308

TABLE IV. NUMBER OF FLARE AND NO-FLARE DATA IN TIME INDEPENDENT TRAINING AND TESTING SETS

Training set (Dec 2002-Dec 2005) (Jun 2009- Dec 2012)			Testing set ( Jun 2006- Des 2008) (Jun 2013-30 Jan 2017)		
Flare	No-Flare	Total	Flare	No-Flare	Total
793	2391	3184	534	1590	2124

**B. Machine Learning using Cross-Validation**

Cross-validation is a method that partitions the input data into subsets so that the learning algorithm can be trained on a subset and internally tested on a different subset. Cross-validation is a useful approach for analysing the prediction performance of machine learning, as it is could help avoid over-fitting. Over-fitting occurs when the learning algorithm performs very well on the training data, but not so well when provided with new data. Different forms of cross-validation exist and the repeated random sub-sampling validation is applied here. This method is based on randomly dividing the data into a number of subsets, which is repeated a number of times so that the learning algorithm is trained and tested on different data. For each repetition, one subset is used for training and the rest are used to evaluate the prediction performance by calculating a number of forecast verification metrics. These measurements are then averaged in order to provide an indication of the effectiveness of the machine learning on the training data [19].

Two separate portions of data are created: a training portion (60%) and a testing portion (40%). The MTF images and their corresponding flare/no-flare classifications from the training portion are fed into the learning algorithm for training purposes. When the training process is completed, the learning algorithm is fed with the MTF images from the testing portion. The learning algorithm attempts to predict their Flare/No-Flare classifications. These predicted outputs are compared with the testing datasets actual classifications using standard forecast verification measures to evaluate the prediction performance of the learning algorithm. Among the prediction measures, HSS is one of the best indicators of the overall performance of a prediction method since it accounts for correct chance forecasts [20]. The cross-validation process is repeated 9 times and the means of the prediction measures are calculated.

**C. Verification Results**

This system generates a prediction in binary form so 0 means no flare and 1 means a flare. In practice, flares occur rarely compared to no-flares events. Various measures are used to evaluate the predictions of the system. These measures are for categorical prediction (Yes or No) and take the binary prediction as an input to evaluate the output of the system. As shown in Table V, the following four criteria are used to investigate the predictions generated by the system.

TABLE V. CONTINGENCY TABLE FOR PERFORMANCE MEASUREMENTS CONTAINING THE FOLLOWING ABBREVIATIONS FOR THE NUMBERS OF PREDICTED TRUE POSITIVES A, FALSE POSITIVES B, FALSE NEGATIVES C, AND TRUE NEGATIVES D

Flare prediction	Flare observations	
	Flare	No- Flare
Flare	a	b
No- Flare	c	d
$n= a+b+c+d$		

- If an MTF image is associated with a flare, and the system prediction is a flare then this successful prediction is a true positive (TP).
- If an MTF image is associated with a flare, but the system prediction is no-flare then this failed prediction is a false positive (FP).
- If an MTF image is not associated with a flare and the system prediction is no-flare then this successful prediction is a true negative (TN).
- If an MTF image is not associated with a flare and the system prediction is flare then this failed prediction is a false negative (FN).

To further evaluate the results we used various prediction verification measures for the 20 minute time window, shown in Table III. The measures used are the Heidke Skill Score (HSS), the percentage corrects (PC), the false alarm rate (FAR), the probability of detection (POD), and the Brier Score (BS). The formulae for these measures are defined in terms of the abbreviations given in Table IV.

The percentage correct measure, PC, is used to calculate the rate of predictions that are correct [13], and is defined as:

$$PC = \frac{(a+b)}{n} \tag{2}$$

The PC rate for the 20 minute time window is shown in Table VI for all the predictions (flare or no-flare) and is 78%.

The Heidke Skill Score (HSS) is a measure showing the improvement of the prediction over random prediction. HSS ranges from -1 (for no correct predictions) to +1 (for very accurate predictions) and a value of zero indicates that the predictions are randomly generated [13]. HSS is defined by:

$$HSS = \frac{2(ad-bc)}{[(a+c)(c+d)+(a+b)(b+d)]} \tag{3}$$

HSS is a really useful measure for verifying systems that seek to predict rare events, as in the present case.

The False Alarm Ratio FAR is the fraction of flare predictions that are wrong. The range of FAR is from 0 (best outcome) to 1 (poorest outcome) [18]. FAR is defined as:

$$FAR = \frac{b}{a+b} \tag{4}$$

TABLE VI. PREDICTION MEASURES ACHIEVED BY APPLYING MACHINE LEARNING AND CROSS-VALIDATION WITH DATASETS COVERING (03 DEC 2002-30 JAN 2017)

SPEC	SENC	QR	FAR	POD	PC	HSS
0.851	0.574	0.136	0.492	0.574	0.787	0.365

The Probability of Detection (POD)  $P_d$ , also known as the Hit Rate (H), measures the probability of a solar flare being correctly predicted by the system [18]. POD is given by:

$$P_d = \frac{a}{(a+c)} \quad (5)$$

It ranges from 0 (poorest outcome) to 1 (best outcome). The  $P_d$  result for this system with a 20 min time window is 0.574.

This process separately uses data covering the complete time range (03 Dec 2002 - 30 Jan 2017). The prediction measures achieved for datasets are shown in Table VI. It can be seen that the good levels of prediction measures are achieved.

## VI. CONCLUSION

This paper has introduced a prediction system that uses a new technology for predicting solar flares from GOES data using deep learning. This is the major contribution of this paper. The system predicts automatically whether a flaring event is going to occur in the next 20 minutes. Different prediction windows were investigated using the QR measure, and the most promising performance was found to be for the 20 minutes prediction window.

The performance of the prediction system introduced here depends on the ability of the deep learning neural network to efficiently classify the MTF images that have been generated to visualise the GOES data. As demonstrated in Table VI all the metrics used to evaluate the prediction performance (POD, FAR, HSS, KSS, and PC) provide fairly good performances. In particular, HSS results prove that the generated predictions are definitely not generated by chance.

The prediction rates for our systems can be improved by exploiting the advanced classification capabilities of machine learning systems. Hence, we believe that it is important to monitor the performance of the system during its initial stages which include comparing the prediction performance with the actual flares reported by NOAA. Evolutionary algorithms may be used to allow the learning algorithms to evolve and provide better optimization.

This work is continuing but we believe the initial results, as reported in this paper, are very encouraging. However, we note that not all flares have pre-flare phases occurring before them, and this could be one of the reasons affecting our predictions. To tackle these causes, our system could be integrated with another statistical or machine learning prediction model (e.g. ASAP<sup>1</sup>).

## REFERENCES

- [1] T. I. Gombosi, D. L. Dezeuw, C. P. T. Groth, K. G. Powell, C. Robert Clauer, and P. Song, "From Sun to Earth: Multiscale MHD Simulations of Space Weather," in *Space Weather*. (2001), Geophys. Monogr. Ser., vol. 125, edited by P. Song, H. J. Singer, and G. L. Siscoe, pp. 169-176, AGU, Washington, D. C., vol. 125, 2013, pp. 169-176.
- [2] T. Colak and R. Qahwaji, "Automated Solar Activity Prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares," *Sp. Weather*, vol. 7, no. 6, p. n/a-n/a, Jun. 2009.
- [3] T. Colak, R. Qahwaji. AUTOMATED PREDICTION OF SOLAR FLARES: Integrating Image Processing and Machine Learning for the Creation of a Hybrid Computer Platform that Provides Real-Time Prediction of Solar Flares. s.l. : LAP, 2010. ISBN-13: 978-3838370309.
- [4] A. David, Falconer, Ronald L. Moore, Abdunnasser F. Barghouty, Igor Khazanov. MAG4 versus alternative techniques for forecasting active region flare productivity. 306-317, s.l. : Space Weather AGU Journal, 2014, Vol. 12. 10.1002/2013SW001024. vol. 125, edited by P. Song, H. J. Singer, and G. L. Siscoe, pp. 169-176, AGU, Washington, D. C., vol. 125, 2013, pp. 169-176.
- [5] O. W. A. Ahmd, "ENHANCED FLARE PREDICTION BY ADVANCED FEATURE EXTRACTION FROM SOLAR IMAGES," University of Bradford, 2011.
- [6] Sunhak Hong, Jaehun Kim, Jinwook Han, Yungkyu Kim I. An Automated Solar Synoptic Analysis Software System. s.l. : American Geophysical Union, Fall Meeting 2012
- [7] C. Chifor, D. Tripathi, H. E. Mason, and B. R. Dennis, "X-ray precursors to flares and filament eruptions," *Astron. Astrophys.*, vol. 472, no. 3, pp. 967-979, Sep. 2007.
- [8] Janki Bhimani, Zhengyu Yang, Miriam Leeser, and Ningfang Mi. "Accelerating Big Data Applications Using Lightweight Virtualization Framework on Enterprise Cloud." s.l. : IEEE, 2017. 978-1-5386-3472-1/17.
- [9] <http://darts.isas.ac.jp/pub/solar/sswdb/goes/xray/>
- [10] Caspi, T. N. Woods, and J. Stone, "A New Observation of the Quiet Sun Soft X-ray (0.5-5 keV) Spectrum."
- [11] NASA, "GMS: SDO EVE Late Phase Flares." [Online]. Available: <https://svs.gsfc.nasa.gov/10817>. [Accessed: 28-Mar-2017].
- [12] Ng, "CS294A Lecture notes Sparse autoencoder." <https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf>
- [13] J. A. Guerra, A. Pulkkinen, and V. M. Uritsky, "Ensemble forecasting of major solar flares: First results," *Sp. Weather*, vol. 13, no. 10, pp. 626-642, Oct. 2015.
- [14] Liu, L., & Wang, Z. (2016). Encoding Temporal Markov Dynamics in Graph. Arxiv, 2.
- [15] S. I. Syrovatskii and S. I., *Comments on astrophysics and Space Physics.*, vol. 4. Gordon and Breach], 1972.
- [16] Adam Coates, Andrew Ng, Honglak Lee. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. : Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. PMLR 15:215-223,.
- [17] <https://github.com/cauchyturing/Imaging-time-series-to-improve-classification-and-imputation/wiki>
- [18] C. C. Balch, "Updated verification of the space weather prediction center's solar energetic particle prediction model," *Sp. Weather*, vol. 6, no. 1, pp. 1-13, 2008.
- [19] Hall, M.A.: Correlation -based feature selection for Machine Learning .PhD Thesis, The University of Waikato, Hamilton, New Zealand.
- [20] Barnes and Leka, 2008 *Astrophys. J.Lett.* 688, L107.

<sup>1</sup> <http://spaceweather.inf.brad.ac.uk/>

# Cyber-Security Incidents: A Review Cases in Cyber-Physical Systems

Mohammed Nasser Al-Mhiqani, Rabiah Ahmad, Warusia Yassin, Aslinda Hassan,  
Zaheera Zainal Abidin, Nabeel Salih Ali, Karrar Hameed Abdulkareem  
Information Security and Networking Research Group (InFORSNET),  
Center for Advanced Computing Technology,  
Faculty of Information Communication Technology,  
Universiti Teknikal Malaysia Melaka  
Melaka, Malaysia

**Abstract**—Cyber-Physical Systems refer to systems that have an interaction between computers, communication channels and physical devices to solve a real-world problem. Towards industry 4.0 revolution, Cyber-Physical Systems currently become one of the main targets of hackers and any damage to them lead to high losses to a nation. According to valid resources, several cases reported involved security breaches on Cyber-Physical Systems. Understanding fundamental and theoretical concept of security in the digital world was discussed worldwide. Yet, security cases in regard to the cyber-physical system are still remaining less explored. In addition, limited tools were introduced to overcome security problems in Cyber-Physical System. To improve understanding and introduce a lot more security solutions for the cyber-physical system, the study on this matter is highly on demand. In this paper, we investigate the current threats on Cyber-Physical Systems and propose a classification and matrix for these threats, and conduct a simple statistical analysis of the collected data using a quantitative approach. We confirmed four components i.e., (the type of attack, impact, intention and incident categories) main contributor to threat taxonomy of Cyber-Physical Systems.

**Keywords**—Cyber-Physical Systems; threats; incidents; security; cybersecurity; taxonomies; matrix; threats analysis

## I. INTRODUCTION

The world accepted that Cyber-Physical Systems (CPSs) connect computers, communication devices, sensors and actuators of the physical substratum, either in heterogeneous, open, systems-of-systems or hybrid. Systems become more interconnected, thereby more complex [1]. Computer networks currently have joined water, food, transportation, and energy as the critical resource for the function of the nationals' economy. Application of CPS can be seen in many forms of industries. The common sector is oil and gas, the power grid manufacturing, defense and public infrastructures are fully relying on the advancement of CPS. Therefore, cyber-physical systems security has become a matter for societal, infrastructures and economic to every country in the world due to the tremendous number of electronic devices that are interconnected via networks communication [2]-[4]. Latest reports have shown that cyber-attacks are aimed to destroy nation's systems that used for country development. CPS starts with by not simply disrupt a single enterprise or damage an isolated machine, but a target to damage infrastructures via

modern dynamics threats [5], [6]. Those types of attacks are able to provide destruction to critical infrastructures system which used in sectors such as defense, finance, health, and the public [7]. To accomplish their goals criminals, activists, or terrorists are mostly looking for new and innovative techniques and targets, so cyber-physical systems currently one of the important targets for the hackers [3]. Increased security risk awareness and appropriately security relevant information management provide an equally important role in the trusted infrastructure maintenance [8]-[10]. This paper discusses some instances of attacks on cyber-physical systems that have occurred in the Organization of Islamic Cooperation (OIC) countries. The diversity of the attacks will be covered and analyzed based on their types and targets. The analysis will allow researchers to clearly understand the nature of the attacks and how they were carried out. A proposed matrix for threats verification and threats taxonomy will be discussed using a modified version of many taxonomies presented in [11]-[14] to classify the threats based on certain factors to enable researchers to analyze them along with their types and targets. The different matrices include types of attack, target sector, intention, impact, and incident categories. This article is structured into seven sections describe cyber-physical system threats from fundamental concept to threats categorization and impact. The following section will provide related work on the issue discussed. The remainder of the paper is structured as follows: In Section 2 reviews and discusses several taxonomies that have been presented to classify the threats based on certain factors. Section 3 provides a clear description of the proposed taxonomy to classify the CPS attacks based on types of attacks, target sector, intention, impact, and incident categories. Also, presents a comprehensive detail regarding the proposed matrix in Section 4. In addition, different CPS incidents surveyed from various sources in Section 5. Section 6 discusses and analyses the incidents by the modified taxonomy. Finally, Section 7 concludes this study.

## II. RELATED WORK

In [12], the author discusses and classifies incidents of cyber-physical attacks based on the sources, sectors, and impact of the incidents. The research paper provides an example of how the standardization of the cyber incidents information collection can be useful for attack victims and aids in understanding the cyber incidents threats towards different

targets. Four dimensions taxonomy proposed in [13] to provides a holistic taxonomy to enable the researchers to deal with inherent problems in the computer and network attack field. The first dimension of the taxonomy covers the attack's vector and the main attack behavior. The second dimension categorizes the attacks based on their targets. The third and fourth taxonomy dimensions categorized the vulnerabilities and payloads, respectively. The framework in [14] describes core components in cyber terrorism. The data is analyzed using a grounded theory approach in which the framework is drawn. The framework defined the cyber terrorism from six perspectives: target, motivation, domain, attack method, perpetrator action, and the impact of the attack. In addition, the proposed framework provided a dynamic method for defining cyber terrorism and describing its influential considerations. Incident analysis security ontology research is presented in [15] and provides a taxonomy which has some similarities to the framework presented in [14], but some aspects have been added in their classification such as action and unauthorized results. In their taxonomy Giraldo et al. [16] categorized cyber-physical systems by focusing on some of the CPS characteristics such as its domains, defenses, attacks, network security, research trends, security-level implementation, and computational strategies.

### III. PROPOSED TAXONOMY

The proposed taxonomy in this paper uses a modified versions of those presented taxonomies in [12]-[15] to classify the attack based on types of attacks, target sector, intention, impact, and incident categories. Each part of the attack will be broken down to the terms shown in Fig. 1 and explained.

#### A. Types of Attacks

**Worm:** in their propagation worm is like viruses with no direction by the network from the attackers. However, unlike viruses, in worms, no interaction is needed from the user for activating their attempt to spread.

**Trojan:** is a type of a program where subversive functionality is added to associate with the existing program.

**Virus:** virus may be defined as a piece of codes that usually attaches itself to another program, and when the program runs it will run with them.

**DDoS:** represents the coordinated attacks on the target system service availability that has been given or a network that is indirectly launched through a number of compromised computing systems.

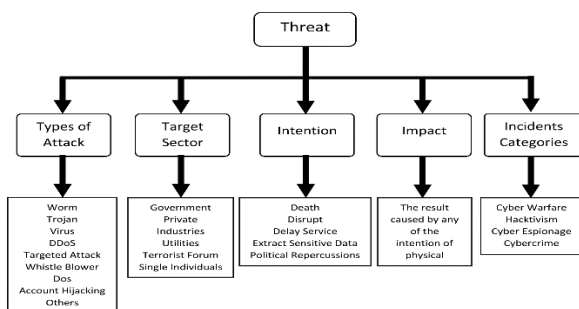


Fig. 1. Threats taxonomy.

**Targeted Attack:** refers to malicious attack which is targeted to a particular individual, software, systems, or company. It might be used to extract information, disrupt operations, or destroy a certain type of data on the target machine.

**Whistleblower:** indicates the disclosure of the information for perceived wrongdoing within the organization, or the risk to individuals or entities that have the ability to effect action.

**Denial of Service:** is defined as an attack that design to disable a network or computer from providing normal services. It is considered to occur only when access to a network or computer resource is intentionally degraded or blocked as a result of malicious action by another user.

**Account Hijacking:** is defined as a process where a particular individual's computer, email, or other account associated with service or a computing device is hijacked or stolen by hackers.

#### B. Target Sector

**Government:** is denoting local or national governments including buildings/housing, emergency services, public benefits, and social services, federal and state governments, tribal governments, military, protection of workers, and environment [16].

**Private:** refer to the part of a country's organization run by individuals and companies, rather than the government.

**Industries:** are the sectors that consist of all equipment and facilities used for producing, processing, or assembling goods [17].

**Utilities:** The utility sector comprises companies such as electric, water, gas, and integrated utility providers [18].

**Terrorist forum:** is the target sector which relates to any terrorist group such as ISIS or Al Qaeda.

**Single Individuals:** is the sector in which the attacker aims to affect the individual users.

#### C. Intention

**Death:** is the loss of human life.

**Disrupt:** change of access, removal access to information or to a victim. Manipulate the permission, e.g., Trojan horse or Denial of Service (DoS) attack. Disruption could be the least invasive of the attack [16].

**Service Delay:** where organizations or companies delay providing services on time due to the problems in the system.

**Extract sensitive data:** where unauthorized or hackers entities secure access to particular data and extract private information [19].

**Political Repercussions:** refer to events whose impact affects the government or the people leading the country.

**Others:** cases that not falling under any of the abovementioned categories.

D. Impact

The impact of the incident describes the incident effect. The impact description requires addressing all the entities affected which include the computer systems, the physical systems which the cyber-physical system interacts with, and the broader impacts on the community and organization [20].

E. Incident Categories

Cyberwarfare (CW): “using cyberspace (by operating within or through it) to attack personnel, facilities, or equipment with the intent of degrading, neutralizing, or destroying enemy combat capability” [21].

Hackivism (H): “is the convergence of the hacking process and activism where hacking refers to the operations that exploit computers in ways that are unusual or often illegal, normally with the help of certain software” [22].

Cyber Espionage (CE): is the arm of the corporate high-tech crime. It mainly involves attacks on companies and institutions and not individuals. Cyber espionage does not always necessarily occur on a large scale [23].

Cyber Crime (CC): Involves the all criminal act which deals with the networks and computers (hacking). Additionally, traditional crimes that are conducted through the Internet are included in cybercrime [24].

IV. PROPOSED MATRIX

The proposed matrix in this study uses two separated matrices i.e., threat matrix analysis and target matrix analysis to collect information that is required in cyber-physical system incidents analysis. The threat matrix analysis (see Table I) contains the associations between the intention and the type of attack, while the target matrix analysis (see Table II) contains the association between the attack target sector and the incident category. When the incidents analysis is initially conducted, threats and target are generated then added to the particular table. The matrix is then populated by adding data which correlates the column of the matrix with the row of the matrix. Finally, the threat matrix data is aggregated using (1) and then presented in Table I. Similarly, the Target analysis matrix data are aggregated using (2) and presented in Table II.

The derived equations 1 for the threat analysis are shown below:

$$\text{ThreatVal}_{ti} = \text{likelihood}_t \times \text{Rate}_i \tag{1}$$

$$\text{Total Score}_t = \sum_{i=1}^6 \text{ThreatVal}_{ti} \tag{2}$$

$$\text{Total Score}_t = \sum_{i=1}^6 (\text{likelihood}_t \times \text{Rate}_i) \tag{3}$$

$$\text{Total score}_t = \text{likelihood}_t \times \sum_{i=1}^6 (\text{Rate}_i) \tag{4}$$

Where:

*t*: Represent the Types of Attacks

*i*: Represents the Intention

The derived equations 2 for the target analysis are shown as below:

$$\text{ThreatVal}_{ti} = \text{likelihood}_t * \text{Rate}_i \tag{1}$$

$$\text{Total Score}_t = \sum_{i=1}^5 \text{ThreatVal}_{ti} \tag{2}$$

$$\text{Total score}_t = \sum_{i=1}^5 (\text{likelihood}_t * \text{Rate}_i) \tag{3}$$

$$\text{Total score}_t = \text{likelihood}_t * \sum_{i=1}^5 (\text{Rate}_i) \tag{4}$$

Where:

*t*: represents the incidents Categories

*i* represents the Target Sector

TABLE I. THREATS ANALYSIS (CORRELATION BETWEEN TYPES OF ATTACK AND INTENTION)

THREAT ANALYSIS								
INTENTION \ TYPES OF ATTACK	Likelihood	Death	Disrupt	Delay Service	Extract sensitive data	Political Repercussions	Others	Total score
Rate of Impact		6	6	3	6	3	3	
Worm	1							
Trojan	1							
Virus	6							
DDoS	1							
Targeted attack	3							
Whistleblower	1							
Denial of Service	1							
Account Hijacking	6							
Others	3							

TABLE II. TARGET ANALYSIS (CORRELATION BETWEEN INCIDENTS CATEGORIES AND TARGET SECTOR)

TARGET ANALYSIS							
TARGET SECTOR \ INCIDENTS CATEGORIES	Likelihood	Government (Gov)	Private	Utilities	Terrorist forum	Single individuals (SI)	Total score
Rate of Impact		6	3	6	3	3	
Cyberwarfare	6						
Hackivism	6						
Cyber Espionage	3						
Cyber Crime	3						

A. Rate the Matrix

In this part of the analysis, we have a list of types of attack that apply to a particular intention and the incidents category



that relates to the target sector. From our litterer review, we can rate the threats based on their impact level and the attack likelihood having occurred. This eases the addressing of the threats by presenting the high-risk ones first and then resolving the other threats.

This method indicates that the threats posed by specific types of attacks are similar to the probability of the threats occurring multiplied by the intention which indicates the consequences to CPS system if the attacks were to occur.

A 0–6 scales can be used for probability where 0 represents the types of attacks that are unlikely to occur and 6 representing those that are mostly occurring. Similarly, a 0–6 scale is used for intention with 0 indicating the intention that has no impact and 6 the intention that causes the highest impact. The same method is applied to the second matrix for the incidents category and the target sector.

## V. SURVEY OF INCIDENTS

We surveyed many different CPS incidents from various sources and provide details of each one to examine how it was conducted. Some of the cyber incidents are explored in this study due to their high impact on daily life. Table III provides a summary of the incidents.

### A. Stuxnet

In 2010, a worm named Stuxnet hit the Iranian nuclear facilities at Natanz. Stuxnet utilized 4 ‘zero-day vulnerabilities’ (vulnerabilities were previously unknown, so there was no time to distribute and develop patches). The worms employed default passwords of Siemens to access the operating systems of Windows that run PCS7 and WinCC programs. They sought out frequency-converter drives manufactured by FararoPaya in Vacon in Finland and Iran. To power centrifuges, these drives were used to be utilized in the uranium 235 isotope concentration. The current electrical frequency to the drivers was altered by the Stuxnet which modified them between low and high speeds that they weren’t designed for [25].

Type of Attack: Root, Worm, Trojan

Target Sector: Military (nuclear industry)

Intention: Disrupt

Incident Categories: CW

### B. Iranian Infrastructure Attack

Cyber attackers disrupted the Internet network in Iran by attacking the country’s infrastructure and communications companies and forcing the Internet to be limited due to the heavy attack. All the attacks were arranged systematically and included nuclear, oil, and information networks [26].

Type of Attack: unknown

Target Sector: Gov

Intention: Disrupt

Incident Categories: CW

### C. Iran Hijacking of US Drone

Iranian specialists in electronic warfare were able to bring down an American bat-wing RQ-170 Sentinel by cutting off its communications links according to an Iranian Engineer working for an Iranian team attempting to unravel the stealth and intelligence secrets of the drones.

Iranians used the “spoofing” technique which considers landing altitudes, longitudinal and latitudinal data accurately causing the drone to land to the wanted location, without needing to crack the remote-control signal and communications from the control center [27].

Types of Attack: spoofing

Target Sector: Military

Intention: captured drone's systems

Incident Categories: CW

### D. Iranian Oil Terminal ‘offline’

A malware attack forced Iran to disconnect its key oil facilities. It is believed that the computer virus targeted the Iranian oil ministry and the national oil company by attacking their internal computer system. As prevention, the equipment at many Iranian different plants such as on the Island of Kharg was disconnected from the internet [11].

Type of Attack: Virus

Target Sector: Gov (Oil Company)

Intention: Disrupt

Incident Categories: CE

### E. Saudi Aramco Attacks

The external source-originated virus targeted the Saudi Aramco Company and infected around 30,000 of its workstations. The company suspected the attack to be the outcome of a virus that had infected individual workstations without influencing the main parts of the network [28]. To prevent further attacks, Aramco was forced to cut off the electronic system from outside access.

Type of Attack: virus

Target Sector: Gov (Oil Company)

Intention: Disrupt

Incident Categories: H

### F. Egypt Maritime Transport Sector

The attacked list comprised the websites of the Presidency, the Armed Forces, the Maritime Transport Sector, the Parliament, the Egyptian Accreditation Council, the Large Taxpayer Center, Ministry of Interior and many others. The attack affected the websites of the Egyptian government [30].

Type of Attack: DDoS

Target Sector: Gov (Transport)

Intention: Delay service

Incident Categories: H

TABLE III. SUMMARY OF INCIDENTS

Year	Country	Title	Type of Attack	Target Sector	Intention	Incident Category
2010	Iran	Stuxnet	Worm, root, Trojan	Military (Nuclear industry)	Disrupt	CW
2011	Iran	Iranian infrastructure and communications companies	unknown	Gov ( infrastructure companies)	Disrupt	CW
2011	Iran	Iran hijacked US drone	spoofing	Military (US drone)	Captured drone's systems	CW
2012	Iran	Iranian oil terminal 'offline'	virus	Gov (Oil company)	Disrupt	CW
2012	Saudi	Saudi Aramco	virus	Gov (Oil Company)	Disrupt	H
2012	Egypt	Maritime transport sector	DDoS	Gov (Transport)	Delay service	H
2012	Syria	Syrian Ministry of Foreign Affairs	unknown	Gov (foreign ministry)	Extract sensitive data	CW
2012	Syria	Secret Assad emails lift lid on life of leader's inner circle	Whistleblowing	Single Individual	Extract sensitive data	H
2012	Qatar	Qatar's RasGas Attack	virus	Private (Oil Company)	Disrupt	H
2013	Saudi	Saudi Arabian Defense Ministry System Breached	Account Hijacking	Gov (military)	Extract sensitive data	CW
2014	Syria	Syrian Hackers Ramp Up RAT Attacks	Targeted attack	Single Individual	Remote PC	CE
2015	Turkey	Attack on Istanbul Airport passport control system	virus	Gov (Airport)	Delayed service	CC
2015	UAE	Energy companies attacked by Trojan Laziok	Trojan	Gov (Energy)	Extract sensitive data	CC
2016	Turkey	Leaks Turkish Police data	Account Hijacking	Gov (Police data)	Extract sensitive data	CW
2016	Saudi	Shamoon 2	Malware	Gov (Industries)	Disrupt	CC
2016	UAE	The Operation Ghoul in UAE	Targeted attack	Industrial and Engineering companies	Extract sensitive data	CC
2017	Turkey	The source of the widespread electricity cuts across Istanbul	unknown	Gov (Transmission & electricity)	Disrupt	CW
2017	Qatar	Qatar News Agency Hacked	Account Hijacking	Gov (website)	Political Repercussions	CC

### G. Syrian Ministry of Foreign Affairs

Around one gigabyte of documents was released by unknown hackers. The documents allegedly represented the internal government emails contents from the Ministry of Foreign Affairs. The publication of the documents was considered as part of the Syria campaign. The published documents comprised all information types, such as scanned copies of Syrian ministers' passports, specifics about an arms transport from Ukraine [30].

Type of Attack: unknown

Target Sector: Gov (foreign ministry)

Intention: Extract sensitive data

Incident Categories: CW

### H. Secret Assad Emails Hacked

The attack targeted to sign into emails of nearest helpers of the president of Syria using a simple and straightforward password of numbers from 1 to 4. Israeli Haaretz site published selected documents from the hacked emails. The documents involved emails between Bouthaina Shaaban the president's media adviser and the press attaché in Syria's UN mission. The emails briefed the president before his interview with Barbara Walters in which the president denied responsibility for his governments' troops killing of civilians in Syria [31].

Type of Attack: Whistleblowing

Target Sector: Gov (President's Email)

Intention: Extract sensitive data

Incident Categories: H

### I. Qatar's RasGas Attack

These attacks have brought down the computers of the RasGas Company due to a virus that hit the computer systems. Qatar RasGas was forced to close the email system and its website. The company's experts in security warned of hackers efforts to hit the energy and oil industry [32].

Type of Attack: Virus

Target Sector: Gov (Oil Company)

Intention: Disrupt

Incident Categories: H

### J. Saudi Arabian Defense Ministry Mail System Breached

A source claimed that Syrian Electronic Army (SEA) received a secret document hacked from the emails of Saudi Arabia's Ministry of Defense involving secret arms deals. The documents were forwarded to the government of Syria [33]. A screenshot was shown to prove the successful attack on the mail system of the ministry.

Type of Attack: Account Hijacking

Target Sector: Gov (military)

Intention: Extract sensitive data

Incident Categories: CW

### K. Syrian Hackers Ramp up RAT Attacks

Ramp up RAT attacks were launched through the social network. Hackers from Syria tried to download remote access Trojans (RATs) into the victim's computers. According to security researchers, they also discovered evidence of rising attacks from Syria [34]. The attackers seemed to take advantage of people's fear of government monitoring in the state. They created fake messages or posts on the social network such as in Skype and Facebook warning users about being attacked and where these messages themselves led to fake AV downloads.

Type of Attack: Targeted

Target Sector: Single Individual

Intention: Remote PC

Incident Categories: CE

### L. Cyber Attack Hits Istanbul Airport

The cyber-attack targeted Istanbul Ataturk Airport specifically the passport control system at the international departure area, and at another airport in Istanbul. As a result, the passport control system shut down, flights were delayed, and passengers waited in lines for hours at the two airports [29], [35].

Type of Attack: Virus

Target Sector: Gov (Airport)

Intention: Delay of services

Incident Categories: CC

### M. Energy Companies Attacked by Trojan Laziok

An Attack called Trojan Laziok attacked the energy sectors. These attacks targeted the Middle East companies especially United Arab Emirates companies, according to Symantec, Trojan Laziok acted as reconnaissance tools that enable the hackers to steal database from the targeted computers. The attacks targeted oil, helium gas and companies through spam emails from the domain money.trans.eu. Microsoft Excel files are attached to the emails with an exploit for the Microsoft Windows Common Controls ActiveX Remote Code Execution Vulnerability. By clicking on the attachments it starts up its infection process. Trojan Laziok hid in the directory: %SystemDrive%\Documents and the other directory Settings\All Users\Application Data\System\Oracle [36].

Type of Attack: Trojan

Target Sector: Gov (Energy sector)

Intention: Extract sensitive data

Incident Categories: CC

### N. Leaks of Sensitive Data from Turkish Police Servers

Hackers known as ROR [RG] released a huge amount of sensitive data belonging to the Turkish National Police database. Around 50 million citizen data was leaked and publicly shared online such as first name, surname, citizenship number, sex, address, and date and place of birth [37].

Type of Attack: Account Hijacking  
Target Sector: Gov (Police Law Enforcement)  
Intention: Extract sensitive data  
Incident Categories: CW

#### O. Shamoon 2 Malware

Three new waves of the destructive Shamoon 2 attacked many companies in Saudi Arabia. Bryan Lee and Robert Falcone “determined that the actors conducting the Shamoon 2 attacks use one compromised system as a distribution point to deploy the destructive Distrack Trojan to other systems on the targeted network, after which the Distrack malware will seek to propagate itself even further into the network” [38].

Type of Attack: DNS Hijacking  
Target Sector: Private (Airlines)  
Intention: Delay service  
Incident Categories: CC

#### P. The Operation Ghoul in UAE

This is named after the Operation Ghoul group was the source of a multiple cyber-attacks that were reported in the United Arab Emirates. What the cyber-hackers did was to send malicious attachments with phishing emails particularly these emails sent to the top managers and some of the middle-level employee of various companies. The phishing emails are appearing to be coming from a local bank with messages that claiming to offer some advice on the payment from their bank. The email contains SWIFT document attachment which contains a malware [39].

Type of Attack: Targeted attack  
Target Sector: Industrial and Engineering companies  
Intention: Extract sensitive data  
Incident Categories: CC

#### Q. The Source of Widespread Electricity Cuts Across Istanbul

A source from the Ministry of Energy in Turkey claimed that critical cyber-attacks caused widespread electricity cuts in the city. It mentioned that many infiltration attempts which the hacker tried on the controlling systems of electricity and transmission were prevented [40].

Type of Attack: Malware  
Target Sector: Gov  
Intention: Disrupt  
Incident Categories: CC

#### R. The Official State News of Qatar Agency Hacked

Qatar announced that the Qatar News Agency (QNA), its national news agency, was hacked and a few articles about sensitive issues published on the website before it went down. The articles focused on the Palestinian-Israeli conflict, relations between Qatar and the Republic of Iran, remarks on Hamas, and negative perspectives on the relationship between Qatar and President Trump. The articles were attributed to Sheikh

Tamim bin Hamad Al-Thani the Emir of the country, leading to Saudi Arabia, the United Arab Emirates, Egypt and Bahrain breaking off all the relations with Qatar in the worst diplomatic crisis to hit Gulf Arab states in decades [41].

Type of Attack: Account Hijacking  
Target Sector: Gov (News Agency)  
Intention: Political Repercussions  
Incident Categories: CC

## VI. ANALYSIS OF INCIDENTS

### A. Analysis of Incidents by Modified Taxonomy

Fig. 2 shows that among all the OIC countries, Iran has the highest number of cyber-physical attacks which are mostly related to political issues in the country, followed by Turkey, KSA, and Syria. The other surveyed countries have between one to two cases of CPS attacks.

Fig. 3 represents the incidents by year for the attacks surveyed in this work. As can be seen, 2012 had the highest number of attacks. That was the year following the Arab Spring in the Middle East and the Israeli-Palestinian conflict in Gaza [42] where the number of incidents in cyber-physical systems increased.

Fig. 4 details the attacks by type. Four cases took advantage of a virus, 3 utilized account hijacking, 1 case each of the other methods, which are, targeted attack, Spoofing, DDoS, and DNS Hijacking, and 2 cases where the method of attack is not defined.

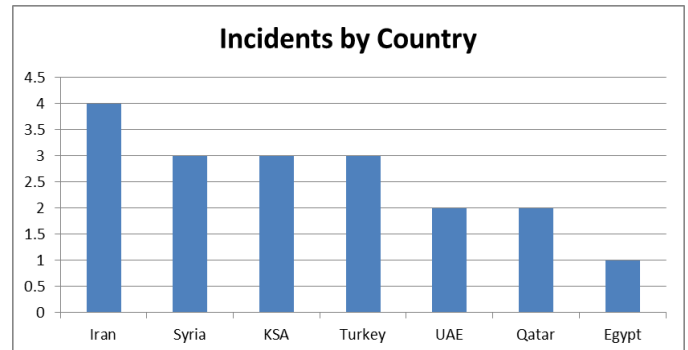


Fig. 2. Incidents by Country.

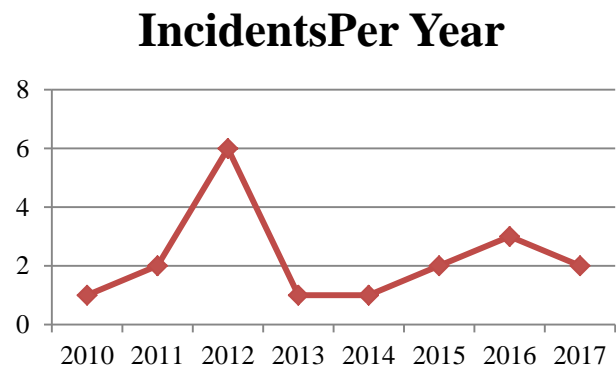


Fig. 3. Incidents by Year.

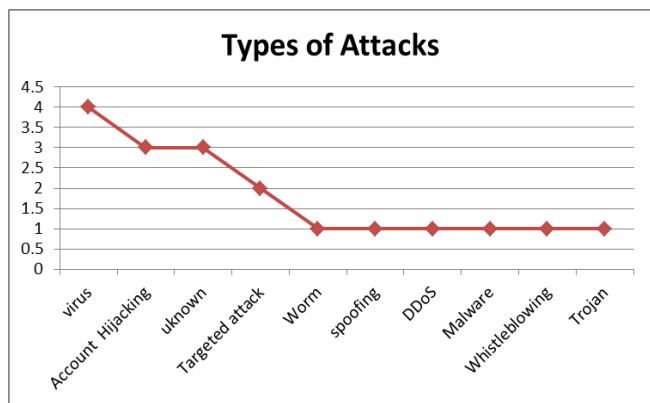


Fig. 4. Types of Attacks.

## Target Sector

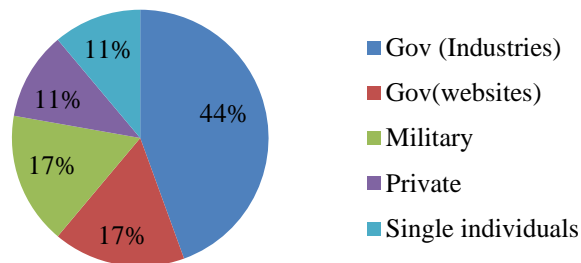


Fig. 7. Target sector.

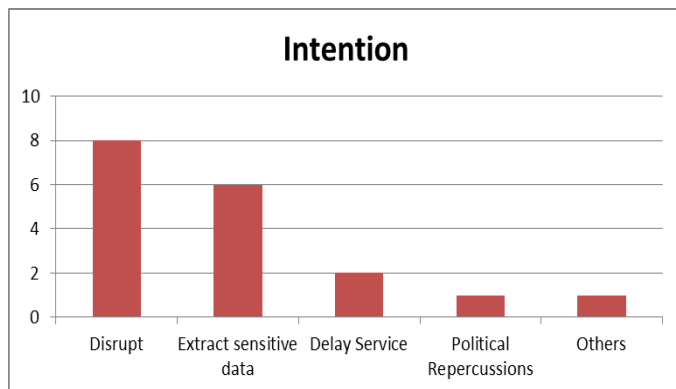


Fig. 5. Intention.

We next look at the intention of these attacks. As shown in Fig. 5, most aimed at disruption and extracting sensitive data, 2 at delays in services, and 1 each at political repercussions and for other intentions.

Fig. 6 represents the categories of the incidents. Cyberwarfare with 8 cases formed the highest category, while 5 incidents were cybercrime, 4 involved Hacktivism, and 1 cyber-espionage.

Fig. 7 shows the attacks by sectors. Most attacks were in the government sector involving the oil industry, transport, and other utilities at 44% of surveyed incidents, government websites (17%), the military (17%), and 11% for both private single individuals.

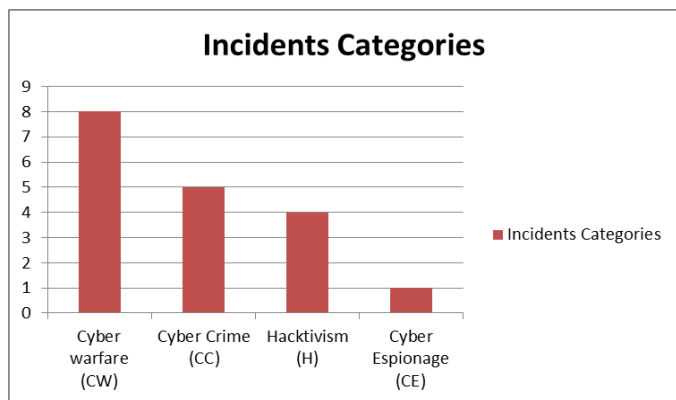


Fig. 6. Intentions categories.

### B. Analysis of Incidents using Matrix

In the threat matrix in Table IV the data is aggregated and then sorted to define the types of attacks and intention relative importance. Since death, disruption, and accessing sensitive information has a strong impact, their ranks are high in the threat matrix especially when the type of attacks have a high probability of occurring many times like virus and account hacking types. The aggregate intention data then added into threat matrix along with the corresponding threat to the types of attacks.

The results of this analysis and the aggregate data in the matrices are used to increase overall awareness of each type of these attacks.

TABLE IV. THREAT ANALYSIS DATA

THREAT ANALYSIS								
INTENTION TYPES OF ATTACK	Likelihood	Death	Disrupt	Delay Service	Extract sensitive data	Political Repercussions	Others	Total score
		6	6	3	6	3	3	
Worm	1	6	6	3	6	3	3	27
Trojan	1	6	6	3	6	3	3	27
Virus	6	36	36	18	36	18	18	162
DDoS	1	6	6	3	6	3	3	27
Targeted attack	3	18	18	9	18	9	9	81
Whistleblower	1	6	6	3	6	3	3	27
Account Hijacking	6	36	36	18	36	18	18	162
Others	3	18	18	9	18	9	9	81

TABLE V. TARGET ANALYSIS MATRIX

TARGET ANALYSIS							
TARGET SECTOR	Likelihood	Government	Private	Utilities	Terrorist forum	Single individuals (SI)	Total score
INCIDENTS CATEGORIES		6	3	6	3	3	
Cyberwarfare (CW)	6	36	18	36	18	18	126
Hackivism (H)	6	36	18	36	18	18	126
Cyber Espionage(CE)	3	18	9	18	9	9	63
Cyber Crime(CC)	3	18	9	18	9	9	63

The data in the target analysis matrix in Table V is similar to the previous matrix which is aggregated and then sorted to define the types of attacks and intention relative importance, while this matrix is aggregated and then sorted to define the incidents category and target sector relative importance. Government services, websites, and utilities have a high impact when their systems are hacked. The likelihood of cyberwarfare and Hackivism occurring in the target analysis is very high since most of the incidents analyzed in our study fall under these two categories.

VII. CONCLUSIONS

The wide uses of CPS nowadays bring some risks and means for cybercriminals to use in their attacks against governments, organizations, or individuals. In this paper, we classified CPS threats based on modified taxonomies in generating organized information for other academics, experts, and researchers. This paper also provides researchers with matrices for studying the threats and enabling them to rapidly identify and correlate key threats involving CPS systems which, in turn, will lead to increased overall awareness of these incidents. However further work though is needed, the first suggestion is to include the study on how to trace the source of incidents, which cover the study of the groups and single individual hackers where the source of incidents come from, and the second suggestion is to study the cyber-physical security detection mechanisms to detect the attacks whether it comes from outsider or insider.

ACKNOWLEDGEMENTS

This project is funded by the Ministry of Higher Education Malaysia under Transdisciplinary Research Grant Scheme (TRGS) with project Number TRGS/1/2016/UTEM/01/3. And

this project referred as TRGS/1/2016/FTMK-CACT/01/D00006 at UNIVERSITI TEKNIKAL MALAYSIA MELAKA

REFERENCES

- [1] I. Friedberg, K. McLaughlin, P. Smith, D. Lavery, and S. Sezer, "STPA-SafeSec: Safety and security analysis for cyber-physical systems," J. Inf. Secur. Appl., vol. 34, pp. 183–196, 2017.
- [2] W. Wang and Z. Lu, "Cybersecurity in the Smart Grid: Survey and challenges," Comput. Networks, vol. 57, no. 5, pp. 1344–1371, 2013.
- [3] C. W. Ten, G. Manimaran, and C. C. Liu, "Cybersecurity for critical infrastructures: Attack and defense modeling," IEEE Trans. Syst. Man, Cybern. Part A Systems Humans, vol. 40, no. 4, pp. 853–865, 2010.
- [4] J. Walker, B. J. Williams, and G. W. Skelton, "Cybersecurity for emergency management," Technol. Homel. Secur. HST 2010 IEEE Int. Conf., pp. 476–480, 2010.
- [5] J. J. Walker, T. Jones, M. Mortazavi, and R. Blount, "CyberSecurity Concerns for Ubiquitous/Pervasive Computing Environments," 2011 Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov., pp. 274–278, 2011.
- [6] N. S. Ali, "A four-phase methodology for protecting web applications using an effective real-time technique," Int. J. Internet Technol. Secur. Trans., vol. 6, no. 4, p. 303, 2016.
- [7] Al-Mhiqani, M.N., Ahmad R., Abdulkareem K. H., Ali N.S., "Investigation Study of Cyber-Physical Systems: Characteristics, Application Domains, and Security Challenges," ARPN Journal of Engineering and Applied Sciences, Vol. 12, No. 22, pp. 6557-6567, 2017
- [8] Ali, N. S., & Shihghatullah, A. S., "Protection Web Applications using Real-Time Technique to Detect Structured Query Language Injection Attacks," International Journal of Computer Applications, Vol. 149, No. 6, pp. 0975-8887, 2016.
- [9] Sridhar, S., Hahn, A., & Govindarasu, M. "Cyber-physical system security for the electric power grid". Proceedings of the IEEE, 100(1), 210-224.
- [10] Ten, C. W., Liu, C. C., & Manimaran, G. . "Vulnerability assessment of cybersecurity for SCADA systems". IEEE Transactions on Power Systems, 23(4), 1836-1846, 2008.
- [11] B. Miller and D. Rowe, "A survey SCADA of and critical infrastructure incidents," in Proceedings of the 1st Annual conference on Research in information technology - RIIT '12, 2012, p. 51.
- [12] M. Kjaerland, "A taxonomy and comparison of computer security incidents from the commercial and government sectors," Comput. Secur., vol. 25, no. 7, pp. 522–538, Oct. 2006.
- [13] S. Hansman and R. Hunt, "A taxonomy of network and computer attacks," Comput. Secur., vol. 24, no. 1, pp. 31–43, Feb. 2005.
- [14] A. Rabiah, Y. Zahari, "A Dynamic Cyber Terrorism Framework," Int. J. Comput. Sci. Inf. Secur., vol. 10, no. Xxx, 2012.
- [15] C. Blackwell, "A security ontology for incident analysis," in Proceedings of the Sixth Annual Workshop on CyberSecurity and Information Intelligence Research - CSIIRW '10, p. 1, 2010.
- [16] J. Giraldo, E. Sarkar, et al., "Security and privacy in cyber-physical systems: A survey of surveys," IEEE Design & Test, 2017.
- [17] EIA, "Fuel Oil and Kerosene Sales - Energy Information Administration," Office of Petroleum and Biofuels Statistics, Office of Energy Statistics, 2013.
- [18] J. R. Klinefelter and T. A. Klinefelter, Minimalist Investor Maximum Profits, 1st editio. Page Publishing Inc, 2015.
- [19] Y. G. B, P. C. Bhaskar, and R. K. Kamat, "Assessing the Guilt Probability in Intentional Data Leakage," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 3, pp. 4075, 2012.
- [20] W. B. Miller, D. C. Rowe, and R. Woodside, "A Comprehensive and Open Framework for Classifying Incidents Involving Cyber-Physical Systems," in IAJC/ISAM Joint International Conference, 2014.
- [21] K. B. K. B. L. G. Alexander, "Warfighting in Cyberspace," JFQ NDU Press, vol. 35, no. 46, pp. 58–61, 2007.

- [22] D. E. Denning, "Activism, Hacktivism, And Cyberterrorism: The Internet As A Tool For Influencing Foreign Policy," in *Networks and Netwars: The Future of Terror, Crime, and Militancy*, 1999.
- [23] P. Warren and M. Streeeter, *Cyber Crime & Warfare: All That Matters*. Hodder & Stoughton, 2013.
- [24] T. Critchley, *High Availability IT Services*. Taylor & Francis, 2014.
- [25] S. D. Applegate, "The Dawn of Kinetic Cyber," in *Cyber Conflict (CyCon)*, 2013 5th Int. Conference, 2013.
- [26] S. Aryan, H. Aryan, and J. A. Halderman, "Internet censorship in Iran : A First look," 3rd USENIX Work. Free Open Commun. Internet, no. August, p. 8, 2013.
- [27] B. W. O'Hanlon, M. L. Psiaki, J. A. Bhatti, D. P. Shepard, and T. E. Humphreys, "Real-time GPS spoofing detection via correlation of encrypted signals," *Navigation*, vol. 60, no. 4, pp. 267–278, 2013.
- [28] B. van N. Barend Pretorius, "Cybersecurity and Governance for ICS/SCADA in South Africa" - The Proceedings of the 10th International Conference on Cyberwarfare and Security, in *The Proceedings of the 10th International Conference on Cyber*, 2015, p. 558.
- [29] Urban, J., "Not Your Granddaddy's Aviation Industry: The Need to Implement Cybersecurity Standards and Best Practices Within the International Aviation Industry". *Albany Law Journal of Science & Technology*, 2017.
- [30] W. S. PENDERGRASS, "What is Anonymous?: A case study of an information systems hacker activist collective movement," 2013.
- [31] J E. Grohe, "The Cyber Dimensions of the Syrian Civil War Implications for FutureConflict," 2015.
- [32] S. K. Venkatachary, J. Prasad, and R. Samikannu, "Economic Impacts of CyberSecurity in Energy Sector : A Review," *Int. J. Energy Econ. Policy*, vol. 7, no. 5, pp. 250–262, 2017.
- [33] P. Bradshaw and P. Bradshaw, *Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation*, vol. 2017.12. University of Oxford, 2017.
- [34] W. R. Marczak, J. Scott-Railton, M. Marquis-Boire, and V. Paxson, "When Governments Hack Opponents: A Look at Actors and Technology," *Proc. 23rd USENIX Secur. Symp.*, pp. 511–525, 2014.
- [35] E. Livanis, "Financial Aspects of Cyber Risks and Taxonomy for the Efficient Handling of These Risks," in *14th International Scientific Conference on Economic and Social Development*, 2016, no. May, pp. 80–87.
- [36] R. de Oliveira Albuquerque, L. J. Garc a-a Villalba, A. L. Sandoval Orozco, R. T. de Sousa J nior, and T. H. Kim, "Leveraging information security and computational trust for cybersecurity," *J. Supercomput.*, vol. 72, no. 10, pp. 3729–3763, 2016.
- [37] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang, "Targeted Online Password Guessing," *Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur. - CCS'16*, pp. 1242–1254, 2016.
- [38] F. O. H. and M. Sulmeyer, "Getting beyond Norms (New Approaches to International CyberSecurity Challenges)," 2017].
- [39] Q. Tao, M. Jiang, X. Wang, and B. Deng, "A cloud-based experimental platform for networked industrial control systems," *Int. J. Model. Simulation, Sci. Comput.*, vol. 9, no. 4, p. 1850024, 2017.
- [40] Y. Biran, J. Dubow, S. Pasricha, G. Collins, and J. M. Borky, "Considerations for Planning a Multi-Platform Energy Utility System," *Energy Power Eng.*, vol. 9, no. 12, pp. 723–749, 2017.
- [41] J. Cordy, "The Social Media Revolution: Political and Security Implications," *NATO Parliam. Assem.*, no. August, p. 10, 2017.
- [42] M. Khalid, " Cyber Attacks: The Electronic Battlefield" .Doha, QatarArab Center for Research and Policy Studies2013.

# Measuring Quality of E-Learning and Desaire2Learn in the College of Science and Humanities at Alghat, Majmaah University

Abdelmoneim Ali Mohamed

Department of Mathematics, College of Science and Humanities at Alghat, Majmaah University, Majmaah 11952, Saudi Arabia.

Faisal Mohammed Nafie

Department of Computer Science, College of Science and Humanities at Alghat, Majmaah University, Majmaah 11952, Saudi Arabia.

**Abstract**—E-learning and Desaire2Learn (D2L) system were used in several higher education institutions; the learning satisfaction depends on the quality of the system applied to serve this issue and its importance in users mind. Therefore, this study, intended to explore the degree of students and Satisfaction of faculty members with the importance and quality of e-learning used and D2L system as a tool for learning some courses. We took a sample of 57 faculty members and 135 students participated in this study. We used two questionnaires as a tool to collect data from participants, one for faculty members and the other for students; both of these questionnaires had the same idea with different questions. We implemented Statistical Package for Social Science (SPSS) to analyze data. The results show that the Satisfaction of faculty members is high with the quality of e-learning and D2L system as a method of teaching, moderate satisfaction with using D2L tools, the result shows there was a positive relationship between e-learning quality and using D2L tools in teaching. But the result record high satisfaction from students towards the quality of e-learning; the D2L system as a method of learning and the result shows there was no statistically significant effect of gender on the D2L system quality. Finally, the study discussed the implications and recommendations of the work.

**Keywords**—E-learning; Desire2Learn; D2L quality; E-learning quality; learning satisfaction

## I. INTRODUCTION

E-learning is an interactive distance learning system that provides the learner according to demand and relies on an integrated digital electronic environment, aimed at building and delivering courses through electronic networks, counseling, guidance, organizing tests, managing and evaluating sources and processes [1].

Technological development has encouraged the interest of distance education and e-learning to take advantage of its features, where students can attend lectures through smart devices, and this saves time and effort for educational institutions as well as students [2], in addition, to improve the educational product and enhancing quality of the learner [3].

There are many systems used in e-learning and distance education, including the D2L system, which requires

training to the use of its tools. In this system, there were obstacles, namely the lack of skilled trainers, the weakness of the Internet services, the inability of faculty members to use some tools such as online room, and the inability of students to use some tools, all of which require continuous training, flexibility, technical support, and system quality denotes two sides of the information system itself, such as processing speed, ease of use, necessary requirements, and navigability. These are important factors that are the responsibility of the technical group, from the inception of the system to its planning and implementation [4].

The purpose of this study was to determine the level of satisfaction of students and faculty members with distance learning by teaching some online courses using e-learning system D2L, which was used to support traditional teaching method, ready to be widely disseminated, especially on theoretical courses that do not require labs.

## II. LECTURE REVIEW

### A. Introduction to E-Learning

E-learning is defined as “an educational system that uses information technology and computer networks to support and expand the scope of the learning process through a range of media, including computers, the Internet and electronic programs” [5].

E-learning is appealing for many reasons. For most, online courses do not require regular attendance at scheduled lectures. Thus, those working full time or who have other responsibilities are drawn to online courses, as is the case with many nontraditional students. Another convenience of online education is the ability for students to learn materials at their own pace [6].

### B. Distance Education

Distance education has become a component of the educational system in most universities; the usual trend in distance education has been that new technologies have been applied to make the independent study more closely resemble the traditional classroom [7]. This occurred by great development in the field of telecommunications and Internet services. The term ‘distance education’ utilized to define instructional delivery that does not restrict students



to being physically present in the same location as an instructor [8].

C. Systems used

Several systems used to serve this purpose. One of these systems is the Desire2Learn system as a new e-learning platform (Horn, Anne, and Sue Owen, 2011), which is an online education management system designed to help faculty members and students interact in online lectures, use online coursework, as well as activities complementary to regular classroom teaching. Staff members were able to provide course materials, dialogue forums, discussion, short online exams, as well as academic and other resources.

The importance of Desire2Learn lies in providing the Library with the opportunity to deliver its service and support, online, as it does in the library's physical spaces [9].

Majmaah University used the D2L system in the educational processes and allocated some courses for distance teaching, this step found great satisfaction from students and staff members.

D. Objectives

The main objectives of this study were to explore the degree of students and faculty members Satisfaction with the importance and quality of e-learning and Desaire2Learn system used as a tool for learning some courses.

III. METHODOLOGY

In this study test staff members and students level of satisfaction towards distance learning and e-learning system (Desaire2Learn) used in the College of Science and Humanities at Alghat in Majmaah University. Two questionnaires were used as the main instrument to collect data from both to determine their satisfaction towards the importance, quality of distance learning and quality of e-learning, which used (D2L) system. The sample taken from faculty members teaching in the College of Science and Humanities at Alghat and students enrolled in the first semester in 2017. The sample of 57 faculty members and 135 students responded to the questionnaire. The faculty members respondents were 31 male (54.4%) and 26 female (45.6%). The student respondents were 76 male (56.7%) and 59 female (43.3%).

The reliability test conducted by a pilot test of the instrument and Cronbach's Alphas for 30 staff members and 20 students instrument items were 0.89 and 0.93 respectively, which indicates the good internal consistency of the items.

IV. RESULTS

Descriptive statistics were utilized to measure respondents (students & faculty members) scores among the importance of distance learning, quality of e-learning and quality of the Desiare2Learn system.

Firstly: faculty member's satisfaction towards Importance of e-learning based on seven items.

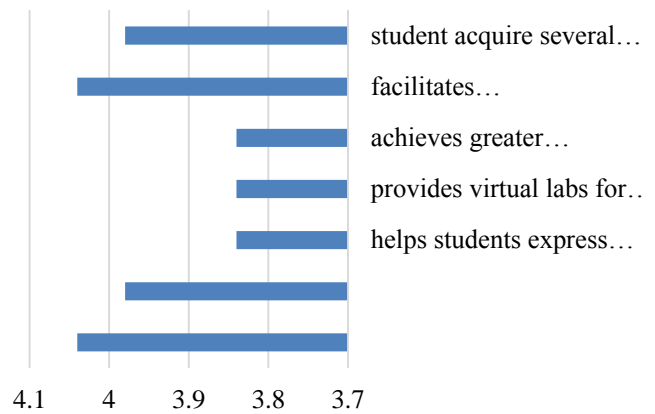


Fig. 1. Descriptive statistics for faculty members Importance in distance learning.

E-learning is a good method of teaching and using modern mechanisms (M=4.04, SD=0.755), increased of electronic communication changed the method of training (M=3.98, SD=0.694), "DL" helps students express themselves in different ways, and gives them easy access to their teachers (M=3.84, SD=0.922) which is earned the lower score of participants, DL provides virtual labs for scientific experiments (M=3.84, SD=0.649), DL facilitates communication between learner and learning resources (M=3.84, SD=0.797), DL facilitates communication between learner and learning resources (M=4.04, SD=0.462) received the highest score for the importance of DL, with DL, the student can acquire several skills (M=3.98, SD=0.719) (see Fig. 1). In all these items, the satisfaction degree on the "Importance in distance learning" is high.

E-learning quality composed of nine items grouped into three variables (see Table I).

The results show that respondent was highly satisfied with "internet services" (M=4.20, SD=0.57), and high satisfaction with both "Improving the educational environment" (M=3.96, SD=0.44) & "Training" (M=3.83, SD=0.43) and respondent were highly satisfied with e-learning quality where the overall score is (M=3.91, SD=0.34).

TABLE I. QUALITY OF E-LEARNING (SATISFACTION OF FACULTY MEMBERS)

Variable	M	SD	Satisfaction
internet services	4.20	0.57	high
Improving the educational environment	3.96	0.44	high
Training	3.83	0.43	high
Quality of e-learning	3.91	0.34	high

Indicator: 1-1.8 very low, 1.81-2.6 low, 2.61-3.4 moderate, 3.41- 4.2 high, 4.21-5 very high

TABLE II. D2L QUALITY (SATISFACTION OF FACULTY MEMBERS)

Variable	M	SD	Satisfaction
flexibility	3.75	0.81	high
Manuals	4.02	0.77	high
Technical support	3.61	0.90	high
Providing training courses	3.96	0.46	high
D2L Quality	3.84	0.55	high

Indicator: 1-1.8 very low, 1.81-2.6 low, 2.61-3.4 moderate, 3.41- 4.2 high, 4.21-5 very high

D2L quality composed of four items (see Table II). The results show that respondent was highly satisfied with the quality of the D2L system in all its variables.

Table III displays the mean and standard deviation scores of faculty member’s usage of D2L tools in teaching. The results show that respondent was moderately satisfied with their using D2L tools in teaching their courses where the overall score is (M=2.92, SD=0.51) by determining each variable of D2L.

The result appeared high satisfied with, “Upload course content and lessons” (M=3.96,SD=0.46) and “Use the Dropbox tool” (M=4.09,SD=1.17), moderate satisfied with “Use the discussion tool” (M=2.79,SD=0.78) and low satisfied with “Use The Group and discussion tool” (M=2.53,SD=0.78), “Use the Quiz tool” (M=2.56,SD=0.95), Using Online Room tool” (M=1.91,SD=0.662) and “Giving lectures remotely” (M=2.92,SD=1.24).

Table IV shows examining the relation between the faculty members using D2L tools in teaching students vs E-learning quality and System quality (D2L). The result show there was a positive relationship e-learning quality and using D2L tools in teaching “(r=0.45, P =0.000), positive relationship system quality (D2L) and using D2L tools in teaching (r=0.40, P =0.002)”.

TABLE III. USING D2L TOOLS (SATISFACTION OF FACULTY MEMBERS)

Variable	M	SD	Satisfaction
Upload course content and lessons	3.96	0.46	high
Use the discussion tool	2.79	0.79	moderate
Use The Group and discussion tool	2.53	0.78	low
Use the Dropbox tool	4.09	1.17	high
Use the Quiz tool	2.56	0.95	low
Using Online Room tool	1.91	0.662	low
Giving lectures remotely	2.58	1.24	low
Uses of D2L tools	2.92	0.51	moderate

Indicator: 1-1.8 very low, 1.81-2.6 low, 2.61-3.4 moderate, 3.41- 4.2 high, 4.21-5 very high

TABLE IV. THE EFFECT OF E-LEARNING QUALITY AND SYSTEM QUALITY (D2L) TOWARDS USING TOOLS (SATISFACTION OF FACULTY MEMBERS) (N=57)

Using satisfaction		
	r	p
E-learning Quality	0.45	0.000
System Quality (D2L)	0.40	0.002

Note. Magnitude: .01 ≥ r ≥ .09 = Negligible, .10 ≥ r ≥ .29 = Low, .30 ≥ r ≥ .49 = Moderate, .50 ≥ r ≥ .69 = Substantial, r ≥ .70 = Very Strong. \*p < .05.

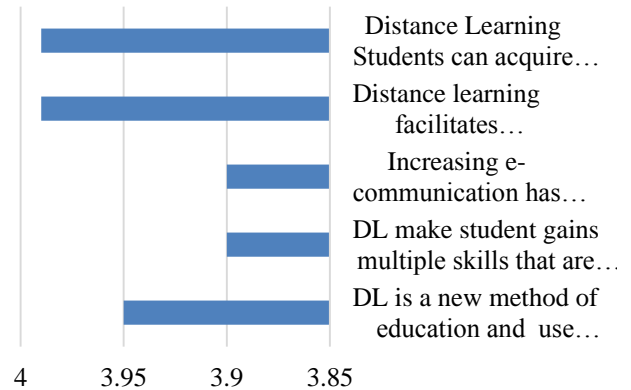


Fig. 2. The importance of distance education in the students' perspective.

Secondly: Students’ satisfaction towards the importance of distance education in the students’ perspective composed of five items (see Fig. 2). All items were of similar average, which indicates to students’ awareness of the importance of distance education and e-learning. The degree of satisfaction is high in all items.

E-learning quality composed of ten items grouped into three variables (see Table V).

The results show that respondent was highly satisfied with “internet services” (M=3.69, SD=0.64), and high satisfaction with both “Improving the educational environment” (M=3.87, SD=0.53) & “Training” (M=3.84, SD=0.59) and respondent were highly satisfied with e-learning quality where the overall score is (M=3.80, SD=0.51).

D2L quality composed of seven items grouped into for variables (see Table VI). The results show that respondent was highly satisfied with the quality of the D2L system in all its variables.

TABLE V. THE QUALITY OF E-LEARNING (STUDENTS’ SATISFACTION)

Variable	M	SD	Satisfaction
internet services	3.69	0.64	high
Improving the educational environment	3.87	0.53	high
Training	3.84	0.59	high
Quality of e-learning	3.80	0.51	high

Indicator: 1-1.8 very low, 1.81-2.6 low, 2.61-3.4 moderate, 3.41- 4.2 high, 4.21-5 very high

TABLE VI. D2L QUALITY (STUDENT'S SATISFACTION)

Variable	M	SD	Satisfaction
flexibility	3.70	0.69	high
Manuals	3.88	0.67	high
Technical support	3.76	0.82	high
Providing training courses	3.91	0.67	high
D2L Quality	3.81	0.54	high

TABLE VII. THE EFFECT OF SYSTEM QUALITY (D2L) TOWARDS USING TOOLS (STUDENT'S SATISFACTION) (N=135)

Quality of e-learning student satisfaction		
	r	p
System Quality (D2L)	0.823	0.000

Table VII shows “there was a high positive correlation between System Quality (D2L) and quality of e-learning student satisfaction,  $r=0.823$ ,  $n=135$ ,  $p=0.000$ ”.

The main effect of gender is the “system quality D2L student satisfaction, (male and female)”. Was found for satisfaction with “Quality of e-learning students satisfaction” was not significant,  $F(1,132)=1.993$ ,  $P<0.16$ .

#### V. CONCLUSION

This study tested the satisfaction of both faculty members and students on the e-learning and Desire2Learn quality. The results reached the following conclusions: First, the results showed the satisfaction of faculty members with the quality of e-learning and its components, which include namely internet services, Improving the educational environment and manuals provided by the Deanship of e-learning at the university, as high satisfies. At the same time, the study shows student satisfaction on the quality of e-learning which includes the same variables “internet services, Improving the educational environment and manuals”, as high satisfies.

The quality of D2L registered high satisfaction in all its variables, namely, flexibility, manuals, technical support and providing training courses for both staff members and student participants. This confirms what was mentioned in the previous study, which concluded that students preferred

Desire2Learn as their technology choice for their online classes [10].

The study showed that the satisfaction of staff members with the use of D2L tools was moderate in general, week specifically in “Quiz tool”, “Online Room”, “Group and discussion tool”, and “Giving lectures remotely” which in the line with of [11] which show that Desire2Learn faculty significantly increased their level of use for all but two of the tools that didn't see a statistically significant increase in faculty usage were the SCORM and “Synchronous Session which corresponded to”, “Online Room”, “Group and discussion tool”.

#### REFERENCES

- [1] Simonson, M., Smaldino, S., & Zvacek, S. M. (Eds.). (2014). Teaching and learning at a distance: Foundations of distance education. IAP.
- [2] Clark, R. C., & Mayer, R. E. (2016). E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. John Wiley & Sons.
- [3] Amasha, M. A., & Alkhalaf, S. (2015). A Model of an E-Learning Web Site for Teaching and Evaluating Online. *ArXiv preprint arXiv: 1501.05578*.
- [4] Machado-Da-Silva, F. N., Meirelles, F. D. S., Filenga, D., & Brugnolo Filho, M. (2014). STUDENT SATISFACTION PROCESS IN VIRTUAL LEARNING SYSTEM: Considerations Based on Information and Service Quality from Brazil's Experience. *Turkish Online Journal of Distance Education*, 15(3).
- [5] Kumar, S. (2014). Ubiquitous smart home system using android application. *arXiv preprint arXiv:1402.2114*.
- [6] Block, A., Udermann, B., Felix, M., Reineke, D., & Murray, S. R. (2006). Achievement and satisfaction in an online versus a traditional health and wellness course (Doctoral dissertation. University of Wisconsin-La Crosse).
- [7] Moore, M. G. (Ed.). (2013). Handbook of distance education. Routledge.
- [8] Riggins, M. E. (2014). Online versus face-to-face biology: A comparison of student transactional distance, approach to learning, and knowledge outcomes. The University of Southern Mississippi.
- [9] Horn, A., & Owen, S. (2011, January). Deakin University Library: an active partner in the implementation of the new generation e-learning platform Desire2Learn. In IATUL 2011: Libraries for an open environment, strategies, technologies, and partnership: Proceedings of the 32nd International Association of Scientific and Technological University Libraries Conference (pp. 1-11). IATUL.
- [10] Chawdhry, A., Karen Pullet, and Daniel Benjamin. "Comparatively assessing the use of Blackboard versus D2L: student perceptions of the online tools." *Issues in Information Systems* 12.2 (2011): 273-280.
- [11] Rucker, R., & Downey, S. (2016). Faculty technology usage resulting from institutional migration to a new learning management system. *Online Journal of Distance Learning Administration*, 19(1), n1.

# TSAN: Backbone Network Architecture for Smart Grid of P.R China

Raheel Ahmed Memon<sup>†\*</sup>, Jianping Li<sup>†</sup>

<sup>†</sup>School of Computer Science & Technology  
University of Electronic Science and Technology China,  
Chengdu, Sichuan Province, China

\*Department of Computer Science, Sukkur IBA University  
Sukkur, Sindh, Pakistan

Anwar Ahmed Memon

Department of Electrical Engineering,  
Mehran University of Engineering & Technology  
Jamshoro, Pakistan

Junaid Ahmed

School of Automation  
University of Electronic Science and Technology China  
Chengdu, Sichuan Province, China.

Muhammad Irshad Nazeer, Muhammad Ismail

Department of Computer Science,  
Sukkur IBA University  
Airport Road, Sukkur, Sindh,  
Pakistan

**Abstract**—Network architecture of any real-time system must be robust enough to absorb several network failures and still work smoothly. Smart Grid Network is one of those big networks that should be considered and designed carefully because of its dependencies. There are several hybrid approaches that have been proposed using wireless and wired technologies by involving SDH/SONNET as a backbone network, but all technologies have their own limitations and can't be utilized due to various factors. In this paper, we propose a fiber optic based Gigabit Ethernet (1000BASE-ZX) network named as Territory Substation Area Network (T-SAN) for smart grid backbone architecture. It is a scalable architecture, with several desired features, like higher coverage, fault tolerance, robustness, reliability, and maximum availability. The use case of sample mapping the T-SAN on the map of People Republic of China proves its strength to become backhaul network of any territory or country, the results of implemented architecture and its protocol for fault detection and recovery reveals the ability of system survival under several random, multiple and simultaneous faults efficiently.

**Keywords**—Smart Grid; TSAN; 1000BASE-ZX ethernet; backbone architecture

## I. INTRODUCTION

The generated data from different sources in a smart grid system is enormous. This data might contain meter readings, real-time price updates, sensor data or other control information. To enable the smart substation system for the exchange of this much huge data is the most critical part of communication infrastructure in smart substations architecture. Though there is no de facto networking standard of smart grid available [1]. The current implementation has the hybrid approaches for consideration, in which each technology carries its own weakness and strengths. Thus an entirely new networking system is needed for interconnected substations [1]. The communication network of an interconnected substation system demands few essential features, such as Reliability, Acceptable Response delay, Scalability, Fault Tolerance, high availability, Wide coverage, and Security.

## A. Reliability

In real-time systems, the communication linkage over a wide area should be based on reliable backbone architecture to enable the timely exchange of messages and commands between the nodes. There could be a number of reasons why a network fails to achieve time critical communication such as; time-out when message delivery delayed because of fault detection and recovery process took longer to resolve, assembling delay, failure of routing protocol and resources failure when any of the responsible hardware resources such as links and other communication devices encounters a physical failures.

## B. Acceptable Response Delay

According to IEC 61850 standards, there are three kinds of control messages exchanged in smart grid communication; Generic Object Oriented Substation Event (GOOSE) deals with critical information such as warning and control signals. Manufacturing Messaging Specification (MMS) used to transfer the substation status information, and Sample Measured Value (SMV) transfer power line current and voltage values measures. GOOSE and SMV are time critical messages need to arrive in < 4ms [2].

## C. Scalability

As per statistics of shared public data by World Bank the power consumption in China has grown from the Year 2004 to 2014 from 1585.83 kWh to 3927.04 kWh, Fig. 1 shows the increasing trend of electric power consumption in last 10 years, this growth would be increasing with increase of population and growth of electronic/electrical devices, private power generation plants and other consumption and generation resources in coming future. This huge growth in the field of smart grids needs to have a scalable architecture to adjust and merge new changes flawlessly without disturbing the existing network.

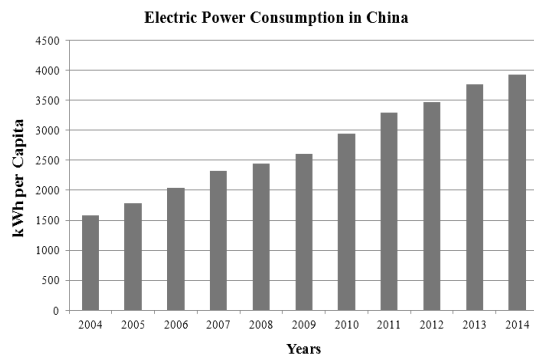


Fig. 1. Growth of electric power consumption of China.

#### D. Fault tolerance

The backbone shall have the ability to tolerate each kind of rising faults in the network, and these faults can be broadly categorized into two: Hardware faults and software faults; Hardware faults are those in which hardware equipment are involved like Network Interface Cards, Routers, Switches, network cables, loose connections, and broken/dead systems, Software faults are; congestion on one line, drop of packets due to interference and suspended systems [3], [4]. A system should be able to cope up both kinds of hardware and software faults.

#### E. High Availability

It is essential for smart grid implementation to ensure the availability of power and communication to the consumers, especially when dealing with issues like latency and security [5]. And the availability is dependent on the robust connectivity of both power and communication networks and timely data exchange between consumer and service provider entity.

#### F. Wide Coverage

The power system doesn't require serving a particular place, an entire country or territory should be covered by a Wide Area Network to serve the scattered locations of the power grid systems. Wide Area Network should be able to bring the real-time measurement data from substations/customers to the Control Centers situated at distances.

#### G. Security

Wide area of the electrical system can be scattered over 1000 of miles, thus the communication system should be able to protect itself from potential attacks in the cyber-physical world. Communication mediums like Wireless technology are the most vulnerable candidates to consider for wide area network because of security.

Thus a backhaul network with large data rate is required which satisfies all the stated requirements of Interconnected Substation System and it should also be a cost-effective solution in implementation and maintenance. Currently the implemented system is using a hybrid approach of wireless and wired technologies for providing this interconnection on the backbone of SDH/SONET, these are not available as a standard technology neither they are satisfying all the

requirements of the smart grid system, Section 2 further covers the details of all the available technologies and their up and downs. In this paper there are two major contributions to consider:

The first one is proposed novel network backbone architecture named as TSAN, it is an inspiration of our already proposed Recursive Scalable Autonomous Fault-tolerant Ethernet (RSAFE) architecture for mission-critical systems [6]. The difference in RSAFE and TSAN is that it involves fiber optic communication, implemented on wide range, and the fault-tolerant protocol is enhanced in TSAN than RSAFE. which is based on Gigabit Ethernet standard 1000BASE-ZX (also referred as GigE) transmission using 1550 nm wavelength ranges the distance of 70 km (at least, while some vendors claim the maximum range of 120 km) over single mode fiber optic cable. The proposed backbone architecture is a hybrid approach of star, mesh and ring topologies. It divides the overall land into manageable Regions and Zones despite considering the boundaries of geographical division (town, city, district, province, etc.) assigned by a territory or country. A zone connects all the small and large substations and control centers in the same network using dual star topology with two Ethernet layer 3 switches/routers. Then two or more than two zones combined in a dual ring topology to form a region; similarly, all the parts of a country are divided into zones and regions and then those regional networks are merged with each other using neighboring zones of respective regions. All the nodes in the proposed network architecture are dually connected, where one link serves as the primary path and other as a standby path. This well-connected hybrid topology aims to provide robustness, reliability, scalability, and wide coverage and high availability to the system.

The second major contribution is Fault Tolerant Ethernet Protocol (FTEP), implementation of FTEP is based on two modules fault detection and fault recovery. In first module, all the substations within a zone send an Aliveness Beat Message (ABM) to their respective control centers for indicating their active path. Each ABM is dispatched at the fixed interval of 2ms from all the substations to the control center. The control centers besides communicating the operational commands also update the routing table according to received network status and forward the changes to the neighboring zones. The second module is executed if any fault has been detected by first module (missing of 2 consecutive ABMs), as discussed in proposed architecture above, there are two links to reach a node in a zone, one is primary (operational) and the other one is secondary (standby), fault recovery module enables the standby link and update the link information by making standby path as default communication path in routing table. The results of proposed backbone architecture and protocol demonstrate the ability of the TSAN to fulfill all the needs and provide a basic and complete suite of solution for Smart Grid backbone network.

Rest of the paper is organized as follows. Section II presents the available technologies for WAN in Smart Grids, Section III describes the proposed TSAN for Smart Grids, Section IV is about experimental setup, Section V presents the results and Section VI concludes.

## II. CANDIDATE TECHNOLOGIES FOR WAN IN SMARTGRIDS

Currently, the smart grid Wide Area Network implementation has eight main technologies for consideration: Power Line Communication (PLC), Digital Subscriber Line, Wireless Mesh, WiMAX, Cellular, Space Communication (Satellite Communication), SDH/SONET and Gigabit Ethernet. These all technologies have their strengths and weaknesses as shown in Table I.

### H. Digital Subscriber Line

The Digital Subscriber Line or DSL have three systems to offer ADSL, HDSL, and VDSL depending on usage it may offer different range and data rates. But due to its reliability and downtime issue, it couldn't be considered for the backbone of huge and real-time communication architecture.

#### I. Power Line Communication

This technology could be a good candidate as its network already exist, but its limitation to transfer of signal across the transformer raises the concern also it experiences heavy noise, interference and involve insecurity issue [7]–[13].

#### J. Wireless Mesh

It's a good candidate; provide a mesh of a multi-hop wireless network. Has a good data rate, can be implemented using 802.11, 802.15 and 802.16. But like almost all other wireless technologies this has also a problem with interference and suffer from noise [7].

#### K. WiMAX

Worldwide interoperability for microwave access is a 4G wireless technology dedicated to the advancement of IEEE 802.16 and series of standards for Metropolitan Area network such as IEEE 802.16-2004, 802.16e. It operates on both licensed and unlicensed frequency bands. However, it's an expensive implementation on a wide range and also requires high power consumption. Furthermore, unfavorable weather conditions can also affect this technology [14], [15].

#### L. Cellular

It's a mobile communication technology; works on radio signals, the network is made up of several radio cells. One of the main advantages of this technology is that its infrastructure is already available [14]. The downside is that the consumer equipment up gradation cost is high and the network is shared with mobile customers, it's not only a security concern but congestion is also expected [16].

#### M. Space Communication or Satellite Communication

The satellite communication is able to provide global coverage even in the rural areas of the country with a data rate of 1 Mbps. It can provide GPS based satellite monitoring and synchronization of any site. It's a cost-effective solution but the limitation of this implementation is that it suffers severely from the weather conditions, which may lead to long round trips [17].

#### N. SDH/SONET

Its large data rate and range convinced fiber optic to be a backbone cable. The up gradation and installing new network could be expensive but the quality of service delivered by fiber

optic is better than all other existing technologies, moreover, it is immune to noises and used in long run [7]. The good side is SDH/SONET provides large and managed network, simple topology, and the resilient ring, which is the most liked and its widely used option. On the downside, the dynamic IP traffic is not optimized here, configured with fixed P2P bandwidth, so the allocated bandwidth is not efficiently utilized and always the unused bandwidth is wasted. Limited topology options like P2P, Linear, and Ring. Inefficient in transferring the multicast traffic, the implementation of the Ring is up to maximum 16 Nodes with coverage of 1200km. SDH/SONET doesn't cover all over the transmission and distribution line communication because of its limited coverage. Thus it may not work as an individual network to cover every corner of a country, the smart grid network would have to depend on any of above-mentioned technologies to gather the data from all areas and send it on to the SDH/SONET backbone network [18].

#### O. Gigabit Ethernet

The Ethernet 1000BASE-ZX Standard is Gigabit over single-mode fiber cable with a wavelength of up to 1550nm, which is able to cover 70 Km distance [19]. The limitation of this technology is that it requires entirely new-dedicated cable installation. In comparison with SONET/SDH 1000BASE-ZX could be a good option in several ways: Efficient utilization of bandwidth in P2P and mesh [20]. Furthermore, it is highly scalable, and no need to provide conversion for synchronization from different wires, cost-effective, mesh topology offers exceptional utilization of bandwidth.

It is clear from comparison given in Table I and discussing pros and cons of PLC, DSL, Wireless Mesh, WiMAX, Cellular, Satellite and SDH/SONET the Gigabit Ethernet could be the potential technology, to provide vast and efficient coverage with good data rate, less interference, less CapEx and OpEx, and moreover an entirely dedicated network.

## III. PROPOSED TSAN ARCHITECTURE

In a country, there could be thousands of substations that send and receive information updates every time to each other using communication network. Thus the backbone communication network of a system is an important part to pay attention and develop the strategic network that should be able to survive longer, stronger, secure and scalable. This paper aims to propose network design architecture and an implementation of a fault tolerant protocol that results to achieve a network according to the needs of smart grids.

### A. Network Architecture

The proposed design is a Gigabit Ethernet network with a combination of mainly three topologies (star, mesh, and ring) this network is providing reliable communication architecture, which can work well with multiple failures, the TSAN divides the overall territory into manageable zones and regions despite considering the boundaries of geographical division such as; town, city, district or a province assigned by a country government. The geographical divisions are named as Zonal Substation Area Network (ZSAN), Regional Substation Area Network (RSAN) and the combination of these two; Territory Substation Area Network (TSAN).

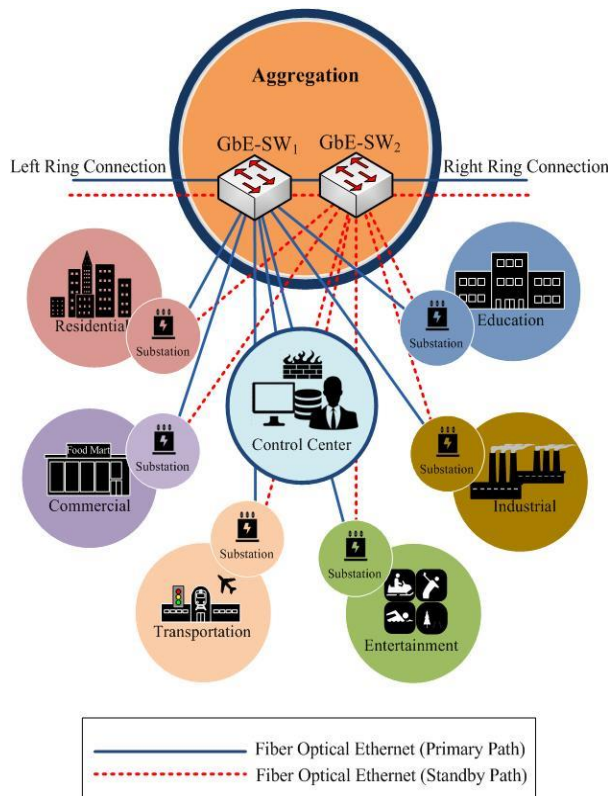


Fig. 2. Proposed substation area network.

**B. Zonal Substation Area Network (ZSAN)**

One Zone is considered as a locality with multiple substations connected together and reporting the locally available control center(s), their communication network is a combination topology of star and mesh, where each substation serves a particular area feeders and report to the dedicated control center. As shown in Fig. 2, a ZSAN provides a primary path the connection in blue and a standby path dotted red connection. Both are used to ensure reliable communication within the network. The standby path here is used only to rescue the condition of failures in a communication network. One ZSAN can connect maximum 64 nodes including large and small substations and the control centers.

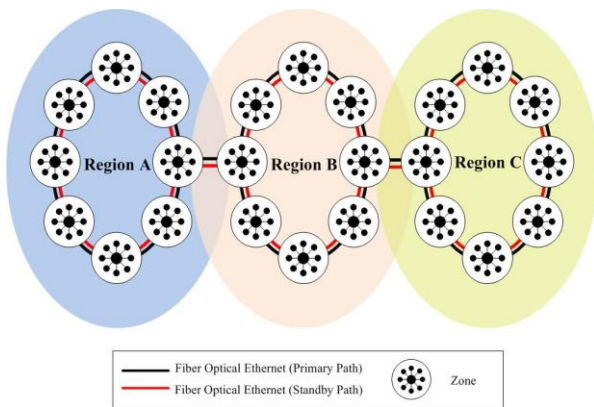


Fig. 3. Proposed territory area network.

**C. Regional and Territory Substation Area Network (RSAN and TSAN)**

As shown in Fig. 3, a region is a huge area connecting different zones in a dual ring network like SDH/SONET; it is named as Regional Substation Area Network (RSAN). Fig. 3 shows three regions, namely, Region A, B, and C; all of these regions combining to form a Territory Substation Area Network (TSAN). The connection in blue is primary communication path and red is standby path. The redundant rings ensure the intra-region connectivity and redundant linkage with other regions are for reliable inter-region connectivity, where, the size and the number of regions depending on country’s geographical features, size, and location.

The proposed network design architecture has several advantages over SDH/SONET:

- a) The architecture is entirely based on Ethernet, so it’s not needed to convert to another communication protocol
- b) It provides a combo topology (Mesh, Star, and Ring), which makes it more reliable than SONET/SDH, and less expensive (in terms of wastage of bandwidth) which is a common issue of SDH/SONET.
- c) It uses COTS products, thus no proprietary hardware needed.

The Fault detection and recovery protocol suit provide the desired results according to IEC 61850 standards.

**D. Fault Tolerant Ethernet (FTE) Protocol**

The FTE protocol is implemented using Aliveness Beat Messages (ABM), where an ABM is a lightweight Ethernet frame periodically sent by all substations to the dedicated Control Centers in their respective zones for indicating their active connections; on the other hand all the control centers receive ABM and maintain their own copy of updated list of routing table. In a zone there could be more than one control centers, thus one of them is selected as a primary control center (PCC), which perform some additional responsibility of communicating their routing table with the PCC of the zones (left and right zones) in the region as shown in Fig. 2..

The FTE protocol working depends on the recipient of ABMs from the all substations in a network to their control centers. We can say that FTE Protocol is divided into 2 different working modules. One is Fault Detection another is Fault Recovery. As shown in Fig. 4(a) where one substation monitors its connections and send ABM to its control center, if any change has taken place in connection information, it will be updated there. The second module of fault recovery is as depicted in Fig. 4(b), which shows a control center is receiving ABMs. In a zone each substation is responsible to send an ABM over a fixed interval of 20ms to their dedicated control center, and if two consecutive ABMs are missed the link would be considered as down and it will change the primary link information with standby link and convey the updated information with all the substations in its own zone and other zones of the same region.

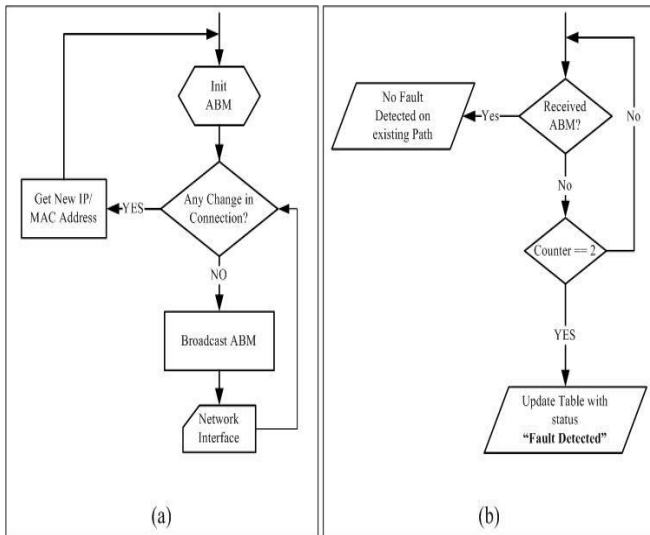


Fig. 4. Sending and receiving of aliveness beat message.

The suggested TSAN and the FTE routing protocol satisfies the requirements of the WAN networks in Smart Grids:

- 1) **Reliability:** Dually Connected, multipath architecture, so even after random multiple failures a substation can still communicate with rest of the world using alternative standby path(s)
- 2) **Scalability:** At any instance of time the new substation or an entire new zone can be added without interrupting the existing network.
- 3) **Fault Tolerance:** The proposed protocol provides efficient recovery from faults within specified time limits.
- 4) **High availability:** This fault detection and recovery and dually connected network provides high availability of the communication network, and have the ability to adjust with a large number of random faults occurrence, which can be a helpful strategy in disaster situations.
- 5) **Wide coverage:** The implementation of proposed scheme is exemplified by mapping the TSAN on the map of the Republic of China in next section.
- 6) **Security:** The dedicated network and firewall protection provide maximum security from unidentified intrusions, which is not possible with other shared networks or any of Wireless technologies.

*E. Use Case of TSAN - People’s Republic of China*

The sample mapping is shown in Fig. 5. Depicting the connectivity example of smart grid substation area network, there are 39 zones where each zone may have maximum connectivity option of 64 substations.

In current scenario, if every subnet consists of 64 substations then this overall network is comprised of 2496 large and small substations. For security each zone have their own firewalls to provide intra and inter-network protection. Fig. 6 shows the clean backbone connectivity rings throughout the country for providing communication network in every corner of the country.

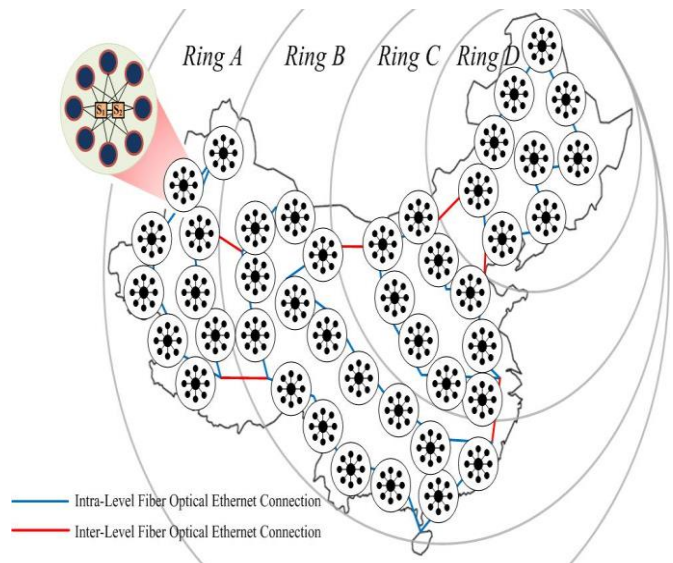


Fig. 5. Sample mapping of TSAN on China map.

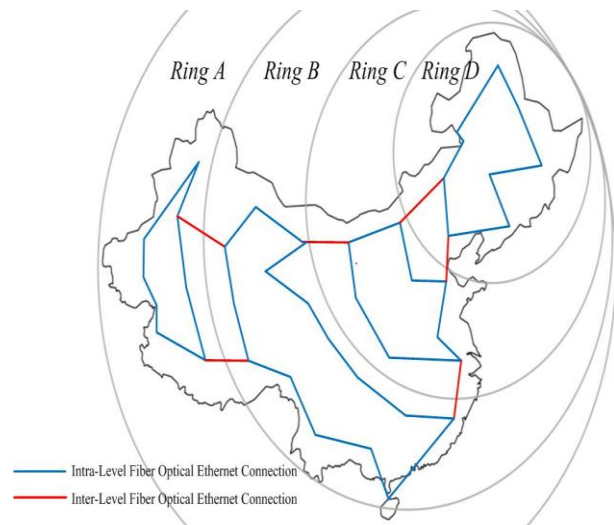


Fig. 6. Ethernet connectivity rings throughout the country.

**IV. EXPERIMENTAL SETUP**

The experiment is done by creating a flatbed setup in a lab, where 12 computers forming 3 regional subnets and connected using Fast Ethernet. The regional subnet connects the entire substation dually in a network, and periodically an ABM is sent via each substation to their dedicated control center for showing aliveness of a link.

As shown in Fig. 7, A Region consists of 4 nodes (3 substations and 1 control center) in this scenario, and 2 switches. Where every node is having 2 NICs named as Eth-A and Eth-B and connects to Switch-A and Switch-B respectively in their zone. The control center besides performing its controlling task is also additionally made responsible to collect Aliveness Beat Messages from the all substations and communicate the status updates with other regions by broadcasting a periodic status updates. Table I gives the details of hardware and software components used in experiment.



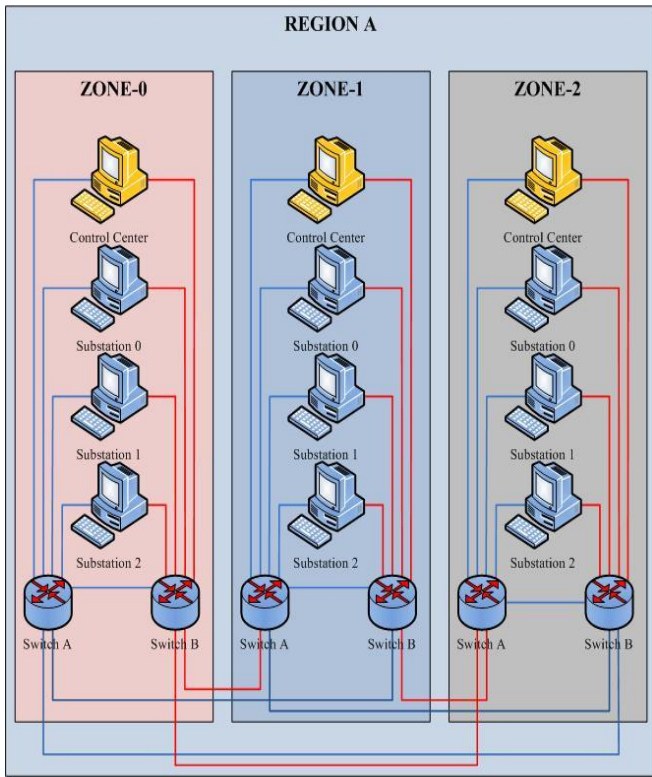


Fig. 7. Experimental setup.

TABLE I. SPECIFICATIONS OF EXPERIMENTAL SETUP

	Components	Description
Hardware Components	Computers	3 x 2.4 GHz Dual Core, 2 GB, 80 GB (Control Centers)
		9 x P-4, 2.8 GHZ Dual Core, 2 GB, 80 GB (Substations)
		Operating System Ubuntu 16.04.2 LTS
	Switch 10/100 BASE-TX	6 x D-Link DES-1226G
	Network Interface Cards	24 x 10/100 Ethernet NICs
Connections	Fast Ethernet	
Software components	Aliveness Beat Message	Initialization of ABM module installed on each substation
	Fault Detection & Recovery	Receiving of ABM and updating routing table modules installed on control centers

V. RESULTS

Several experiments were conducted to validate the functionality of the suggested scheme TSAN. Figure 8, shows our theoretical assumptions of aliveness message and fault detection which is unrealistic when it comes into action as shown in Fig. 9 shows the spikes as variations in number of aliveness messages that reflects the actual time of fault occurrence.

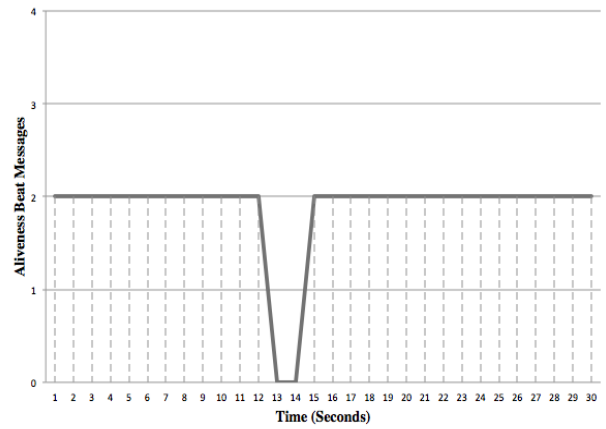


Fig. 8. Theoretical assumption of aliveness beat message.

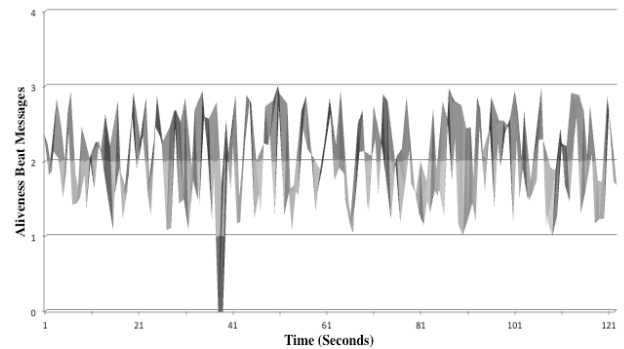


Fig. 9. Theoretical Assumption of Aliveness Beat Message.

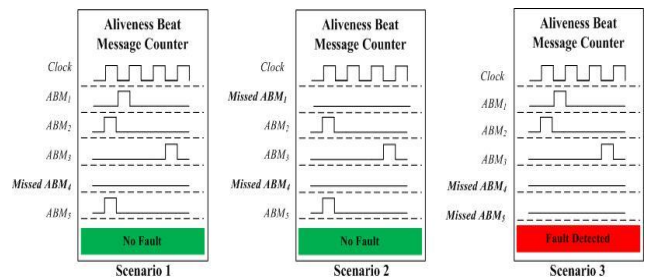


Fig. 10. Different scenarios of aliveness beat messages.

As shown in Scenario 1, Fig. 10 the ideal situation is that every substation sends an aliveness message exactly at the interval of every 2 seconds to the control center and if there is a link or device or any kind of failure occurred somewhere it should take 4.4 seconds to detect and recover that fault.

But practically it varies as Scenario 2 and 3 of Fig. 10, the Substation A sends aliveness message to Control Center and got a failure, but the detection took longer than expected time but it cannot exceed the time of three heartbeat messages. Equation 1 defines overall latency in the network.

$$Time_{TL} = [(F_{Gap} + F_{Size}) \times N] + SW_{Latency} + MD_{Latency} \quad (1)$$

In (1),  $Time_{TL}$  is total latency,  $F_{Gap}$  is the time interval between frames (ideal time is 2ms),  $F_{size}$ , is size of the frame,  $N$

number of nodes in a zone,  $SW_{Latency}$  is the operational latency of switch,  $MD_{Latency}$  latency involved by used cable medias.

For checking the sustainability of our proposed system we have created 8 different link failures on different time intervals as shown in Fig. 11, these multiple random failures are recovered in a timely manner and the network is made again stable before introducing new fault. As mentioned above, this experiment involves 9 substation PCs and 3 Control Centers PCs, not only substations but also control centers may encounter physical faults in a network, so every substation sends aliveness message to their control center and control center exchanges the status updates with other networks, that exchange is considered as the aliveness signal of a regional area network.

Thus the result in Fig. 11 contains 0 to 8 substations and 9 to 11 control centers. From these 8 failures, 3 failures were control center failures and 5 are the substation failures. And after each fault, the linkage information is successfully updated within a limited time. From (1), three main parameters  $N$ ,  $SW_{Latency}$  and  $MD_{Latency}$  are responsible for causing a delay in communication; the resulting spikes can be seen in Fig. 11.

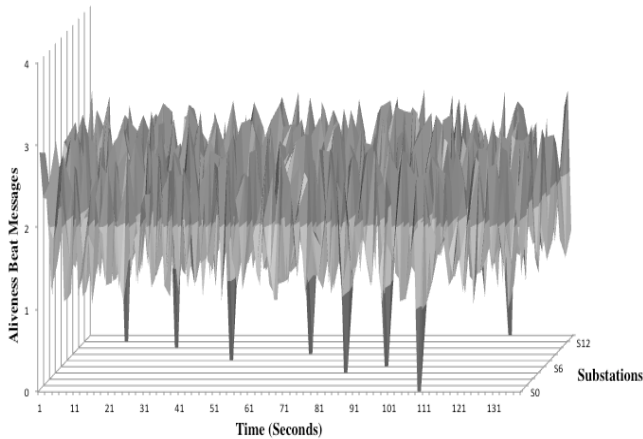


Fig. 11. Multiple random link failures at different time intervals.

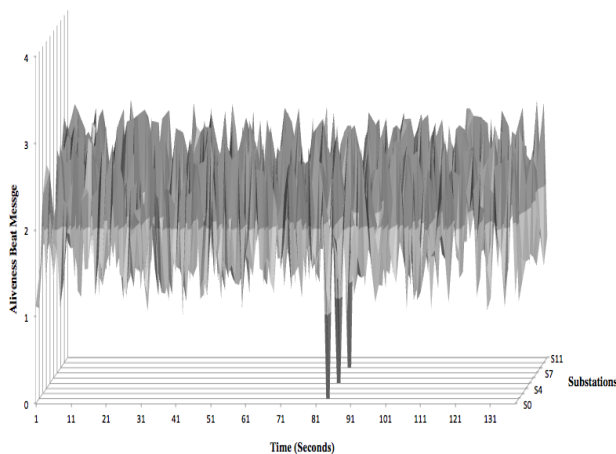


Fig. 12. Simultaneous link failures and successful recovery.

And the results shown in Fig. 12 are presenting the simultaneous multiple failures; these are detected and recovered exactly in the same manner as random faults. Every node is still made approachable in the network in the very short duration of time. This proves that our system can tolerate limited multiple and simultaneous link failures, and still, the network is completely connected and approachable from everywhere.

## VI. CONCLUSION AND FUTURE WORK

We implemented Ethernet-based network architecture for Territory Area Network of Smart Grid Equipment and a protocol for detection and recovery of physical failures in the network system. The beauty of using Ethernet is it doesn't require extra effort of conversion of the protocol stack during transmission because the Territory Area Network is entirely implemented using Ethernet. We modeled the approach on the map of China considering 2496 Substations and control centers throughout the country to show the applicability of the scheme in big countries. The validity of proposed scheme and protocol is supported by a flatbed experiment, which has 3 zones, each zone with 3 substations and 1 control center in it. The experimental result shows that the system is able to detect and recover faults within acceptable time delay. The system is also able to handle multiple random and simultaneous faults within a specified limit, which is quite sufficient for supporting during rescuing condition of Smart Grid system. We also plan to work on our proposed protocol for reducing the fault detection and fault recovery time in future.

## ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61370073), the National High Technology Research and Development Program of China (Grant No. 2007AA01Z423), the project of Science and Technology Department of Sichuan Province.

## REFERENCES

- [1] D. Al Abri, A. S. Malik, M. Albadi, Y. Charabi, and N. Hosseinzadeh, "Smart Grid," in Handbook of Climate Change Mitigation and Adaptation, Cham: Springer International Publishing, 2017, pp. 1465–1501.
- [2] F. Gianaroli, A. Barbieri, F. Pancaldi, A. Mazzanti, and G. M. Vitetta, "A novel approach to power-line channel modeling," IEEE Trans. Power Deliv., vol. 25, no. 1, pp. 132–140, 2010.
- [3] D. Wu, S. Member, and C. Zhou, "Fault-Tolerant and Scalable Key Management for Smart Grid," IEEE Trans. Smart Grid, vol. 2, no. 2, pp. 375–381, 2011.
- [4] D. Wu and C. Zhou, "Fault-tolerant and scalable key management for smart grid," IEEE Trans. Smart Grid, vol. 2, no. 2, pp. 375–381, 2011.
- [5] C. H. Lo and N. Ansari, "The progressive smart grid system from both power and communications aspects," IEEE Commun. Surv. Tutorials, vol. 14, no. 3, pp. 799–821, 2012.
- [6] R. A. Memon, J. P. Li, and F. Shah, "Autonomous fault detection and recovery system in large-scale networks," in 2016 13th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2016, pp. 285–288.
- [7] M. Kuzlu and M. Pipattanasomporn, "Assessment of communication technologies and network requirements for different smart grid applications," in 2013 IEEE PES Innovative Smart Grid Technologies Conference, ISGT 2013, 2013.
- [8] N. Ginot, M. A. Mannah, C. Batard, and M. Machmoum, "Application Of Power Line Communication For Data Transmission Over Pwm Network," Tsg, vol. 1, no. 2, pp. 178–185, 2010.

- [9] C. Konaté, A. Kosonen, J. Ahola, M. Machmoum, and J. F. Diouris, "Power line communication in motor cables of inverter-fed electric drives," *IEEE Trans. Power Deliv.*, vol. 25, no. 1, pp. 125–131, 2010.
- [10] A. Kosonen and J. Ahola, "Communication concept for sensors at an inverter-fed electric motor utilizing power-line communication and energy harvesting," *IEEE Trans. Power Deliv.*, vol. 25, no. 4, pp. 2406–2413, 2010.
- [11] Z. Marijic, Z. Ilic, and A. Bazant, "Fixed-Data-Rate Power Minimization Algorithm for OFDM-Based Power-Line Communication Networks," *IEEE Trans. Power Deliv.*, vol. 25, no. 1, pp. 141–149, Jan. 2010.
- [12] N. Andreadou and F. N. Pavlidou, "Modeling the noise on the OFDM power-line communications system," *IEEE Trans. Power Deliv.*, vol. 25, no. 1, pp. 150–157, 2010.
- [13] J. Zhang and J. L. Meng, "Robust Narrowband Interference Rejection for Power-Line Communication Systems Using IS-OFDM," *IEEE Trans. Power Deliv.*, vol. 25, no. 2, pp. 680–692, 2010.
- [14] A. Usman and S. H. Shami, "Evolution of communication technologies for smart grid applications," *Renewable and Sustainable Energy Reviews*, vol. 19, pp. 191–199, 2013.
- [15] F. Z. Yousaf, K. Daniel, and C. Wietfeld, "Performance evaluation of IEEE 802.16 WiMAX link with respect to higher layer protocols," in *Proceedings of 4th IEEE International Symposium on Wireless Communication Systems 2007, ISWCS, 2007*, pp. 180–184.
- [16] C. Gungor et al., "Smart Grid Technologies: Communication Technologies and Standards," *Ind. Informatics, IEEE Trans.*, vol. 7, no. 4, pp. 529–539, 2011.
- [17] M. Kuzlu, M. Pipattanasomporn, and S. Rahman, "Communication network requirements for major smart grid applications in HAN, NAN and WAN," *Comput. Networks*, vol. 67, pp. 74–88, 2014.
- [18] C. Huang, "Studies of Uncertainties in Smart Grid: Wind Power Generation and Wide-Area Communication," University of Tennessee, Knoxville, 2016.
- [19] F. R. De Souza and M. R. N. Ribeiro, "An optical performance monitoring method for Carrier Ethernet networks using oam continuity check messages," *Photonic Netw. Commun.*, vol. 23, no. 1, pp. 74–82, 2012.
- [20] Y. Katsuyama, M. Hashimoto, ... A. U.-I. S. on, and U. 2008, "Proposal of Bi-Directional ROADM for use in Regional IP-over-CWDM Networks," *Int. Symp. Comput. Networks*, pp. 148–154, 2008.

# Data Synchronization Model for Heterogeneous Mobile Databases and Server-side Database

<sup>1</sup>Abdullahi Abubakar Imam, <sup>2</sup>Shuib Basri, <sup>3</sup>Rohiza Ahmad, <sup>4</sup>Abdul Rehman Gilal

<sup>1,2,3</sup>Department of Computer and Information Sciences, Universiti Teknologi PETRONAS Malaysia

<sup>1</sup>Computer Science Department, Ahmadu Bello University, Zaria-Nigeria

<sup>4</sup>Department of Computer Science, Sukkur IBA University, Pakistan

**Abstract**—Mobile devices, because they can be used to access corporate information anytime anywhere, have recently received considerable attention, and several research efforts have been tailored towards addressing data synchronization problems. However, the solutions are either vendor specific or homogeneous in nature. This paper proposed Heterogeneous Mobile Database Synchronization Model (HMDSM) to enable all mobile databases (regardless of their individual differences) and participate in any data synchronization process. To accomplish this, an experimental approach (exploratory and confirmatory) was employed. Also existing models and algorithms are classified, protracted and applied. All database peculiar information, such as trigger, timestamp and meta-data are eliminated. A listener is added to listen to any operation performed from either side. To prove its performance, the proposed model underwent rigorous experimentation and testing. X<sup>2</sup> test was used to analyze the data generated from the experiment. Results show the feasibility of having an approach which can handle database synchronization between heterogeneous mobile databases and the server. The proposed model does not only prove its generic nature to all mobile databases but also reduces the use of mobile resources; thus suitable for mobile devices with low computing power to proficiently process large amount of data.

**Keywords**—Heterogeneous databases; data synchronization; mobile databases; mobile devices; NoSQL database; relational databases

## I. INTRODUCTION

Heterogeneity of mobile databases, complexity in mobile applications development [1] as well as the mobile devices themselves has engineered several obstructions in data synchronization. Data Synchronization (DS) can be defined as record exchange between two different databases [2]. It is the system that establishes the movement of data between the mobile device and the server-side databases [3]. On the other hand, A heterogeneous database is an automated (or semi-automated) system that has disparate data model for the local nodes, Operating System, DBMS to present user with a single, unified query interface [4], [5]. Based on this, numerous works have been conducted to address the DS concerns with different techniques. Amongst them are [6], [7] who proposed Synch Algorithm using Message Digest (SAMD) and [8] who introduced a stateful DS, all for the purpose of minimizing the load on the mobile devices. Also, [9] suggested a target based algorithm which always initiate the synchronization process from the target database.

In this paper we consider a variation of the DS problem with slightly different approach from the above. It focuses on the heterogeneity concept where several databases (regardless of their individual differences) connect and exchange data seamlessly. These differences do not stop at only the database versions or vendors but also different DBMS and data model.

At first the approach eliminates the use of any database dependent information such as timestamp, trigger and meta-data. It also pushed the highest percentage of operation (computations) to the server for calculations and conflict resolution, thus relieving the mobile devices. In addition, JSON technology was considered for data packaging and transfer as a flat file which has no bond to any mobile database. Moreover, the synchronization process is always initiated by the mobile device. This is because mobile devices cannot stay connected to the network all the time so, the server cannot know which device is online before engaging on any synchronization process. On the side of starting the synch event, anything in the mobile device can be set to trigger the synchronization event such as on-boot-up, on-button-click, and on-application-start. It is worthy to mention here emphatically and unequivocally that our approach is flexible, customizable and extendable, it is not close-ended solution, rather, it gives a blue print on how to setup a synchronization environment for the heterogeneous mobile databases and server-side database. The approach adopts the use of message digest to encode messages before transmission and decode upon arrival at the destination.

Message Digest (MD) is also called cryptographic hash, hashes or hash function [10]. It takes a message as input and produces a fixed-length output. The output is normally smaller in size than the input (original message) which is generally referred to as message digest, fingerprint or hash value [11].

The rest of the paper is structured as follows. Section II discusses the related works. Section III explains the adopted method. In Section IV, the proposed model is presented and elucidated. The results are discussed in Section V. Finally, Section VI concludes the paper with future focus.

## II. RELATED WORKS

As said in the introduction of this paper that, Data Synchronization is a record exchange between two different databases or coherently keeping replicated copies of a data-set [2]. Database synchronization on the other hand, can be a one-way or a two-way process, and can be real time or periodic mode, namely, *Synchronous and Asynchronous* [12]. Based on

these, varieties of data synchronization solutions are provided to enable mobile device databases seamlessly communicate with the server database. Some of these solutions are discussed below, starting with factors that negatively affect the synchronization process.

Since synchronization process occur frequently for mobile devices that house variety of unlike databases with dissimilar data-models and also have a number of limitations such as storage space and processing capacity, the factors that influence the processing speed, allow conflicts, as well as prevent solution generalization in terms of database vendors need to be carefully explored and addressed. Therefore, we retrieved and compartmentalized factors from existing works which we are believed to have significant negative impact to an effective synchronization. The factors and their dependencies are illustrated in Fig. 1.

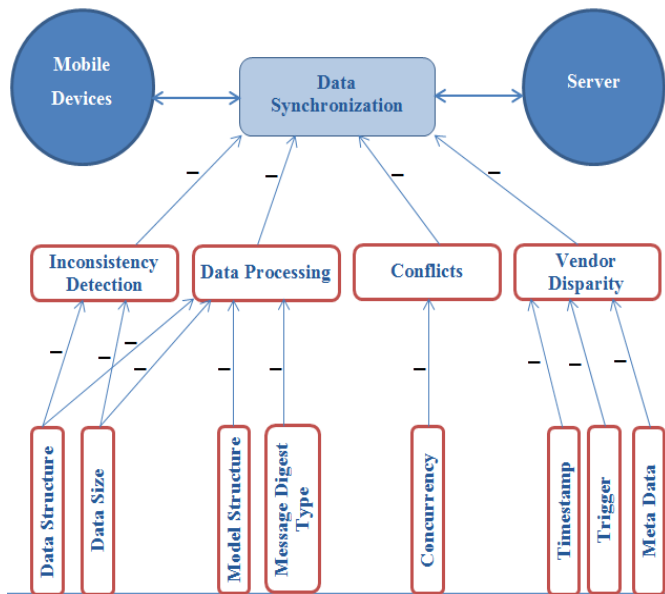


Fig. 1. Factors influencing data synchronization.

There are three layers in the figure above (Fig. 1). Starting from the bottom, the lowermost layer contains the factors that directly influence the main factors shown in the second layer, which in the end persuade or induce data synchronization process as a whole. Looking at the factors above, it is believed that, level two (second layer) has the potential to directly affect, negatively, the data synchronization process. Several approaches have been provided to subside the effect of these factors. Based on the scope of this study, only the factors that affect mobile database heterogeneity will be our focus of this research. The approaches that focus on the same or related concept are painstakingly selected and discussed below in accordance with the main factors (level 2, Fig. 1. above).

Referring to vendor disparity factor (in Fig. 1), in distributed databases systems and mobile databases, a solution is considered to be vendor specific if it is based on a particular functionality or feature that is not standard across all database vendors that may wish to participate in data synchronization at any given time [5].

In consideration of the above, several approaches that are vendor specific as well as database category specific such as RDBMS only are described. In [7], [13], and [14], the standard SQL query as certified by the ISO was adopted in their solutions to enable cross platform synchronization without having any limitation. However, this does not make it fully independent to all vendors because it is applicable only to RDBMS category of databases. Other databases, such as Analytical Databases, Operational Databases, FlatFile, XML etc. are not included in the solution. Whereas in [15], a model was developed to independently establish communication between the mobile devices and the server; the model's independence makes it adoptable by any system or platform. Nevertheless, the solution is based on RDBMS only which operate on a particular data model. It also has some table structures that must be adopted by both parties that wish to communicate. In addition, a given function ( $M=h(H)$ ) is used to generate message digest that must be the same for both side to be able to decode the encoded data.

However, many solutions for mobile data synchronization happened to be vendor specific such as the solution in [14] which is based on Microsoft SQL Server and [16] whose solution is solely dependent to MySQLite. Furthermore, others like [2], [3], and [16] voted timestamp database feature as a means of determining the most current state of the data on either side of the databases. So if the timestamp of A is higher than the timestamp of B, A is considered the most up-to-date data and it is synchronized with B. Another database feature that is used by [16] is Trigger which is used to trigger an event in case of any inconsistency that is discovered using the timestamp database feature and thus making all the above not suitable for databases that are fully heterogeneous in nature. Other solutions such as [17] and [18], have great synch techniques suitable for server to server communication only.

Having said that it can be concluded that the above solutions are vendor specific or peculiar to RDBMS only. This is because some proposed solutions use vendor dependent functions such as time stamp, trigger or database dependent information like metadata. To be precise, both vendors of the mobile database and the server-side database should be identical or the same entity.

Furthermore, some solutions are dependent to a particular mobile database vendor only. Such solutions are in most cases independent of the server-side database vendor and operate on a separate synchronization server [7], [13]. That is to say, the solution must match the mobile databases for a synchronization to take place. For example, when a programmer is to develop or modify existing mobile application, some particular vendor's libraries for mobile device databases have to be embedded to the solution that resides in the synch server (like AnySyn in Fig. 2) for an effective synchronization.

As a result of these constraints, the flexibility, adaptability as well as extensibility of mobile business systems have been noticeably declined. The problems above need to be tackled as we are heading to the environment where mobile devices will be further diversified and their databases (DBMS) will be heterogeneous in nature [19].

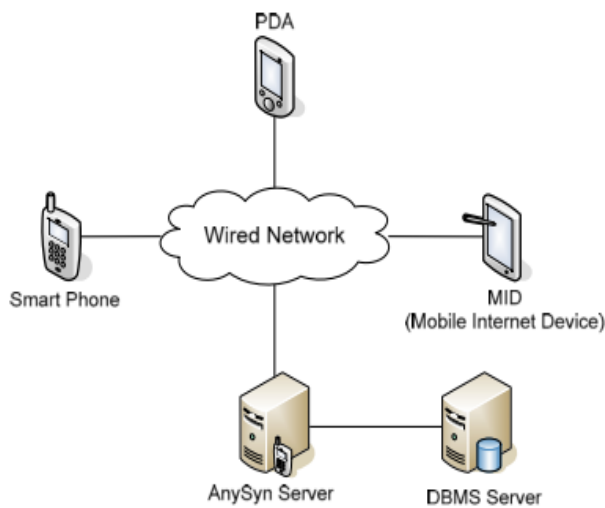


Fig. 2. Model development steps.

### III. METHODOLOGY

The purpose of this research is to study the state of the art in the context of mobile databases with respect to data synchronization as well as to propose a generic model that can be adopted by heterogeneous mobile databases. The structural flow and factors that persuade and affect the generality of any synchronization solution need to be painstakingly identified and empirically validated. As a result, it is necessary to adopt a method that allows studies to be carried out in real life context. Out of the five software engineering methods discussed by [20], a case study method was found to be the most suitable for this research.

According to [21], a case study is “an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident.” This type of method explains, comprehensively, how and why certain phenomena occur. To derive new hypotheses and build theories, exploratory case study is adopted as initial investigations of some phenomena while confirmatory case study is used to test existing theories [22]. The clear understanding that confirmatory case study reveals can be useful in reconciling between rival theories.

The results obtained when Case Study Method (CSM) is applied are more valid than when controlled experiment is applied [20]. This is because the variables under study are measured from the real world context.

#### A. Invention Method

In our proposed synchronization method, although the communication is bidirectional, we consider mobile devices as the clients and the server as the master. This implies that both *Send-In* and *Send-Out* process are initiated at the client side. This is because the clients cannot stay connected all the time [13], thereby making it difficult for the server to know which among the numerous clients is actually online and ready to receive a package.

Verifications are done at the beginning of the synchronization process to confirm whether there is need for

the synchronization and also at the end to verify the successful completion of the synchronization process; we aim to decrease the number of tuples retrieved from the source database that already match their counterpart in the target database.

For each successful transmission, a copy of the hashed data is saved in the temporary repository which can later be used to know whether synchronization is fully or partially completed.

In any synchronization process, the source database horizontally organizes the total order of records in the source database, summarizes the tuples using the hash function, and for the same range of tuples, retrieves the equivalent hash summary from the target database. If the summaries match then we assume both the target and source databases have the same content for the selected range. If the summaries do not match then the same range of records are retrieved from the source database, summarize and send to the target database.

In comparison, this method differs from the existing methods in the following ways:

1) Temporary Repository (TR): For each successful transmission a copy of the generated message digest is saved in TR until the process is successfully completed and is removed thereafter.

2) Embedding Data Extraction Formula (DEF) that was proposed by [23] into the proposed solution: Network might fluctuate during the synchronization process and the data might have been partly transmitted. To avoid starting the process a fresh, the DEF is used to extract only the records that failed.

3) Process Initiator: In some methods, synchronization process begins from the target database. The target database can be either client or master database [9]. Whereas in others such as [2], [14] and [15], the process is initiated by the owner of changes, i.e. the database that is altered would be the initiator of the process. While in our method, considering the fact that mobile devices have no stable connection [24], they are given the responsibility to initiate the synchronization process when they are online. The process can be sending to the server or receiving from the server.

4) Data Bank: Keeps the synchronization history. This addresses many issues such as resolving conflicts and comparing whether the source and target databases have identical content before initiating the synchronization process.

#### B. Model Development Process

Based on our synchronization procedure, after analyzing the information retrieved from the literature as well as outlining the major strength and weaknesses of each of the existing solutions, the model that aims to mitigate those weaknesses is developed following the three steps as depicted in Fig. 3.

At first, we identified the relevant elements for the proposed model such as entities and attributes from the existing solutions some which strongly guided our selection criteria for the appropriate model structure. Also properties that are unique and common are identified and aggregated.

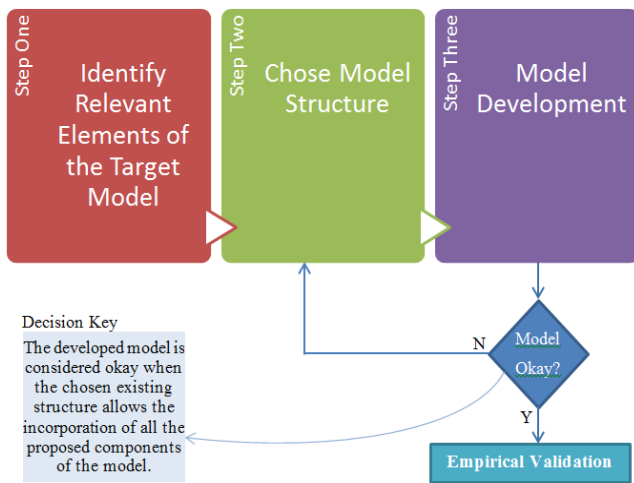


Fig. 3. Model development steps.

Secondly, to choose the appropriate model structure to be adopted, several factors were painstakingly considered such as the most adopted structure, the one that is closely independent to database vendors, the one that considered the utilization of mobile resources like CPU and Memory. Also, date of release is one of our major factors which determine the most appropriate model to adopt. Considering the trend in publications, each proposed model is an upgrade of the existing ones; therefore, the most recent (latest) model would have covered some of the loopholes of its predecessors, thereby providing a strong guide for the development of the proposed model structure.

Thirdly, reconciliation and construction of the proposed model is considered for heterogeneous mobile databases. It should be noted that, the construction of the proposed model that based on the existing models is iterative in nature, therefore, this process involves going back to step two (choose model structure) until we find the structure that best suites our approach or answer our research questions.

### C. Data Analysis

To effectively analyze the data generated from both the proposed model and the existing model, two different mobile devices with different specifications were used for the proposed model as well as the existing model. Besides, a single computer was considered as a server to house the central repository. Additionally, the same network was used and at almost the same time, which means, there was no big interval in the network speed for all the trials. Having big interval may result to network inconsistencies which will in the end affect the accuracy of our measurements. The specifications of the devices involved as well as the network itself are as follows.

#### 1) Empirical Validation Tools

In this section, the devices used for validating the proposed model are described. 1) First Mobile Device: ASUS phone brand was utilized with Wi-Fi of 7.10 and battery of 2000mAp. It runs on Android 4.4. 2) Second Mobile Device: iPhone 6 was employed which runs on iOS 10.3.2 operating system with Wi-Fi version of 802.11 and battery of 1810mAh. It also has 32 GB of memory and 1 GB of RAM. 3) Server-side Computer: HP laptop intel® processor, Core™ i7 was

deployed which runs on windows 10 with CPU speed of 4.20GHz and 8.00GB of RAM. 4) Network: Ralink wireless network was used with 802.11bgn the network uploads at 3,364,303 and downloads at 58,105,411. Also its speed was at 54.0mbps and 98% of signal quality.

#### 2) Statistical Tool

In this study, we adopted the use of Chi<sup>2</sup> test as our analytical tool to analyze the data generated from both the proposed model and the existing model. As for the level of significance,  $\alpha = 0.05$  (5%) was used to indicate a 5% risk of concluding that a difference exists when there is no actual difference. the following section presents the proposed model.

## IV. PROPOSED MODEL

This chapter extensively presents the proposed model. Based on the findings and the shortcomings identified from the literature and a review conducted by [25], a generic model is proposed to address some of the untouched areas with respect to data synchronization between mobile device databases and the server-side database.

The ultimate goal of the model is to synchronize server's database with mobile devices heterogeneous databases that are remotely interlinked with utmost consideration on the usage of resources of the mobile device. Primarily, the specific concern is to avoid database vendor dependency approach and provide a solution that is heterogeneous in nature which can be used to synchronize data between mobile databases and server-side database, such that all categories of mobile databases can participate in the synchronization process regardless of their individual peculiarities.

The proposed model comprises of several components which when combined produce a complete working model. In this section, we started by introducing the three (3) architectures, followed by the overall concept and finally elaborate the major components of the architectures one after the other where applicable.

### A. Architectures

The architecture section of the proposed model is categorized into three different sections. Each section performs different tasks or responsibilities from the other. First section presents *Send-In Synchronization Architecture*, while section two put forward the *Send-Out Synchronization Architecture*. Finally, Server-side Synchronization Architecture is introduced which illustrates the activities of record bank entity situated on the synchronization participating server.

#### 1) Synchronization Architecture (Send-In)

The process of *Send-In* begins from the mobile devices to avoid information broadcast from the server as the current practice (see Fig. 4 below).

Referring to Fig. 4 above, the mobile devices listen for any changes made on the server, if there is any, the mobile sends request for update along with relevant parameters (authentication, data required). At this point the server takes the charge, thus reducing the burden on the mobile. The server receives request, do the comparison between the record bank and the server private repository. The latest version of the records is then sent to the mobile device as requested.

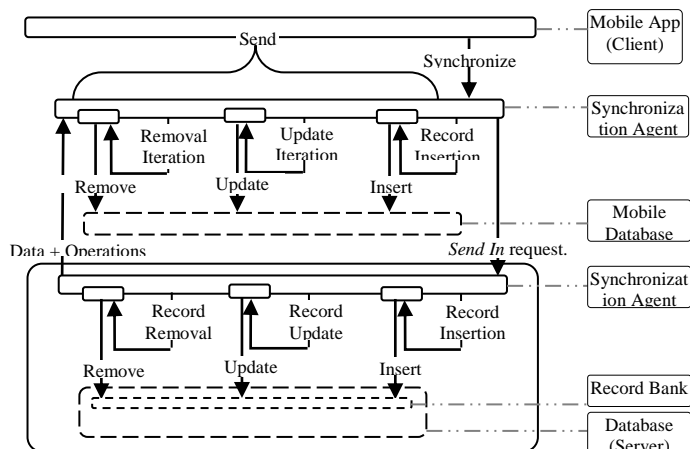


Fig. 4. Send in sequence diagram.

### 2) Synchronization Architecture (Send-Out)

In the *send-out* phase of the architecture, the mobile devices create, modify or delete records and need to notify the server about the changes made. The scenario is depicted in Fig. 5 as follows:

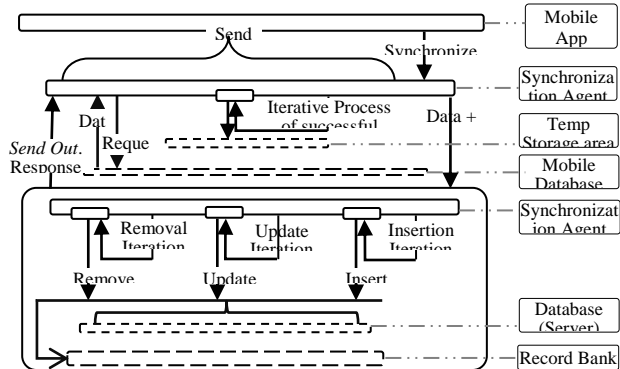


Fig. 5. Send out sequence diagram.

To start the send-out process, the mobile device retrieves the affected data using the formula proposed by [23]. The content is then hashed and sent to the server. Each successful transfer of record is cataloged in the mobile device temporary storage area [23] to monitor the synchronization status. The server does the comparison upon receipt of the data and applies appropriate operation.

### 3) Server-side Synchronization Architecture

One server can have multiple clients (mobile devices), each of them sends and receives records from the record bank of the server. For the server to have most up-to-date records in its private repository, there is need to (from time-to-time) synchronize with the record bank since clients communicate with the record bank regularly. The process runs periodically to check for any discrepancies between the data in the records bank and the records of the server. If there are changes, an update operation is applied to the server. Consequently, other clients that require the same updates will eventually see the alert and proceed for synchronization. The process is illustrated in Fig. 6 below.

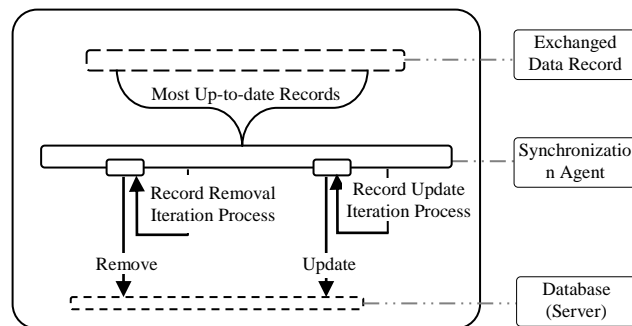


Fig. 6. Record bank to server synchronization.

## B. Overall Concept

The overall concept of this model is similar to the existing synchronization solutions; where unlimited client devices connect with the database of the server in order to synchronize data as shown in Fig. 7 below:

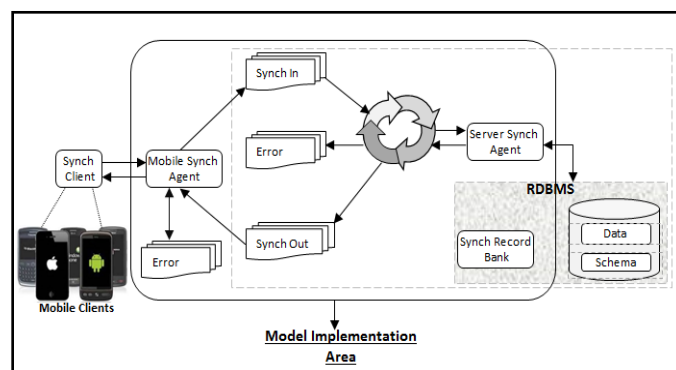


Fig. 7. Overall concept.

With regards to the clients, they are the mobile devices consisting of different types of mobile databases with a light weight storage area, mobile applications that are used to create and manipulate records on the move, and a module where the proposed model is implemented for an effective synchronization. Conversely the server is a computer system that consist of an agent where part of the synchronization model is implemented, Synchronization Records Bank (SRB) where the histories of all synchronization processes are kept regardless of their status (success, failure or removed) in order to track history and to resolve conflict in the future.

## C. Components of the Architecture

Regardless of the architectural categorization, each of the earlier discussed architecture cannot work alone. Meaning, they must be merged together to form a complete architecture. Therefore, the merged architecture has the following components:

### 1) Table Structure (Record Bank)

As one of the synchronization staging areas, record bank is a server-side located repository. It keeps the history of data exchanged between the devices and the server. Due to its size and computations (comparisons) involved, it's placed on the server, since the mobile devices have limited storage space. The structure of the repository is as follows:



a) *Record Owner*: It is an attribute that uniquely identifies the actual client that created the record. It is used to differentiate who amongst clients have the most up-to-date record in case one record is being used by many clients.

b) *Private Key*: It is the primary key of a record on the mobile. It is used to differentiate records on the client side.

c) *Public Key*: It is a unique identifier of a record on the server. That is to say, is the primary key of the records on the server. It is used to determine which record is the most recent and also resolve conflicts between the data in record bank and the actual data on the server.

d) *Records Message Digest*: It's the message digest generated by the hash function at run time where the business data is the input. Because it is produced at run time, it is considered to be the most recent version of the record.

e) *Flag*: It's an attribute that records the synchronization outcome (success or failure). The flag is 0 when there is no error in the process and 1 otherwise.

f) *Active*: This is where the status of a record is stored. It records 0 if a given entry is no longer in use (removed) or 1 if it's still active. Note that, an entry is not completely deleted, instead, is archived for future reference.

#### 2) Data Extraction Formula

Data Extraction Formula (DEF), is a formula proposed by [23] for the purpose of extracting the records that only matter for synchronization. This formula does the comparison between records and retrieve only the affected records which will be used as an input of the following hashing formula.

#### 3) Message Digest Formula

Message digest formula is a formula that is used to compute and produce message digest for transmission to the target database.

$$h = H(M)$$

The input of this formula is the output of the DEF explained above. But for the DEF to be able to extract data correctly it requires the following storage space on the mobile device.

#### 4) Temporary Storage Area (Client)

Temporary storage area is part of the DFD explained above. It save any successfully transmitted record so as to ease the process of locating a starting point for the DEF. it is also explained in [23].

#### D. Synchronization Procedure

In this section, we explain and demonstrate the procedure that our proposed model follows to synchronize data between one database to another. Each of the following figures (Fig. 8, 9, 10 and 11) depicts a particular task that might be assigned to it during the synchronization process. The first (Fig. 8) is the main procedure that, at some point, branches to link to its sub-procedures for a separate responsibility.

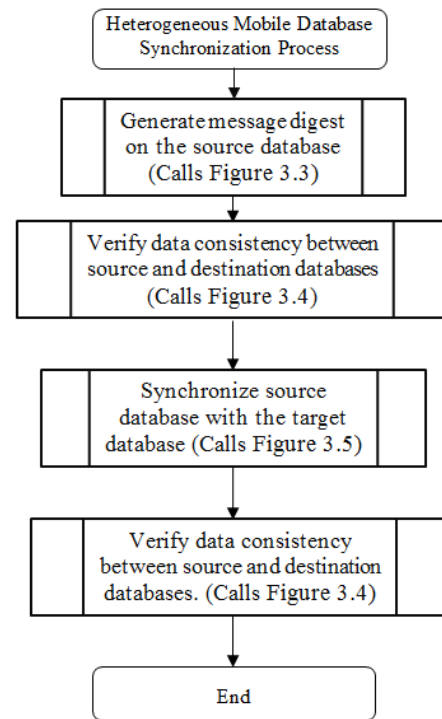


Fig. 8. Data synchronization using HMDS model.

There are four significant procedures involved in the synchronization process which starts with generation of message digest, verification of the inconsistencies between the source and the target databases, perform synchronization, and finally verify the consistencies between the two databases. At the outset, message digest generation is explained.

#### 1) Message Digest Generation

The process of generating the message digest is the same for both the source and the target databases. However, to minimize the burden on the mobile devices and also keep the history of all performed synchronizations, a hashed copy of each dataset is kept in the record bank of the server after any successful synchronization. The saved hash of any completed synchronization can be later used to resolve conflicts between two or more different clients that are meant to manipulate the same dataset. Please refer to Fig. 9 below:

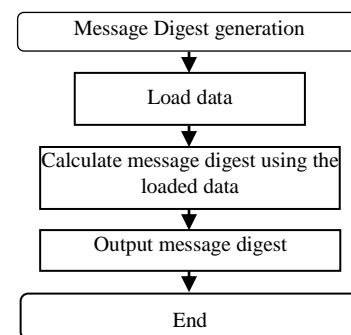


Fig. 9. Message digest generation.

In this phase, the first activity is to load the data that is to be hashed, after that, a hashing formula is applied to the loaded data which calculates the message digest, and finally the computed message digest is produced for the first and final verifications as well as synchronization process.

### 2) Verification of inconsistencies between the target and the source databases

This is the second procedure which the proposed model follows to verify whether the records of both the source and the target databases have identical values. This process confirms if the data to be synchronized is not available in the target databases. The figure below depicts the verification process:

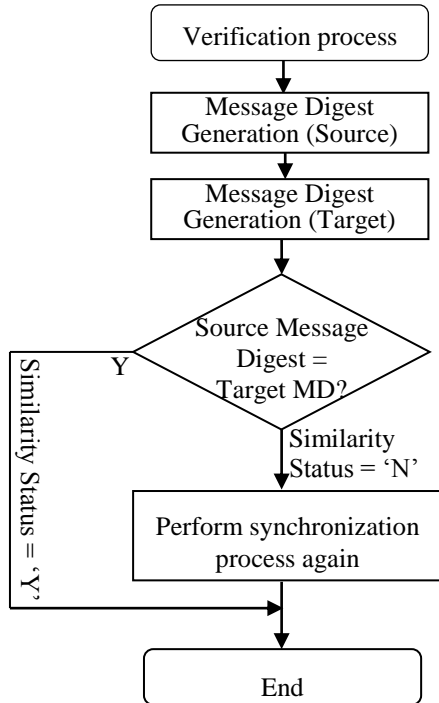


Fig. 10. Verification process.

When the verification process starts, a cryptographic representation of the data in both databases (source and target) are produced which are further compared to see whether the two defined data ranges are different. If they are the same, the process ends there, otherwise, the synchronization is performed to fill the identified missing information in the target databases as explained in the next section below.

### 3) Synchronization Process

In this phase, after all necessary verifications have been made and confirmed that, there is need for the synchronization to take place, the following process is called to administer the changes accordingly.

The process begins with comparing the two generated hashes (the source and the target) if the verification was not called in the main procedure. This might be possible when there is more data to be synchronized immediately after the first assignment. After the comparison, if the records are the same, it calls for more data, otherwise the synchronization process continues.

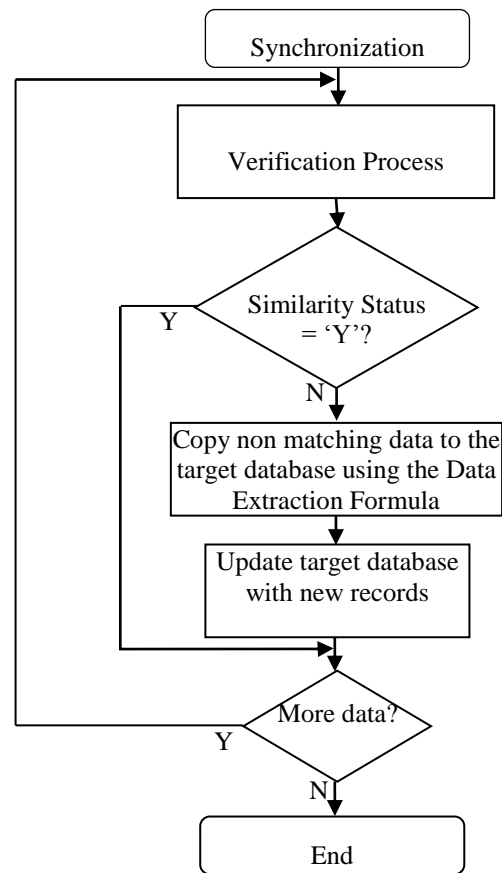


Fig. 11. Synchronization process.

Using data extraction formula, the nonmatching records are copied to the target database, thereafter; appropriate action is applied to the copied records. If there are more data to synchronize the process is repeated, else, the process is ended. The verification is repeated in this phase because there is need to localize the verification at some points such as when there is more data to be synchronized immediately after the completion of the assigned tasks. Meaning the loop within the phase should be maintained until the data to be synchronized is exhausted.

### 4) Final Verification Process

This time, the verification process is to confirm the status of the most recent (just completed) synchronization whether is successfully completed or an error occurred during the process. If there was an error, the synchronization process is repeated, otherwise, the process is tagged to at rest, which means the process is disabled for now until there are more changes from either side.

The process uses the same diagram as shown in Fig. 10. Checking and comparison of data is repeated at multiple points due to the need to confirm that there is no more data to be synchronized before putting the entire process at rest. Putting the process at rest after a successful synchronization greatly minimizes the consumption of batteries and other mobile valuable resources such as CPU and Memory.

V. RESULTS AND DISCUSSION

This section presents and discussed in details the results of the proposed model and its counterpart. After subjecting the proposed model to a proper implementation, thorough testing was conducted to ascertain its capability and reliability in different aspects. These aspects are in-line with our goal in making data synchronization possible between the different mobile database vendors and the server-side database. At first the hypothesis are presented and tested, and results are produced for the null and alternative hypothesis. For clarity and easy reference, results are discussed immediately after they are presented.

The data obtained from the proposed and existing model were analyzed using the Chi<sup>2</sup> test which produces the P value used to measure whether the null hypothesis (H<sub>0</sub>) should be accepted or rejected. As for the level of significance, α = 0.05 (5%) was used to indicate a 5% risk of concluding that a difference exists when there is no actual difference. Considering the formulated hypothesis, two-tailed comparison was considered. The tow-tailed test allows the comparison process to be fair to all the participating groups. Meaning, the proposed model could be better than the competitor’s model or vice versa.

After examining the process repeatedly, the proposed model yields outstanding, profound and remarkable improvements from the existing solutions, mostly in the utilization of mobile resources due to the incorporation of DEF [23]. It also showed the prospect in multiple mobile database vendors’ involvement in the synchronization process. The results are categorized and presented based on the aim of this study (mobile database heterogeneity).

Database heterogeneity refers to having different databases of different data model, DBMS, Vendors, and OS. Since there are varieties of mobile database vendors, a solution that neglect their individual differences and permit standard uniformity is required to be able to synchronize data limitlessly. The results of the proposed model in this regard are thereby presented in two scenarios: the first being the records exchange from Mobile Device Databases (MDD) to Server Side Database (SSD) while the second takes the opposite direction. In this context therefore, we aim to provide a general solution that can be used for data synchronization regardless of the aforementioned individual differences. One (latest) of the existing solutions was used to measure our proposed solution based on the following hypothesis.

**H<sub>0</sub>:** Database dependent information such as time-stamp, trigger or stored procedure have no impact in making any solution heterogeneous, i.e. vendor specific.

**H<sub>1</sub>:** Database dependent information such as time-stamp, trigger or stored procedure have impact in making any solution heterogeneous, i.e. vendor specific.

To answer these hypotheses, experiments are conducted and results were analyzed in two different scenarios. Scenario one (1) shows and discussed the results obtained from both the models when records are sent from Mobile Device Database (MDD) to Server Side Database (SSD), while scenario 2

present and discussed the results when records are sent from SSD to MDD. At first we begin by presenting scenario 1.

A. Scenario 1: Data Exchange Possibility from MDD to SSD

Since our proposed solution is bidirectional (send to the server and receive from the server) we started answering the hypothesis by initiating a communication from the Mobile Device Databases (MDD) to Server Side Database (SSD). Table I summarizes the trials that has the highest available number of records.

TABLE. I. SCENARIO 1 MDD (SQLITE & XML) TO MYSQL SSD

	SQLite	XML
Number of Records	10000	10000
Proposed Model AST (ms)	5.74	35.46
Competitor Model AST (ms)	9.93	-
Data received by SSD using the PM	True	True
Data received by SSD using the CM	True	False

Table I shows the possibility of receiving data and the Average Synchronization Time (AST) of both the proposed model and the competitor model. Looking at the Proposed Model (PM), apart from being able to receive the data sent from different mobile database vendors, it is also faster. While the Competitor Model (CM) was only able to receive data sent from SQLite because both SQLite and MySQL have the same data model and use the same DBMS, but for the case of XML, the process couldn’t be completed. Table II below shows the trails at multiple levels.

TABLE. II. SCENARIO 1 MDD (SQLITE & XML) TO MYSQL SDD

	SQLite	XML
Trial 1 Number of Records	500	500
Trial 1 Proposed Model TST (ms)	39010.1	181052.7
Trial 1 Competitor Model TST (ms)	31011.3	-
Trial 1 data received by SSD using PM	True	True
Trial 1 data received by SSD using CM	True	False
Trial 2 Number of Records	2000	2000
Trial 2 Proposed Model TST (ms)	48035.7	232311.2
Trial 2 Competitor Model TST (ms)	45501.1	-
Trial 2 data received by SSD using PM	True	True
Trial 2 data received by SSD using CM	True	False
Trial 3 Number of Records	5000	5000
Trial 3 Proposed Model TST (ms)	63210.3	291492.9
Trial 3 Competitor Model TST (ms)	69224.8	-
Trial 3 data received by SSD using PM	True	True
Trial 3 data received by SSD using CM	True	False
Trial 4 Number of Records	7000	7000
Trial 4 Proposed Model TST (ms)	71563.4	325143.4
Trial 4 Competitor Model TST (ms)	72100.0	-
Trial 4 data received by SSD using PM	True	True
Trial 4 data received by SSD using CM	True	False
Trial 5 Number of Records	10000	10000
Trial 5 Proposed Model TST (ms)	79461.3	364660.7
Trial 5 Competitor Model TST (ms)	91433.2	-
Trial 5 data received by SSD using PM	True	True
Trial 5 data received by SSD using CM	True	False

Talking of the Table II above, scenario 1 shows the possibility of synchronizing data to the server with a number of trials in both the

SQLite and XML databases using the Proposed Model (PM). In trial 1 of the scenario 1, the Mobile Device Database (MDD) was able to effectively synchronized data to Server

Side Database (SSD) for both the SQLite and XML databases. This achievement did not stop in trial 1 only, but across the remaining trails with different number of records, whereas, in the same trials, the Competitor’s Model (CM) was able to synchronize data with SQLite database only. This is a clear indication that, the proposed model can be embraced by several database vendors, regardless of their individual difference because the model considers the interception areas rather than focusing on their individual differences.

Furthermore, the Total Synchronization Time (TST) taken for the proposed model to synchronized data to the server was 39010.1(ms) at first trial and 31011.3 (ms) for the competitor model in the same trial. The increase continued to correspond to the number of records in the trials diagonally with around 5.5(s).

**B. Scenario 2 Data Exchange Possibility from SSD to HMDD**

In scenario 2, the opposite direction of the synchronization was considered where Server Side Database (SSD) sends records to Mobile Device Databases (MDD). Table III shows the summary of the trials conducted in Table IV.

TABLE III. SCENARIO 2 MYSQL SDD TO MDD (SQLITE & XML)

	SQLite	XML
Number of Records	10000	10000
Proposed Model ADAT (ms)	3.98	4.73
Competitor Model ADAT (ms)	4.23	-
Data received by MDD using the PM	True	True
Data received by MDD using the CM	True	False

Table above shows that, the records sent by the server was received by Mobile Device Databases (SQLite and XML) using the Proposed Model (PM). While using the Competitor Model (CM), only SQLite was able to receive the data. Also, the Average Data Arrival Time (ADAT) was lower using the PM. Table below presents the results of the 5 trials.

As the case of scenario 2, the second direction of the synchronization is considered where the server sends records to its clients. Using the proposed model, both SQLite and XML databases were able to receive data composed and sent by the server crosswise, in all trials. While the competitor model behaved in contrast, where only the SQLite did received the records. This is because the competitor’s model was based on SQL queries while others use database dependent information such as timestamp and trigger in the cause of synchronization, which eliminates some database vendors that do not have such techniques or mechanisms embedded or do not belong to RDBMS category at all.

In addition, it can be seen in both the scenarios 1 and 2 above that, using the Proposed Model (PM), the average time taken to synchronize records using SQLite is way less than the time taken with XML database even though they both send and receive data. This is because, in SQLite, multiple rows carry a fixed number of columns identifiers unlike in XML where multiple tags are used to wrap each record and group of records [26].

TABLE IV. SCENARIO 2 MYSQL SDD TO MDD (SQLITE & XML)

	SQLite	XML
Trial 1 Number of Records	500	500
Trial 1 Proposed Model DAT (ms)	21023.4	24663.8
Trial 1 Competitor Model DAT (ms)	26001.3	-
Trial 1 Data received by MDD using PM	True	True
Trial 1 Data received by MDD using CM	True	False
Trial 2 Number of Records	2000	2000
Trial 2 Proposed Model DAT (ms)	30331.2	32415.9
Trial 2 Competitor Model DAT (ms)	34311.2	-
Trial 2 Data received by MDD using PM	True	True
Trial 2 Data received by MDD using CM	True	False
Trial 3 Number of Records	5000	5000
Trial 3 Proposed Model DAT (ms)	38096.9	39736.1
Trial 3 Competitor Model DAT (ms)	44709.7	-
Trial 3 Data received by MDD using PM	True	True
Trial 3 Data received by MDD using CM	True	False
Trial 4 Number of Records	7000	7000
Trial 4 Proposed Model DAT (ms)	46543.1	48280.7
Trial 4 Competitor Model DAT (ms)	47016.3	-
Trial 4 Data received by MDD using PM	True	True
Trial 4 Data received by MDD using CM	True	False
Trial 5 Number of Records	10000	10000
Trial 5 Proposed Model DAT (ms)	52206.4	55113.9
Trial 5 Competitor Model DAT (ms)	59236.8	-
Trial 5 Data received by MDD using PM	True	True
Trial 5 Data received by MDD using CM	True	False

For example, in SQLite, if you have 10000 rows of records and have 5 columns then you would have 5 columns identifies, one for each column. However, for the same number of records using XML, you would have 100,000 wrappers (that is to say, 10,000 rows \* 5 records par row \* 2 opening and closing tags). This adds so much load to the data, thus make heavy for mobile devices to manipulate easily.

Chi<sup>2</sup> test was used to analyze the above data that states the possibility of synchronizing records between mobile heterogeneous databases. Since our data in this case is TRUE or FALSE, a statistical tool that will allow the probabilities to be counted and aggregated is selected which works as follows.

TABLE V. CHI<sup>2</sup> TEST DATA FROM SCENARIO 1 AND 2

	Number of True	Number of False	Grant Total
Proposed Model	20	0	20
Competitor Model	10	10	20
Grant Total	30	10	40

Table V shows the data (True and False count) retrieved from scenario 1 and scenario 2 as presented in Table II and Table III. Therefore, the (column total \* row total) /grant total formula was used to calculate the expected values for the data presented in Table V. The results of the computation are as shown in Table VI.

TABLE VI. EXPECTED VALUE RESULTS

	Number of True	Number of False	Grant Total
Proposed Model	15	5	15
Competitor Model	15	5	10
Grant Total	20	5	25

After computing the Expected Values (EV) as indicated above, the Actual Values (AV) and the EV were included in the Chi<sup>2</sup> test formula to obtain the probability value. On the other hand, 0.05 was set to be the alpha ( $\alpha$ ) value. These values can be used to either accept or reject the null hypothesis.  $H_0$  can be only rejected if the probability value is less than the alpha value. Results of the analysis shows that, the  $p$  value is 0.00026073 for both the two scenarios, which is less than the alpha ( $\alpha$ ) value of 0.05.

Based on this therefore, the null hypothesis is fully rejected since the probability value is less than the alpha value. The outcome shows the possibility of sharing data across multiple mobile databases when database dependent information such as timestamp, triggers and Meta data are excluded in the solution. Solutions that adopt any of these techniques are thereby considered vendor specific or solution that is homogenous in nature.

## VI. CONCLUSION

We have presented a model for purpose of addressing the problem of data synchronization between the heterogeneous mobile devices databases and server-side database with significant consideration to the limitations of the mobile devices such as memory, CPU, power supply and continuous network fluctuations. Based on the goals of this study, experimental method which allows study to be carried out in a real life context was considered to be the most suitable for this research. This method was selected out of the five methods discussed by Easterbrook [20] for the empirical software engineering research. The study explored and investigated numerous solutions from the existing literature where various incredible research contributions were found. However, mobile database heterogeneity was uncared for in spite of its great importance. Thus hinders other types of databases to participate in the synchronization process since they were not considered as part of the solution in the first place.

Based on the review outcome, existing solutions properties were identified which guided the construction of the proposed model. To empirically validate the proposed model, a prototype was developed which implemented the model in a real-life context. Also one latest existing solution was implemented for the purpose of performance analysis.

The proposed model further weighs against the existing model to mark the improved areas. Results indicate that the objectives of this study have been achieved where the proposed model proved feasibility of engaging multiple mobile databases in a synchronization process; thus delivering substantial evidence to repudiate the null hypothesis. Moreover, the proposed model displayed some strength in the synchronization speed and also the utilization of the mobile resources.

Looking at the unique intensity that the competitor model and proposed model offer, there is need to consider the significance of heterogeneity and resource consumption when making the decisions between the models. The actual potency of the proposed model lies in the aforementioned variables. The study has provided a clear benchmark that can be used to compare these models when adopting a synchronization solution for mobile devices. Unstructured data are another key

important component that will be given due consideration in the nearby future since mobile devices are now one of the major sources of big data [27].

## ACKNOWLEDGEMENT

This paper/research was fully supported by Ministry of Higher Education Malaysia, under the Fundamental Research Grant Scheme (FRGS) with Ref No of: FRGS/1/2015/ICT01/UTP/02/1. Any opinions, findings, and conclusions stated in this paper are those of authors and do not necessarily reflect those of the MOHE.

## REFERENCES

- [1] M. Nayebe, B. Adams, and G. Ruhe, "Release Practices for Mobile Apps -- What do Users and Developers Think?," 2016 IEEE 23rd Int. Conf. Softw. Anal. Evol. Reengineering, pp. 552–562, 2016.
- [2] D. Sethia, S. Mehta, A. Chodhary, K. Bhatt, and S. Bhatnagar, "MRDMS-Mobile Replicated Database Management Synchronization," 2014 Int. Conf. Signal Process. Integr. Networks, pp. 624–631, 2014.
- [3] J. Sedivy, T. Barina, I. MOrOzan, and A. Sandu, "MCSync – Distributed , Decentralized Database for Mobile Devices," IEEE 2012, pp. 1–5, 2012.
- [4] M. F. Qaisrani, "Types of Distributed Database Management System," Benazir Bhutto Shaheed University, 2014. [Online]. Available: <http://www.slideshare.net/TAHAROC/types-of-data>.
- [5] G. Thomas, G. R. Thompson, C.-W. Chung, E. Barkmeyer, F. Carter, M. Templeton, S. Fox, and B. Hartman, "Heterogeneous Distributed Database Systems for Production Use," ACM Comput. Surv. - Spec. issue Heterog. databases, vol. 22, no. 3, pp. 237–266, 1990.
- [6] M. Choi, E. Cho, D. Park, J. Bae, C. Moon, and D. Baik, "A Synchronization Algorithm of Mobile Database for Ubiquitous Computing," Fifth Int. Jt. Conf. INC, IMS IDC, NCM 2009., p. pp.416,419, 25-27, 2009.
- [7] M. Choi, E. Cho, D. Park, C. Moon, and D. Baik, "A database synchronization algorithm for mobile devices," IEEE Trans. Consum. Electron., vol. 56, no. 2, pp. 392–398, May 2010.
- [8] B. S. Ramya, S. B. Koduri, and M. Seetha, "A Stateful Database Synchronization Approach for Mobile Devices," Int. J. Soft Comput. Eng., vol. 2, no. 3, pp. 316–320, 2012.
- [9] M. Ahluwalia, R. Gupta, A. Gangopadhyay, and M. Mcallister, "Target-Based Database Synchronization," in Proceedings of the 2010 ACM Symposium on Applied Computing, 2010, pp. 1643–1647.
- [10] H. Preston and M. Narayan, "Message digest based data synchronization," US 09/896,321.
- [11] P. Bottorff, C. L. Allen, A. Hudson, and M. R. Krause, "Distributed database synchronization," US 12/911,356.
- [12] L. Zhenyu, C. Zhang, and L. Zunfeng, "Optimization of Heterogeneous Databases Data Synchronization in WAN by Virtual Log Compression," Futur. Networks, 2010. ICFN '10. Second Int. Conf., pp. 98–101, Jan. 2010.
- [13] V. Balakumar and I. Sakthidevi, "An Efficient Database Synchronization Algorithm for Mobile Devices Based on Secured Message Digest," 2012 Int. Conf. Comput. Electron. Electr. Technol. [ICCEET] Messag., pp. 937–942, 2012.
- [14] T. A. Alhaj, M. M. Taha, and F. M. Alim, "Synchronization Wireless Algorithm Based on Message Digest ( SWAMD ) For Mobile Device Database," 2013 Int. Conf. Comput. Electr. Electron. Eng. Synchronization, pp. 259–262, 2013.
- [15] J. Domingos, N. Sim??es, P. Pereira, C. Silva, and L. Marcelino, "Database synchronization model for mobile devices," in Iberian Conference on Information Systems and Technologies, CISTI, 2014.
- [16] G. P. Zaia, C. R. C. Messias, R. G. Eduardo, and C. J. Olivete, "MySQLite Sync: Middleware for stored data synchronization in mobile devices and DBMSs," 2014 XL Lat. Am. Comput. Conf. 2 agente, pp. 1–7, 2014.

- [17] A. A. Imam, S. Basri, and R. Ahmad, "Synchronization Algorithm for Remote Heterogeneous Database Environment.pdf," in *Advances in Intelligent System and Computing*, 2014, pp. 55–65.
- [18] A. Stage, "Synchronization and replication in the context of mobile applications," in *ICFN '10. Second International Conference on*, 2012, p. 98,101.
- [19] H. Chen, J. Yu, C. Hang, B. Zang, and P. Yew, "Dynamic software updating using a relaxed consistency model," *Softw. Eng. IEEE Trans. on*, Vol. 37(5), pp. 679–694, 2011.
- [20] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting Empirical Methods for Software Engineering Research," *Guid. to Adv. Empir. Softw. Eng.*, pp. 285–311, 2008.
- [21] R. K. Yin, "Case Study Research: Design and Methods.," Sage, 2002.
- [22] D. Perry, A. Porter, and L. Votta, "Empirical Studies of Software Engineering: A Roadmap," *Int. Conf. Softw. Eng.*, pp. 345–355, 2000.
- [23] A. A. Imam, S. Basri, and R. Ahmad, "Data Extraction Formula for Efficient Data Synchronization between Mobile Databases and Server-side Database," in *International conference on Computer and Information Science (IEEEC 2016)*, 2016.
- [24] N. Banivaheb, "Mobile Databases," Slide Presentation, 2012. [Online]. Available: [http://www.cse.yorku.ca/~jarek/courses/6421/F12/presentations/Mobile-Databases\\_Presentation.pdf](http://www.cse.yorku.ca/~jarek/courses/6421/F12/presentations/Mobile-Databases_Presentation.pdf).
- [25] A. A. Imam, S. Basri, and R. Ahmad, "Data Synchronization Between Mobile Devices and Server-side Databases: A Systematic Literature Review," *J. Theor. Appl. Inf. Technol.*, vol. 81, no. 2, pp. 364–382, 2015.
- [26] J. Fong, H. K. Wong, and Z. Cheng, "Converting relational database into XML documents with DOM," *Inf. Softw. Technol.*, vol. 45, no. 6, pp. 335–355, 2003.
- [27] B. B. Mehta, "First Credit Seminar Presentation on " Privacy and Big Data : Issues and Challenges", 2014.

# Data Mining Techniques to Construct a Model: Cardiac Diseases

Noreen Akhtar, Muhammad Ramzan Talib, Nosheen Kanwal

Department of Computer Science  
Government College University  
Faisalabad, Pakistan

**Abstract**—Using echocardiography flexible Transthoracic Echocardiography reported data set detecting heart disease by using mining techniques designed prediction model the data set can develop the reliability of analysis of cardiac diseases by echocardiography, using eight iterative and interactive steps consisting Knowledge Discovery in Database (KDD) methodology including from 209 patients with echocardiography to extracting the data important mode of action Transthoracic Echocardiography inspection report. This study used data from Faisalabad Institute of Cardiology study from 2012 to 2015. All models exposed the results of J48 decision tree, naïve bayes classifier and neural network that has extraordinary classification precision and predictive of heart disease cases are generally comparable. However, J48 model predictive classification accuracy shows of 80% based on the true positive rate ratio and performance slightly better. This study shows to predict heart disease cases and People can be used the results of our study to make more consistent diagnosis of cardiac disease and to help them as a support tool for cardiac disease specialists.

**Keywords**—Knowledge Discovery in Database (KDD); data mining; decision trees; neural networks; Bayesian classifier; heart disease

## I. INTRODUCTION

Heart disease causes higher mortality rate in our Pakistan. In our country the male and female having the age 65-year-old they are facing the heart disease. Data mining technology technique is used to decrease cardiac disease in entirely over the world. In this study, researcher can easily identify heart diseases by skillfully doctor through extreme risk factors. To choose the best predictive method researcher use various data mining techniques to predict cardiac diseases at this end. The Manimekalai [1] says that different risky aspects in the manner that smoking, high blood pressure, diabetes, obesity did not increase heart diseases.

### A. Discovering Knowledge for KDD process

Now a days, more information or data but lack of knowledge have health department. The researchers use the huge volume of information to the medical prediction of heart diseases treatment by Knowledge Discovery in Database (KDD).

In Fig. 1, researchers used a knowledge Discovery Database (KDD) approach to develop predictive models made by transthoracic echocardiography for predicting heart disease cases based on measurements. The data mining consists of nine steps that project's life cycle.

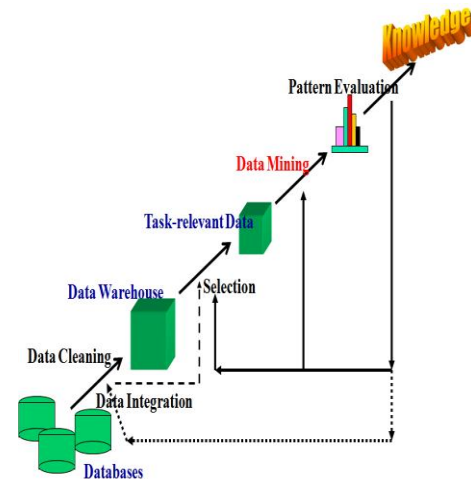


Fig. 1. Discovering knowledge for KDD process.

The researcher uses the various data mining techniques and compare with the same dataset to predict these techniques. Then I had to choose the best techniques to predict [2] Soni *et al.* says that Men and women almost equally affected low income countries out of the current CVD death of 82%. Low income society is unequally impressed. CVD is the most common disease about 2360 million folks died by CVD. Mostly heart disease and hit lead to social death. The major will be Eastern Mediterranean region percentage present. Southeast Asian lifestyle, work culture and eating habits change because the death clang increased the most. As a result of which changes social lifestyle reduces the disease.

The aim of this paper is extracting a data set bad or good key mode or feature. For heart disease diagnosis, we choose and identify the more relevant attributes. Decision tree, Neural Network and Bayesian classification compared to predict heart disease cases. With the help of domain experts, we chose the model to explain and analyze the results.

## II. REVIEW OF LITERATURE

Shafique, et al. [3] studied the data mining is the region that reviews which implies that data and knowledge are helpful from past information. There are various strategies for information mining. Data mining can be utilized as a part of various regions including medical utilize. Heart or cardiovascular disease is a hot topic in the global health care industry. Chandna [4], explained the health of the professional data mining is effective in forecasting disease. The number of

test numbers must be from the patient to detect the necessary conditions for the diseases in any case, utilizing certainties mining innovation can lessen the quantity of exams required. Lakshmi, et al. [5] conducted heart diagnosis dirty disease is an important medical and annoying task. Healthcare department is commonly believed that “information rich” and “knowledge-poor”. Alizadehsani, et al. [6] conducted experiment Cardiovascular disease is often very rare and is the important reason of decease. The fundamental sort of these sicknesses as Coronary Artery Disease (CAD) and the determination is essential. It has many side effects that are expensive and angiography is more accurate CAD Coronary Artery Disease diagnostic method. The existing research from the patient to collect data using several characteristics and the use of different data mining algorithms to achieve high precision side effects cost of the method. Gamberger et al. [7] studied that with the aid of information mining model Intelligent Heart Disease Forecasting System (IHDFS) innovation workmanship for example, decision trees, naïve Bayes and neural system. Chaurasia and Pal [8] explained that the death of the history of the largest study shows that heart disease has gradually become the world’s number one killer. Death age group occurs from 25 to 69 years old about 25% due to heart disease. In the event that all age bunches are incorporated coronary illness represents around 19% of all passings. Muthukaruppan and Er [9] explained that Particle Swarm Optimization (PSO) founded fuzzy master framework aimed at the finding of Coronary Artery Disease (CAD). The framework is outlined in light of informational indexes of Cleveland and Hungarian coronary illness. Yeh et al. [10] studied that acquired 493 legitimate examples from expectation and conduct programs that cerebro vascular ailment and embraced three order calculations, decision trees, Bayesian classifier and BP neural system, to construct an arrangement demonstrate individually. Hand et al. [11] explained that artificial neural network is a highly parametric statistical model has attracted considerable attention in recent years. In the artificial neural network is a highly parameterized fact that they are actual springy so that they are correctly functioning with irregular model insignificant. Pham et al. [12] conducted an experiment to decision tree algorithms have been utilized as a part of numerous applications arrangement for example, comfort medication assembling and creation money related investigation, stargazing and sub-atomic science.

Khemphila and Boonjing [13] explained that given meaning tree “which can be used to divide a large number of structures through over-application of simple sequence records gathered to decrease continuously record set decision-making rules. The KDD procedure demonstrate embraced in this examination along these lines as indicated by [14] Han and Kamber, sub-class is to locate a work of art (or process reason) depict and recognize information projects or thoughts keep in mind that end goal to foresee motivation behind the question class of the model can be utilized Its class tag is unidentified. Weka table is the first country into a set of data preprocessing algorithms and machine learning tools. It includes almost all popular algorithms. Its design allows you to quickly try new methods in a flexible way the existing method [15] Frank, et al. The data mining goal standard data

collection strategy play no role. This is a lot of data mining statistical data where data is frequently used effective strategies to answer specific questions and collect different types of the method. Data mining is frequently called “secondary” information and for this reason investigation [16] Hand, et al. KDD focuses on data from including how data is stored and access, how the algorithm is extended to large data sets known to the whole process of knowledge discovery still operate effectively, how to solve interpretation and visualization of results and how to effectively support and overall modeling Robot Interaction [17] Senes Applied, says the attention of this paper by using data mining tools and techniques, particularly development of analytical models which can be identified in the situation of general predictive cardiac diseases classification technology. The experiments have been conducted, on the data which was collected from the Faisalabad Institute of cardiology hospital from 2012 to 2015.

### III. MATERIAL AND METHOD

The purpose of this study by applying classification techniques to detect heart disease and attempting to build up a forecast displays in view of decision tree, neural system and Bayes classifier. In this paper, researcher has done citation valuable information from the heart hospital Faisalabad Institute for the collection of Faisalabad institute of cardiology data including 4 years of validity [2012-2015] data cleaning, data selection, data conversion and data mining. Where in the presence of this paper was realized, and different prediction methods were used for the age of disease data in each step the value of chest pain, resting blood pressure, blood sugar and different steps resting electrocardiogram result, maximum speed of the heart rate, exercise angina, diseases and display the capability of data mining technology to predict the values.

#### A. Data Pre-Processing Steps

##### 1) Data Cleaning

At this stage, we have to recover the missing data from the large amount of the datasets. Researches clean the data remove the data redundancy and recovered the missing values of the data. We had prepared the data according to appropriate format for data mining.

##### 2) Data Selection

In this step, the applicable analytical data is determined from the data set to be retrieved. The second data compression technique applied to the data set is the attribute selection.

##### 3) Normalization

The data is scale within small range for example 1 to 0 or 0 to 1 and fall in only small range.

##### 4) Attribute Construction

The new attribute is built in the dataset and add the new attribute in the given set that is used for mining data.

#### B. Knowledge Discovery

There are many data mining techniques that are used for statistical data mining and techniques for example outlier analysis, clustering, prediction and classification and association rule.



C. Outlier Analysis

Information libraries can contain general conduct or occasionally utilized information model of the information question. These information objects are special case ranges. It is first applied to the early removal of outliers to avoid its impact on other mining methods.

D. Clustering

In this research paper we have used K-means clustering. K-means clustering is come in unsupervised learning. The k-means clustering is used to grouping the data on the base of similarity.

E. Classification

Sub class is used to describe and differentiate the data to find the class/concept is to use the model for predicting the object class and its class label is unknown process models. The classification models are IF-THEN rules, J48 decision tree, Neural Network and Naïve Bayes and can be expressed in these forms.

F. Prediction

Forecast has been complicated in quite a lot of attention given the success of forecasting business setting. However, predictions of the time related data missing, or increasing/decreasing trends are more frequently mentioned. The main purpose is to use past values of larger numbers to consider future possible values.

IV. RESULTS AND DISCUSSION

Four experiments managed for this paper and we done all observations in both cases is considered that contains all 8 and containing other attribute sand 4 one of the selected attributes. With four experiments and eight different scenes a total of eight models of development work.

A. Experiment

1) Performance Measure for J48 Experiment

The first purpose of the experiment was to evaluate a J48 performance class unpruned tree to predict heart disease and investigate the properties of selected effect. In this Table I, first of all we select 8 attributes after completion of the all attribute experiment and then start the selected 4 attributes.

Algorithm In a first aspect of containing 209 instance of the training set has a complete run 8 attributes spent 0.45 seconds to build the model and model size of tree generated by the tree 50 times 30 leaves

TABLE. I. CONFUSION MATRIX

Model	Confusion-Matrixes		
	Positive Predicted	Negative Predicted	Actual results
J48:unpruned with attributes	66	26	Positive
	24	93	Negative
	Yes Predicted	No Predicted	Actual results
J48:unpruned with Selected attributes	77	15	Positive
	28	89	Negative

As shown in Table II, the model correctly identified 66 patients who were enrolled in 92 patients with heart disease and the remaining 26 were identified by errors that were

disease free and in fact these had a disease. This result gives the model of 0.756 Precision rate. The better model is to determine the negative cases as a model of TN rate is 0.78 correctly identify 92 patients were 117 patient who had no heart disease and the remaining 25 were identified have the disease but he had not actually.

TABLE. II. DETAIL PERFORMANCE OF J48 EXPERIMENTS

Models	Accuracy	TP-Rate	FP-Rate	Precisions	F-Measure	ROC-Area
J48 Unpruned with all attributes	76%	0.756	0.252	0.756	0.756	0.828
J48 Unpruned with selected attribute	79%	0.794	0.194	0.804	0.795	0.771

For precision score model labeled as belonging to class positive patients 79% (a) determining a real belong to the class affirmative (YES) and marked as belonging to the class-negative patients 76% (no) is not really a real negative part of the class (no). With 80.4% of the average precision it is in a very successful pattern for each class to retrieve the relevant values. With the 0.795 F-measured values it can be concluded that the accuracy and model recall rates are significantly balanced.

The results of this experiment show that a J48 decision unpruned tree algorithm is highly capable of when a prediction of heart disease. In addition the results show the impact of attributes select the classification accuracy, the size of the decision tree and the complexity of the model.

2) Naïve Bayes Classifiers

In this Table III, we predict the heart disease through Naïve Bayesian classifier and assess the performance of the experiment. In the third experiment, two scenarios are considered first we take all attributes 8 and the other we take selected 4 attributes.

In the first embodiment of the algorithm for solving the 209 instance of the complete set of training run 8 points and the attributes of the model execution time of 0.04 seconds. In a second embodiment the algorithm contained 209 one instance selected 4 attributes and a complete run on a training set of the model execution time of 0.00 seconds.

TABLE. III. CONFUSION MATRIX OF NAÏVE BAYES

Models	Confusion Matrixes		
	Positive Predicted	Negative Predicted	Actual results
Naïve-Bayes with attributes	67	25	Positive
	18	99	Negatives
	Yes Predicted	No Predicted	Actual results
Naïve Bayes with Selected attributes	75	17	Positive
	29	88	Negatives

As shown in Table IV, the overall classification accuracy of the model than all similar experiments performed better properties but it is still more than the success of the more. The model correctly identified 163 (77%) patients for the 209

embodiment who heart disease and the remaining 46 (22%) is determined to be error from the disease-free charges but they actually had the disease. This result gives the model TP rate of 0.78. This model is better in the case of determining the negative because the model of TN rate is 0.74 pass through correctly identified 76 patients performed 92 Li who had no heart disease and the remaining 16 were identified have the disease but he had not actually.

TABLE IV. DETAIL PERFORMANCE OF NAÏVE BAYES

Models	Accuracy	TP-Rate	FP-Rate	Precisions	F-Measures	ROC-Area
Naïve-Bayes with attribute	80%	0.798	0.217	0.798	0.799	0.872
Naïve-Bayes with selected attribute	77%	0.780	0.210	0.788	0.781	0.827

For precision score model, labeled as belonging to class positive patients 78% (a) determining a real belong to the class affirmative (YES) and marked as belonging to the class-negative patients 74%(no) is not really a real negative part of the class (no). With 74% of the average precision it is in a very successful pattern for each class to retrieve the relevant values. 0.78 F-measured values can be concluded that the accuracy and model recall rates are significantly balanced. In here, the better naive Bayes model selected property.

3) Neural Network

This experiment was designed to explore the ability of the neural network to predict the disease. Neural carried out by a multi-layer perception network algorithm is selected experiments.

In Table V, a first embodiment of the algorithm 209 run instances of complete training set 8 points and the attributes of the algorithm taken 0.56 seconds to build the model and super over 3 of 5 bell to produce confusion matrix. In the second embodiment of the algorithm for solving the 209 instance selected 4 complete attributes operation training set and the 0.17 seconds to build the models and super over 2 of 5 minute to produces confusion matrix.

TABLE V. CONFUSION MATRIX OF NETWORK EXPERIMENT

Models	Confusion_Matrixes		
	Positive predicted	Negative predicted	Actual results
Neural-network with attribute	66	26	Positives
	22	95	Negatives
Neural-network with Selected attribute	77	15	Positives
	29	88	Negatives

In Table VI, all 8 of the first attributes of neural network model correctly classified 160 (76.55%) of the instance while Example 49(23.45%) class. The overall accuracy of the velocity model is successful models discussed so far. The model correctly identified 64-patients performed 92 who heart disease and the remaining 28 were identified errors are free from the disease and they actually had the disease. This result

gives the model of 0.766 Purpose price rate. The model was determined in the negative case the better TN rate model was 0.820 correctly identified 96 patients who were enrolled in 117 patients who had no heart disease and the remaining 21 had identified the disease while they did not actually.

Models	Accuracy	TP-Rate	FP-Rate	Precisions	F_Measures	ROC - Area
Neural-network with attribute	75%	0.767	0.248	0.766	0.765	0.845
Neural-network with selected attribute	78%	0.789	0.200	0.798	0.790	0.797

For precision score model labeled as belonging to class positive patients 78% (a) determining a real belong to the class affirmative (YES) and marked as belonging to the class-negative patients 75%(no) is not really a real negative part of the class (no). With 79.8% of the average precision it is in a very successful pattern for each class to retrieve the relevant values. With the 0.790 F-measured values it can be concluded that the accuracy and the recall rate of the model are significantly balanced. The result shows that the neural networks model of the selected properties better than the whole property. The classification accuracy rate increased from 75% to 79.8%. Moreover the execution time decreased significantly from 0.56 to 0.16 seconds.

TABLE VI. DETAIL PERFORMANCE OF ALL IMPLEMENTED ALGORITHMS

Algorithms	Accuracy (%)	TP-rate	FP-rate	Precision	F-measu	ROC-Cur	Time: (sec)
J48-Decision Tree-pruned with all	77.04%	0.771	0.240	0.771	0.771	0.818	0.05
J48-Decision Tree-pruned with selected attribute	79.5%	0.795	0.193	0.805	0.794	0.772	0.04
J48-Decision Tree-un-pruned with all attribute	75.61%	0.757	0.251	0.757	0.757	0.827	0.46
J48-Decision Tree un-pruned with selected	79.43%	0.795	0.193	0.805	0.794	0.772	0.01
Multilayer-perceptron with all	76.56%	0.767	0.248	0.766	0.766	0.844	0.57
Multilayer-perceptron with selected	78.96%	0.788	0.201	0.799	0.791	0.798	0.17
Naïve-bayes with all attribute	79.91%	0.798	0.217	0.798	0.797	0.872	0.03
Naïve-bayes with selected attribute	77.98%	0.781	0.211	0.787	0.780	0.828	0.01

In Table VII, all sub-class algorithms have almost as high as 80% of the remarkable accuracy and precision of a minimum score of 76%. Naïve Bayes classifier to achieve the highest accuracy in the all property (80%) while a Naïve Bayes classifier to achieve a selected attribute it is a 78% of the sub-class accuracy followed On the other hand simple two implement a decision tree classifier score and the entire group selected attribute properties lowest sub class accuracy which were 75% and 77%.

## V. CONCLUSION

Known information mining and Knowledge disclosure (KDD) expressions is utilized to extract the learning (mode) from an extensive number of information acquired is helpful for a given application or information data. From the generated knowledge of the user can determine and meet our requirements. For detecting the heart disease classification and prediction techniques developed in this study. The main aim of this paper is to diagnose heart disease and prevent attacks on people. To this end we use three different monitoring machine learning algorithm to build the model to facilitate the people. Different oversight algorithm is a decision tree classification algorithm using a Bayesian classifier and neural networks 3.8.1 of Weka machine learning software. For predicting heart disease, we have collect the heart patient data from Faisalabad Institute of cardiology contain 209patients, from 2012 to 2015. We use three constructing supervised machine learning algorithms, for example naïve Bayes is plain on j48 and Multilayer Perceptron Weka 3.8.1 machine learning to run the learning software. We established model tests or diagnosed heart disease by pretreatment of chest echocardiographic data sets. All sub-class algorithms have almost as high as 80% of the remarkable accuracy and precision of a minimum score of 76%. Naïve Bayes classifier to achieve the highest accuracy in the all property (80%) while a naïve Bayes classifier to achieve a selected attribute it is a 78% of the sub-class accuracy followed on the other hand simple two implement a decision tree classifier score and the entire group selected attribute properties lowest sub class accuracy which were 75% and 77%.

## VI. FUTURE WORK

With respect to future work the specialists intend to lead progressively extra exploratory informational data and algorithms to enhance sub class precision and have the capacity to assemble the model sort of particular expectation of heart illness. To improve the model further research should be carried out using a classification accuracy of different sub class algorithms such as Support Vector Machines (SVM) and rule induction. Most of the experiments carried out this study the default parameters used to implement the algorithm further studies should use a different set of parameters to carry out, in

order to increase strength and ability to predict model to expand.

## REFERENCES

- [1] Manimekalai, K. (2016). Prediction of Heart Diseases using Data Mining Techniques. International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol, 4.
- [2] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.
- [3] Shafique, U., Majeed, F., Qaiser, H., & Mustafa, I. U. (2015). Data mining in healthcare for heart diseases. International Journal of Innovation and Applied Studies, 10(4), 1312.
- [4] Chandna, D. (2014). Diagnosis of heart disease using data mining algorithm. (IJCSIT) International Journal of Computer Science and Information Technologies, 5(2), 1678-1680.
- [5] Lakshmi, K., Krishna, M. V., & Kumar, S. P. (2013). Performance comparison of data mining techniques for predicting of heart disease survivability. International Journal of Scientific and Research Publications, 3(6), 1-10.
- [6] Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., . . . Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 111(1), 52-61.
- [7] Gamberger, D., Lavrač, N., & Krstačić, G. (2003). Active subgroup mining: a case study in coronary heart disease risk group detection. Artificial Intelligence in Medicine, 28(1), 27-57.
- [8] Chaurasia, V., & Pal, S. (2013). Early prediction of heart diseases using data mining techniques. Caribbean Journal of Science and Technology, 1, 208-217.
- [9] Muthukaruppan, S., & Er, M. J. (2012). A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Systems with Applications, 39(14), 11657-11665.
- [10] Yeh, D.-Y., Cheng, C.-H., & Chen, Y.-W. (2011). A predictive model for cerebrovascular disease using data mining. Expert Systems with Applications, 38(7), 8970-8977
- [11] Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining: MIT press.
- [12] Pham, B. T., Bui, D., Prakash, I., & Dholakia, M. (2016). Evaluation of predictive ability of support vector machines and naive Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. J Geomat, 10, 71-79.
- [13] Khemphila, A., & Boonjing, V. (2010). Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. Paper presented at the Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on.
- [14] Jaiwei, H., & Kamber, M. (2006). Data mining: concepts and techniques. ed: Morgan Kaufmann San Francisco.
- [15] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2005). Weka. Data mining and knowledge discovery handbook, 1305-1314.
- [16] Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining: MIT press.
- [17] Sen, A. K., Patel, S. B., & Shukla, D. (2013). A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level. International Journal of Engineering and Computer Science, 2(9).

# Fuzzy Logic based Approach for VoIP Quality Maintaining

Mohamed E. A. Ebrahim, Hesham A. Hefny  
Computer and Information Science Department  
Institute of Statistical Studies and Research (ISSR)  
Cairo University  
Giza, Egypt

**Abstract**—Voice communication is an emerging technology and has great importance in our routine life. Perceptual, Voice over Internet Protocol quality is an important issue for VoIP Apps services because VoIP Apps require real-time support. Many network factors (packet loss, packet delay, and jitter) affect to VoIP quality, to achieve this objective we used an approach based on Fuzzy Logic. We configure Resource Reservation Protocol application to control Token Bucket Algorithm and the simulation experiments are carried out with Opnet. In addition, compare Token Bucket with and without Quality of Service for measure network factors. In this paper, building Fuzzy Token Bucket System consists of three variables (Bandwidth Rate, Buffer Size, and New Token) in order to improve Token Bucket Shaper output variable (New Token) by Fuzzy Stability model for Voice over IP quality maintaining.

**Keywords**—Voice over Internet Protocol (VoIP); Fuzzy model System (FMS); Fuzzy Token Bucket Algorithm (FTBA); Resource Reservation Protocol (RSVP); Quality of Service (QoS)

## I. INTRODUCTION

VoIP (pronounced as voyp), also known as IP Telephony, is the real-time transmission of voice signals using the Internet Protocol (IP) over the public Internet or a private data network. In simpler terms, VoIP converts the voice signal from your telephone into a digital signal that travels over the Internet as in [1]. (RSVP) Resource reservation is a network protocol that enables IP-based applications to obtain particular (QoS) for data flows. It should be considered as a protocol that delivers QoS requests to the nodes along the data flow path by maintaining appropriate states in these nodes to provide the requested service as in [2], [15]; therefore, VoIP uses a combination of Real-Time Transport Protocol and User Datagram Protocol over IP. UDP, an unreliable service provides no guarantees for delivery and no protection from duplication using IP to transport messages between endpoints in an Internet. RTP, used in conjunction with UDP, provides end-to-end (ETE) network transport functions for applications transmitting real-time data, such as video or voice Apps over network services as in [3], [17]. (RSVP) Resource reservation uses Token Bucket algorithm to maintain QoS attribute. In this paper, we simulate Token Bucket with QoS attribute and

compare results by Token Bucket without QoS attribute then use Fuzzy Token Bucket to enhance this scenario. We use Fuzzy logic to get the way to solve the problems facing VoIP to this day and focused on one these problems Token Bucket in order to improve them. The authors used different ways to improve VoIP, we used RSVP protocol for Token Bucket with applying on Matlab for Fuzzy Logic by using Opnet program results for real-time, and this is what distinguishes our paper from the rest of the articles.

This paper is organized as follows: Section 2 provides an overview of related work. Section 3 shows VoIP Based Token Bucket Rate Section 4 presents fuzzy logic based Token Bucket models. Section 5 presents the proposed models to Comparison of results with and without FL. Section 6 describes the simulation environment. Section 7 discusses the derived results. Section 8 presents the conclusion and the future work.

## II. RELATED WORK

Many types of research had been done for overcoming challenges VoIP QoS, in order to improve performance QoS. They had used different simulators to achieve their goal. In [4], the authors adapted VoIP schemes based on Adaptive multi-rate codec mode to match voice quality to available network bandwidth, the authors focus on using fuzzy logic with the Adaptive multi-rate codec to enhance Priority QoS.

Fuzzy token bucket scheme is compared with token bucket scheme based on two parameters: Average Delay and Throughput was presented in [5], for high-speed ATM networks.

The proposed to carry voice calls over IP networks can generate network congestion due to the weak supervision of the traffic-incoming packet, queuing and scheduling. The authors of [6] presented an approach for using the fuzzy inference system to classify the queuing incoming packet (voice, video, and text); that can reduce recursive loop and starvation.

The authors of [7] used a real-time fuzzy algorithm to estimate the strength of the line echo component of the voice quality in VoIP networks with using Fuzzy Logic to maintain a high level of MOS value in cases of network congestion.

The authors of [8] used four types of different mechanisms, Jump Window, Exponentially Weighted Moving Average, Leaky Bucket and modified Fuzzy Leaky Bucket techniques to be identified, analyzed and simulated by the traffic parameters for peak rate, mean rate and the burst time, which characterize the source behavior.

In [17], improve the quality of signal transmission in video surveillance system based on IP network, the authors focus on one parameter bandwidth only to reduce the time delay of the data packets.

The proposed to introduce a new class of forwarding error correction (FEC) codes for VoIP communications which support different recovery delay depending on the channel conditions. The authors of [18] presented to Experiments over real-world packet traces further show performance gains of DD codes in terms of perceptually motivated ITU-T G.107.E-model.

Optimization of VoIP network performance based on voice call routing and network reorganization the authors of [16] used reorganization to meet the requirements of VoIP networks deployment at their base has been proposed and focus on VoIP algorithm.

The authors have made an excellent effort to improve performance VoIP QoS; we participated in the effort to reach a new research point by using fuzzy logic based on different rules for three scenarios to reach the best results not achieved in previous works. This work presents a set of ideas combined as simulate VoIP network with and without QoS based on RSVP Protocol by Opnet and configure RSVP and building fuzzy interface system (FIS) based on three variables by Matlab and compare results as in Section 5.

### III. VOIP BASED TOKEN BUCKET RATE

Attributes describing Resource Reservation Protocol (RSVP) Parameters set by the application are defined in two objects: The QoS Attribute Configuration object and the Application Attribute Configuration object. To run an RSVP simulation, both objects must be included in the scenario. Bandwidth (bytes/sec) specifies the amount of traffic generated by the application at the IP level (including TCP/UDP and IP headers). This value is set to be the token bucket rate in flow specification of Path and Reservation messages. When the reservation is made using this flow specification, this value is set as the reserved bandwidth for the session according to [9], [15].

Because the controlled load service does not precisely control packet delay, any device implementing the controlled load service should not penalize bursts of packets from an application. It should be possible to buffer bursty data. The amount of data that should be buffered can be configured using the Buffer Size (bytes) attribute. This value is used as "bucket size" and is set in Path or Reservation messages for the session. When a reservation is made, the Buffer Size value is the size of the buffer created for data of a particular flow (each queue in the implementation is defined by a bandwidth

and buffer size). The amount of data traffic sent over all time periods should not exceed  $r * T + b$  where  $r$ , is the **token bucket rate** which is the value of **Bandwidth** attribute,  $b$  is **token bucket size** which is the value of **Buffer Size**, and  $T$  is the measurement interval in seconds as in Opnet modular documentation, 2014 [10], [3].

Token Bucket Algorithm	
(r, Max Tokens)	
-	Generate r tokens every time unit If number of tokens more than Max Token, reset to Max Tokens
-	For an arriving packet: enqueue r
-	While buffer not empty and there are tokens: send a packet and discard a token $\sigma = \text{Max Tokens} \ \& \ \rho = r / \text{time unit.}$ <i>What does a router need to support stream:</i>
( $\sigma_1, \rho_1$ ) ... ( $\sigma_k, \rho_k$ )	
-	Buffer size $B > \sum \sigma_i$
-	Rate $R > \sum \rho_i$ <i>Admission control ( at the router )</i>
-	Can support ( $\sigma_k, \rho_k$ ) if
-	Enough buffers and bandwidth $R > \sum \rho_i$ and $B > \sum \sigma_i$

### IV. FUZZY LOGIC BASED TOKEN BUCKET MODELS

The fuzzy logic predictor predicts the token bucket rate required based on the average buffer size rate and available bandwidth rate according to [11]. The fuzzy inputs are the Average Buffer Size Rate 'BufferSize' and Available Bandwidth Rate 'BandwidthRate'. The output is the New Token Bucket Rate denoted by 'NewToken'.

The linguistic values are given below:

$$\text{BufferSize} = \{VL, L, M, H, VH\}$$

$$\text{BandwidthRate} = \{VL, L, AL, BA, AV, AA, BH, H, VH\}$$

$$\text{NewToken} = \{VL, L, BA, AV, AA, H, VH\}$$

Where, the input BufferSize variable is divided into five fuzzy subsets as shown in Table I.

TABLE I. FIS BUFFERSIZE PARAMETERS

VL	Very Low
L	Low
M	Medium
H	High
VH	Very High

These values are normalized in the range of [0,1]. We use triangular membership functions, S-shaped and Z-shaped functions based on the simplicity of these kinds of functions as Mathworks, 2016 [12] mentioned, is as in (1) and Fig. 1. Triangle-Shaped Membership Function:

$$f(x; a, b, c) = \max(\min(\frac{x-a}{b-a}, \frac{c-x}{c-b}), 0) = \quad (1)$$

Where, the parameters a and c locate the "feet" of the triangle and the parameter b locates the "peak".

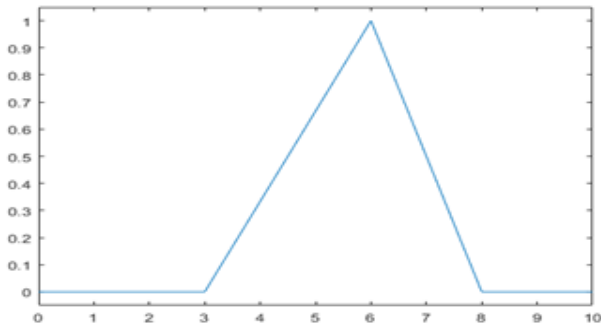


Fig. 1. Triangle trimf, P = [3 6 8] [12].

After declaring input and output variables, the membership functions are plotted i.e. the range is defined between 0 and 1 in a normalized form according to [13], [5].

There are five membership functions for FIS BufferSize variable. The value of the variable is increasing from very low to very high. The lower value is VL and the higher value is VH. Therefore, the name of the variable is given according to the strength of variable. These are shown in Table II as in [8].

TABLE II. MEMBERSHIP FUNCTION PARAMETERS

Name	Type	Parameter
VL	zmf	0 0.25
L	trimf	0.0 0.25 0.5
M	trimf	0.25 0.5 0.75
H	trimf	0.5 0.75 1
VH	smf	0.75 1

As shown in Fig. 2, the input variable 'BufferSize' has five membership functions. 'BandwidthRate' was described in Fig. 3, nine membership functions and in Fig. 4, the output variable 'NewToken' has seven membership functions which are of three types: Membership function VL is of Z – shaped, membership function VH is of S-shaped and from L to H triangular shaped membership function. The input and output variables also have different membership functions. The linguistic variables are those variables whose values are words rather than numbers. Much of fuzzy logic may be viewed as a methodology for computing with words rather than numbers according to [14], [8].

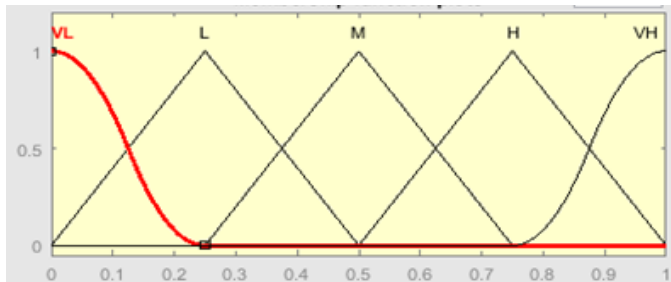


Fig. 2. Membership function for BufferSize.

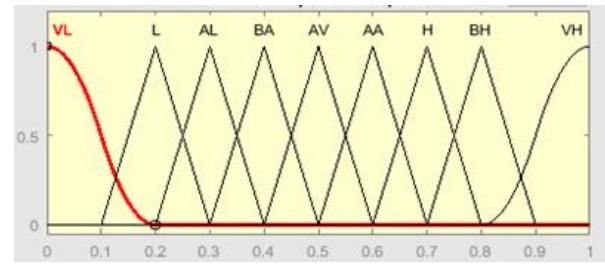


Fig. 3. Membership function for bandwidth rate.

Use Fuzzy sets and fuzzy operators as the subjects and verbs of fuzzy logic to form rule. These “if-then rule” statements are used to formulate the conditional statements that include FL. A single fuzzy if-then rule i.e. assumes the form If (BufferSize is L) and (BandwidthRate is VL) then (New Token is VL); where L and VL are linguistic values defined by fuzzy sets on the ranges (universes of discourse) Low and Very Low respectively. The if-part of the rule “BufferSize is L and BandwidthRate is VL” is called the antecedent or premise, while the then part of the rule “NewToken is VL” is called the consequent or conclusion.

The fuzzy engine has 45 rules that relate the two inputs with the fuzzy output. The constructions of the rules are based on logical reasoning of how the system can track bandwidth usage. It is the normalized form of inputs and output according to [20].

TABLE III. FUZZY CONDITIONAL RULES FOR THE POLICER

```
ruleBlock->addRule(fl::Rule::parse("if BandwidthRate is M and BufferSize is VL then NewToken is L",engine));
ruleBlock->addRule(fl::Rule::parse("if BandwidthRate is M and BufferSize is L then NewToken is L", engine));
ruleBlock->addRule(fl::Rule::parse("if BandwidthRate is H and BufferSize is L then NewToken is M", engine));
ruleBlock->addRule(fl::Rule::parse("if BandwidthRate is H and BufferSize is H then NewToken is H", engine));
ruleBlock->addRule(fl::Rule::parse("if BandwidthRate is VH and BufferSize is L then NewToken is H",engine));
```

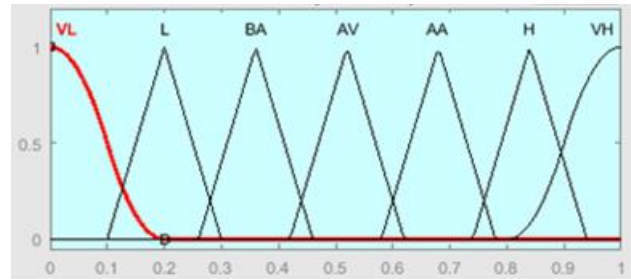


Fig. 4. Membership function for NewToken.

It is observed that the z-axis variable smoothly varies from 0 to 1, i.e., the new token generation rate is varying in continuous form. In a token bucket with threshold scheme, there was a sudden change in the token generation rate, and this problem is overcome in fuzzy token bucket scheme. Hence this fuzzy token bucket scheme works more efficiently than token bucket scheme. Fuzzy conditional rules for the policer are given in Table III.

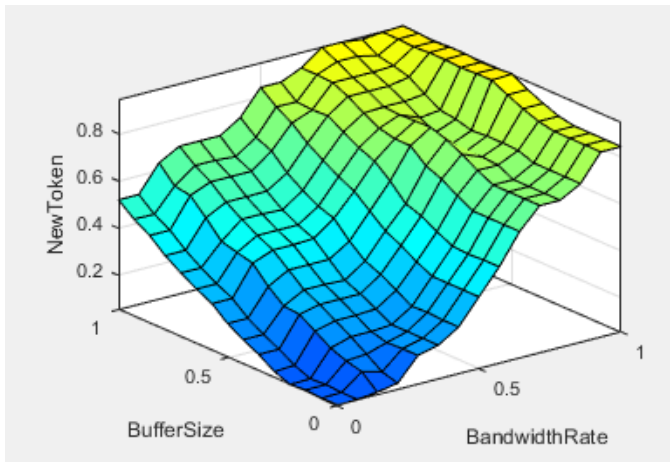


Fig. 5. Surface viewer fuzzy token bucket.

From Fig. 5, the fuzzy token bucket is clear that whenever increase bandwidth rate and buffer size rate the result new token bucket high quality. We conclude from this the strength Fuzzy Token Bucket Shaper depends on bandwidth rate and buffer size rate, and thus lower rate Packet Delay, jitter and highest quality scale Mean Opinion Score (MOS). And here we have succeeded in experience Fuzzy logic to define the criteria on which depend for maintaining quality voice over internet protocol.

V. COMPARISON OF RESULTS WITH AND WITHOUT FL

The proposed fuzzy models were tested by OPNET Modeler 14.5 and MATLAB R2016a fuzzy toolbox. Three scenarios were used, first scenario VoIP with QoS, second scenario VoIP without QoS, and third scenario compare the previous result with VoIP based fuzzy logic token bucket. After what we succeeded in identifying the causes of weakness Token Bucket Shaper, we assume Bandwidth Rate and Buffer Size in perfect condition. The proposed explain different factors End-to-End Delay, Delay Variation, Received and sent packet traffic and Mean Opinion Score MOS scale between VoIP with and without Token Bucket and VoIP with fuzzy Token Bucket as in ITU-T Recommendation [19].

In this step, use high Bandwidth Rate and Buffer Size for getting New Token is performed well and work outstanding as shown in Fig. 6.

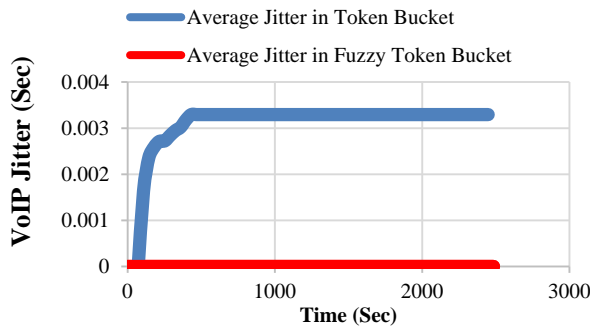


Fig. 6. Comparison of jitter with and without fuzzy logic.

$$JITTER = (T4 - T3) - (T2 - T1) \tag{2}$$

Fig. 6 compares the Jitter by using a fuzzy token bucket and non-fuzzy token bucket, the time is shown on x-axis and Jitter rate is shown at the y-axis. Blue line depicts Average Jitter in a token bucket and red line depicts the Jitter by the fuzzy token bucket. Therefore, the fuzzy token bucket is performed very well compared with a non-fuzzy token bucket.

Fig. 7 compares the Delay variation by using a fuzzy token bucket and token bucket with and without QoS. The time is shown on the x-axis and Delay variation rate is shown at the y-axis. Blue line depicts Delay variation with a token bucket, brown line depicts Delay variation without QoS, and red line depicts the Delay variation by the fuzzy token bucket. The result is no Delay variation by the fuzzy token bucket.

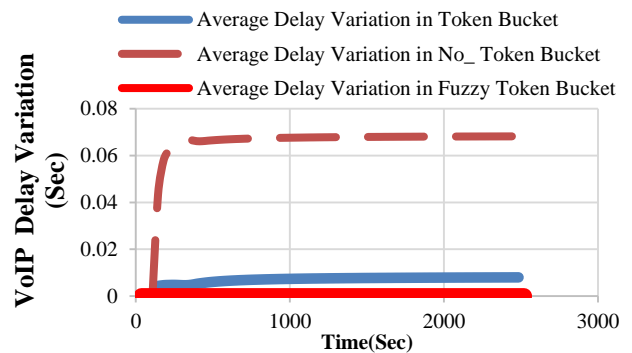


Fig. 7. Comparison of delay variation with and without fuzzy logic.

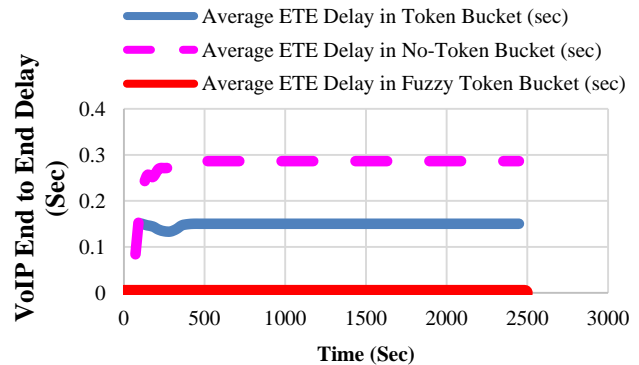


Fig. 8. Comparison of end-to-end (ete) delay with and without fuzzy logic.

$$DELAY = (NETWORK\_DELAY + ENCODING\_DELAY + DECODING\_DELAY + COMPRESSION\_DELAY + DECOMPRESSION\_DELAY) \tag{3}$$

Fig. 8 compares the total voice packet delay called End-to-End Delay by using a fuzzy token bucket and token bucket with and without QoS, the time is shown on the x-axis and End-to-End Delay rate is shown at the y-axis. Light Blue line depicts Delay variation with a token bucket, Fuchsia line depicts End-to-End Delay without QoS, and red line depicts the End-to-End Delay by the fuzzy token bucket. The result is no ETE Delay by the fuzzy token bucket.

Fig. 9 compares the Mean Opinion Score (MOS) by using a fuzzy token bucket and token bucket with and without QoS, the time is shown on the x-axis and MOS scale is shown at the y-axis. Red line depicts MOS by the fuzzy token bucket, Purple line depicts MOS with a token bucket, and Blue line depicts MOS without QoS. The result is high accuracy for VoIP.

Fig. 10 compares the data loss by using a fuzzy token bucket and non-fuzzy token bucket the time is shown on x-axis and data loss rate is shown at the y-axis. Blue line depicts data loss with a token bucket and red line depicts the data loss by the fuzzy token bucket. The result is no Packet Loss by the fuzzy token bucket.

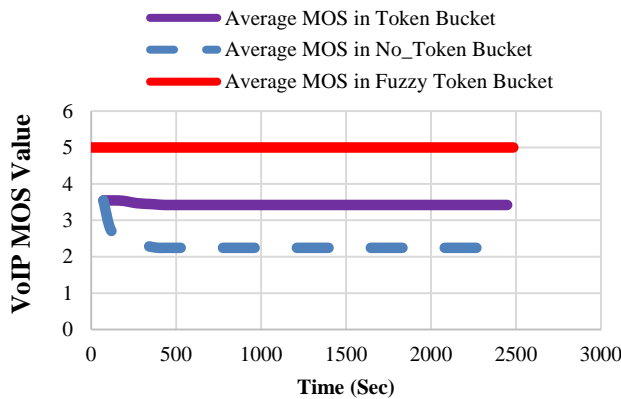


Fig. 9. Comparison of MOS with and without fuzzy logic.

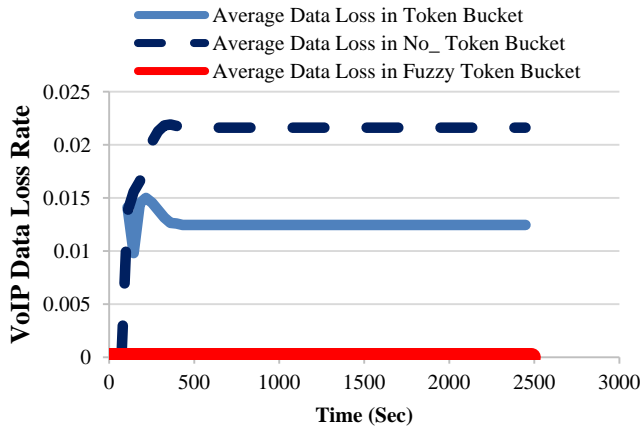


Fig. 10. Comparison of data loss with and without fuzzy logic.

## VI. DISCUSSION

In Fig. 6 comparison between Average Jitter by using a fuzzy token bucket and non-fuzzy token bucket as is evident in the graph. Use fuzzy model has a clear effect on the new token bucket and give rise to zero rates of Jitter. In addition, try another test fuzzy logic token bucket with delay variation parameter in Fig. 7. The result was in favor of fuzzy token bucket after comparison of Delay variation with and without QoS token bucket then add a fuzzy token bucket. Then we repeated another comparison by the end-to-end delay in Fig. 8. The result was clear in graph packet ETE Delay in fuzzy token

bucket much better than packet ETE Delay in a token bucket without a fuzzy model. In Fig. 9 we compare the Mean Opinion Score (MOS) by using a fuzzy token bucket and token bucket with and without QoS The result in the graph is high score 5 value in favor of fuzzy token bucket while QoS token bucket is 3.4 value and token bucket without QoS is 2.2 value and also in Fig. 10. We use packet Data loss parameter for comparison data loss in a fuzzy token bucket and non-fuzzy token bucket, the result was no data loss with a fuzzy token bucket, and here it clearly shows that fuzzy model has an effect on token bucket algorithm shaper. We have succeeded in proving Fuzzy logic define the criteria on which depend for maintaining quality voice over internet protocol (VoIP).

## VII. CONCLUSIONS AND FUTURE WORK

This thesis has discussed a proposed approach for improving VoIP quality. The Objective of thesis was developed an approach for VoIP QoS, we used fuzzy logic control system for improve QoS by apply fuzzy token bucket and depend on Fuzzy Stability model (FTBS) to determine the optimal way to get new token bucket do not cause any factors as (jitter – Delay – packet loss) and compare with multi scenarios for parameters Jitter, End to End Delay, Delay Variation, packet traffic and Mean Opinion Score MOS scale between VoIP with and without QoS and VoIP with fuzzy Token Bucket. And was the result of research whenever improving bandwidth rate and buffer size rate the result a new token bucket at the same rate of improvement and vice versa. Therefore, Bandwidth should be sufficient for traffic since insufficient bandwidth may decrease QoS for the flow and buffer size also. Future work will be to analyze the working of VoIP on mobile 5th generation internet. The performance of fuzzy predictor for packet traffic rate in 5th generation mobile networks will also be a future scope of the project.

## ACKNOWLEDGMENT

First and foremost, I would like to express my deepest sense of gratitude to my supervisor Prof. Dr. Hesham Ahmed Hefny who has been supporting my studies using his vast knowledge and skill in many areas, guiding my thesis and constant involvement in guiding me towards my goal, also for a lot of kindness, patience and effectual support. My search could not have been finished reasonably without insightful advice from him.

My thanks and appreciation to all other ISSR professors for many things I learned from them during my study in ISSR.

Finally, I'm deeply grateful to my family, especially my wife, and my sons for their love, support, and patience during writing the thesis.

## REFERENCES

- [1] Joe Hallock, "A Brief History of VoIP" Evolution and Trends in Digital Media Technologies-COM538, University of Washington. 26<sup>th</sup> November 2004.
- [2] Nicos A. Antoniou, "Experimental Study of RSVP over ATM" San Diego State University, 2001.



- [3] Chintan Vaishnav, "Voice over Internet Protocol (VoIP): The Dynamics of Technology and Regulation" The Massachusetts Institute of Technology, USA, 12<sup>nd</sup> May 2006.
- [4] E. Jammeh, I. Mkwawa, L. Sun, E. Ifeakor, "Type-2 fuzzy logic control of PQoS driven adaptive VoIP scheme" *Electronic Letters*, **46**(2), United Kingdom 21<sup>st</sup> January 2010.
- [5] Anurag, A. "Fine Tuning of Fuzzy Token Bucket Scheme for Congestion Control in High-Speed Networks" In: IEEE Second International Conference on Computer Engineering and Applications, 1, pp. 170–174. March 2010.
- [6] Suardinata, Bin Abu Bakar, N. Suanmali, N. "comparison process long execution between PQ algorithm and new fuzzy logic algorithm for VoIP" *International Journal of Security and Its Application (IJNSA)*, **3** (1), January 2011.
- [7] Oyetade Durojaiye, Elizabeth. N. Onwuka, "Voice Quality Evaluation of a Call Using Fuzzy Logic" *International Journal of Network and Communication*, **2**(2): 7-12, 2012.
- [8] Ming-Chang Huang, Seyed Hossein Hosseini, K. Vairavan, and Hui LAN, "Fuzzy Congestion Control and Policing in ATM Networks" *International Journal of Engineering (IJE)*, **3**(1), 2012.
- [9] Lixia Zhang, Stephen Deering, Deborah Estrin, Scott Shenker, Daniel Zappala, "RSVP: A New Resource Reservation Protocol" *IEEE Network Magazine*, September 1993.
- [10] Opnet modular documentation, 2014.
- [11] Demet Dilekci, Conrad Wang, Jiang Feng Xu. "The Analysis and Simulation of VoIP" *ENSC 427 Communication Networks*, Spring 2013.
- [12] <http://www.mathworks.com/help/fuzzy/trimf.html>
- [13] Anurag Aeron, C. Rama Krishna, Mohan Lal. "Performance Evaluation of Fine Tuned Fuzzy Token Bucket Scheme for High-Speed Networks" *CCSIT, Part2, CCIS 132*, PP. 126-136 Springer-Verlag Berlin Heidelberg 2011.
- [14] R. Alcalá, J. Casillas, O. Cordón, F. Herrera, S. J. I. Zwiry, "Techniques for Learning and Tuning Fuzzy Rule-based system for Linguistic Modeling and their Application" *CICYT TIC96-0778 Spain* 2010.
- [15] Esmat Mirzamany, Aboubaker Lasebae, and Orhan Gemikonakli "Using aggregated RSVP in nested HMIPv6" *IEEE Conference Publications*, 2012.
- [16] Alexander Soloviev, Victor Bondarenko, "Optimization of VoIP network performance based on voice call routing and network reorganization" *IEEE*, 2017.
- [17] Zhou Lin, Chen Yingmei, Li Zhen, He Zhuzhen, "An Improved Video Monitoring System Based on RSVP Protocol" *IEEE*, 2015.
- [18] Ahmed Badr, Ashish Khisti, Wai-tian Tan, Xiaoqing Zhu, John and G. Apostolopoulos, "FEC for VoIP using dual-delay streaming codes" *IEEE*, 2017.
- [19] ITU-T International Telecommunication Union Recommendation: Traffic control and congestion control in B-ISDN Section 7.2.7, defines traffic shaping as a traffic control mechanism, I.371, March 2004.
- [20] Hamdy, A.M., Sayedahmed, Hefny, H.A., Fahmy, I.M.A. "Improving Multiple Routing in Mobile Ad Hoc Networks Using Fuzzy Models" *Springer International Publishing AG* 2018. *Conference on Advanced Intelligent Systems and Informatics 2017, Advances in Intelligent Systems and Computing* 639. 2017.

# Conceptual Modeling of a Procurement Process

## Case study of RFP for Public Key Infrastructure

Sabah Al-Fedaghi  
Computer Engineering Department  
Kuwait University  
Kuwait

Mona Al-Otaibi  
Information Technology Department  
Ministry of Finance  
Kuwait

**Abstract**—Procurement refers to a process resulting in delivery of goods or services within a set time period. The process includes aspects of purchasing, specifications to be met, and solicitation notifications as in the case of Request For Proposals (RFPs). Typically, such an RFP is described in a verbal ad hoc fashion, in English, with tables and graphs, resulting in imprecise specifications of requirements. It has been proposed that BPMN diagrams be used to specify requirements to be included in RFP. This paper is a merger of three topics: 1) Procurement development with a focus on operational specification of RFP; 2) Public key infrastructure (PKI) as an RFP subject; and 3) Conceptual modeling that produces a diagram as a supplement to an RFP to clarify requirements more precisely than traditional tools, such as natural language, tables, and ad hoc graphs.

**Keywords**—Procurement; RFP; public key infrastructure; conceptual modeling; diagrammatic representation

### I. INTRODUCTION

Procurement refers to “a careful, usually documented process resulting in delivery of goods or services within a set time period” [1]. In project management the process includes aspects of purchasing, specifications to be met, and solicitation notifications. Procurement, also known as purchasing and supply, “is amongst the key links in the supply chain and as such can have a significant influence on the overall success of the organization” [2]. Without loss of generality the present study focuses on the first phase of the procurement process, which includes needs specification and construction of the request for proposal (RFP).

#### A. Problem and solutions

Typically an RFP is described in a verbal ad hoc fashion, in English, with tables and graphs, resulting in imprecise specifications of requirements. Challenges of the traditional RFP approach include difficulty in holding vendors accountable, and contract management issues that often result in massive change requests and overruns [3].

Organizations that are in the process of developing a Request for Proposal (RFP) have often looked to existing sources for ideas on *how to phrase language* to cover a specific topic. They are often disappointed to learn that the search for RFP language examples is a time-consuming exercise that involves searching across multiple publications that may or may not include the topical information that they seek. (Italics added)

According to [4], it is quite common to see RFPs with requirements that are very broad, derived from a vendor’s list of features, or copied from another organization’s RFP. Among their suggested remedies is to prepare diagrams of the RFP process. “Model your business process graphically. Business process diagrams (or models) are excellent at showing gaps in the process or errors in your understanding” [4]. They particularly recommend Swim Lane diagrams.

Hadrian and Evequoz [5] enumerate the main difficulties in RFP requirements specification:

- Expressing precisely what will be needed (i.e., specific requirements and attaching requirements to specific parts in a process).
- Expressing requirements in a standardized form.
- Tracing requirements coming from different sources

In general, according to Hadrian and Evequoz [5], a methodology to produce more precise requirement specifications would be helpful for all stakeholders. Requirements should be unambiguous and validated by business users. Hadrian and Evequoz [5] proposed use of BPMN diagrams [6] to specify requirements to be included in Request for Proposals. BPMN is an International standard for process documentation that bridges the gap between business and IT people.

Similarly, we propose applying a conceptual model (the Flowthing Machine, FM) that can be used to facilitate creation of RFP specifications. This can then be used by all stakeholders in the process, since FM is a conceptual model that can be understood without substantial knowledge of technical details. Hence, the aim in the next section is to demonstrate that FM can be utilized as a tool for a comprehensive expression of what is needed. It is understood that, initially, developing an RFP entails a certain amount of guesswork about details. An advantage of FM is that the drawing can be modified fairly easily as details evolve.

#### B. Additional problem: Communication among stakeholders

An additional problem in requirements specification for an RFP is related to communication among stakeholders. In a government RFP [7], it is stated that,

The assumptions, assessments, statements and information contained in this RFP, may not be complete, accurate, adequate or correct. Each Bidder should, therefore, conduct

their own investigations and analysis and should check the accuracy, reliability and completeness of assumptions, assessments and information contained in this RFP and obtain independent advice from appropriate sources.

A general aim of this paper is to introduce a modeling language that expresses the technical parts of the RFP in a “neutral” representation that facilitates communication among stakeholders.

Public key infrastructure (PKI) is intentionally selected as the content of RFP because “all of the books or Web sites on the subject either assume that you already know all about PKI or they use so many big words that they are hard for a beginner to understand” [8]. PKI is suitable as a test case for communication among stakeholders by providing a non-technical language that underlies the RFP.

A *neutral* (i.e., independent of whatever technology is used) *representation*, mentioned previously, is a product of the FM conceptual model. This paper considers the topic of conceptual *modeling* in order to demonstrate its advantages in the field of software engineering for procurements. Consequently, this paper is a merger of three areas of study:

- 1) Procurement development with a focus on operational specification of RFP.
- 2) Public key infrastructure (PKI) as an RFP subject.
- 3) Conceptual modeling that produces a diagrammatic description as a supplement to the RFP for clarifying requirements in a more precise manner than traditional tools such as natural language, tables, and ad hoc graphs.

### C. Conceptual modeling

Twenty years ago, modeling of systems was viewed as a great discovery for accelerating resolution to challenges to manufacturing industries by 2020 [9]. One major scientific area that embraces modeling is software engineering. Software is everywhere in the infrastructure and affects all fields of life. Software engineers deal with more complex problems than any other engineering discipline [10]. Decades of work on software abstraction have helped gain intellectual control over systems of ever-increasing complexity. This has motivated adopting a modeling approach throughout the software development process with tools such as UML and SysML.

According to Armstrong [3], the traditional RFP process involves a phased approach similar to a waterfall: a requirements specification phase, system requirements phases, a design phase, and an implementation phase. Requirements specification is a basic phase in software lifecycle system development. Software engineers have put much effort into the process of transforming requirements into software architecture, including creating a *text description* of the envisioned system as well as creating *models*. The key problem is to give an unambiguous, easy to understand description of a system and how it works. “We can do so with English descriptions; but such descriptions are often cumbersome, incomplete, ambiguous and can lead to misunderstandings” [11].

Armstrong [3] recommended incorporating Agile into an adaptive collaborative development process, *significantly*

*leveraging UML for modeling*, using a comprehensive traceability strategy, and automatically generating RFPs. In the first iteration of the process, “a business *use case model* that include[s] coarse-grained business workflow diagrams (activity diagrams) [and] business use case outlines” [3]. Later the process would incorporate development of UML collaboration diagrams for business use cases, and class diagrams for business participant responsibilities.

Douraid et al. [2] modeled the procurement process at the operational level by using UML to describe the static and dynamic behavior of the system [12]. “UML is not restricted to modeling software. It is also used for business process modeling, systems engineering modeling and representing organizational structures. It is a general-purpose modeling language that includes a graphical notation used to create an abstract model of a system. It is designed to specify, visualize, construct, and document software-intensive systems [2]”.

### D. Approach

The aim of this paper is to supplement the RFP with a model, i.e., diagrams that express how the features and services of PKI would logically operate in the proposed system. Such an approach is not new, and the following is an illustrative example.

In requests for proposals by the Judicial Council of California [13], proposers must respond to “Use Case Scenarios with a narrative response describing how their product features and or services will excel or be challenged in addressing these use case scenarios.” An example (supported by a diagram) of such a use case is as follows:

A person, business or government agency brings a document to the clerk’s office. The clerk records the document in the Case Management System (CMS) and receives a case number from the CMS (either for an existing case or as a newly filed case). A cover sheet is produced that contains the information that will be used as index values for this document. The cover sheet and document will be scanned into the Document Management System.

The authors [13] provide a sample diagram of the PKI process accompanying an RFP showing how the agency conceives the workings of the PKI system. This does not impose a rigid method; rather it is an initial “solution” to the problem that the agency tries to solve; and the bidder can respond with a counter model that is a modification or replacement of this conceptualization (see Fig. 1).

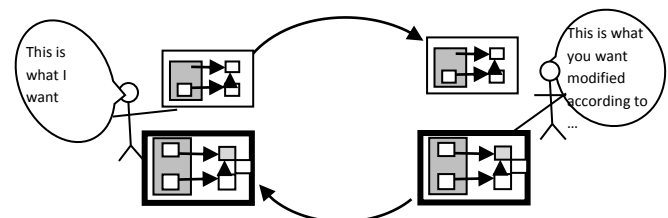


Fig. 1. Diagram showing how the system works.

## II. FLOWTHING MACHINE (FM)

This section briefly reviews the FM model that forms the foundation of the theoretical development in this paper; however, the example given here is a new contribution.

### A. Basic notions

The FM model (see [14–23]) is a diagrammatic schema that uses *flowthings* (hereafter, *things*), defined as *what can be created, released, transferred, received, and processed*, by means of stages in a flow machine (Fig. 2). *Things* begin to flow through the stages of the machine when they are created by the machine or imported from other machines.

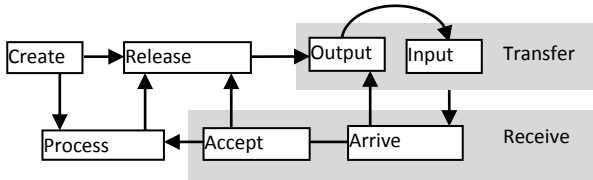


Fig. 2. Flow machine.

*Flow* here entails transition or realization of change as well as movement and positioning. **Create** is the emergence of a thing in the system from outside it. The rest of the flow is succession from one stage to the next. Such flows are specified in analogy to drawing traffic flows on a city map. There, as will be discussed later, *dynamic* flows are shown in terms of *events* that describe the behavior of the system, when the streets of the city become streams of flow of cars, pedestrians, etc.

The point here is that the *flow* is often thought of as physical movement, but in FM, it can be much more than that. It is a notion that captures the *conceptual* movement of thought, sensation, being, and doing. The modeler builds a conceptual construct and also a conceptual “movement”; we call it flow. Thus, a physical house *flows* from one sphere (e.g., a *class* in UML terminology) to another when there is a change in ownership from a person to a certain bank, and a car on an assembly line *flows* to robots and workers *simultaneously* when it is processed, e.g., one fixes glass while another puts on tires, etc. Flows might be fast or slow, parallel or sequential, physical or digital (e.g., uploading software) or mental (e.g., inspecting finished products), or comprise only creating, only processing, etc.

The stages in Fig. 2 can be described as follows:

**Arrive:** A thing reaches a new machine.

**Accept:** A thing is approved to enter a machine. If arriving things are always accepted, *Arrive* and *Accept* can be combined as a **Receive** stage.

**Process** (change): The *thing* goes through some kind of transformation that changes its “state” without creating a new thing.

**Release:** A thing is marked as ready to be transferred outside the machine. Note that things can be released from a given system without being transferred, as in the case of sent emails waiting for a damaged channel to be fixed.

**Transfer:** The thing is transported somewhere from or to outside the machine.

**Create:** A new thing is born (created) in a machine.

Flow machines use the notions of *spheres and subspheres*. These are constructs (mental conceptions) of machines and submachines. Multiple machines can exist in a sphere if needed. A sphere can be a person, an organ, an entity (e.g., a company, a customer), a location (a laboratory, a waiting room), a communication medium (a channel, a wire). A machine is a subsphere that embodies the flow; it itself has no subspheres. This sphere notion is taken from cognitive linguistics where an *idea* is treated as complex units associated with other entities or other forms of association. “A door, for example, also connotes a door knob, a key hole, a door jamb, etc.” [17].

FM also utilizes the notion of *triggering*. Triggering is the activation of a flow, denoted in the machine diagrams by a *dashed arrow*. It is a dependency relationship among flows and parts of flows. A flow is said to be triggered if it is created or activated by another flow (e.g., a flow of electricity triggers a flow of heat), or activated by another point in the flow. Triggering can also be used to initiate events such as starting up a machine (e.g., by remote signal). Multiple machines can interact by triggering events related to other machines in those machines’ spheres and stages.

### B. Example

Douraid et al. [2] introduced a model for generally depicting a procurement process, including supplier management, inventory management, and invoicing and delivery procedures. Their set of conceptual and UML models was designed for use in constructing a simulation framework for a procurement process. “The behavioral aspect is captured from activity and state diagrams to characterize the dynamic side of our approach” [2]. Fig. 3 and 4 show partial views of their state and activity diagrams.

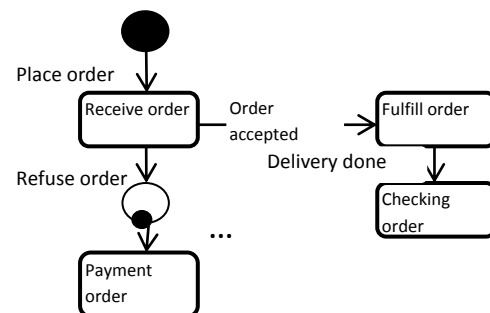


Fig. 3. Order state diagram (redrawn, partial from [2]).

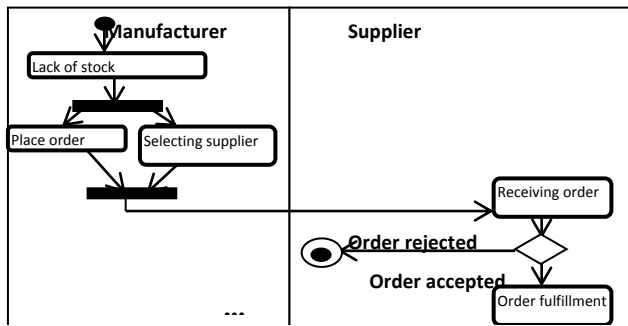


Fig. 4. Supplier-manufacturer relationship activity diagram (redrawn, partial from [2]).

Fig. 5 shows the corresponding FM representation of this supplier-manufacturer relationship. First, the storage of the manufacturer (circle 1 in the figure) is processed (checked), and if there is a lack of stock (2) then this *triggers*,

- Generating data, e.g., item name, quantity (3), and
- Selecting a supplier (4)

Accordingly, these two *things* flow to an ordering management procedure (5) that triggers the creation of an order (6).

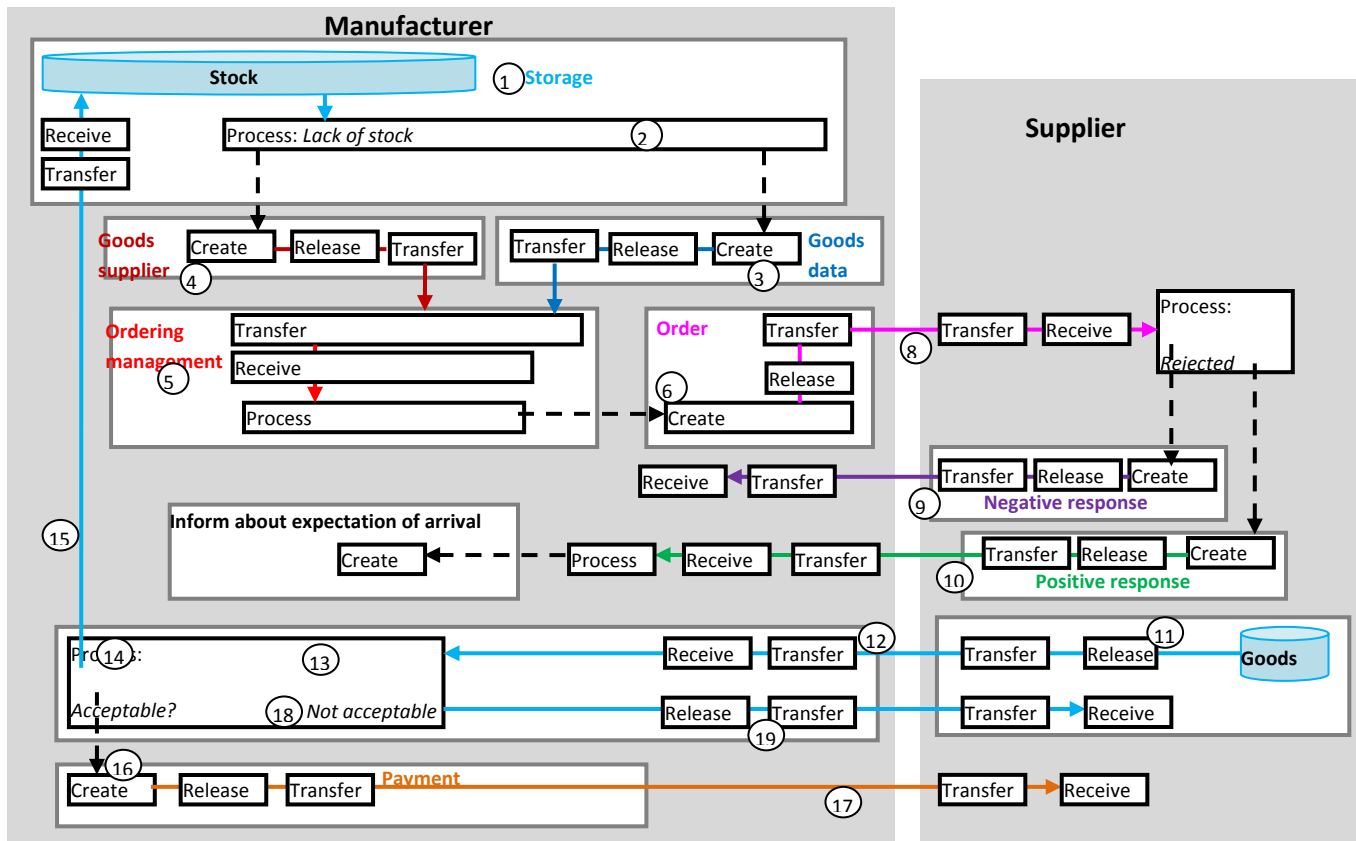


Fig. 5. FM representation of the example

This manual process increases costs and time, and impede the benefits of a fully electronic workflow. Digital Signatures provide a solution for creating legally enforceable electronic

The order flows to the supplier (8) where it is processed.

- If the order is rejected, a negative response is sent back (9).
- If the order is accepted, a positive response is sent (10). Additionally, the goods are released (11) and sent to the manufacturer (12).

There the goods are processed (13).

- If acceptable, (14) they are sent to storage (15). Additionally, a payment is made (16) and sent to the supplier (17).

If the goods are not acceptable (18), they are returned to the supplier (19).

### III. CASE STUDY: PUBLIC KEY INFRASTRUCTURE

The aim of eGovernance is to automate government operations, business processes, and service delivery online. As a result, electronic documents are infiltrating every aspect of the government workflow. Difficulties arise when a signature authorization is needed that requires a physical signature.

records while eliminating the need to print documents for signing.

A digital signature can be used to authenticate the identity of the sender of a message or the signer of a document. Here we assume general knowledge of public key cryptography since a digital signature requires a key pair: the *Public* and *Private Keys*.

The private key is retained by the owner and the public key is public for everyone. Information encrypted by a private key can be decrypted only by means of the corresponding public key. Because of our case study, this paper focuses on certificate authorities (CAs) instead of such approaches as web of trust and simple public key infrastructure.

*Public Key Infrastructures* is a support system for usage of public key cryptography [24]. It includes all hardware, software, people, policies, and procedures for creating and handling digital certificates and manages public-key encryption. This is accomplished through (i) providing digital signatures with (ii) verification of the ownership of public keys. Common PKI functions include issuing certificates, revoking certificates, storing and retrieving certificates. Enhanced functions include time-stamping and policy-based certificate validation.

#### A. How to create a digital signature

In a digital signature, a process called "hashing" converts the data to what is called a message digest which is encrypted with the private key to produce the digital signature that is appended to a document.

*Example* (from [8]): Suppose that **I** need to send **you** an e-mail message. Assume that the message does not need to be encrypted, but that what is needed is as follows (see Fig. 6):

- Assurance that the message came from **me**.
- Verification that the message was not intercepted and altered in transit.

Assume that the message is: The check is in the mail.

1) **I** produce a non-reversible hash of the message. That is, I create a hash by adding together the ASCII values of each character in the message:  $84 + 104 + 101 + 32 + 99 + 104 + 101 + 99 + 107 + 32 + 105 + 115 + 32 + 105 + 110 + 32 + 116 + 104 + 101 + 32 + 109 + 97 + 105 + 108 + 46 = 2180$ . The hash 2180 is non-reversible because there is no way that we produce from 2180 the message: The check is in the mail.

2) The hash is appended to the end of the message: The check is in the mail.

3) **I** use my private key to encrypt the hash value 2180 and append it to the end of the message before I transmit it to you.

4) When **you** receive the message, you calculate the message's hash by using the same algorithm that was used to produce the hash in the first place. If **you** calculate the same value as the hash value that is appended to the end of the message, then you can be sure that the message has not been altered in transit.

5) **You** use my public key to decrypt the hash value. If you are able to do this successfully, then you know beyond doubt that I am the one who encrypted the hash value.

#### B. Certificates

A PKI is based on things called certificates that are issued by the Certificate Authority (CA) and serve as digital identification. Certificates associate users with their public keys. They can be created by way of software, and we limit our interest in this paper to a standard that defines the format of public key certificates required in the case study that will be discussed later.

We assume here that the CA generates the public and private keys for the user. The public key has to be signed by the CA, where:

1) The CA uses a hash algorithm to generate the so-called digest.

2) The digest is encrypted with a private key. The result is a digital signature.

3) The CA then makes the digitally signed certificate available for download to the person who requested it.

In general the Public Key Infrastructure works as follows:

A user applies for a certificate with his public key at a registration authority (RA). The latter confirms the user's identity to the certification authority (CA) which in turn issues the certificate. The user can then digitally sign a contract using his new certificate. His identity is then checked by the contracting party with a validation authority (VA) which again receives information about issued certificates by the certification authority [25].

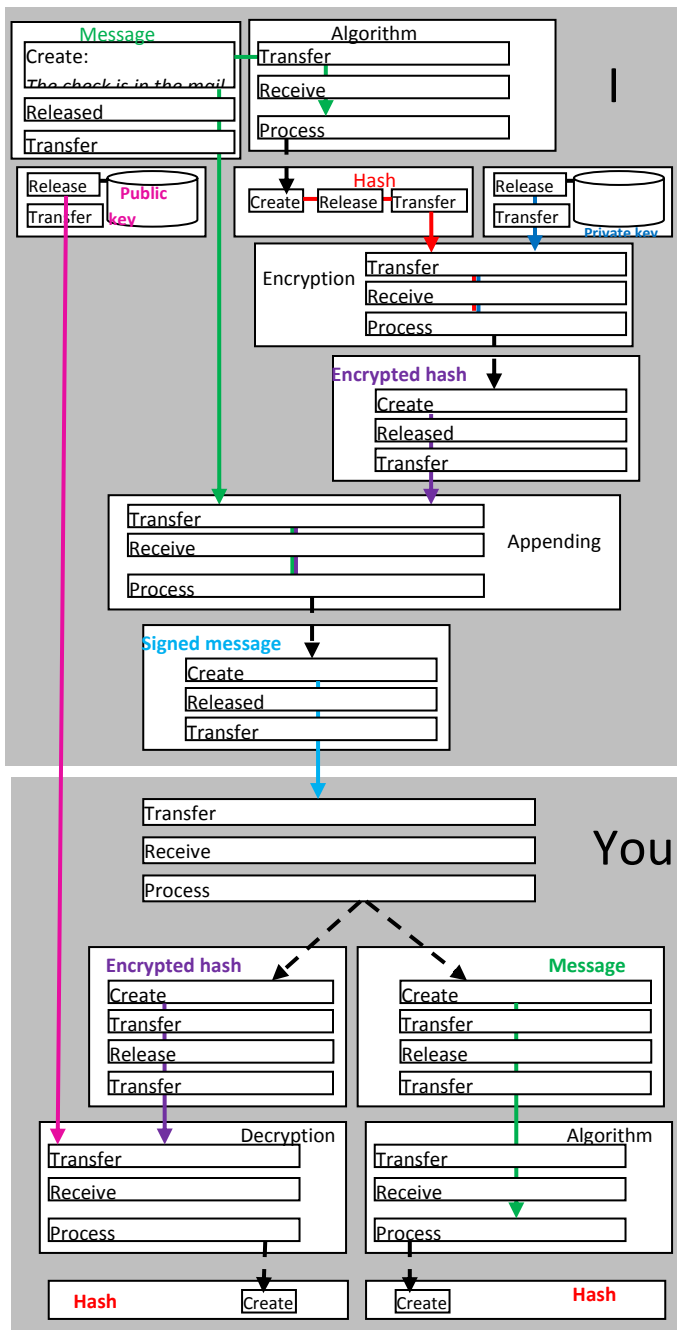


Fig. 6. Example that illustrates a digital signature.

#### IV. RFP CASE STUDY

The case study discussed in this section involves a government agency that seeks the services of a bidder specialized in Enterprise Public Key Infrastructure (PKI) services.

##### A. General Description of the RFP

The RFP contains 59 pages, including a section on the Current Environment with a general view of existing infrastructure, mainframe, and network base IT infrastructure. Of interest in this paper is the section where CA/RA functional and technical requirements are described. In the

RFP, the section titled Certificate Issuance and PKI Lifecycle Management is a mix of textual description and diagrams. The diagrams are mostly textbook illustrations such as the one shown in Fig. 7.

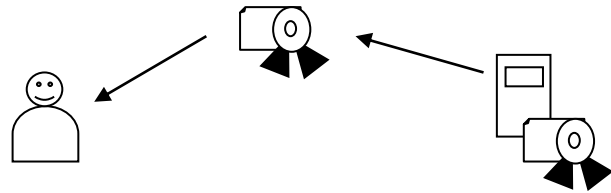


Fig. 7. Example of a diagram used in the RFP (redrawn).

A sample text is the following.

##### Certificate Authority

The key generation and certification services must be used with a Registration Authority (RA) Server. The CA Server is a PKI Server including:

- Consists of CAs with their own certificate signing keys and other parameters from one Server instance
- Provides simplified server-side key generation and client-side key generation
- Provides RSA certificate signing with keys of 1024, 2048, 4096 bits

##### Certificate Validation

Proposed OSCP Server must have an advanced x.509 certificate Validation Authority server that fully conforms to the IETF RFC 6960 standard. It is approved for use by US federal agencies for HSPD-12 implementations.

##### B. A Justification for Incorporating FM as a Supplement to the RFP

Even though it is clear that the main objective of the project is “to identify and implement the most appropriate PKI solution that fulfills the [Agency’s] requirements to improve the security, accuracy, and agility of its IT Infrastructure,” it is unclear what these requirements are. We will focus here on parts that describe digital signatures. Searching all instances of “signature” in the RFP, we copied the following requirements directly from the RFP text:

- Requesting and embedding timestamp responses, requesting and, requesting and embedding OSCP responses, PDF permissions, and server-side archiving of signed documents to disk.
- Creating own PKI systems for **Digital Signature** issuance and Staff logical access Smart card.
- Signing Server should be complete solution for creating and verifying **digital signatures** on document, web form or transaction.
- Server must provide autonomous and irrefutable proof of time for transactions, documents and **digital signatures**.

- Prove when a **digital signature** was applied by the signer so that its validity can be verified in the long-term, even after revocation of signer's digital credentials.
- PKI can provide robust user authentication and strong **digital signatures**.
- The USB should include **digital signatures** and encryption.
- Signing Server can create and verify all common **signature** formats.
- A **signature** service should have the flexibility to be integrated with any application either on the web or a local workstations. It should easily integrate the signing process into the business workflow.
- **Signature** services should be made obtainable for multiple devices and scenarios. It should work on the principle of 'Anytime, Anywhere, Any device' access. The signature capability should be integrated with client applications to allow for documents, emails, data, etc., to be easily signed by their intended signatories.
- **Signature** service should support What You See Is What You Sign (WYSIWYS).
- PDF and Document **signature** should provide visible signatures.

We point out the crucial role of Requirements Specifications within an RFP as the main basis for evaluation by bidders and for the challenges associated with gathering and specifying requirements. In general, according to Hadrian and Evequoz [5], while the legal basis that governs public procurements gives precise guidelines, there is a lack of clear instructions regarding the form and necessary content of a request for proposal.

## V. FM DESCRIPTION OF PUBLIC KEY INFRASTRUCTURE

This section includes a conceptual model of how the required system registers users, issues PKI certificates, and is used by the employees of the agency. It includes conceptual components that include hardware (e.g., servers), software, and manual operations.

### A. Issuing of Certificates

Fig. 8 shows the FM representation of digital signature and certificate issuing under the PKI framework.

### *Application for certificate*

An employee (circle 1) chooses the option (2) to request a digital certificate through his/her account. The request flows (3) to the web interface server dedicated to the PKI system, then to (4) the server of the cryptographic service provider (5). The request process (5) triggers creation of the key (6), including a public key (7) and a private key (8).

### *Registration Authority (RA)*

An RA verifies the identity of employees requesting their digital certificates to be stored at the CA. RA functions include the processes of collecting user data and verifying user identity, which is then used to register a user.

Accordingly, the created key flows to the server of RA (9) to be processed to stamp it with a validation period (10) and to verify the employee's identity.

### *Certificate Authority*

Then, the RA passes the keys with their validation information to the Certificate Authority (CA) system (11). The CA combines validation data (12) and the public key (14) with other information (identity proof, name of CA, and serial number) to create the Digital Key Certificate (15). The private key (13) is kept separately for later encryption of signed documents.

Accordingly, the digital key certificate (15) and the private key (13) are stored in the Database (Repository) (16) to be ready for the employee's use. The database is a secure location in which to store and index keys. An acknowledge-1 is sent to the employee to inform about creating and storing the digital certificate. The acknowledge-1 instructs the employee on the next step, which is to request (18) digital signature creation (19).

### *Digital Signature Creation*

The digital signature request is received (20), triggering turning ON the Signing Hardware Attached (iPad) (21) to enable the employee to input his/her signature (22) through the scanner (23). The scanner (23) sends the image (24) of the signature to the PKI system server (20). The image is hashed using a special hash algorithm (25). The created hash (26) flows (27) to be combined with the private key (28) which was sent (29) earlier. The hash and private key (30) are encrypted to trigger the signature (31) along with the digital key certificate (32) to flow together (33) to the database (repository) (16) to be stored, producing an acknowledge-2 (35) that flows to the employee (36).



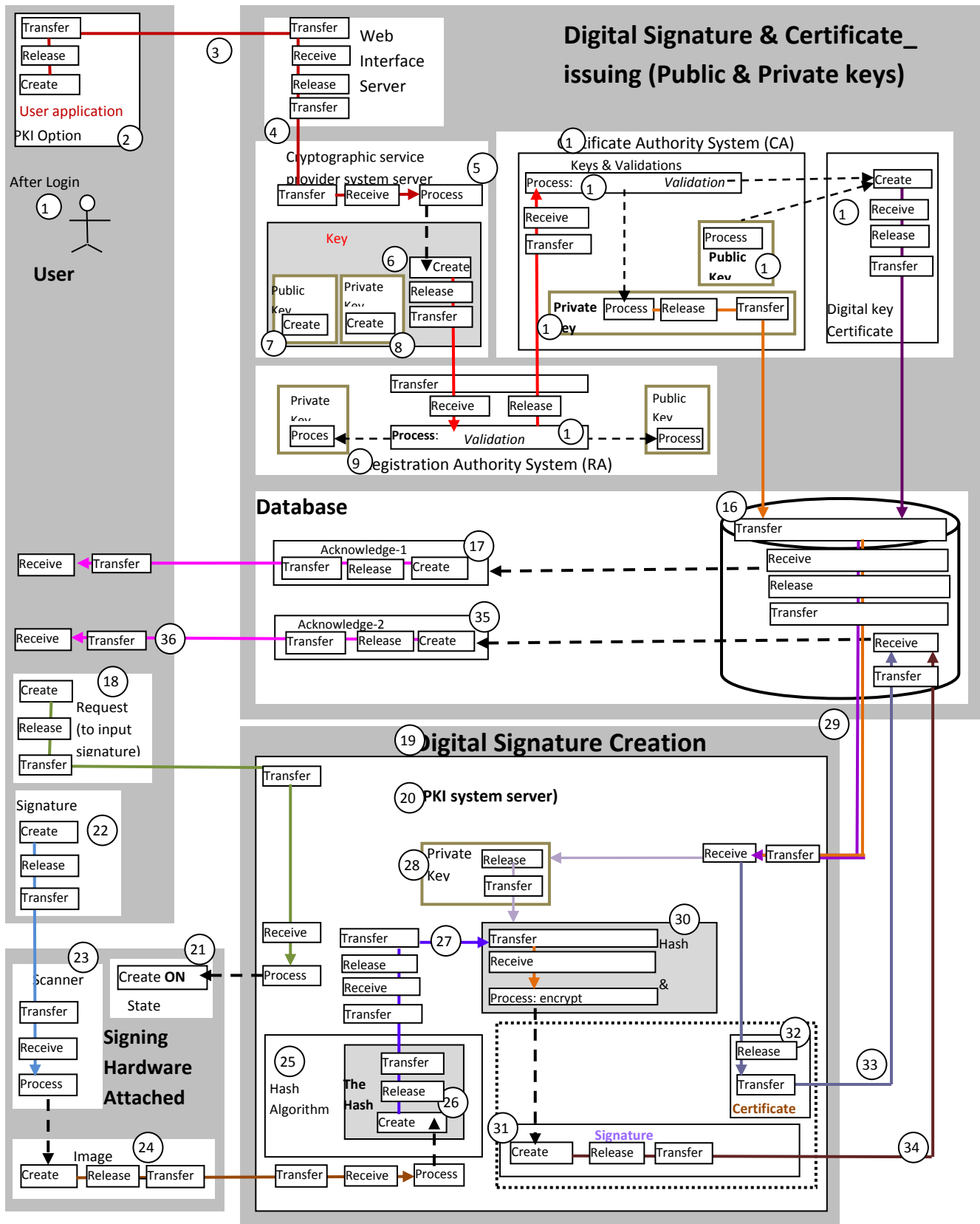


Fig. 8. FM description of the digital certificate as conceptualized by the agency extracted from the RFP and general knowledge of the subject.

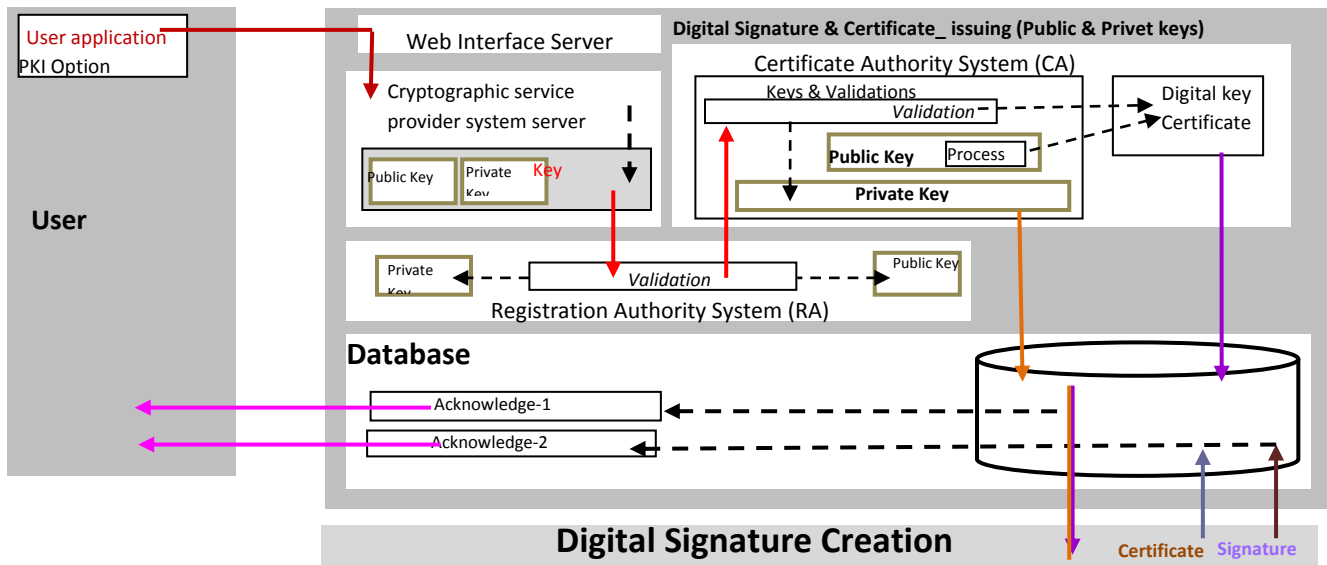


Fig. 9. Simplification of the upper part of Fig. 8 by deletion of stages.

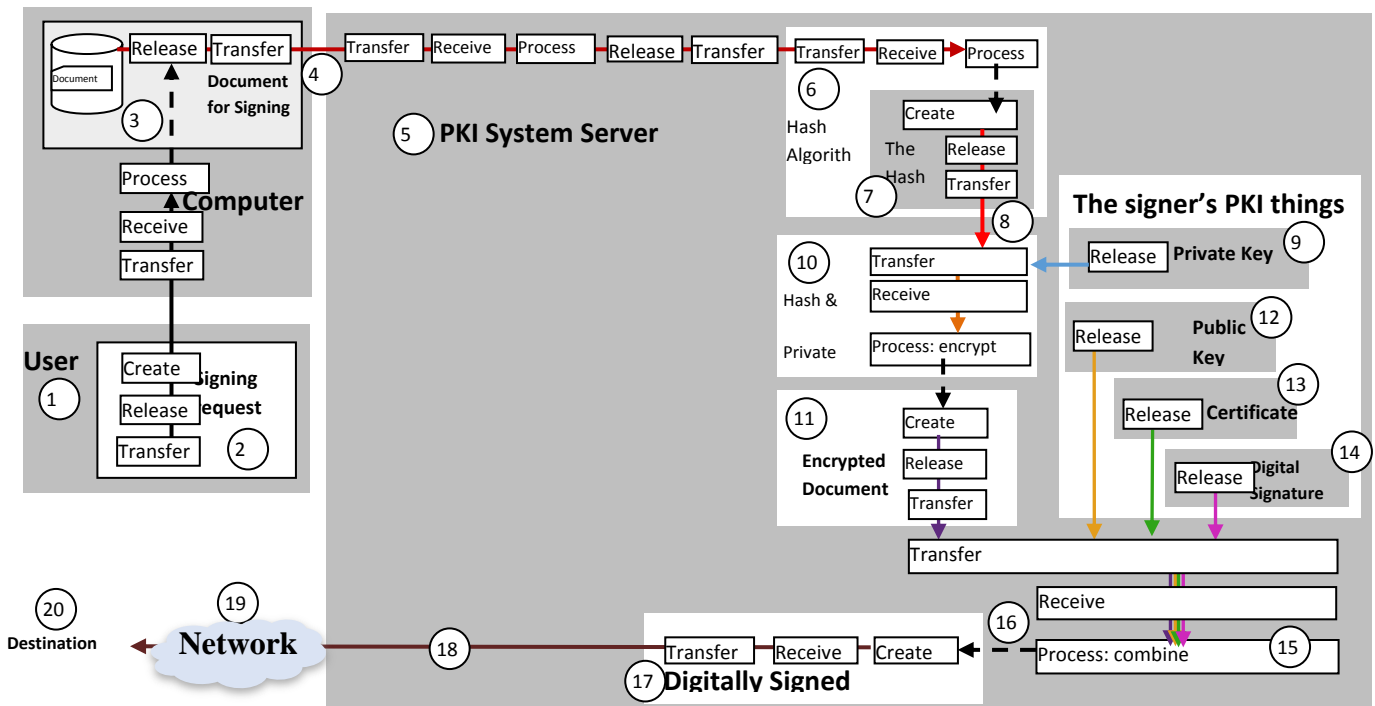


Fig. 10. FM description of the process of digitally signing a document.

Fig. 8 provides a basis for communication and explanation. Multilevel simplifications of the figure can be made for different purposes such as presentations for high-level technical management. For example, Fig. 9 shows the figure simplified after all depiction of stages has been omitted.

### B. Digitally Signed Document

Signing a document digitally is modeled in Fig. 10. A user (circle 1) selects to request (2) signing a document (3) which is already stored on the user's computer. The document flows (4) to the PKI system (5) to be processed using a hash algorithm (6). The created hash (7) flows (8) to be combined with the private key (9) in the CA repository.

The hash & private key (10) are encrypted to create an encrypted document (11). Then, the encrypted document (11) is combined (15) with the other signer's PKI Objects (the public key (12), the certificate (13), and the signature (14)) to create the Digitally Signed Document (16). The digitally signed document (17) is sent (18) through the network (19) to its destination (20).

### C. Decrypting the Received Document

As shown in Fig. 11, a user (Recipient) (circle 1) selects to request (2) decrypting a received digitally signed document (3) that is already loaded on the recipient's computer. The document (3) flows (4) to the PKI system (5) to be processed

(6). Processing separates the encrypted document (7) from the signer's PKI certificate (10), which contains the public key (8) and the digital signature (9). Using the public key (8), two decrypt operations (11 and 12) are applied to the encrypted document (7) and the digital signature (9). Decryption (11) triggers creating the document (13) to be hashed (14) in order to create the hash (15), additionally decryption (12) triggers creating the hash (16). The two hashes (15 and 16) are compared (17) (equal or not) to verify the sender's identity and validate his or her signature.

#### D. Additional General Specifications

General specifications can be superimposed (in their correct places) on the FM diagrams, including:

- (CA) Server specifications such as using a web services interface like XML/SOAP.
- Supporting of X.509 standard.
- Providing RSA certificate signing with, say, 4096 bits
- Supporting several hash algorithms, e.g., SHA-1, SHA-2

The diagrams can also be expanded to include:

- Backup
- Time Stamp Authority.

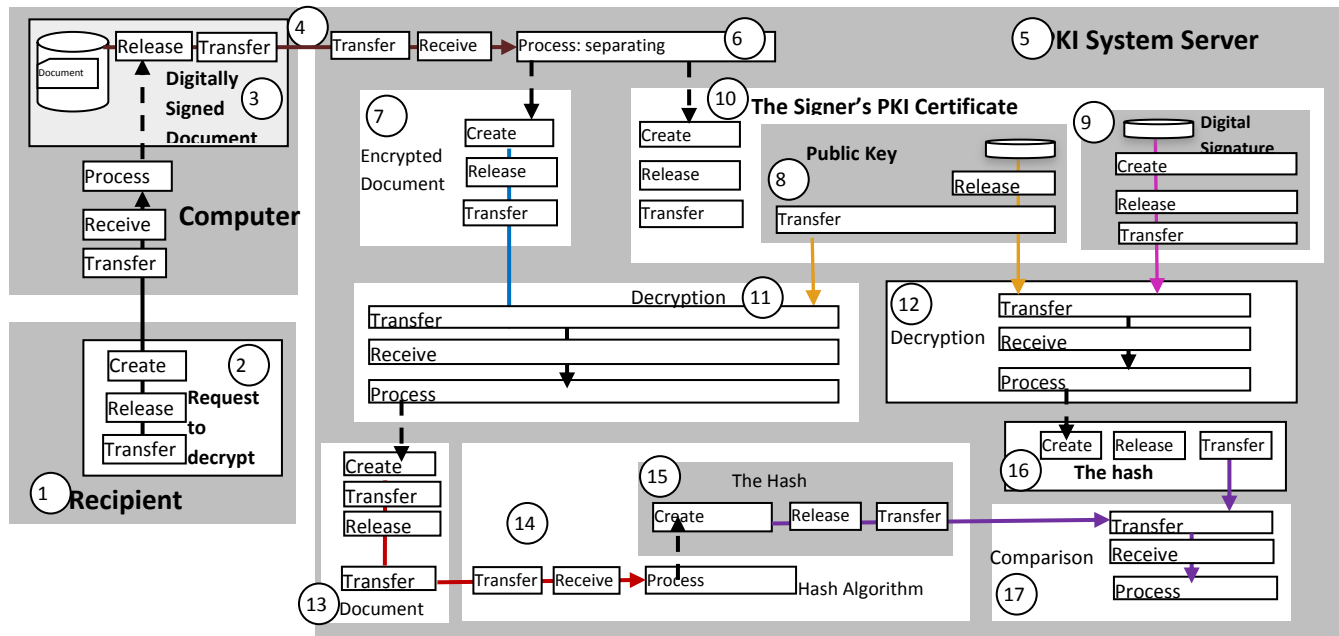


Fig. 11. FM description of the process of decrypting a document.

## VI. CONCLUSION

This paper has introduced a diagrammatic conceptual representation (FM) as a tool for the specification of requirements in RFPs. The FM model includes basic elements of things, their flows, and their stages, within spheres that overlap with other spheres. FM is applied to a sample case study of RFP for public key infrastructure (PKI). The results indicate the following:

- 1) FM is viable as a modeling tool that complements RFP.
- 2) FM lends itself as a theoretical base for defining requirements in procurements.

The complex FM diagrams may present difficulties; however, some solutions to visual complexity have already been implemented in many engineering systems (e.g., aircraft and high-rise building schemata) through multilevel simplifications, as we did in Fig. 9. The details can be lumped together by omitting stages and unifying flows in the model. Nevertheless, the underlying FM schema remains the reference for any further usage such as analysis and documentation.

Further research will work on other types of RFPs. Many issues remain to be clarified; however, this paper demonstrates potential feasibility of the approach.

#### REFERENCES

- [1] J. Mak, "What is procurement?" (No date) [Online]. Available: [http://www.rfpsolutions.ca/articles/Jon\\_Mak\\_IPPC6\\_What\\_is\\_Procurement\\_3Mar2014.pdf](http://www.rfpsolutions.ca/articles/Jon_Mak_IPPC6_What_is_Procurement_3Mar2014.pdf)
- [2] Douraid, S. L. Elhaq and H. Ech-cheikh, "A conceptual and UML models of procurement process for simulation framework," IJCSI Int. J. Comput. Sci. Issues, vol. 9, Issue 6, No. 1, November 2012, pp 120-127.
- [3] Armstrong, "Iterative RFP process," SEI Conference on the Acquisition of Software Intensive Systems, Jan. 1, 2004. Available: [SEI\\_CASIS\\_2004\\_IterativeRFPProcessMgmt.pdf](http://www.sei.cmu.edu/SEI_CASIS_2004_IterativeRFPProcessMgmt.pdf)
- [4] H. Black & Company, "How to prepare better RFP requirements lists for IT success," CaseWare, 2017 [Online]. Available: <https://www.caseware.com/us/2017/04/27/prepare-better-rfp-requirements-lists-success?lang=es>
- [5] Hadrian and F. Evequoz, "CARES: requirements specification with BPMN 2.0 in WTO procurement," Institut d'Informatique de Gestion, HES-SO Valais-Wallis, 2014 [Online]. <http://publications.hevs.ch/index.php/attachments/single/974>
- [6] Specification of Business Process Modeling Notation version 2.0 (BPMN 2.0) [Online]. <http://www.omg.org/spec/BPMN/2.0/PDF>

- [7] Department of Local Government Punjab, Request for Proposal: Implementation of e-Governance in Local Government, Volume I: Functional, Technical and Operational Requirements, July 2017 [Online] Available: <http://pmidc.punjab.gov.in/wp-content/uploads/2017/07/eGovRFPVolumeI13July2017.pdf>
- [8] B. Posey, "A beginner's guide to Public Key Infrastructure: PKI can help keep your network secure, but it can be a hard concept to understand," September 15, 2005 [Online]. <http://www.techrepublic.com/article/a-beginners-guide-to-public-key-infrastructure/>
- [9] Board on Manufacturing and Engineering Design, Preparing for 2020, Visionary Manufacturing Challenges for 2020, Committee on Visionary Manufacturing Challenges, Commission on Engineering and Technical Systems, National Research Council, National Academy Press, Washington, D.C, 1998.
- [10] R. Kazman, "Computing the next 50 years: software engineering," IEEE Computer, vol. 50, Issue 7, July 2017.
- [11] J. Mylopoulos and S. Easterbrook, Conceptual Modeling, 2003 [Online] Available: <http://www.cs.toronto.edu/~jm/340S/PDF2/CM2.pdf>
- [12] Object Management Group (OMG), Unified modeling language: Superstructure and infrastructure, 2009.
- [13] Judicial Council of California, Administrative Office of the Courts, Request for Proposal, Administrative Office of the Courts (AOC), Judicial Branch Enterprise, Document Management System, RFP# FIN122210CK, January 13, 2011.
- [14] S. Al-Fedaghi and M. Alsulaimi, "Re-Conceptualization of IT Services in Banking Industry Architecture Network," 7th IEEE International Conference on Industrial Technology and Management (ICITM 2018), Oxford University, Oxford, United Kingdom, March 7-9, 2018.
- [15] S. Al-Fedaghi and M. BehBehani, "Modeling Banking Processes," 2018 International Conference on Information and Computer Technologies (ICICT 2018), DeKalb, IL, USA | March 23-25, 2018.
- [16] S. Al-Fedaghi and A. Esmaeel, "Modeling Digital Circuits as Machines of Things that Flow," 2018 International Conference on Mechatronics Systems and Control Engineering (ICMSCE 2018), Amsterdam, Netherlands, February 21-23, 2018.
- [17] S. Al-Fedaghi and H. Alahmad, "Integrated Modeling Methodologies and Languages," ACM 12th International Conference on Ubiquitous Information Management and Communication, Langkawi, Malaysia, January 5-7, 2018.
- [18] S. Al-Fedaghi, "Diagramming the class diagram: toward a unified modeling methodology," Int. J. Comput. Sci. Inform. Sec., vol. 15, no. 9, Sept. 2017.
- [19] S. Al-Fedaghi, "Context-aware software systems: toward a diagrammatic modeling foundation," J. Theor. Appl. Inform. Technol. vol. 95, no. 4, 2017.
- [20] S. Al-Fedaghi, "How to create things: conceptual modeling and philosophy," Int. J. Comput. Sci. Inform. Sec., vol. 15, no. 6, June 2017.
- [21] S. Al-Fedaghi, "Securing the security system," Int. J. Sec. Appl., vol. 11, no. 3, pp. 95-108, 2017.
- [22] S. Al-Fedaghi, "Conceptual modeling in simulation: a representation that assimilates events," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 10, pp. 281-289, 2016.
- [23] S. Al-Fedaghi, "Heraclitean ontology for specifying systems," Int. Rev. Comput. Softw. (IRECOS), vol. 10, no. 6, 2015.
- [24] S. Choudhury, K. Bhatnagar, and W. Haque, Public Key Infrastructure Implementation and Design, M&T Books, 2002.
- [25] H. Yang, "PKI Tutorials – Herong's Tutorial Examples, What Is PKI (Public Key Infrastructure)?" 2016 [Online]. Available: <http://www.herongyang.com/PKI/About-PKI-Tutorial-Book.html>

# Average Link Stability with Energy-Aware Routing Protocol for MANETs

Sofian Hamad, Salem Belhaj  
Northern Border University  
Arar – Saudi Arabia

Muhana M. Muslam  
Al-Imam Muhammad Ibn Saud Islamic University  
Riyadh, Saudi Arabia

**Abstract**—This paper suggests the A-LSEA (Average Link Stability and Energy Aware) routing protocol for Mobile Ad-hoc Networks (MANETs). The main idea behind this algorithm is on the one hand, a node must have enough Residual Energy (RE) before retransmitting the Route Request (RREQ) and declaring itself as a participating node in the end-to-end path. On the other hand, the Link Life Time (LLT) between the sending node and the receiving node must be acceptable before transmitting the received RREQ. The combination of these two conditions provides more stability to the path and less frequent route breaks. The average results of the simulations collected from the suggested A-LSEA protocol showed a fairly significant improvement in the delivery ratio exceeding 10% and an increase in the network lifetime of approximately 20%, compared to other re-active routing protocols.

**Keywords**—*Mobile Ad-hoc Network (MANET); routing protocol; energy aware; link life time; AODV*

## I. INTRODUCTION AND MOTIVATION

During normal operation of reactive routing protocols, the routing path between a source and a destination must be discovered before data packets transmission [1]. The routing process in MANETs requires that mobile nodes cooperate together to effectively direct traffic between communicating pairs [2]. The availability of the node is crucial for applying such cooperation. Indeed, its absence affects the state of active connections in its neighborhood. In MANETs, several factors can affect the availability of nodes and cause link breaks, such as interference, obstacles, mobility, and node residual energy (lifetime of the battery).

Two factors will be considered in this paper as being the main contributors to link breaks in MANETs, namely, 1) mobility; and 2) Remaining Energy (RE) of the mobile node.

Regarding the mobility of MANET nodes and the limitation of their power supplies, mobile nodes are considered as energy-constrained devices; this factor has an impact on the availability of the nodes as well as on the network lifetime. Besides, the routing control messages consume a significant amount of the node's battery [3]. Likewise, the mobility of nodes in MANETs is one of the main features of that cause frequent changes in the network topology and therefore increase the probability of link failures and route breaks. As a result, link failures cause the nodes to begin a path maintenance process to find alternate routes. Nevertheless, finding a new path requires a lot of bandwidth, consumes nodes batteries, and adversely effects on network performance by adding re-routing delays and routing overhead. Thus, the routing process should seek only the best

routes ensuring long-term stability and sustainability, by taking into account nodes mobility and their residual energy.

The current research paper suggests a new path discovery algorithm using the RE of the nodes and their LLT. The main idea underlying the suggested algorithm is to transmit RREQ packets on stable links across nodes with sufficient RE and acceptable LLT among the participating nodes on the route. Thus, the suggested protocol can be deployed for MANETs with most existing on-demand routing schemes such as AODV [4], as well as DSR [5].

## II. STATE OF THE ART ON LINK STABILITY AND ENERGY AWARE PROTOCOLS

In the MANETs literature, various LLT estimation methods already exist. Some of them are based on Received Signal Strength (RSS) [6], while other methods, predict the LLT using the location information of the nodes forming the links. Furthermore, several routing algorithms [7] exploit nodes RE and LLT as primary routing metrics to enable the selection of the best end-to-end (e2e) path for transmission, in terms of stability and energy saving.

In the following, the existing routing algorithms implementing the concept of RE and LLT will be discussed.

### A. RSS-Based Routing Protocols

In [8], authors use Received Signal Strength (RSS) as the basic routing metric defining the quality of a link, which varies between two mobile nodes in accordance with a predefined Signal Strength Threshold (SST); it decreases when the RSS between the communicating nodes is lower than the predefined SST and increases in the opposite case.

Furthermore, the research paper [9] raises the Signal Stability-based Adaptive (SSA) routing protocol. In this method, the links are grouped according to the RSS metric. The route discovery mechanism consists on classifying neighbouring nodes connections into two groups: Weakly Connected (WC) and Strongly Connected (SC) links. This classification is carried out by the receiving nodes according to the RSS of the neighbouring nodes when they send the Route Request packet. During the transmission phase, SSA can go through WC links, causing path breaks.

The routing protocol based on the signal strength suggested in [10] first uses the previously established path for packet transfer. Subsequently, it modifies the established path to the strongest RSS.

In [11], the authors proposed a local link management mechanism for OLSR [12]. They use a multi-layer mechanism based on RSS, which makes it possible to decide on the quality of the link; if it is improved or degraded. Moreover, The OLSR RFC [12] describes the hysteresis method dealing with packet loss. This technique anticipates the link breaks to strengthen link management and consequently improves the performance of the network.

### B. Locatio -Based Routing Protocols

The Geographical Positioning System (GPS) [13] is used by most routing protocols using location to obtain motion information about nodes in the network such as direction, coordinates, and speed. In [14], the authors propose a route lifetime and a link prediction algorithm based on node location and motion information. They supposed that all the clocks of the network nodes are synchronized with the GPS clock itself. Thus, the connection time between a couple of nodes can be calculated using (1), if the motion parameters of the two connected nodes are known (direction, coordinates and speed). The idea used in [14] is to estimate the Link Expiration Time (LET) of the path at every hop, which makes it possible to estimate the e2e Route Expiration Time (RET), defined as the minimum LET of the links concerned in an e2e path. Then the route with the highest RET is chosen as the best path.

$$LET = \frac{-(x+v) + \sqrt{(x^2+y^2)v^2 - (xz-vy)}}{x^2 + y^2} \quad (1)$$

Where,

$$x = a_i \cos \theta_i - a_j \cos \theta_j ,$$

$$v = b_i - b_j$$

$$y = a_i \sin \theta_i - a_j \sin \theta_j ,$$

$$z = c_i - c_j$$

$\theta_i, \theta_j, a_i$  and  $a_j$  are respectively the nodes i and j movement directions and velocity.

The authors of [15] present three algorithms (HARP1, HARP2 and HARP3) grouped under the Heading-direction Angles Routing Protocol. In all these algorithms, LET was obtained by applying Equation (1). In order to obtain the angle ( $\theta$ ), the authors used a different solution from that presented in [14].

The authors of [16] proposed a new approach using a stability function as a selection criterion of the main path based on the computation of the degree of mobility of a node relative to its neighbour.

In [17], the authors proposed a new mechanism for deciding which node should retransmit the received RREQ as a function of the distance of the RREQ transmitter. In this sense, they proposed two protocols: Furthest Candidate Neighbours for Rebroadcasting the RREQ (F-CNRR) and Closest- CNRR (C-CNRR). In case of F-CNRR they allow only the far nodes to rebroadcast the RREQ to gain more coverage area. On the contrary, In case of C-CNRR only the closest nodes to the transmitter of the RREQ will rebroadcast the received RREQ.

The authors of [18] propose the protocol LPBR (Location Prediction-Based Routing). The basic idea behind the LPBR protocol is to include for each node its location and its mobility information in the RREQ packet before transmitting it. When the RREQ packet reaches the destination node, all the collected motion parameters, such as mobility and direction, will be saved in its routing table. This information will be used by the destination node when a route fails to estimate the actual location of the desired node based on the previously collected information.

### C. Energy-Aware Routing Protocols

The key concept of the Energy-Aware routing protocols is to properly manage the node's energy consumption to extend the network lifetime. In this sense, the Minimum Battery Cost Routing (MBCR) protocol is suggested in [19], where an e2e path is selected based on the RE summation criteria of all nodes participating in the individual path. However, the trouble with such a technique is that it can choose an e2e path including weak residual energy nodes, which can then cause frequent path breaks. To remedy the shortcomings of the MBCR protocol, the Max-Min Battery Cost Routing protocol (MMBCR) selects a path having nodes with a maximum of RE relative to the other nodes of the network. This approach uses the minimal mobile node RE to evaluate each MMBCR path. Afterwards, the destination node chooses the highest value for each path and sends back the RREP to the source node.

However, the author of [20] introduces the Conditional Maximum Battery Capacity Routing protocol (CMMBCR) that attempts to extend the lifetime of the nodes by selecting only paths containing nodes with a battery power greater than a predefined threshold.

The Authors of [21] succeed to conserve the power of the mobile node by using the uni-cast packet to find any route to destination node rather than using the broadcast message. This technique helps in reducing the consumption of the mobile node and imposes lower overhead.

The Improved-AODV is proposed in [22] to treat selfish nodes in the network, using the remaining power and a new technique that records the nodes acceptance for helping in relay the data. In addition, to extend the network lifetime, I-AODV selects network nodes with significant residual energy and therefore a high probability of data transmission.

In [23], a bandwidth-based energy-efficient routing protocol is proposed to save energy and extend the network lifetime. The suggested algorithm measures the RSS and exploits it to evaluate the bandwidth using a specific dB-to-bandwidth table. Moreover, this method proposes to use RSS variation to evaluate link lifetimes and predict the amount of data that could be transferred.

In our previous work [24], Fixed-Link Stability and Energy-Aware (F-LSEA) protocol is suggested. In this method, the RREQ message will be forwarded only if it satisfies the link lifetime and the residual energy conditions for the transmitter nodes, according to their fixed thresholds of RE and LLT.

### III. METHODOLOGY

In the following, the design of the suggested protocols and their variants will be presented and discussed.

#### A. Problem Statement

As discussed earlier, there are two main reasons leading to link breaks: a node dies from depletion of its battery and a node that leaves the coverage area of the radio range of its neighbouring node.

Fig. 1 illustrates the effect of LLT on the network, where there are six nodes including source node "S" and destination node "D". Each link maintains a link lifetime value defining the quality of the link connecting two communicating neighbouring nodes. The source node "S" broadcasts the RREQ to all the nodes of the network. Node 1 and Node 2 will receive this RREQ, and then register the node "S" as the reverse path in their routing table. Then they will rebroadcast the RREQ because it is assumed that "D" does not exist as a valid entry in their routing table. Similarly, node 3 and node 4 receive the RREQ packets respectively from node 1 and node 2, which will be registered as a reverse path for "S", in the routing table of nodes 3 and 4.

Then, these last nodes will rebroadcast the RREQs accordingly. Moreover, the node 3 receives and discards the duplicate RREQ received from the node 4. The RREQ packet sent by node 3 reaches the destination node "D" and finally prepares to reply with the RREP packet.

The route (D, 3, 1, S) is considered as a reverse path from the destination node "D" to the source node "S".

According to the lifetime value between nodes D and 3, equal to 5 seconds, the RREP packet sent by D will successfully reach node 3. For the same reason, the RREP sent by node 3 will reach node 1, because the link lifetime between nodes 1 and 3 is equal to 2 seconds.

Unfortunately, the link between nodes 1 and S may be broken due to the weakness of the link (when receiving the RREQ, LLT equals 0.5 sec.), even if the node S receives the RREP packet returned by node 1. Admittedly, the weakest link will be broken after some transmissions due to the movement and the speed of nodes 1 and S; as the LLT between two nodes depends on their movements and speeds (1).

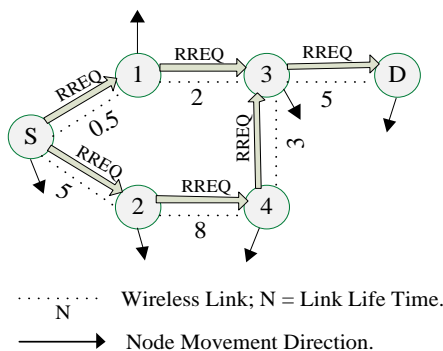


Fig. 1. An example illustrating the effect of Link Lifetime (LLT).

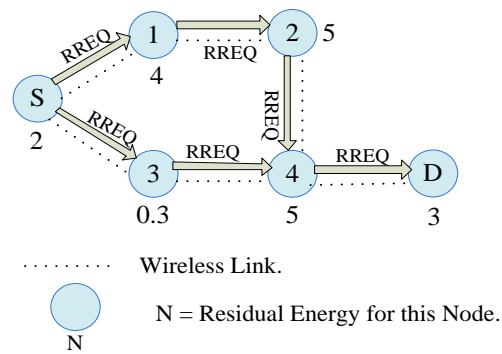


Fig. 2. An example illustrating the effect of Residual Energy (RE).

Similarly, Fig. 2 illustrates the same network as shown in Fig. 1, but involves the nodes RE on the network instead of the LLT. In this case, if the node S selects the route (S, 3, 4, D) to transmit data, the link will be broken after the transmission of some packets due to the low RE of node 3, which is equal to 0.3.

#### B. Preliminaries

In order to enhance the route discovery mechanism, by only allowing nodes that can check specific conditions, two routing protocols have been suggested, namely, 1) F-LSEA; and 2) A-LSEA.

The suggested protocols use the Equation (1) to calculate the link lifetime (LLT). For example, if the LLT value between two nodes in the network equals to 3, it means that the link connecting them will be broken after 3 seconds. Furthermore, the mobile nodes can easily get their RE.

#### C. Fixed-LSEA routing protocol

The main purpose of this section is to enhance the e2e route discovery mechanism every time a node tries to reach a destination node for which it has no entry in its routing table. As described in the simulation setup (Section IV), the RE and the LLT thresholds have been set to specific values. When using the F-LSEA protocol, if no path has been previously defined between two nodes, the source node broadcasts the RREQ to its neighbouring nodes. Upon receipt of the RREQ, any neighbouring node must verify two necessary conditions before rebroadcasting. The first is to compare its residual energy with the fixed threshold. If it is below the threshold, the current RREQ will be rejected. Otherwise, the node goes to the second necessary condition. The second check is to compare its link lifetime with the fixed threshold that has been pre-defined. If it is below the threshold, the RREQ will be rejected; otherwise, it will be retransmitted. Both conditions must be checked before the neighbouring node transfers the received RREQ.

The proposed F-LSEA protocol aims to achieve efficiency and simplicity. This made it possible to differentiate this protocol from its precedents. In fact, F-LSEA receives a RREQ packet on any node and therefore decides to retransmit or not the RREQ according to its RE and LLT.

On the other hand, in previous protocols such as [19]-[21], all the nodes rebroadcast all the received RREQ packets, and enable the destination node to choose a route according to the

received RREQ packet. This route includes nodes with acceptable LLT and a high RE level, where LLT and RE are used as metrics respectively. Thus, concerning the F-LSEA protocol, the following fundamental question was raised: why does a node have to transmit a RREQ when its LLT with the sender of the RREQ is about to be broken and the reply can never reach the RREQ transmitter?

Moreover, redundant RREQs packets cause more overhead when only one route will be selected at the end. Unlike the previous work, the F-LSEA protocol removes most of the redundant paths from the beginning by choosing the best paths.

Fig. 3 illustrates an ad-hoc network topology to better understand the F-LSEA protocol. The network consists of five nodes, where each node is identified by an address (number inside the circle). The value under each node represents the RE, while the number below the links defines the respective LLTs. In this example, we define the value of the RE and LLT thresholds equal to 3. The source node “S” wishes to communicate with the destination node “D” by using other intermediate nodes (node 1, node 2 and node 3) in an ad-hoc network.

Suppose that the source node “S” does not have any route in its routing table to reach the destination node “D”. Then, the source node will broadcast a RREQ packet to all its neighbour nodes. For the classic AODV, the receiving nodes (1, 2 and 3 in our case) will rebroadcast the RREQ packet, if there is no valid route that exists to reach the destination.

In the case of our F-LSEA protocol, on the one hand, node 1 verifies the first necessary condition with respect to its LLT value with the source node “S” (knowing that the LLT threshold = 3 seconds). On the other hand, if the first condition is satisfied, it checks the second energy-related condition (threshold RE = 3 Joule). If the node’s RE value is insufficient (less than threshold), the node rejects the received RREQ packet.

The same steps will be applied to the other intermediate nodes (node 2 and 3). The node 3 receives the packet RREQ and realizes that it’s RE level is greater than the threshold 3. Only, by checking the second condition, the node 3 detects that its link lifetime with the source node “S” is low (LLT = 2 < 3), so it rejects the RREQ packet.

In this diagram of the example (Fig. 3), only the node 2 can retransmit the received RREQ packet because it satisfies the two conditions relative to RE and LLT.

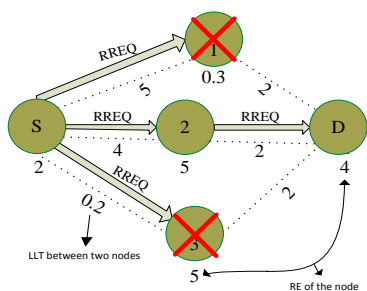


Fig. 3. Operating scenario of the F-LSEA protocol.

The decision of each node receiving the RREQ packet relied on the following algorithm (Algorithm 1).

**Algorithm 1: F-LSEA**

```

1. Let S be the RREQ source/forwarding node.
2. Let N={n1, n2, n3,..., n|N|} be the neighbour nodes of S
3. Let L= {L1, L2,..., L|N|} be the links between S and all its' neighbours | Ln ∈ L is a link between S and the neighbouring node nn, ∀ nn ∈ N.
4. Let LLT be the link life time associated with each link Ln ∈ L.
5. Let RE be the residual energy of each neighbour nn ∈ N.
6. Let α and β be the threshold LLT and RE for any link Ln ∈ N, for any neighbor nn ∈ N.
7. for i= 1 to |N|
    //at each RREQ recipient neighbouring node
8.     if ( LLT ≥ α ) and ( RE ≥ β )
        Forward RREQ
10.    else
11.        Drop RREQ
12.    end if
13. next i
    
```

As can be observed in Algorithm 1, this verifies for each node and at each reception of RREQ, whether the LLT and the RE satisfy the requirements of the predefined thresholds (α and β). At line 9 of the algorithm, the RREQ is forwarded if both conditions are satisfied, otherwise it will be rejected (line 11). The same actions are also illustrated using the flowchart in Fig. 4.

**D. Average-LSEA routing protocol**

The stability of the link at each hop is guaranteed by the F-LSEA algorithm because the decision of forwarding/discarding the RREQ is taken at each receiver node upon receiving the RREQ based on the LLT relative to the RREQ sender and its RE. Nevertheless, the forwarding or discarding the RREQs is entirely based on a specific node RE and LLT thresholds. The fixed thresholds, on the one hand, and the decision-making of the receiver, on the other hand, are not sufficiently flexible under these conditions of fixed thresholds.

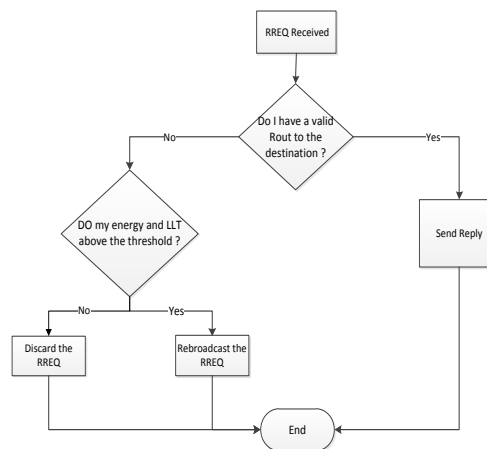


Fig. 4. Flow Chart illustration for the F-LSEA algorithm for each node.



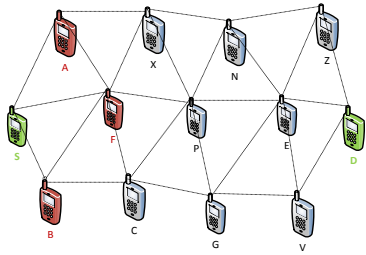


Fig. 5. Example of an isolated node.

This inflexibility occurs for two principal reasons: First of all, one can fall in the case where all the neighbouring nodes of the receiving node have their RE and LLT lower than the predefined thresholds.

As shown in Fig. 5, the isolation of the receiving node occurs due to the inability of neighbouring nodes to rebroadcast the received RREQ packet to the next hop.

As shown in the example of Fig. 5, the source node “S” sends a RREQ asking for “D” as destination. After receiving the RREQ at nodes A, B and F, they check if there are routes to reach the destination node “D”. If yes, a RREP packet will be send to the source node “S”. Otherwise, they will proceed with the decision-making process to determine whether or not to forward the RREQ packet to their neighbours.

In such cases, the B, F and A nodes verify their residual energy and link lifetime parameters with the source node “S”, from which the RREQ has been received. Suppose that one of the RE or LLT parameter of all neighbouring nodes of S is less than the predefined threshold. In that case, there is no possible scenario for node “S” to propagate its RREQ across the network. This issue persists in the suggested Fixed-LSEA protocol.

Therefore, the Average-LSEA routing protocol is suggested in order to overcome this deficiency of Fixed-LSEA.

Unlike the F-LSEA protocol, the A-LSEA is particularly based on the average values of the RE and LLT parameters, which can be calculated using the following methods:

- A node periodically sends a "hello" message to its neighbour nodes. Each node receiving this message responds with a modified "hello" message including its RE and its coordinates (x, y), as illustrated by Fig. 6. In this way, each node can sum the RE of its neighbouring nodes to obtain the average RE (REavg), using (2).

Adding the coordinates (x, y) of the node to the "hello" message allows the receiving node to calculate its LLT with that node (Sender of the Hello).

- Each node in the network can receive LLTs from all its neighbour nodes. In order to compute the mean LLT (LLTavg), the considered node sums its LLT value with the LLT values of the other neighbouring nodes and divides the obtained total by the number of neighbours, using (3).

IP Address	Sequence Number
Hop Count	Lifetime
POS <sub>x</sub>	POS <sub>y</sub>
Residual Energy	

Fig. 6. Modified "hello" message including the node RE and its coordinates (x, y).

Consider the A-LSEA protocol, when receiving RREQ by any node; it consults its routing table searching a route for the current request. If the result of the query is negative, it means that there is no existing route in its routing table; it calculates the average values  $RE_{avg}$  and  $LLT_{avg}$  of all its neighbours. Then, it compares its LLT and RE parameters with those of the calculated averages: If the LLT and the RE are respectively greater than or equal to  $LLT_{avg}$  and  $RE_{avg}$ , the node forwards the RREQ. Otherwise, it will be discarded.

As described in Fig. 5, the source node “S” looks for an e2e path to reach the destination node “D”, obviously; “S” broadcasts the RREQ to all its neighbouring nodes. First of all, it should be noticed that source node S is the origin of the RREQ packet and that it will always broadcasts the RREQ to its neighbours. So the RREQ originator node (that is to say the source node of the e2e path) is eliminated from the A-LSEA protocol. Therefore, the source node will always follow the path discovery mechanism of reactive routing protocols, such as AODV. All other network nodes forward the RREQ packet in accordance to A-LSEA protocol as described below.

### Algorithm 2: A-LSEA

**Input:** Local information on neighbours' RE and LLT.

**Output:** A stable end-to-end routing path.

1. Let S be the RREQ source/forwarding node.
2. Let  $N = \{n_1, n_2, n_3, \dots, n_{|N|}\}$  be the neighbour nodes of S
3. Let  $L = \{\lambda_1, \lambda_2, \dots, \lambda_{|N|}\}$  be the links between S and all its neighbours |  $\lambda_n \in L$  is a link between S and the neighbouring node  $n_n, \forall n_n \in N$ .
4. Let  $LLT_{Avg}$  be the averaged link lifetime associated with all neighbouring node links and the set threshold.
5. Let  $RE_{Avg}$  be the averaged residual energy of all neighbouring nodes, including source S..
6. for  $i = 1$  to  $|N|$   
//at each RREQ recipient neighbouring node
7. if  $(LLT \geq LLT_{Avg})$  and  $(RE \geq RE_{Avg})$
8. Forward RREQ to all neighbours.
9. else
10. Drop RREQ
11. end if
12. next i

Suppose “F” is the candidate node that decides about the routing of the RREQ packet according to the A-LSEA algorithm. To generalize:

Let  $N = \{N_1, N_2, N_3 \dots | N\}$  the set of neighbouring nodes of F.

Let  $T = \{N\} \cup S$  represent the collection of neighbouring nodes with the source/forwarding node, and let  $RE = \{RE_1, RE_2, RE_3, \dots, RE_{|N|}\}$  the REs of the N neighbouring nodes.

The average RE of all neighbouring nodes is calculated using (2):

$$RE_{Avg} = \sum_{i=1}^{|T|} \frac{RE_i}{|T|} \quad (2)$$

Likewise, let  $(LLT_1, LLT_2, \dots, LLT_{|N|})$  the LLTs between the node “F” and its neighbouring nodes. The average LLT is obtained using the following (3):

$$LLT_{Avg} = \sum_{i=1}^{|N|} \frac{LLT_i}{|N|} \quad (3)$$

As described in Fig. 5, if “F” receives the RREQ packet from node “S” and  $LLT_{(S-F)}$  represents the LLT between these two nodes. The node “F” will then compare its residual energy and  $LLT_{(S-F)}$  with respectively the average residual energy and  $LLT_{Avg}$ , of its neighbours. If both parameters residual energy and  $LLT_{(S-F)}$  of the node “F” are, respectively, greater than or equal to the calculated  $RE_{Avg}$  and  $LLT_{Avg}$ , (3) and (2), the node “F” rebroadcasts the RREQ packet otherwise, it will be rejected (Algorithm 2).

#### IV. PERFORMANCE EVALUATION

The NS2 network simulator [25] is used to evaluate the performance of the suggested A-LSEA protocol. Simulation parameters, scenarios, performance measures and results are presented and discussed in the following sections.

##### A. Simulation Environment

The configuration of the MAC layer within the implemented simulation runs on the IEEE 802.11 Distributed Coordination Function (DCF) [26]. The transmission range of any node was fixed at 250 meter and the bandwidth set at 2 Mbps. To gather the result a different scenario was carried out using 100 nodes to simulate the network, these node disseminated in area of 600 meter<sup>2</sup>. The mobility of the nodes was simulated using Random Waypoint [27]. In this model, each node begins its movement with a randomly selected velocity, chosen within the interval [5 m / s, 30 m / s], from its current location to a random location. The simulation time of each scenario lasts 600 seconds. All tests used the same fixed packet size of 1 Kilobyte using Constant Bit Rate (CBR) as the flow type, generated at a constant interval rate of 4 packets per second. Also, 15 flows have been scheduled and configured to randomly select a source node and a destination node for the simulation period.

Finally, the LLT was set at 2 seconds and the initial values of the REs ranged from 1 to 4 Joules.

##### B. Simulations Results and Discussion

In the following subsections, three routing protocols that are A-LSEA, F-LSEA and AODV have been compared and analyzed for their performance in the wireless network.

###### 1) Packet Delivery Ratio

As shown in Fig. 7, the combined effects of RE and LLT affect the data delivery ratio. Indeed, the curve shape provided by the suggested protocols (F-LSEA and A-LSEA) both give a better average delivery ratio than that of the AODV. This is mainly due to the e2e routes returned by F-LSEA and A-LSEA protocols, which have a longer route lifetime and are more stable than AODV. These protocols regard paths with nodes having the highest RE levels and an acceptable LLT, by performing localized and distributed algorithms represented respectively by Algorithm 1 and 2.

However in the case of the AODV protocol, network nodes are unable to capture the REs and LLTs of their neighbours, and are therefore unable to discern the best from the worst links. Thereby, AODV blindly scatters the RREQ packets in the network and can therefore provide paths with faulty individual links, resulting in greater packet loss. In addition, A-LSEA protocol outperforms F-LSEA due to the flexibility of the average values of LLT and RE with respect to the state of the nodes and the state of the network.

###### 2) Network Life Time

Consider the F-LSEA protocol, Fig. 8 shows the lifetime of the network, which increases with the increase of the energy threshold. Increasing the energy threshold (from 1 to 4) can prevent node whose energy is below this level from forwarding the RREQ. This prevents many nodes from forwarding the RREQ and, as a result, to save their energy and thus improve the network lifetime. Moreover, the simultaneous exchange of several RREQs actually causes the premature death of the nodes, and consequently, they can no longer belong to the network. The elimination of nodes in the F-LSEA protocol, however, saves energy of the network by saving the energy of the nodes, preventing them from transmitting and receiving RREQs. In addition, the choice of more stable paths when using the F-LSEA protocol generates little overhead for path maintenance and therefore less energy consumption.

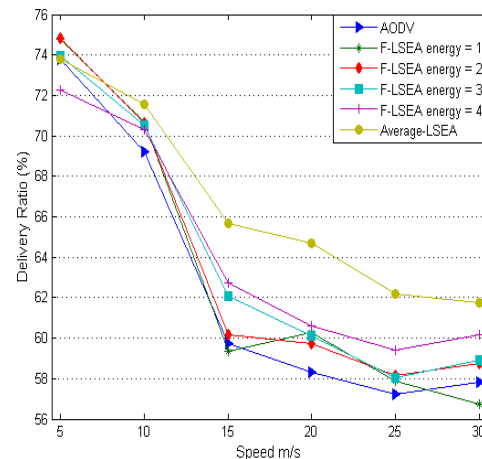


Fig. 7. Delivery Ratio vs. Speed.

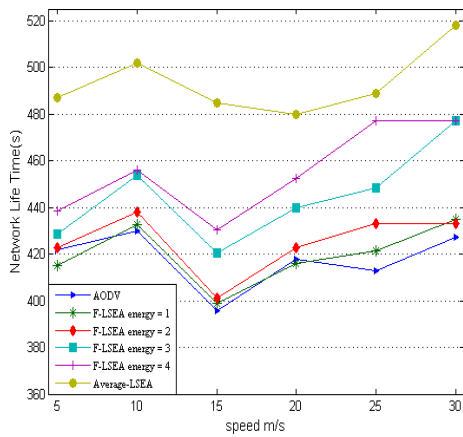


Fig. 8. Network lifetime vs. Speed.

Some RREQ packets may be considered useless in the case where the path created by these RREQ packets is broken, and thus the node becomes unreachable because of the poor link quality or low power.

Reduce the number of useless RREQ packets sent over the network allows the nodes to save energy and thereby increase the network lifetime. The A-LSEA algorithm performs better than the F-LSEA (LLT = 2 and RE = 1, 2, 3 or 4) because of the flexible average thresholds, whereas for F-LSEA, the threshold values are fixed. This flexibility in A-LSEA protocol results in more stable links compared to F-LSEA, and thus a better total network lifetime.

### 3) End-To-End Delay

As can be seen in Fig. 9, the average packet delay observed in the case of AODV is higher than the protocols suggested in the majority of cases. However, in some cases, the average delay of the suggested protocols is higher.

The main reason is that the suggested algorithms choose a link based on its quality and the remaining battery lifetime. As the e2e path returned by the suggested protocols may have more hops compared to AODV, the packets may experience further delays due to more transmission and queues along the path.

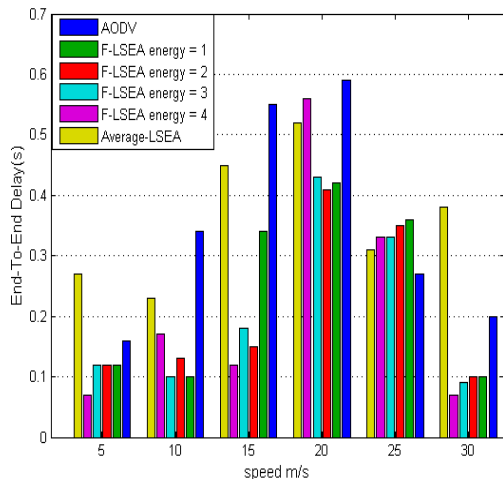


Fig. 9. e2e Delay vs. Speed.

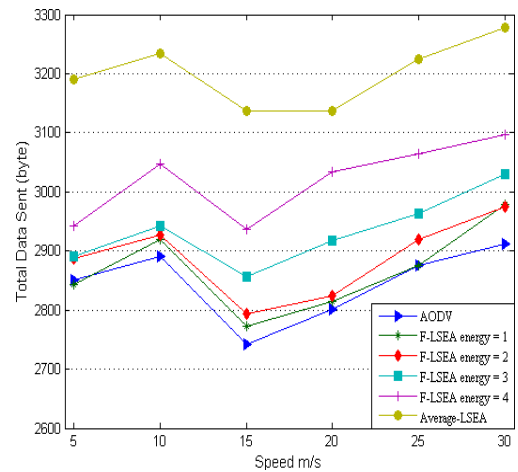


Fig. 10. Total Data Sent vs. Speed.

### 4) Total Data Sent

As shown in Fig. 10, which illustrates a comparison between the suggested algorithms and AODV in terms of data sent, A-LSEA performs significantly better than the F-LSEA and AODV protocols.

The main reason is that for the path discovery phase, A-LSEA algorithm uses an appropriate mechanism to estimate link quality and RE, by adopting their mean values. This more stable path choice result in less link breaks, and therefore the total amount of data sent is higher than in the case of other algorithms. In addition, F-LSEA performs better (for RE varying from 1 and 4) than AODV because of the prevention of many nodes to transmitting the RREQ. Likewise, increasing the RE threshold causes more RREQ packet drops.

### 5) Total Received Data

All the nodes forming the network consume power during the transmission and reception of the data and, simultaneously, the power consumption increases as the distance separating the transmitter and the receiver increases. For these reasons, A-LSEA performs significantly better than F-LSEA and AODV, as illustrated by Fig. 11.

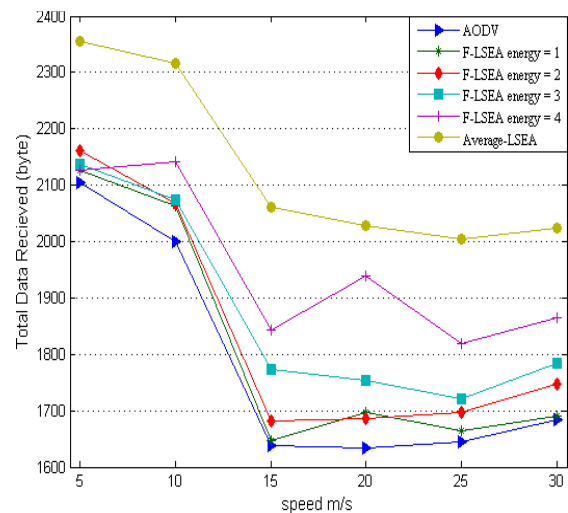


Fig. 11. Total Data bytes Received vs. Speed.

Indeed, the A-LSEA protocol returns the e2e paths that are able to maintain the stability of the route and to provide moderate distances among the nodes; likewise, the main factors affecting the calculation of LLT and RE between two nodes among the nodes that are concerned by the calculation of the route. Clearly, the total amount of data bytes received decreases for all observed protocols as the speed of the nodes increases. Moreover, for the same simulation parameters, the F-LSEA algorithm (for RE varying from 1 to 4) outperforms AODV in terms of the total amount of data bytes received. As shown in Fig. 11, raising the fixed threshold results in lower RREQs meeting the necessary conditions (fewer RREQ packets sent results in fewer RREQ packets received), which directly impacts the total number of received RREQ packets, leading to the effective reception of more data.

### 6) Total Data Drop

This section compares the suggested algorithms against AODV in terms of total data loss. As illustrated in Fig. 12, the A-LSEA protocol has a significant low data loss rate compared to those reported by F-LSEA and AODV protocols. As mentioned earlier, this is due to more stability of paths provided by A-LSEA protocol and therefore, the network will experience less congestion.

Indeed, in the case where the network is congested, it is very likely that the nodes reject more packets than in the case of a low traffic network. In addition, potential interference among nodes during data transmission, collisions, and long queues have the greatest impact on dropping data packets. Furthermore, a very slight variation in the A-LSEA behavior can be reported with respect to F-LSEA and AODV, as the speed of the node increases. This is because A-LSEA generates less RREQ across the network, which results in more channel free time, shorter queues, and fewer collisions. On the one hand, the A-LSEA protocol differs from the AODV and F-LSEA protocols in that it provides more stable routes. These routes are more likely to last long before launching a new path discovery instance, thus reducing overhead on the network. On the other hand, the paths established by the AODV protocol does not last long enough and it is therefore inevitable to search for a new path in a short period of time, which leads to more overhead on the network. Correspondingly, for the F-LSEA protocol (for RE varying from 1 to 4), an already established route will only last for the time set by the predefined threshold before initiating the discovery process of a new path, resulting in more overhead.

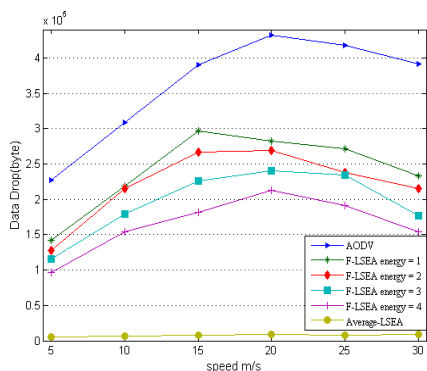


Fig. 12. Total Data Drop vs. Speed.

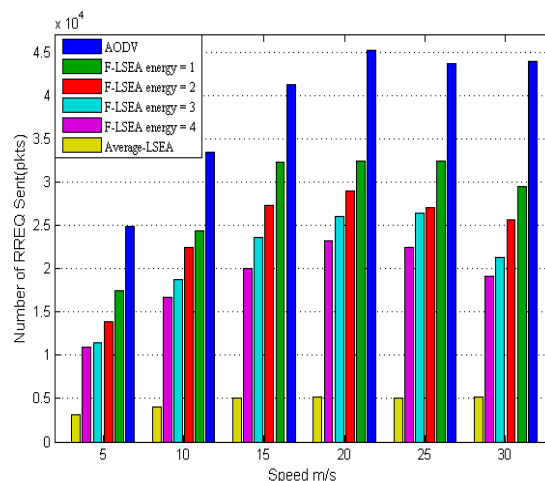


Fig. 13. RREQ Sent vs. Speed.

It should be noticed that during the increase of the predefined threshold, for the F-LSEA algorithm, the overhead decreases because there is less retransmission of RREQ packets.

### 7) Number of RREQs sent

As can be seen in Fig. 13, the number of RREQs sent over the network has a considerable effect on the performance of all the routing protocols considered during this work. Indeed, raising the number of RREQ packets circulating on the network leads to more heavily loaded communication. Moreover, nodes dissipate more power when sending or receiving RREQ packets, which impacts the network lifetime, overhead, the effective data bytes sent, the effective data bytes received and the delivery rate of the effective data sent.

It is also noticed that A-LSEA outperforms the F-LSEA (for RE varying from 1 to 4) and AODV methods. Consequently, the fixed RE threshold for the F-LSEA algorithm plays an important role in reducing the number of the RREQ packets retransmitted over the network; decreasing the fixed threshold value increases RREQ packets routing and vice versa.

### 8) Average Throughput

A comparison of the average throughput of the studied protocols is shown in Fig. 14.

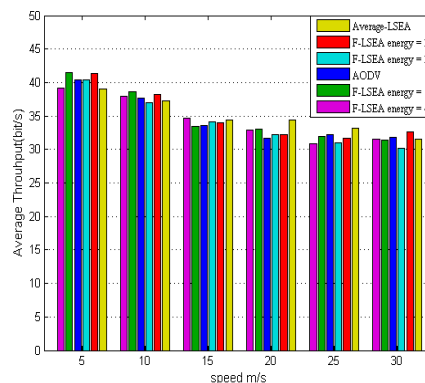


Fig. 14. Throughput vs. Speed.

It is noticed that the throughput of all protocols decreases with the node speed increases. This is obviously due to the fact that the mobility of the nodes favors the breaks of connections and thus leads to more re-initialization of the e2e routes, which results in a reduction of the flow.

Comparatively, the average throughput remains relatively the same in the three considered routing protocols. The suggested schemes have therefore improved the other metrics while maintaining convergent average throughputs.

## V. CONCLUSION AND FUTURE WORK

In reactive routing protocols, the route discovery process can consume a lot of network resources due to the dissemination of RREQs to find the path to a destination node. In addition, because of the mobile nature of the nodes in MANETs, stable path selection is extremely important. This article suggests two protocols for dealing with the flood phenomenon that exists in the reactive routing of MANETs. In suggested solutions, only specific nodes are allowed to forward the received RREQs. On the one hand, the decision to include nodes in an e2e path is based on their residual energies (RE). On the other hand, the proposed protocols guarantee stable paths by considering link lifetime (LLT) between two nodes.

Future work consists in finding optimal value for the LLT and the RE with respect to the decision to rebroadcast the received RREQ packet or not, rather than using a fixed or average values threshold.

## ACKNOWLEDGEMENT

Financial support for this study was provided by deanship of Scientific Research, Northern Border University under grant no. (7167-SCI-2017-2-7-F).

## REFERENCES

- [1] Garg, N., Aswal, K., & Dobhal, D. C. (2012). A review of routing protocols in mobile ad hoc networks. *International Journal of Information Technology*, 5(1), 177-180.
- [2] Chlamtac, I., Conti, M., & Liu, J. J. N. (2003). Mobile ad hoc networking: imperatives and challenges. *Ad hoc networks*, 1(1), 13-64.
- [3] Bheemalingaiah, M., Naidu, M. M., Rao, D. S., & Vishvapathi, P. (2017, January). Performance Analysis of Power-Aware Node-Disjoint Multipath Source Routing in Mobile Ad Hoc Networks. In *Advance Computing Conference (IACC), 2017 IEEE 7th International* (pp. 361-371). IEEE.
- [4] Perkins, C. E., Ratliff, S., Dowdell, J., Steenbrink, L., & Mercieca, V. (2016). Ad Hoc On-demand Distance Vector Version 2 (AODVv2) Routing. Internet Draft (Standards Track), Mobile Ad hoc Networks Working Group. Available at <http://tools.ietf.org/html/draft-ietf-manet-aodvv2-13>.
- [5] Salem, A. O. A., Samara, G., & Alhmiedat, T. (2017). Performance analysis of dynamic source routing protocol. *arXiv preprint arXiv:1712.04622*.
- [6] Ouyang, R. W., Wong, A. K. S., & Lea, C. T. (2010). Received signal strength-based wireless localization via semidefinite programming: Noncooperative and cooperative schemes. *IEEE Transactions on Vehicular Technology*, 59(3), 1307-1318.
- [7] Bolla, R., Bruschi, R., Davoli, F., & Cucchietti, F. (2011). Energy efficiency in the future internet: A survey of existing approaches and trends in energy-aware fixed network infrastructures. *IEEE Communications Surveys & Tutorials*, 13(2), 223-244.
- [8] Athanasiou, G., Korakis, T., Ercetin, O., & Tassioulas, L. (2009). A cross-layer framework for association control in wireless mesh networks. *IEEE Transactions on Mobile Computing*, 8(1), 65-80.
- [9] Dube, R., Rais, C. D., Wang, K. Y., & Tripathi, S. K. (1997). Signal stability-based adaptive routing (SSA) for ad hoc mobile networks. *IEEE Personal communications*, 4(1), 36-45.
- [10] Wang, S. Y., Liu, J. Y., Huang, C. C., Kao, M. Y., & Li, Y. H. (2005, March). Signal strength-based routing protocol for mobile ad hoc networks. In *Advanced Information Networking and Applications, 2005. AINA 2005. 19th International Conference on* (Vol. 2, pp. 17-20). IEEE.
- [11] Huang, D. W., Lin, P., & Gan, C. H. (2008). Design and performance study for a mobility management mechanism (WMM) using location cache for wireless mesh networks. *IEEE Transactions on Mobile Computing*, 7(5), 546-556.
- [12] Jacquet, P., Muhlethaler, P., Clausen, T., Laouti, A., Qayyum, A., & Viennot, L. (2001). Optimized link state routing protocol for ad hoc networks. In *Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International* (pp. 62-68). IEEE.
- [13] Misra, P., & Enge, P. (2006). *Global Positioning System: signals, measurements and performance second edition*. Massachusetts: Ganga-Jamuna Press.
- [14] Su, W., Lee, S. J., & Gerla, M. (2000). Mobility prediction in wireless networks. In *MILCOM 2000. 21st Century Military Communications Conference Proceedings* (Vol. 1, pp. 491-495). IEEE.
- [15] Gerharz, M., de Waal, C., Frank, M., & Martini, P. (2002, November). Link stability in mobile wireless ad hoc networks. In *Local Computer Networks, 2002. Proceedings. LCN 2002. 27th Annual IEEE Conference on* (pp. 30-39). IEEE.
- [16] Moussaoui, A., Semchedine, F., & Boukerram, A. (2014). A link-state QoS routing protocol based on link stability for Mobile Ad hoc Networks. *Journal of Network and Computer Applications*, 39, 117-125.
- [17] Hamad, S., Belhaj, S., & Muslam, M. M. (2017). Smart Selection of Candidate Neighbors for Efficient Route Discovery in MANETs. *Journal of Applied Sciences*, 17, 126-134.
- [18] Meghanathan, N. (2008). A location prediction-based reactive routing protocol to minimize the number of route discoveries and hop count per path in mobile ad hoc networks. *The Computer Journal*, 52(4), 461-482.
- [19] Singh, S., Woo, M., & Raghavendra, C. S. (1998, October). Power-aware routing in mobile ad hoc networks. In *Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking* (pp. 181-190). ACM.
- [20] Toh, C. K., Cobb, H., & Scott, D. A. (2001). Performance evaluation of battery-life-aware routing schemes for wireless ad hoc networks. In *Communications, 2001. ICC 2001. IEEE International Conference on* (Vol. 9, pp. 2824-2829). IEEE.
- [21] Gelenbe, E., & Lent, R. (2004). Power-aware ad hoc cognitive packet networks. *Ad Hoc Networks*, 2(3), 205-216.
- [22] Chen, C. W., Weng, C. C., & Kuo, Y. C. (2010). Signal strength based routing for power saving in mobile ad hoc networks. *Journal of Systems and Software*, 83(8), 1373-1386.
- [23] Xu, Y., Heidemann, J., & Estrin, D. (2001, July). Geography-informed energy conservation for ad hoc routing. In *Proceedings of the 7th annual international conference on Mobile computing and networking* (pp. 70-84). ACM.
- [24] Hamad, S., Noureddine, H., & Al-Rawashidy, H. (2011, October). Link stability and energy aware for reactive routing protocol in mobile ad hoc network. In *Proceedings of the 9th ACM international symposium on Mobility management and wireless access* (pp. 195-198). ACM.
- [25] Fall, K., & Varadhan, K. (2007). *The network simulator (ns-2)*. URL: <http://www.isi.edu/nsnam/ns>. Last visit December 2017.
- [26] Bianchi, G. (2000). Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE Journal on selected areas in communications*, 18(3), 535-547.
- [27] Bettstetter, C., Hartenstein, H., & Pérez-Costa, X. (2004). Stochastic properties of the random waypoint mobility model. *Wireless Networks*, 10(5), 555-567.

# Agent based Architecture for Modeling and Analysis of Self Adaptive Systems using Formal Methods

Natash Ali Mian<sup>1,2</sup>

School of Computer Science, National College of Business  
Administration and Economics, Lahore<sup>1</sup>  
School of Computer and Information Technology,  
Beaconhouse National University, Lahore<sup>2</sup>

Farooq Ahmad

Department of Computer Sciences, Comsats Institute of  
Information Technology, Lahore<sup>3</sup>

**Abstract**—Self-adaptive systems (SAS) can modify their behavior during execution; this modification is done because of change in internal or external environment. The need for self-adaptive software systems has increased tremendously in last decade due to ever changing user requirements, improvement in technology and need for building software that reacts to user preferences. To build this type of software we need well establish models that have the flexibility to adjust to the new requirements and make sure that the adaptation is efficient and reliable. Feedback loop has proven to be very effective in modeling and developing SAS, these loops help the system to sense, analyze, plan, test and execute the adaptive behavior at runtime. Formal methods are well defined, rigorous and reliable mathematical techniques that can be effectively used to reason and specify behavior of SAS at design and run-time. Agents can play an important role in modeling SAS because they can work independently, with other agents and with environment as well. Using agents to perform individual steps in feedback loop and formalizing these agents using Petri nets will not only increase the reliability, but also, the adaptation can be performed efficiently for taking decisions at run time with increased confidence. In this paper, we propose a multi-agent framework to model self-adaptive systems using agent based modeling. This framework will help the researchers in implementation of SAS, which is more dependable, reliable, autonomic and flexible because of use of multi-agent based formal approach.

**Keywords**—Formal methods; self-adaptive systems; agent based modeling; feedback loop; Petri nets

## I. INTRODUCTION

As the complexity has increased, hence, existing approaches do not suffice the requirements of modeling, managing and developing software systems. This has motivated the research community to explore new dimensions in software engineering and integrate other fields like biology, psychology; nature inspired computing, robotics, artificial intelligence and more. The change in the way the software is used needs that it has the capability of self-adaptation [1] which is one of the hot areas of research since last decade.

SAS [2] are capable of modifying their behavior due to change in environment at run time. Modeling of these types of systems is either very difficult or not possible by the use of conventional software engineering approaches. One of the major difference in requirement engineering is that the 'shall' statements become 'may' statements when developing a SAS that has the capability to adapt in accordance with the external

environment [1]. Uncertainty is one of the most certain thing in modeling and development of SAS. [2]. This aspect motivates the practitioners and researchers to use multiple existing approaches or develop new approaches [3] to handle uncertainties of the system [4].

There has been a lot of research in SAS including software engineering, requirements engineering, software architectures, middleware, component-based development and programming languages [5]. In addition to these some research has been done in other areas of Computer Science which includes fault-tolerant computing, biologically inspired computing, multi-agent systems, distributed artificial intelligence and robotics [6].

Formal methods are very effective and concrete mathematical techniques and methods in specifying, modeling, verifying and developing systems. Formal methods have been majorly applied in modeling of SAS [7]. Application of formal methods for verification, model checking and theorem proving is less for SAS [8]. To utilize the formal methods according to its strengths, there is a need to apply them in validation and verification [9] of SAS, this will consequently produce systems that are more reliable and tested [6] at an early stage of development [10].

Use of agents in modeling system [11] that have capacity and capability to adapt to a new behavior at runtime has been very effective and efficient [12]. Agents [13] help the systems to perform all the tasks autonomously and efficiently, this increases the overall productivity of the system. We use agents to perform the tasks autonomously with well-defined and concrete rules which have been developed, analyzed and tested by use of formal methods. More specifically Petri nets will be used to model all these agents.

In this paper we propose an initial architecture of the system that will use the strengths and rigor of formal methods, autonomous working of agents and the effectiveness of feedback loop to model a self-adaptive system [14]. This model will further be extended for distributed systems where most of the components will be reused with some additional components like distributed feedback loop manager, distributed agent manager and distributed application manager will be added. Fig. 1 depicts the scope of the work and the gives an idea of the proposed integration. This is an initial attempt to propose an architecture which integrates agents, formal methods and feedback loop. This paper is introduced in

Section I, literature review is given in Section II, Section III introduces SAS, Section IV gives an overview of feedback loop, Section V elaborates formal methods, followed by the proposed architecture in Section VI, and finally we conclude the paper and give pointers to future work.

## II. RELATED WORK

Life patterns have changed, consequently having a huge impact on working environment [15] and the way software is used in ever changing environment [4]. Improvement in technology, change in working environment, improvement in technology, increase in storage capacity, need of high processing, and availability of variety of data is causing fundamental change in software development, testing and performance [8]. This gives rise to systems that can adapt to the working environment; these systems are categorized as SAS. Development of SAS has increased the flexibility of software systems, however, this has also led increased the complexity of systems resulting in a lot of challenges [16] for software engineering community [4]. Methods as well as processes of engineering software systems have also evolved for development of SAS [17].

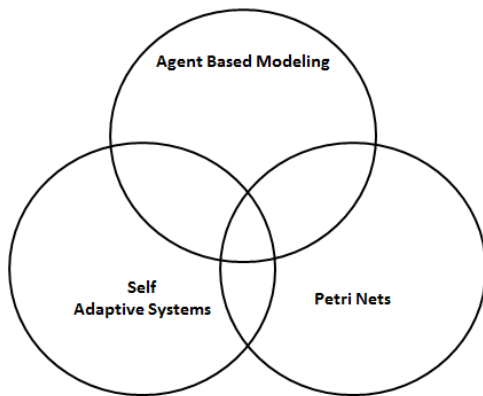


Fig. 1. Scope of work.

One of the major problems in the current approaches is that focus on quality of the output is much less [18], this may be due to the variety of problems, uncertainty and changing environment [15] that had to be handled by SAS. Many model based approaches [19] have been proposed in literature which address the problem of requirement engineering in SAS. A context aware methodology is proposed in literature, this approach performs the adaptation at run time considering the context, not only this but a complete mechanism of verification [20] and validation is also proposed, the focus of this work is done on the basis of image processing algorithms [21]. Goal based [22] and requirement driven architecture [1] for systems are proposed in literature which can adapt themselves to a better configuration by monitoring and analyzing the current actors in the system [23].

Non-functional requirements play vital role in self-adaptive approach which takes information that is not easily identifiable and overwhelmingly against the static nature of information [24]. Although, functional requirements are also important, but non-functional requirements have more infringement in software development and software quality

assurance, using self-adaptive approach [20]. One of the major problems in the current approaches is that focus on quality of the output is much less, this may be due to the variety of problems, uncertainty and changing environment [15] that had to be handled by SAS. Many model based approaches [25] have been proposed in literature which address the problem of requirement engineering in SAS. A context aware methodology which performs the adaptation at run time considering the context, not only this but a complete mechanism of verification [20] and validation is also proposed. A lot of work has been done in identifying the key areas of research, challenges faced, architecture problems, design techniques available, implementation issues in engineering [26] of SAS [10]. Many papers discuss the importance of adaptation and propose architecture based adaptation [25], goal based adaptation [26], feature oriented adaptation [4], parameter based adaptation, requirement driven adaptation [1] and much more. Software agents have been used to model [27] and implement the adaptation process.

It has been observed that formal methods has mostly been used in modeling of SAS [7] and not in model checking and theorem proving which are major strengths of formal methods, hence, the need to apply formal methods for these is positively required to make the overall process of designing the SAS more reliable [6]. A combination of formal and semi-formal methods is also used in modeling of SAS [8] and the results have been very encouraging [6]. There have been a few studies where formal methods are used successfully in model checking [9] and domain specific languages [14] and design patterns [3] are proposed for development of SAS.

## III. SELF-ADAPTIVE SYSTEMS

SAS can alter their behavior during operation [4]. These systems fall under the category of context aware systems [28]. Adjusting as per needs of the user at run-time is one of the major strength of these systems [29]. The adaptations that these systems perform during executions are not included in the requirements for which these systems are developed [24]. This variability makes the development of these systems challenging as the development team has to plan for the uncertainties that may arise at run time [30], [15]. Hence, major part of the requirement engineering has to be completed at run time [31]. Not only the requirement engineering, but testing is also done at run time, all this is done by use of feedback loops [32]. To enhance the efficiency and reliability of these systems, all these steps are performed autonomously by agents [33].

Almost all major systems that exist today have the capability of adaptation; however, in some systems the adaptations are planned at design time and in other it is done at run time. In case, the adaptations are implemented at design time, the systems are categorized as simple adaptive systems and when the adaptation is done at run time, the systems are classified as self-adaptive [34].

## IV. FORMAL METHODS

Formal methods are mathematical tools and techniques that are used in analysis and modeling of different hardware and software systems [35]. Additionally they help us in

validation and verification of systems at an early stage of development [9]. These methods are reliable and help us in analyzing, modeling, reasoning and testing the systems. As these methods are based on concrete mathematical principles, hence the reliability of systems that are developed using formal methods is increased many folds [36]. Strength of these methods is that they can be used in combination with existing software development methods. Most of these methods have specification languages that are based on first and second order predicate calculus, temporal logic, algebraic theory and graph theory. Sets, sequences, relations, functions, mappings and state machines are the foundation of formal modeling techniques. Due to the use of precise mathematical symbols the effectiveness of these methods is much more than conventional methods.

Formal methods are supported by a variety of case tools which help in model development, model checking and simulating the overall scenario. We have successfully modeled a small self-adaptive system by use of Petri Net which is a formal method and is based on graph theory. The tools help in development of concrete model efficiently and reliably. Hence, formal methods are very effective in modeling of complex system like self-adaptive systems. These methods have already been successfully applied in development of many complex industrial systems and many safety critical systems.

### V. FEEDBACK LOOP (MAPE-K LOOP)

Feedback loop comprises of four major steps which are monitor, analyze, plan and execute, this loop is also referred as MAPE-K Loop [6]. Each phase is further divided in to further sub-phases and multiple strategies are used for design and implementation of each phase. Formal methods have the capability to model all phases with success and reliability.

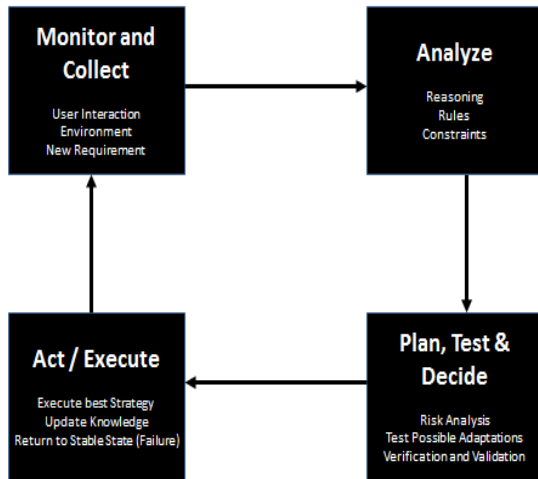


Fig. 2. MAPE-K feedback loop.

In the monitor phase the input is received from the external environment, and after performing the initial transformation of inputs it is checked against the existing requirements. In case a match is found, no adaptations are performed and the requirement is executed. However, if the set of input are new then analysis of inputs is performed which is followed by the

planning and testing phase. It is to be noted that the possible adaptations, testing and execution is done at run time. The execution is executed by the system effectors. In a situation where the proposed adaptation is not successful, the loop starts again and this process continues iteratively [37] till a final adaptation is executed [14]. A simplified version of MAPE-K feedback loop is shown in Fig. 2.

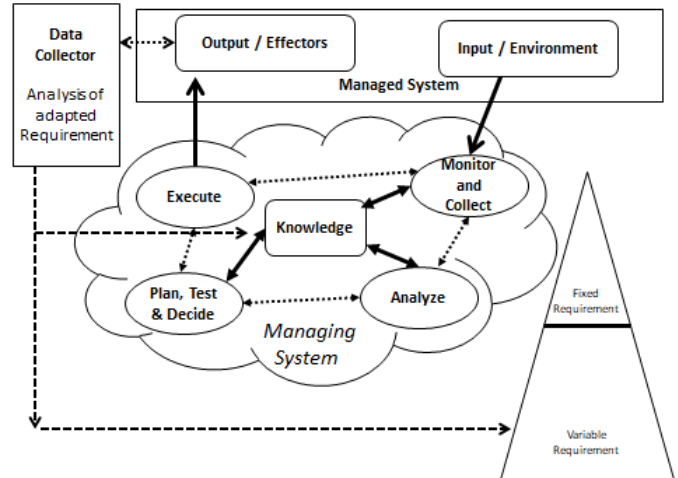


Fig. 3. Proposed architecture for modelling SAS.

### VI. PROPOSED ARCHITECTURE

We have proposed an architecture which not only focuses on performing the adaptation at run time but it also keeps a track of the results of adaptation when it is executed. As given in Fig. 3, the overall system is divided in to four major parts namely managed system, managing system, data collector and requirement pyramid. The top most part covers the input and output channels of the system and is classified as managed system in our architecture. Generally, input is given by sensors and is received by monitor agent. Monitor agent performs the transformation and converts the inputs in to a form that can be further analyzed. Another step that is performed by the monitor phase is to check the set of inputs against the existing knowledge. In case, the input from the environment matches any existing requirement, no adaptation is performed and the system acts according to the existing requirement. If the inputs do not match any of the existing requirements, then it is passed to the next phase, which analyzes the input according to the system goals, existing requirements, system objectives and overall preferences of the user. All this is available in the knowledge repository. Once the analysis is completed, all the data is forwarded to the next agent which is the plan, test and decide agent. Here the requirements are mapped to the nearest match, fuzzy rules are applied and possible adaptations are proposed. After formulating a few possible adaptations, these are tested on the criterion given in knowledge base, additionally; the capability of system effectors is also checked. For instance, an adaptation to take an aerial route to destination will fail the test for a car. Once the planning and testing is completed, one proposed adaptation is finalized and sent to the next agent which transforms the proposed execution in to the form that can be understood by the output channels. The process does



not end here, the data collector agent continuously monitors the managed system during its execution and results of execution are recorded. We may have two possible scenarios here, either the proposed adaptation has been successful or it has ended up in failure. In both cases the data is recorded and knowledge base is updated with a flag of success or failure, the successful adaptation is also recorded in the variable requirement part. In case of failure, the process is repeated iteratively till a final goal is achieved. This is kept for future enhancement in system and to make sure that all capabilities of system are available in the requirement set of the system.

An important contribution of this work that the system is designed in way that performs the adaptation at run-time, monitors the quality of output produced by the proposed adaptation and regular update of the knowledge and requirement base. All these modules will be analyzed, modeled and verified using formal methods.

## VII. CONCLUSION AND FUTURE WORK

This research has two major contributions, firstly we have proposed an integration of formal methods, agent based modeling and SAS for successfully analyzing, testing and implementing the systems that have the capability to adapt at run time. Secondly, an overall architecture of the complete system is given, which includes four major components. It is to be noted that we have successfully modeled the first phase using Petri Nets, the results have been very encouraging and the complete system will be analyzed, modeled, simulated, verified and tested by using formal methods. The given architecture gives a concrete base for the researchers and practitioners to implement systems that have the capability to adapt during execution. This is the first step toward development of a multi-agent autonomous formal model for self-adaptive system. We have successfully applied Petri nets to model feedback loop [12] and the work will be extend for a complete model for SAS using formal methods.

This architecture will be further be extended for distributed systems where the variability of inputs in more and there are multiple feedback loops at each node. Further, each agent will be implemented and the task will be further sub divided in to multiple agents where each agent will be designed to perform an atomic task.

### REFERENCES

- [1] G. Tallabaci and V. E. Silva Souza, "Engineering adaptation with Zanshin: An experience report," in ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems, 2013, pp. 93–102.
- [2] F. Kneer and E. Kamsties, "A framework for prototyping and evaluating self-adaptive systems - A research preview," in CEUR Workshop Proceedings, 2016, vol. 1564.
- [3] Y. Abuseta and K. Swesi, "Design Patterns for Self Adaptive Systems Engineering," *Int. J. Softw. Eng. Appl.*, vol. 6, no. 4, pp. 11–28, 2015.
- [4] N. Esfahani and S. Malek, "Uncertainty in Self-Adaptive Software Systems," in Lecture Notes in Computer Science, 2013, pp. 214–238.
- [5] C. Krupitzer, F. M. Roth, S. Vansyckel, G. Schiele, and C. Becker, "A survey on engineering approaches for self-adaptive systems," *Pervasive Mob. Comput.*, vol. 17, no. PB, pp. 184–206, 2015.
- [6] D. G. D. La Iglesia and D. Weyns, "MAPE-K Formal Templates to Rigorously Design Behaviors for Self-Adaptive Systems," *ACM Trans. Auton. Adapt. Syst.*, vol. 10, no. 3, pp. 1–31, 2015.
- [7] N. Khakpour, S. Jalili, C. Talcott, M. Sirjani, and M. Mousavi, "Formal modeling of evolving self-adaptive systems," in *Science of Computer Programming*, 2012, vol. 78, no. 1, pp. 3–26.
- [8] M. Luckey and G. Engels, "High-quality specification of self-adaptive software systems," in *ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems*, 2013, pp. 143–152.
- [9] P. Arcaini, E. Riccobene, and P. Scandurra, "Formal Design and Verification of Self-Adaptive Systems with Decentralized Control," *ACM Trans. Auton. Adapt. Syst.*, vol. 11, no. 4, pp. 1–35, 2017.
- [10] R. De Lemos et al., "Software engineering for self-adaptive systems: A second research roadmap," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7475 LNCS, pp. 1–32.
- [11] C. Macal and M. North, "Introductory tutorial: Agent-based modeling and simulation," in *Proceedings - Winter Simulation Conference*, 2015, vol. 2015-Janua, pp. 6–20.
- [12] N. A. Mian and F. Ahmad, "Modeling and Analysis of MAPE-K loop in Self Adaptive Systems using Petri Nets," vol. 17, no. 12, pp. 158–163, 2017.
- [13] M. I. Tariq, S. Tayyaba, M. U. Hashmi, M. W. Ashraf, and N. A. Mian, "Agent Based Information Security Threat Management Framework for Hybrid Cloud Computing," vol. 17, no. 12, pp. 57–66, 2017.
- [14] F. Krikava and P. Collet, "A Reflective Model for Architecting Feedback Control Systems," in *Proceeding of the 2011 International Conference on Software Engineering and Knowledge Engineering*, 2011, p. 7.
- [15] F. D. Macías-Escrivá, R. Haber, R. Del Toro, and V. Hernandez, "Self-adaptive systems: A survey of current approaches, research challenges and applications," *Expert Systems with Applications*, vol. 40, no. 18, pp. 7267–7279, 2013.
- [16] B. H. C. Cheng et al., "Software Engineering for Self-Adaptive Systems: A Research Roadmap," *Softw. Eng. Self-Adaptive Syst.*, pp. 1–26, 2009.
- [17] M. Amoui, M. Derakhshanmanesh, J. Ebert, and L. Tahvildari, "Achieving dynamic adaptation via management and interpretation of runtime models," *J. Syst. Softw.*, vol. 85, no. 12, pp. 2720–2737, 2012.
- [18] J. C. Muñoz-Fernández et al., "Capturing ambiguity in artifacts to support requirements engineering for self-adaptive systems," in *CEUR Workshop Proceedings*, 2017, vol. 1796.
- [19] S. Kounev et al., "The Notion of Self-aware Computing," in *Self-Aware Computing Systems*, 2017, pp. 3–16.
- [20] M. Ahmad, N. Belloir, and J. M. Bruel, "Modeling and verification of Functional and Non-Functional Requirements of ambient Self-Adaptive Systems," *J. Syst. Softw.*, vol. 107, pp. 50–70, 2015.
- [21] Y. Brun et al., "A Design Space for Self-Adaptive Systems," *Softw. Eng. Self-Adaptive Syst. II SE - 2*, vol. 7475, pp. 33–50, 2013.
- [22] B. H. C. Cheng, P. Sawyer, N. Bencomo, and J. Whittle, "A goal-based modeling approach to develop requirements of an adaptive system with environmental uncertainty," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5795 LNCS, pp. 468–483.
- [23] F. Dalpiaz, P. Giorgini, and J. Mylopoulos, "Adaptive socio-technical systems: A requirements-based approach," *Requir. Eng.*, vol. 18, no. 1, pp. 1–24, 2013.
- [24] N. Bencomo, K. Welsh, P. Sawyer, and J. Whittle, "Self-explanation in adaptive systems," in *Proceedings - 2012 IEEE 17th International Conference on Engineering of Complex Computer Systems, ICECCS 2012*, 2012, pp. 157–166.
- [25] J. Cámara et al., *Self-aware computing systems: Related concepts and research areas*. 2017.
- [26] N. A. Qureshi, A. Perini, N. A. Ernst, and J. Mylopoulos, "Towards a continuous requirements engineering framework for self-adaptive systems," in *2010 First International Workshop on Requirements@Run.Time*, 2010, pp. 9–16.
- [27] D. B. Abeywickrama, N. Bicocchi, and F. Zambonelli, "SOTA: Towards a general model for self-adaptive systems," in *Proceedings of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE*, 2012, pp. 48–53.

- [28] Q. Liu, S. Wu, D. Wang, Z. Li, and L. Wang, "Context-Aware sequential recommendation," in *Proceedings - IEEE International Conference on Data Mining, ICDM, 2017*, pp. 1053–1058.
- [29] B. Cilogluligil and M. M. Inceoglu, "User Modeling for Adaptive E-Learning Systems," in *ICCSA, 2012*, pp. 550–561.
- [30] J. Andersson, R. De Lemos, S. Malek, and D. Weyns, "Modeling dimensions of self-adaptive software systems," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5525 LNCS, pp. 27–47.
- [31] L. Gherardi and N. Hochgeschwender, "Poster: Model-based Run-time Variability Resolution for Robotic Applications," in *Proceedings - International Conference on Software Engineering, 2015*, vol. 2, pp. 829–830.
- [32] Y. Brun et al., "Engineering self-adaptive systems through feedback loops," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5525 LNCS, pp. 48–70.
- [33] J. Levinson et al., "Towards fully autonomous driving: Systems and algorithms," in *IEEE Intelligent Vehicles Symposium, Proceedings, 2011*, pp. 163–168.
- [34] S. Sucipto and R. S. Wahono, "A Systematic Literature Review of Requirements Engineering for Self-Adaptative Systems," in *Journal of Software Engineering*, vol. 1, no. 1, 2015, pp. 55–71.
- [35] Y. Zhao, Z. Yang, and D. Ma, "A survey on formal specification and verification of separation kernels," *Frontiers of Computer Science*, vol. 11, no. 4, pp. 585–607, 2017.
- [36] S. M. Edgar and S. A. Alexei, "Power and limitations of formal methods for software fabrication: Thirty years later," *Informatica (Slovenia)*, vol. 41, no. 3, pp. 275–282, 2017.
- [37] G. Su, T. Chen, Y. Feng, D. S. Rosenblum, and P. S. Thiagarajan, "An iterative decision-making scheme for markov decision processes and its application to self-adaptive systems," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9633, pp. 269–286.

# Reverse Engineering State and Strategy Design Patterns using Static Code Analysis

Khaled Abdelsalam Mohamed, Amr Kamel  
Faculty of Computers and Information, Cairo University  
Giza, Postal Code: 12613, Egypt

**Abstract**—This paper presents an approach to detect behavioral design patterns from source code using static analysis techniques. It depends on the concept of Code Property Graph and enriching graph with relationships and properties specific to Design Patterns, to simplify the process of Design Pattern detection. This approach used NoSQL graph database (Neo4j) and uses graph traversal language (Gremlin) for doing graph matching. Our approach, converts the tasks of design pattern detection to a graph matching task by representing Design Patterns in form of graph queries and running it on graph database.

**Keywords**—Reverse engineering; source code analysis; design patterns; static analysis; graph matching; Gremlin; Joern; Neo4j

## I. INTRODUCTION

Software as an artifact is not static. Software is in continuous change. During the software lifetime, there are a number of sources of change that affects it, e.g., bug fixing, new features added, requirements changes or technology changes. To make such changes, the assigned developers should have a good understanding of the software internals s/he is going to change. Typically, a team of developers implements applications. Sometimes, the developer who is assigned to change the application is not a member of the original development team, even if the developer was one of the team, it is unlikely that he knows every little detail of the software implementation. Here comes the importance of having a complete documentation of the software, so all development team have required insight of the software.

The documentation always has many problems. An extreme problem is that documentation may be lost, so the developer will need to start understanding the software from scratch, although this not always the case, the most probable problem of documentation is not being synchronized with the application. If a developer depends on this outdated documentation s/he will get wrong understanding of the software at hand, which will be an obstacle for the developer to accomplish the task.

As the documentation goes out of sync, it will be a source of problems. If a developer starts from outdated document and makes his changes without reflecting changes in the documentation continuously, the significance of the documentation will diminish over time, eventually the documentation will be useless. One reason of such problem is that the job of updating documentation is a tedious task for the developers.

A lot of research effort has been done in the field of gaining insight of legacy software and knowing the intentions of software code. In addition, it is considered one of the important reverse engineering research fields. One of the different approaches for gaining understanding of legacy software is to extract design patterns out of the source code; design patterns [1] describe high quality practical solutions to recurring programming problems.

Design patterns are a toolbox of reusable solutions and best practices that have been refined over many years to a compact format. Design patterns do not describe specific algorithms or data structures like linked list or variable length arrays, which are traditionally implemented in individual classes. As each design pattern has a specific intention, detecting them out of source code can lead to understanding the usage of different parts of the software, design patterns provide a coherent map that leads the developers through the design of the software analyzed.

This document is divided into four sections. In Section II, The different approaches used for detecting design patterns are presented. In Section III, structural similarities and behavioral differences between State design pattern and Strategy design pattern are presented. In Section IV, The different techniques and frameworks used for doing static analysis to source code are presented. In Section V, Our approach for detecting design patterns in source code using graph enrichment and static analysis techniques is presented. Finally, Section VI, offers the conclusions and future work.

## II. RELATED WORK

The architecture design of software highly affects its quality. The high quality software follows design patterns. The mining of design patterns can be helpful in understanding and knowing design decisions in legacy systems [2]-[5].

The design pattern recovery is considered one of the hot topics in reverse engineering research field [2], [6], [7]. There are many approaches used in literature to recover design patterns from source code to facilitate software maintenance [8], [9] and program comprehension [10]-[12]. The techniques used in literature can be classified based on two factors [13], the type of analysis and the search methodology.

### A. Analysis Type

Based on the analysis type, the pattern recovery approaches can be classified [13] into structural analysis, behavioral analysis and semantic analysis.

Structural analysis [14] are based on recovering inter-class relationships such as class inheritance, association, composition, modifiers of classes and methods, method parameters, etc. They focus on recovering structural design patterns such as Proxy, Decorator and Adapter, but they completely miss the behavioral aspects of design patterns.

Behavioral analysis [15] focuses on the execution behavior of the program. These approaches are based on dynamic analysis, machine learning and static analysis techniques to extract behavioral aspects of the pattern. Supplementing behavioral analysis by structural analysis techniques helps in recovery of identical or weak-structure patterns where structural analysis fails.

Semantic analysis approaches supplements both structural and behavioral analysis approaches to reduce the false positive rate of recognition of design patterns. The semantic analysis approach uses the naming convention of classes and methods in recovering different roles inside design patterns.

### B. Searching Techniques

Based on the searching techniques, the pattern recovery approaches can be summarized as follows:

#### 1) Database queries

In this approach, the source code is first transformed to an intermediate representation such as (ASG, AST, XMI, metadata and UML structures etc.) then SQL queries are used to extract information from a specific representation.

#### 2) Constraint Resolver

The approach [14] used by The PTIDEJ team is a multilayered approach, where design motifs are described as constraint systems where each role is represented as a variable. Relationships among roles are represented as constraints among these variables.

#### 3) XPG formalism and parsing

This approach [6] used a technique where SVG (scalable vector graphics) format is used as an intermediate representation of source code and design patterns are represented using a visual language. Patterns are recovered using a visual language parsing technique by mapping visual language grammar of the patterns with the graph representation. The advantages of these approaches are the visualization and good precision, but are limited only to structural design patterns.

#### 4) UML structures and matrices techniques

Metrics techniques [16]-[18] compute program metrics such as generalization, aggregation, association, etc. from different representations of source code and then a number of techniques are used to compare metric values of each design pattern definition with source code metrics. These techniques are computationally efficient because of search space reduction through filtration.

### III. STATE VS STRATEGY DESIGN PATTERNS

State and Strategy design patterns are two interesting patterns, as both of them have the same structure although each of them have a different behavior.

Balanyi and Rudolf [19] stated that during their process of pattern formalization, they found an interesting problem, which is that both State and Strategy patterns have identical structure, and the differences between them are in motivation and intention that they could not formalize.

Aikaterini et al. [20] proposed a method to automatically transform/refactor source code to comply with the Strategy design pattern. Their method complements JDeodorant [21] that focuses mainly on the State pattern, by taking into account behavioral properties of the Strategy design pattern during candidate selection phase.

Von Detten and Platenius [22] used dynamic analysis to analyze the runtime behavior of the system. First static analysis is used to detect the structure of a design pattern, the detected classes, methods are annotated, then during the dynamic analysis phase the behavior of annotated classes, and methods are traced during the software execution. For each pattern candidate, a number of traces are generated. A behavioral analysis algorithm assess if traces of each pattern candidate conform to the corresponding behavioral pattern. If most of the traces of a candidate match the behavioral pattern, the candidate pattern is accepted and if the most of traces do not match then the candidate pattern is rejected.

Hummel and Burger [23] mentioned that the class diagram of the strategy and state patterns are identical from the class diagram perspective. In addition, the main difference between them resides in who controls the change of state or strategy. The state implementations have control over state changes themselves, but for strategy pattern, the client is responsible for the changes of the applied strategy.

Uchiyama et al. [24] used an approach of metrics and machine learning technique to detect design patterns. This work was interested in distinguishing between State patterns from Strategy pattern. They firstly using various metrics and their machine learning identify the roles and secondly detect patterns as structure of those roles.

The below class diagram (Fig. 1) shows the structure of the Strategy Design Pattern.

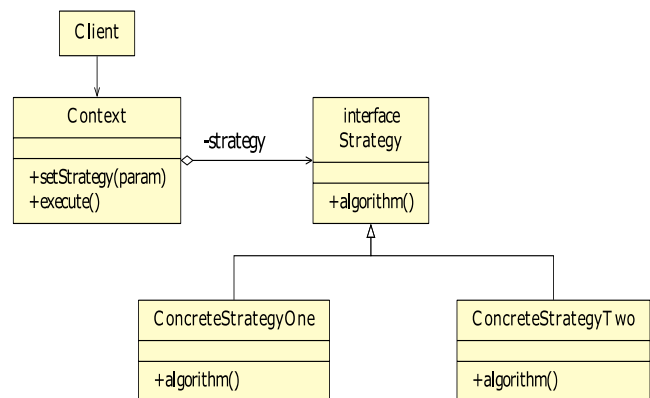


Fig. 1. Class diagram of strategy design pattern.

### A. Structural Characteristics of Strategy Design Pattern

- 1) *Classes*: Client, Context, Strategy, ConcreteStrategy.
- 2) *Use*: Client uses Context
- 3) *Aggregation*: Context aggregates Strategy.
- 4) *Inheritance*: More than one ConcreteStrategy inherits Strategy.
- 5) *Abstract Method*: Strategy contains an abstract method.
- 6) *Method Overriding*: ConcreteStrategy(ies) override Strategy Abstract Method "algorithm".

The below sequence diagram (Fig. 2) show the behavior of Strategy pattern:

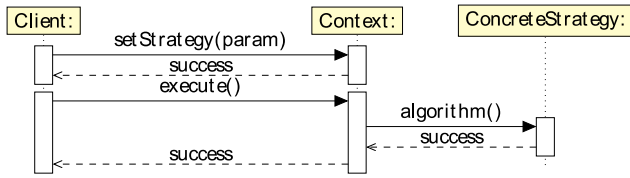


Fig. 2. Sequence diagram of strategy design pattern.

### B. Behavioral Characteristics of Strategy Design Pattern

- 1) *Call A*: A call from "Client" to "Context", to set the required strategy.
- 2) *Object Creation*: "Context" created "ConcreteStrategy", based on value sent by "Context".
- 3) *Call B*: A call from "Context" to the abstract method implementation in "ConcreteStrategy".
- 4) *Call C*: A call from "Client" to "Context", to start execution.
- 5) *Call D*: A call from "Context" to "ConcreteStrategy", to perform the algorithm.

## IV. TECHNIQUES AND FRAMEWORKS

Static analysis [25] is the most frequently used approach for code analysis, dynamic analysis needs that the source code to be runnable; however, static analysis can be used for incomplete source code.

A number of techniques are used for doing static analysis for source code. Next, a number of most common static analysis techniques are presented.

### A. Techniques

#### 1) Call Graph

Call Graph [26] represents the possible callers at each call site in each function. Call Graph is a directed that represents the relationships between the program's functions. There is a wide range of algorithms for call graph construction [26], e.g. RTA, 0-CFA and SCS.

#### 2) Control Flow Graph

Control flow [27] graph, is a directed graph where nodes represent program statements, and edges represent flow paths from one program statements to another.

Constructing CFG can be for source code or byte codes, for example JavaPDG [28], constructed the Control Flow Graph from byte code using the following steps:

- a) Get all instructions/statements of the method.
- b) Create a node that represent method entry.
- c) Make a link between entry node and first instruction/statement.
- d) Create a node that represent method exit.
- e) Get reference to last instruction/statement.
- f) Make a link between last instruction and method exit node, if last instruction/statement is not "Return".
- g) Make a reference to previous and current instruction and Loop all instructions.
- h) If previous instruction type is not ("CP" or "JU" or "Return"), make link between pre and cur instructions.
- i) If current instruction type is ("CP" or "JU"), make a link between cur instruction and all jump labels.
- j) If current instruction of type "Return", make a link between cur instruction and method exit node.

#### 3) Dominator Tree

A dominator tree [29] is a graph  $G = (V, E, r)$ , Where V is the set of vertices, E is the set of edges and r is the root node of the graph. Every node except root node in the graph has a unique immediate dominator. If two nodes "v" and "w" in the dominator tree, and "v" is the ancestor of "w", then "v" dominates "w". Node "v" dominates "w" if all paths from the entry node to "w" contains "v". In addition, "w" post-dominates "v" if all paths from "v" to exit node contains "w".

The dominator tree is computed from Control Flow Graph. By having CFG and DT, control dependence graph can be derived.

#### 4) Control Dependence Graph

Control Dependence Graph [30] is a merge between Control Flow Graph and Dominator Tree, It can be defined as a directed graph "G", it has two unique entries, entry node "START" and exit node "STOP". For any node "N" there exists a path from "START" to "N" and from "N" to "STOP". So, node "Y" control dependent on "X" iff:

- There is a directed path "P" from "X" to "Y", which contains node "Z", where "Y" post-dominates "Z".
- "Y" doesn't post-dominates "X".
- Node "V" is post-dominated by "W", if every directed path from "V" to "STOP" contains "W".

#### 5) Data Dependence Graph

A data dependence graph (DDG) for every method is calculated by tracking data flows on its CFG. A definition-use chain, i.e., one instruction assigns a value to an abstract variable, usually represents a data flow and the other instruction uses the value. Reaching-definition and upward-exposed-uses analyses are conducted following the steps:

Analyze the effect of each instruction in terms of its variable definition and use sets.

Iteratively propagate the information over the CFG;

During each iteration, inspect whether there is any unknown definer/assigner of the variable(s) used in each instruction, and update its information sets accordingly.

Once the information propagation ends (no changes are found), the data dependences between instructions is calculated by the definition-use chain analysis.

#### 6) Program Dependence Graph

A PDG [30] is defined as a labeled, directed graph that maps out control dependences and data dependences between elements in a program.

#### 7) System Dependence Graph

A system dependence graph (SDG) [31] is a generalization of PDG and contains one procedure dependence graph (pDG) for each method.

#### 8) Code Property Graph

A single representation alone to represent the source code in insufficient. Fabian et al. [32] combines three representations into a unified data structure. In [32], author introduced a new concept of Code Property Graph which models ASTs, CFGs and PDGs as property graphs.

Fabian et al. [32] showed that common types of vulnerabilities can be modeled as a traversal of code property graph, also by importing code property graph into a graph database, makes traversals can be executed efficiently on large code base.

A code property graph is a property graph  $G = (V, E, \lambda, \mu)$  constructed from AST, CFG and PDF of source code:

$$V = V_A,$$

$$E = E_A \cup E_C \cup E_P,$$

$$E = \lambda_A \cup \lambda_C \cup \lambda_P \text{ And}$$

$$E = \mu_A \cup \mu_C,$$

### B. Frameworks

Here, a number of frameworks that provides implementations, for different static analysis techniques, first works on binary level, and the second works on source code level.

#### 1) JavaPDG

JavaPDG [28] implements static dependence analysis for Java Virtual Machine (JVM) bytecode. The tool parses the bytecode of a Java program, computes the SDG and related graphs, and stores the data for each program in a database. JavaPDG includes tools for visualizing the graphs it produces and for exporting the data in the JSON format. Additionally, users are able to query the output using SQL by utilizing Apache Derby. The analysis process takes as input the compiled class files of a Java program, and yields a SDG and related graphs as the final output.

The steps for building SDG are as follows:

##### a) Preprocessing

In the SDG, one PDG vertex represents each instruction. Artificial entry and exit vertices for every method are added to the graph to represent the start and end of the method, respectively. A vertex is added for every call-site as its actual-output parameter if the callee method has any return value.

##### b) Inter-procedural Analysis

An SDG is a collection of interconnected pDGs, each of which is composed of the CDG and DDG for a method. The static call graph of a program is used to investigate communications between methods. Based on the call graph, three types of inter-procedural control and data dependences are computed.

The output SDG is a labeled, directed graph consisting of multiple PDGs. Besides the SDG, JavaPDG outputs some additional information, including:

- Static structure of a program that describes classes, fields, methods, and relationships among them.
- Variable information that contains the name, type and scope of every class field, object field and local variable (including formal input parameter).
- Control flow graphs and dominance trees that are constructed during dependence analysis and share the same vertices as in the SDG.
- A static call graph whose vertices correspond to Java methods and whose edges represent potential caller-callee relationships indicated in the program.

#### 2) Joern

Joern [33] is a platform for robust analysis of C/C++ code. It generates code property graphs; code property graph [32] consists of code's syntax, control-flow, data-flow and type information. These graphs are then stored in Neo4J database. By this, it is possible to do code mining through running search queries formulated in the graph traversal language Gremlin.

Joern platform [33] consists of three components joern(-core), python-joern and joern-tools. Joern(-core) is the main component, it takes the source code and parses it, creates a code property graphs [32] and finally, import the graphs into Neo4j database. Python-joern is a python interface to Joern database. It provides a number of utilities for the common operations of traversing code property graphs. Joern-tools is a collection of command line tools that makes using python-joern utilities possible from the shell.

#### 3) Gremlin

Gremlin [34] is the graph traversal language of Apache TinkerPop. Gremlin is a functional, data-flow language that enables users to succinctly express complex traversals on (or queries of) their application's property graph.

Gremlin recently appeared in a number of works such as Model-to-Model transformation [35], modeling and discovering vulnerabilities in source code [32].

## V. APPROACH

In our approach, detecting behavioral design patterns from source code using static analysis techniques is chosen. Program

Dependence Analysis, Control Dependence Analysis and Data Dependence Analysis are applied on source code to be able to capture the behavioral characteristics of design patterns.

In our approach, The detection problem is represented as a graph matching problem, the source code graphs is stored in a graph database e.g. Neo4j, and the design pattern features are extracted by running graph matching queries against the database where the source code graphs are saved.

In our approach, Joern platform [33] is used to analyze the source code and save the analyzed source code in graph database.

Joern platform is designed mainly for the detection of vulnerabilities in code. Therefore, Joern is mainly interested in C++ code at functions level and not interested in Object Oriented interactions between classes.

As inheritance between classes forms an important information that is required during the process of detecting design patterns, a minor change to Joern platform is made to store the parent class of each class during the parsing step of the analysis process.

The following figure (Fig. 3) shows a high-level view of the steps of our approach:

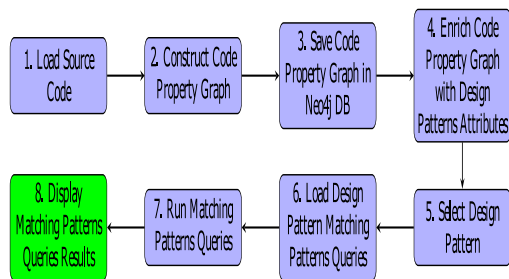


Fig. 3. Approach steps.

The steps of our approach is as follows:

- 1) Load Source Code
- 2) Generate Code Property Graphs [32] of the source code under investigation
- 3) Insert the Code Property Graph into a graph database.
- 4) Enrich Generated Code Property Graph with properties and relations between nodes to simplify the graph matching steps.
- 5) Decide the design pattern(s) to detect.
- 6) Load the list of features (structural and behavioral) that represent the design pattern.
- 7) Load the corresponding graph matching query for each design pattern features
- 8) For each design pattern, run features detection queries using Gremlin language [34] against the Enriched Code Property Graphs in the Neo4j.
- 9) Inspect detected features and decide if design pattern instance is found or not.
- 10) Display results.

Our approach enriches the Code Property Graph with a number of properties and relations between vertices to make the phase of detecting State and Strategy patterns straight forward. Once these relations and properties are constructed, the pattern detection graph matching algorithms for State and Strategy patterns are used to detect State and Strategy patterns from the Enriched Code Property Graph.

To express the capabilities of our approach, differentiating between State and Strategy design patterns is selected, as they are identical from the structural perspective but differs from the behavioral and run time perspective. Our approach show that differentiating between these patterns is possible, while still using static analysis and no dynamic analysis is needed.

The enrichments required for differentiating between State and Strategy Patterns are listed:

1) *Methods to Classes*: C++ class methods can be defined outside its class, in such case Joern tool does not link between the class and its member method, so a link between methods and their classes is created.

The steps are as follows:

- a) List all methods that their names contains symbol “:”.
- b) Split the method name into two parts, class name and method name.
- c) Search for class with the same name of first part of full method name.
- d) Make a link of type “IS\_CLASS\_OF” between the class and method.

2) *Inheritance*: An inheritance relation between each super class and their subclasses (Fig. 4).

The steps used to construct the inheritance relationship:

- List all classes that their base class name not equals to “<unnamed>”.
- For each class in the list, get the class’s base class name.
- Search for a class that its name equals to child’s base class name.
- Make a link of type “INHERITS” between the class and its base class.

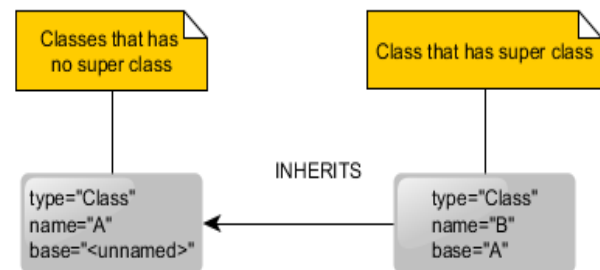


Fig. 4. Inheritance relation.

3) *Abstract (Virtual) Methods*: Abstract methods can have two types, one that has no body definition (Pure Virtual), second that *has* body definition and declared with virtual keyword as modifier (Fig. 5).

The steps to mark a method declaration as virtual are as follows:

- Get list of node that are of type "Decl"
- Extract nodes that contain brackets, as indication that they are methods declaration.
- Extract nodes that do not have body definition.
- Mark extracted nodes as abstract.

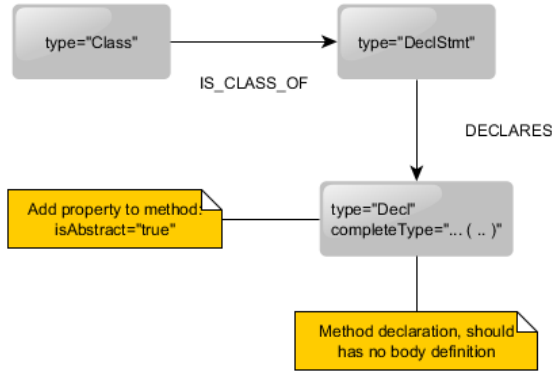


Fig. 5. Pure abstract method.

The steps to mark a method with body definition as virtual are as follows (Fig. 6):

- Get list of node that are of type "Function".
- Traverse to the return type of the function.
- If return type contains keyword "virtual", then this function is marked as abstract.

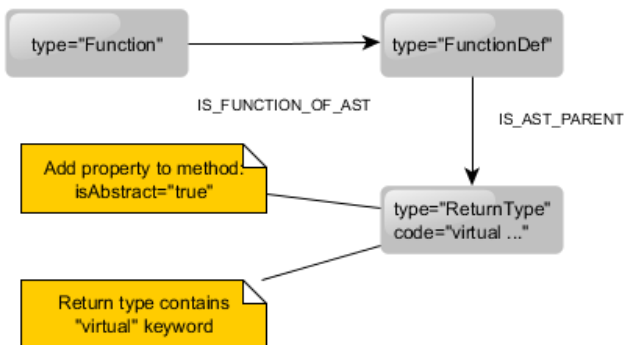


Fig. 6. Abstract method.

4) *Class Aggregates Class*: A new relation between two classes are created if one aggregates the other (Fig. 7).

The steps to construct aggregation relation between two classes are:

- Get all declaration statements for each class e.g. "Class A".
- For each declaration statement extract class name from its "baseType".

- Search for the class with same name extracted in previous step "Class B".
- Create a link between that represents the aggregation between "Class A" and "Class B".

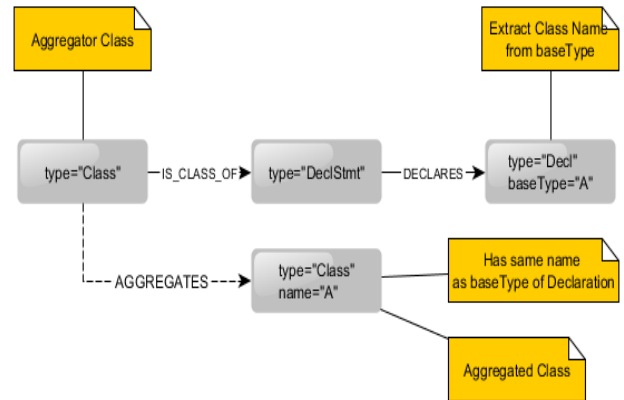


Fig. 7. Class aggregates class.

5) *Method Creates Class*: A relation between a method and a class is created, if a method creates a class (Fig. 8).

The steps to create relation between class and the method that creates it are:

- Get all method statements that contains "new" statement.
- Extract class name from the new statement.
- Search for a class that has the same name of step b.
- Create a link of type "Create" between the Method and the Class.

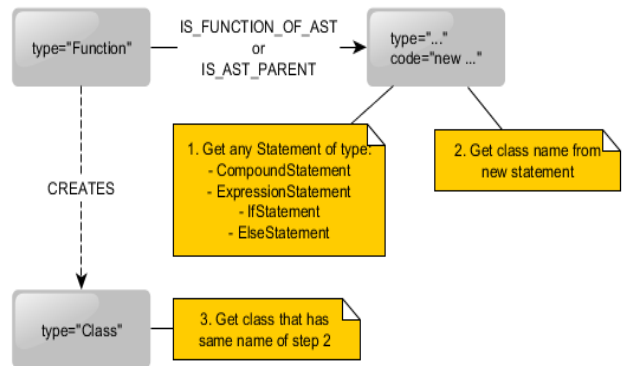


Fig. 8. Method creates class.

6) *Method Overrides Method*: A new relation between two methods, if one method overrides the other method (Fig. 9).

- Get list of all functions and declaration statements.
- Get list of classes of step "a".
- Get list of classes that are super class of classes in step "b".
- Get list of all functions and declaration statements that are abstract of classes in step "c".
- Filter list of step "d", which Subclass method name should be equals to Superclass method name of step "c".



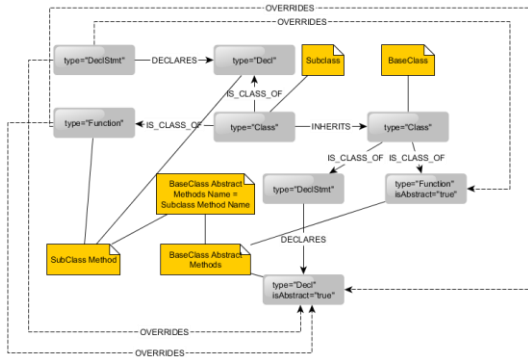


Fig. 9. Method overrides method.

7) *Method Calls Method*: A relation between two methods are created if one *method* creates the other.

The steps to construct these new relations (Fig. 10) are:

- a) Get list of all method “Caller Methods”.
- b) Get list of classes of step “a”, and keep method that are not related to classes e.g. main method.
- c) Get list of all statements of type “Callee” of step “a” “Call Sites”.
- d) Get list of nodes of types “PtrMemberAccess”, “MemberAccess”, or “Identifier” that are linked to step “c”.
- e) Get list of nodes that are linked to nodes of type “PtrMemberAccess” or “MemberAccess” of step “d” with in-edge of type “USE”.
- f) Get list of nodes of type “Parameter”, “Decl”, or “IdentifierDeclStatement” and having in-edge of type “DEF” from step “e”.
- g) Keep nodes of type “Identifier” or “Symbol” that are not in step “f” but have declaration in classes of caller methods of step “a”.
- h) Loop lists of steps “f” & “g”.
- i) Get callee method name.
- j) Get callee class name.
- k) Get node represented by class and method names (Callee Method).
- l) Create a link between Caller Method and Callee Method (Step “h.iii”).
- m) Create a link between Call Site (Step “c”) and Callee Method (Step “h.iii”).

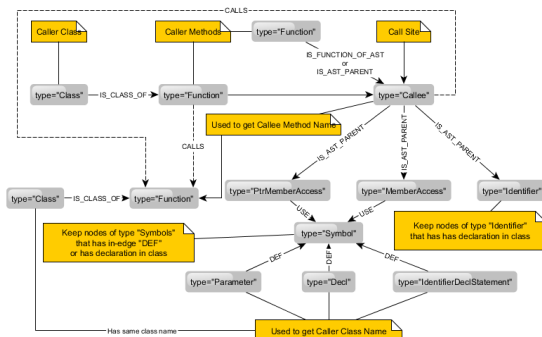


Fig. 10. Method calls method.

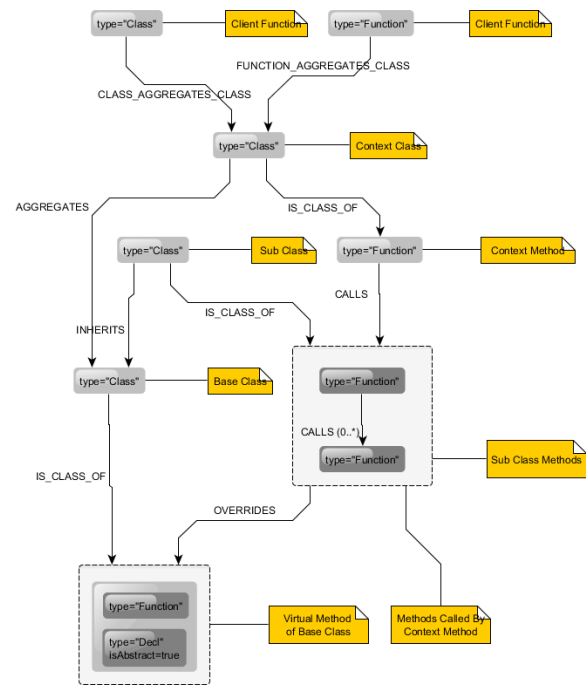


Fig. 11. State and strategy design pattern candidates.

After enriching phase, the code graph is ready for the detection phase, the detection phase for State and Strategy design patterns is divide into two steps, first step to detect candidates that can be State or Strategy (Fig. 11), this step captures the structure of these design patterns. The second step is for differentiating between State and Strategy patterns.

The steps for deciding if a candidate is a State or a Strategy design patterns are:

- a) Loop each candidate.
- b) Get symbols that used by Context Class to aggregate Base Class.
- c) Get methods that use the symbols from step (b).
- d) Check if methods from step (c) includes sub classes methods from pattern candidates.
- e) If step (d) is true then the candidate is a State design pattern.
- f) Check of methods from step (c) includes the client method from pattern candidates.
- g) If step (f) is true, then the candidate is a Strategy design pattern.

## VI. CONCLUSION AND FUTURE WORK

In this work, an approach is presented for detecting design patterns in source code, by representing the source code in form of a special graph named Code Property Graph [32], using Joern platform [33]. In addition, our approach is shown to able to differentiate between State and Strategy design patterns, which are identical from structural perspective, but differs at run time, using advanced static analysis techniques without the need to use run time dynamic analysis. The code property graph is enriched by constructing new properties and

relationships between vertices of the graph, the enrichments done by our approach presented a number of techniques to transform graphs from the functions paradigm level to the level of object oriented paradigm, so that code graph is ready for object oriented analysis and design patterns detection.

In this work, C++ code is used, because Joern platform currently supports C++, in our future work we will work on supporting Java programming language, to be able to compare our results with other approaches, as most design pattern detection benchmarks are java based [4], [11], [36], [37]. Our approach is not dependent on a specific language for enrichment and design pattern detection, as it depends on manipulating the code graph directly at run time before running the detection algorithms, which depends on the code graph also.

In future work, a catalogue of all relationships and properties of design patterns will be created, to enrich the code graph with these relationships and properties as a step before pattern detection step, so a catalogue containing a one to one mapping between a design pattern and its graph query will be available.

Design pattern can have more than one variant [38], in our future work, more than one graph definition to each design pattern will be supported, and detection algorithm will search for all different variants of design patterns to increase the true positive rate of our detection approach. Constructing graphs using design pattern concepts as relationships between vertices will make adding new design patterns or new variants of the design patterns more easily and user friendly.

#### REFERENCES

- [1] Vlissides, John and Helm, Richard and Johnson, Ralph and Gamma, Erich, Design Patterns: Elements of Reusable Object-Oriented Software, Addison-Wesley Pub Co, 1995.
- [2] Costagliola, Gennaro and De Lucia, Andrea and Deufemia, Vincenzo and Gravino, Carmine and Risi, Michele, "Design pattern recovery by visual language parsing," in Software Maintenance and Reengineering, 2005. CSMR 2005. Ninth European Conference on, 2005.
- [3] Dong, Jing and Zhao, Yajing and Peng, Tu, "A review of design pattern mining techniques," International Journal of Software Engineering and Knowledge Engineering, vol. 19, no. 06, pp. 823--855, 2009.
- [4] Fontana, Francesca Arcelli and Caracciolo, Andrea and Zanoni, Marco, "DPB: A benchmark for design pattern detection tools," in Software Maintenance and Reengineering (CSMR), 2012 16th European Conference on, 2012.
- [5] L. Wendehals, "Improving design pattern instance recognition by dynamic analysis," in Proc. of the ICSE 2003 Workshop on Dynamic Analysis (WODA), Portland, USA, 2003.
- [6] De Lucia, Andrea and Deufemia, Vincenzo and Gravino, Carmine and Risi, Michele, "Behavioral pattern identification through visual language parsing and code instrumentation," in Software Maintenance and Reengineering, 2009. CSMR'09. 13th European Conference on, IEEE, 2009.
- [7] Dong, Jing and Zhao, Yajing and Sun, Yongtao, "A matrix-based approach to recovering design patterns," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 39, no. 6, pp. 1271--1282, 2009.
- [8] Ng, Janice Ka-Yee and Guéhéneuc, Yann-Gaël and Antoniol, Giuliano, "Identification of behavioural and creational design motifs through dynamic analysis," Journal of Software: Evolution and Process, vol. 22, no. 8, pp. 597--627, 2010.
- [9] Fulop, Lajos Jenó and Ferenc, Rudolf and Gyimothy, Tibor, "Towards a benchmark for evaluating design pattern miner tools," in Software Maintenance and Reengineering, 2008. CSMR 2008. 12th European Conference on, 2008.
- [10] Guéhéneuc, Yann-Gaël and Antoniol, Giuliano, "Demima: A multilayered approach for design pattern identification," IEEE Transactions on Software Engineering, vol. 34, no. 5, pp. 667--684, 2008.
- [11] Kniesel, Gunter and Binun, Alexander and Hegedus, Peter and Fulop, Lajos Jenó and Chatzigeorgiou, Alexander and Guéhéneuc, Yann-Gaël and Tsantalis, Nikolaos, "DPDX--Towards a Common Result Exchange Format for Design Pattern Detection Tools," in Software Maintenance and Reengineering (CSMR), 2010 14th European Conference on, 2010.
- [12] Kniesel, Gunter and Binun, Alexander, "Standing on the shoulders of giants--a data fusion approach to design pattern detection," in Program Comprehension, 2009. ICPC'09. IEEE 17th International Conference on, 2009.
- [13] Rasool, Ghulam and Streitfeldt, Detlef, "A survey on design pattern recovery techniques," IJCSI International Journal of Computer Science Issues, vol. 8, no. 2, pp. 251--260, 2011.
- [14] De Lucia, Andrea and Deufemia, Vincenzo and Gravino, Carmine and Risi, Michele, "An Eclipse plug-in for the detection of design pattern instances through static and dynamic analysis," in Software Maintenance (ICSM), 2010 IEEE International Conference on, 2010.
- [15] Binun, Alexander and Kniesel, Günter, "Joining forces for higher precision and recall of design pattern detection," CS Department III, Uni. Bonn, Germany, Technical report IAI-TR-2012-01, 2012.
- [16] Guéhéneuc, Yann-Gaël and Guyomarc'h, Jean-Yves and Sahraoui, Houari, "Improving design-pattern identification: a new approach and an exploratory study," Software Quality Journal, vol. 18, no. 1, pp. 145--174, 2010.
- [17] Antoniol, Giuliano and Fiutem, Roberto and Cristoforetti, Luca, "Design pattern recovery in object-oriented software," in Program Comprehension, 1998. IWPC'98. Proceedings., 6th International Workshop on, 1998.
- [18] von Detten, Markus and Becker, Steffen, "Combining clustering and pattern detection for the reengineering of component-based software systems," in Proceedings of the joint ACM SIGSOFT conference--QoSA and ACM SIGSOFT symposium--ISARCS on Quality of software architectures--QoSA and architecting critical systems--ISARCS, 2011.
- [19] Balanyi, Zsolt and Ferenc, Rudolf, "Mining design patterns from C++ source code," in Software Maintenance, 2003. ICSM 2003. Proceedings. International Conference on, 2003.
- [20] Christopoulou, Aikaterini and Giakoumakis, Emmanouel A and Zafeiris, Vassilis E and Soukara, Vasiliki, "Automated refactoring to the Strategy design pattern," Information and Software Technology, vol. 54, no. 11, pp. 1202--1214, 2012.
- [21] Tsantalis, Nikolaos and Chatzigeorgiou, Alexander, "Identification of refactoring opportunities introducing polymorphism," Journal of Systems and Software, vol. 83, no. 3, pp. 391--404, 2010.
- [22] Von Detten, Markus and Platenius, Marie Christin, "Improving Dynamic Design Pattern Detection in Eclipse with Set Objects," in In Proceedings of the 7th International Fujaba Days, 2009.
- [23] Hummel, Oliver and Burger, Stefan, "Analyzing source code for automated design pattern recommendation," in Proceedings of the 3rd ACM SIGSOFT International Workshop on Software Analytics, 2017.
- [24] Uchiyama, Satoru and Kubo, Atsuto and Washizaki, Hironori and Fukazawa, Yoshiaki, "Detecting design patterns in object-oriented program source code by using metrics and machine learning," Journal of Software Engineering and Applications, vol. 7, no. 12, p. 983, 2014.
- [25] García-Ferreira, Iván and Laorden, Carlos and Santos, Igor and Bringas, Pablo García, "A survey on static analysis and model checking," in International Joint Conference SOCO'14-CISIS'14-ICEUTE'14, 2014.
- [26] Grove, David and DeFouw, Greg and Dean, Jeffrey and Chambers, Craig, "Call graph construction in object-oriented languages," ACM SIGPLAN Notices, vol. 32, no. 10, p. 108--124, 1997.
- [27] F. E. Allen, "Control flow analysis," ACM Sigplan Notices, vol. 5, pp. 1--19, 1970.

- [28] Shu, Gang and Sun, Boya and Henderson, Tim AD and Podgurski, Andy, "JavaPDG: A new platform for program dependence analysis," *Software Testing, Verification and Validation (ICST), 2013 IEEE Sixth International Conference on.* IEEE, 2013.
- [29] Lengauer, Thomas and Tarjan, Robert Endre, "A fast algorithm for finding dominators in a flowgraph," *ACM Transactions on Programming Languages and Systems (TOPLAS)*, vol. 1, no. 1, pp. 121-141, 1979.
- [30] Ferrante, Jeanne and Ottenstein, Karl J and Warren, Joe D, "The program dependence graph and its use in optimization," *ACM Transactions on Programming Languages and Systems (TOPLAS)* 9.3, pp. 319-349, 1987.
- [31] Horwitz, Susan and Reps, Thomas and Binkley, David, "Interprocedural Slicing Using Dependence Graphs," *ACM Transactions on Programming Languages and Systems*, vol. 12, no. 1, p. 26-61, 1990.
- [32] Yamaguchi, Fabian and Golde, Nico and Arp, Daniel and Rieck, Konrad, "Modeling and discovering vulnerabilities with code property graphs," *Security and Privacy (SP), 2014 IEEE Symposium on*, pp. 590-604, May 2014.
- [33] "Joern Website," [Online]. Available: <http://www.mlsec.org/joern/>.
- [34] "The Gremlin Graph Traversal Machine and Language," [Online]. Available: <https://tinkerpop.apache.org/gremlin.html>. [Accessed 2017].
- [35] G. a. S. G. a. C. J. Daniel, "Mogwai: a framework to handle complex queries on large models," in *Research Challenges in Information Science (RCIS), 2016 IEEE Tenth International Conference on*, 2016.
- [36] Tsantalis, Nikolaos and Chatzigeorgiou, Alexander and Stephanides, George and Halkidis, Spyros T, "Design pattern detection using similarity scoring," *IEEE transactions on software engineering*, vol. 32, no. 11, 2006.
- [37] Y.-G. Guéhenéuc, "P-mart: Pattern-like micro architecture repository," *Proceedings of the 1st EuroPLOP Focus Group on Pattern Repositories*, 2007.
- [38] Bayley, Ian and Zhu, Hong, "Formal specification of the variants and behavioural features of design patterns," *Journal of Systems and Software*, vol. 83, no. 2, pp. 209-221, 2010.

# OpenMP Implementation in the Characterization of an Urban Growth Model Cellular Automaton

Alvaro Peraza Garzón

Instituto Tecnológico de Mazatlán  
Universidad Autónoma de Sinaloa  
Sinaloa, México

René Rodríguez Zamora

Universidad Autónoma de Sinaloa  
Instituto Tecnológico de Mazatlán  
Sinaloa, México

Wenseslao Plata Rocha

Facultad de Ciencias de la Tierra y el  
Espacio  
Universidad Autónoma de Sinaloa  
Sinaloa, México

**Abstract**—This paper presents the implementation of a parallelization strategy using the OpenMP library, while developing a simulation tool based on a cellular automaton (CA) to run urban growth simulations. The characterization of an urban growth model CA is shown and it consists of a digitization process of the land use in order to get all the necessary elements for the CA to work. During the first simulation tests we noticed high processing times due to large quantity of calculations needed to perform one single simulation, in order to minimize this we implemented a parallelization strategy using the fork-join model in order to optimize the use of available hardware. The results obtained show a significant improvement in execution times in function of the number of available cores and map sizes, as a future work, it is planned to implement artificial neural networks in order to generate more complex urban growth scenarios.

**Keywords**—Cellular automata; parallel programming; simulation models; OpenMP; urban growth

## I. INTRODUCTION

The evolution in the land use of the territory is a fundamental element in our society, since it manifests different variables that affect our daily life, for example, accessibility to different points of interest within the city, slopes of the land, etc. This evolution has gained interest mainly fueled by the different environmental problems especially those in urban areas [1]. Thanks to the advances in the computing field and the development of important analytical tools such as Geographic Information Systems (GIS) or simulation models, the study of the changes taking place in metropolitan areas has been promoted [2]. The analysis of the environmental alterations that result from these changes and the development of new planning instruments, has caused that different disciplines, specifically the Artificial Intelligence (AI), approaches from a computer and mathematical point of view to give alternative solutions to this problem [3].

Numerous modeling tools have emerged in recent years. In the case of urban growth, the models based on cellular automata (CA) are the most widely used [4]. Regression models, artificial neural networks (ANNs), multi-criteria evaluation techniques (MCE), and still incipient, agent-based models (ABM) can also be found.

The CA based models are oriented fundamentally towards the representation of the attributes of a given geographic region

in a two-dimensional lattice, in which a neighborhood radius is defined and a certain rule of evolution is applied in order to define the behavior of the CA. With the use of these models it has been possible to generate territorial scenarios prospectively [5]. To generate these scenarios, a characterization of a CA is needed, this has different components, such as the size of the study area, maps of urban uses, map scales, neighborhood radius, evolution rules, slopes, and others geographical factors [6].

The developing of a CA based simulation tool to generate territorial scenarios prospectively in order to implement future simulation techniques, bring us to address some challenges. One of them was, the huge amount of mathematical operations needed in one single simulation, because the complexity of the algorithm to do such operations results to be exponential.

One key calculus in the whole simulation process is, the transition potentials (TPs) of each cell in the map, these TPs show the probability of a cell to change from one state to another. The amount of these TPs have a direct impact on the computation cost needed to perform the mathematical calculations.

To optimize these calculations, we enhanced sequential algorithms with parallelization strategies in order to maximize computational hardware. The library OpenMP (Open Multi-Processing), widely used in parallel programming, helps to implement a parallel strategy called fork-join. This allows to take advantage of hardware resources for the execution of processes in shared memory [7].

The present work aims to implement the fork-join strategy to speed up the necessary TPs calculations and to compare the results against the first sequential algorithm used in the simulation.

The base maps for the experiments were generated from the study area of Culiacan, México. Being the faster growing city in the State of Sinaloa, we plan to use the simulation tool to understand the dynamics of the urban changes and to forecast for planning urban development as a future work.

The remainder of this paper is structured as follows. All material and methods such as, the study area, digitation process, CA model and OpenMP are defined in Section II. Calculus of transition potentials for each pixel using the fork-join model are explained in Section III. Also proposed

implementations and experiments are analyzed. Finally, the study is concluded with future research directions.

## II. MATERIAL AND METHODS

### A. Study Area and its Digitization

The municipality of Culiacán is located in the central region of the State of Sinaloa (Fig. 1), forming part of the northwest of Mexico. The corresponding coordinates are:  $24^{\circ} 48'15''$  "N (north latitude) and  $107^{\circ} 25'52''$  "O (longitude west), with an altitude of 54 meters above sea level. The city of Culiacán concentrates 81% of the population of the municipality that in the last 20 years has registered a very significant territorial and demographic growth, according to the last census of the National Institute of Statistics and Geography (INEGI), with population of around 800,000 inhabitants. In 1980 the city had an urban area of 5,163 hectares, by 1990 it increased to 7,377 hectares and by 2001 there were 9,800 hectares. This growth occurred in a disorderly, that is, anarchic way under the protection of political leadership resulting in the city currently having more than 275 neighborhoods, most of them formed in common lands, ecological reserved areas, places without feasibility of utilities due to its topographic composition [8].

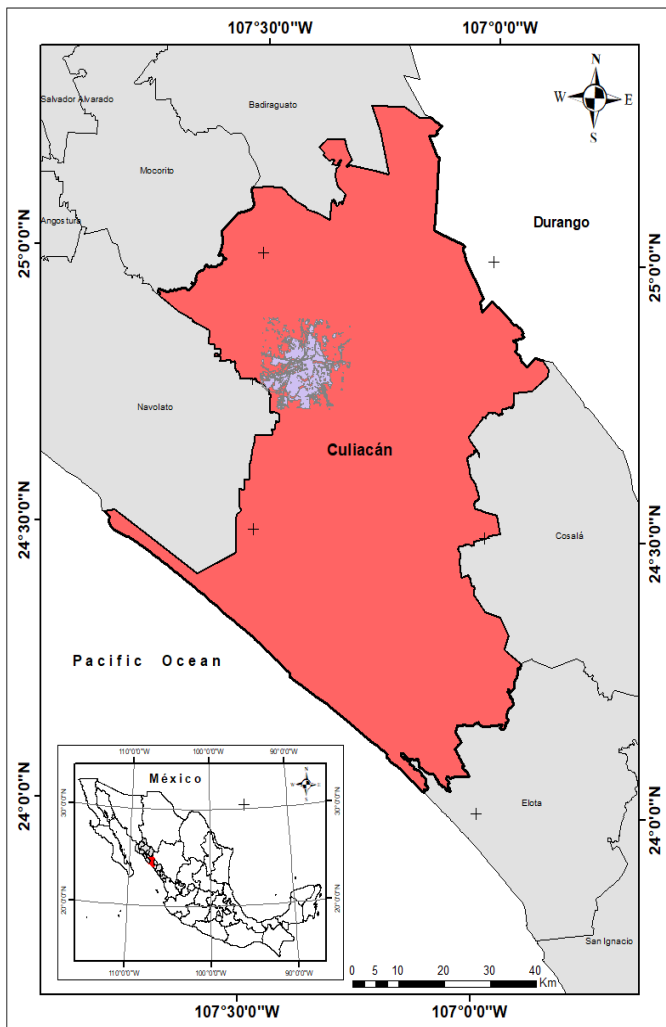


Fig. 1. Culiacán Sinaloa, México.

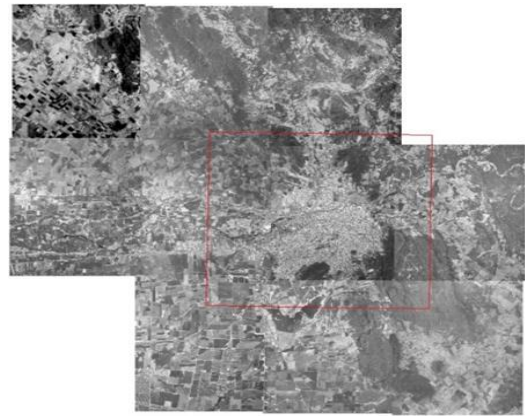


Fig. 2. Urban area of Culiacán 1997 and the study polygon (in red).

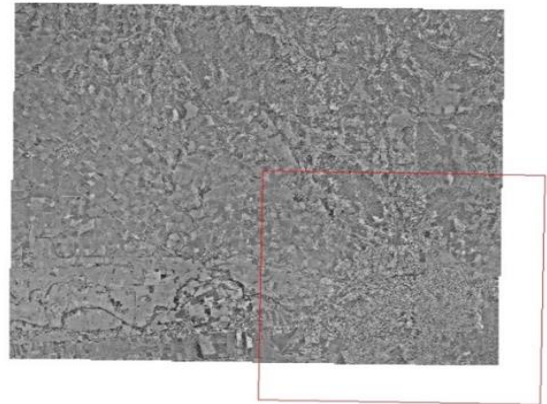


Fig. 3. Urban area of Culiacán 2004 and the study polygon (in red).

The digitization of the study area consisted in the generation of vector cartography over an orthophoto mosaic of the study area (Fig. 2 and 3). We worked with orthophotos (GeoTIFF) in the urban area of 1997 on a scale of 1: 20000, and in 2004 on a scale of 1: 10000, projected in WGS 84 / UTM 13N. The Geographic Information Systems (GIS) used were ArcMap® for vector maps, and IDRISI Selva ® for raster maps.

The digitization process (Fig. 4), urban land uses were classified in order to generate vector maps for each of them.

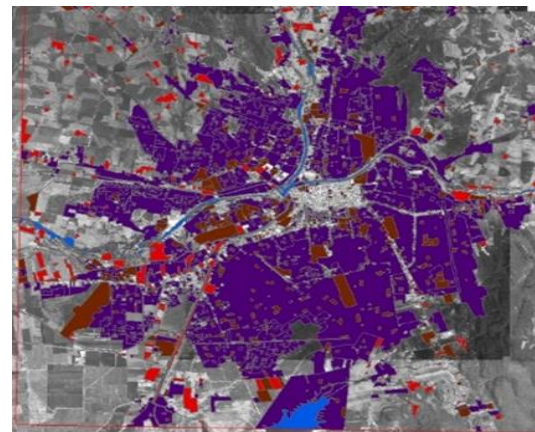


Fig. 4. Process of digitization of the urban area on the orthophoto of 1997.

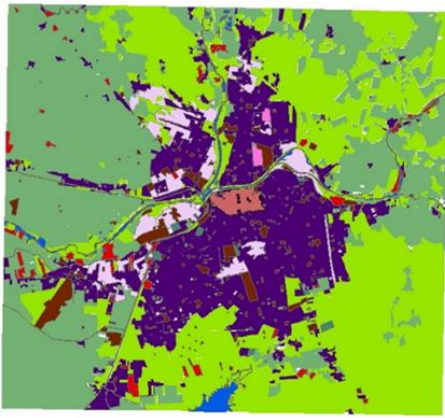


Fig. 5. Raster map 1997 Culiacan city.



Fig. 6. Raster map 2004 Culiacan city.

TABLE I. URBAN USE CLASSES

Urban Land Use	Value in Raster Map
Residential	1
Commerce	2
Industrial	3

All generated maps from the study area were converted from vector to raster format, with the option “to raster image” placed in the module IDRISI Database Workshop.

The resulting raster maps (Fig. 5 and 6) and the classification of urban uses (Table I). Raster maps are the input for the CA model.

To manage map files we used GDAL (Geospatial Data Abstraction Library). This is a library of free use for the reading and writing of geospatial data providing low-level functions that allow the manipulation of raster files.

### B. Cellular Automaton Model and TP

The fundamental idea in CA Models is that the state of a cell at any given time depends on the state of the cells within its neighborhood in the previous time step, based on a set of transition rules [9]. The CA model used in this investigation is the one proposed by R. White (2), is a constrained cellular

automata for high-resolution modelling of urban land-use dynamics [10].

As previously mentioned (Section I), the CA models are oriented towards the representation of the attributes of a given geographic region in a two-dimensional lattice, raster maps provides these data format to the CA.

A raster map can be represented formally by an array of real values. This matrix is represented as  $A = \{a_{ij}\}$  of order  $m \times n$  such that  $0 \leq i \leq m, 0 \leq j \leq n$  where each element  $A = [a_{ij}] \in \mathbb{R}$ .

A neighborhood filter matrix (1) is required to analyze each element  $A = [a_{ij}]$ , this neighborhood is formally represented  $B = \{b_{ii}\}$  of order  $n \times n$  such that  $0 \leq i \leq n$  where each element  $B = [b_{ii}] \in \mathbb{Z}$ .

$$B = \begin{bmatrix} b_{i-1,j-1} & b_{i-1,j} & b_{i-1,j+1} \\ b_{i,j-1} & b_{i,j} & b_{i,j+1} \\ b_{i+1,j-1} & b_{i+1,j} & b_{i+1,j+1} \end{bmatrix}_{3 \times 3 \text{ neighbour}} \quad (1)$$

The neighborhood filter is used to calculate the transition potential from state  $h$  to  $j$  for each element  $A = [a_{ij}]$ . The calculation methodology is detailed below:

$$P_{hj} = v s_j a_j (1 + \sum_{k,i,d} m_{kd} I_{id}) + H_j \quad (2)$$

Where,

$P_{hj}$ : Is the transition potential of state  $h$  to state  $j$ .

$v$  : Stochastic perturbation term.  $v = 1 + [-\ln(\text{random})]^x$ .

( $0 < \text{random} < 1$ ), and  $x$  allows you to adjust the size of the disturbance

$s_j$ : represents the suitability of the state of the cell.

$a_j$ : Euclidean distance from the cell to the nearest road.

$m_{kd}$ : Calibration matrix, contains the weights of each cell as a function of its state  $k$  and distance  $d$ .

$$I_{id} = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases}$$

$i$  is the index of the cell in the current neighborhood,  $k$

The transition potential  $P_{hj}$  of each cell  $A_{ij}$  is calculated only if the suitability of the objective state  $s_j > 0$ . That is, for each cell (pixel) in the map, its transition potential will be calculated except for those in which its suitability is equal to zero. For the neighborhood calculation, the calibration matrix  $m_{kd}$  gives each neighbor cell  $b_{ii}$  a weight based on its state and distance (subscript  $d$  formula 1) concerning the analyzed cell  $a_{ij}$ . The nearby neighbor cells will generally have a higher weight, positive values are taken for an attractive effect and negative for repulsive effect, these values tend to decrease as the distance increases between the analyzed cell and its neighbor, this is called Distance Decay Effect. When analyzing the neighbor cells, the  $I_{id}$  component helps to filter (multiplying by 1) cells with the same state.

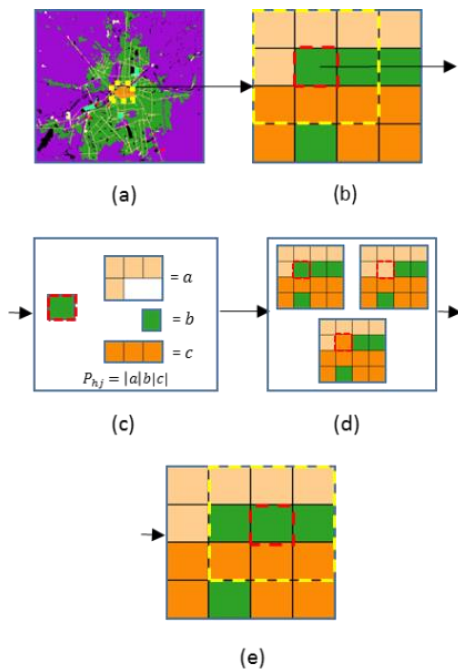


Fig. 7. Transition potential process.

Visually, in Fig. 7(a) we have the urban land use map, in Fig. 7(b), the neighbor is set to 3x3 around the analyzed cell. Fig. 7(c) calculates the transition potential  $P_{hj}$  of each cell from its current state  $h$  to a desired state  $j$ , the higher is selected. For this case we set as the higher to urban use. Fig. 7(d) analyzed cell change its value to the higher urban use. Fig. 7(e) shows observation window moves to the next cell.

An epoch has been completed when the last cell of the map is calculated. A simulation may require one or more epochs. If we take into account that this calculation must be done for each pixel of the map, we find a problem of computational complexity  $O(2^n)$ , this means, larger size of the input maps would increase the execution time of the simulation exponentially.

To handle this complexity, it was necessary to define a strategy to streamline the calculations, and this has been achieved with the development of programming modules in which parallel programming models are used by the OpenMP library.

### C. Openmp the Fork-Join Model

OpenMP is a shared memory application programming interface, provides functions to facilitate shared memory parallel programming, and it is intended to be suitable for implementation on SMP architectures, OpenMP is based on the fork-join model [11], [12].

Under fork-join model, a program starts with a single execution thread, this is named as the initial thread. When a parallel directive (`#pragma omp parallel`) is executed in a current thread, it will create a group of threads (fork) called, parallel region. In this region, every thread can collaborate with the other threads. At the end of the directive, the parallel region terminates (join), and the initial thread is the only which continues. Fork-join model shown in Fig. 8 [11].

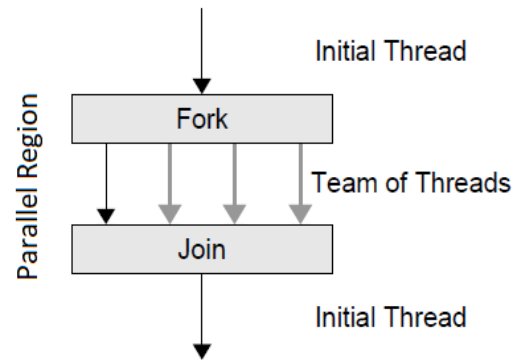


Fig. 8. Fork-join model.

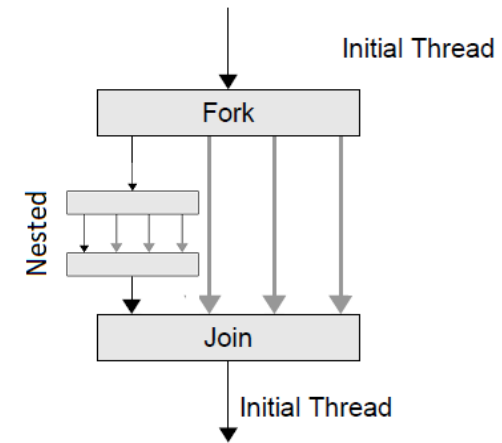


Fig. 9. Fork-join model with nested parallel region.

In addition, if required, OpenMP has the ability to create a parallel region inside another (nested), therefore, it is possible to divide a task as much as necessary and as much as the hardware allows it to, as shown in the Fig. 9.

### III. PROPOSED ALGORITHM AND EXPERIMENTS

To create an OpenMP program from a sequential one, we must first to identify sequence of instructions that may be executed concurrently by more than one processor [11].

We identified the calculus of transition potentials  $P_{hj}$ , as the portion of sequential code which can be parallelized in order to do the mathematical operations using more than one processors' core.

Fig. 10 shows a schematic of the implementation of the strategy to carry out the calculation of two urban uses. In the raster map, an observation window is defined, that window is analyzed in two cores, the transition potentials are calculated ( $x_1$  and  $x_2$ ), one per core, the higher is selected and assigned as a new value to the cell.

For  $n$  urban uses, the basic idea is, for each cell  $A_{ij}$  we must calculate their  $P_{hj}$  from the current state  $h$  to objective state  $j$  where  $j = n$  urban uses. Fig. 11 illustrates the process.

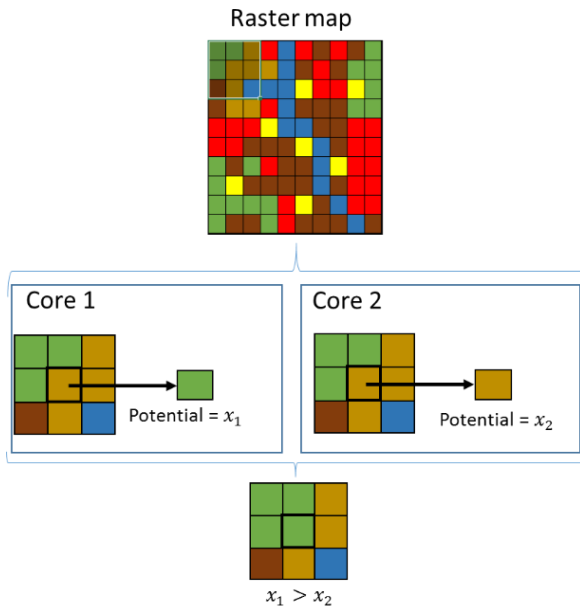


Fig. 10. Calculation of the transition potential (one per core) of a cell to two possible uses.

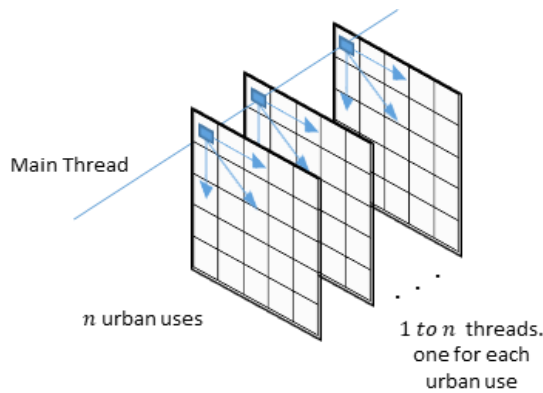


Fig. 11. Calculation of the transition potential for  $n$  urban uses.

```

1: 0 to epoch;
2: (0,0) to mapSize (nxm);
   {
3: pixel = cell(nxm);
4: nCores = nDinamicUses;
5: omp_set_num_threads(nCores);
6: #pragma omp parallel
   {
7:     j = omp_get_thread_num();
8:     vectorP [j]=calculateP
(pixel,j);
   }
9: newMap(nxm) = hPotential (vectorP);
   }

```

Fig. 12. Algorithm to calculate transition potentials.

The proposed algorithm (Fig. 12) to calculate the transition potentials is: 1) The epochs are defined. 2) The loop is set from the first cell to the last one. 3) The value of the current pixel is obtained. 4) The number of processor cores to be used is established (based on the number of dynamical uses). 5) The directive `omp_set_num_threads(nCores)` is used to establish the number of threads and the quantity of processor cores to use. 6) The parallel region `#pragma omp parallel` is initialized. 7) The thread number `j = omp_get_thread_num()` is identified. 8) The result of the potential of the analyzed cell is assigned to the potential vector, once the calculation has been completed in all cores. 9) The highest calculated potential is assigned to the analyzed cell.

The implementation was performed on an HP ProLiant ML350 G6 server, 12 GB RAM, 2 Intel Xeon E5645 processors (2.40 GHz), Linux Centos 6.9 operating system.

### A. Experiments

Three conditions were considered for the experiments: 1) resolutions of raster maps from the study area. 2) Number of epochs for each resolution. 3) Times from sequential and parallel algorithm.

Table II summarizes the maps used.

Times from these 3 sets of resolutions were measured using the sequential algorithm and the one with the fork-join model.

TABLE II. RASTER maps RESOLUTION

resolution		pixel size
cols	rows	(meters)
397	366	50
9,925	9,150	25
19,850	18,300	1

TABLE III. RUNNING TIMES FOR EACH RESOLUTION

pixel resolution		pixel size	Execution times (minutes)	
cols	rows	(meters)	Sequential	OpenMP
397	366	50	0.08	0.02
9,925	9,150	25	53.24	13.01
19,850	18,300	1	214.1	54.69

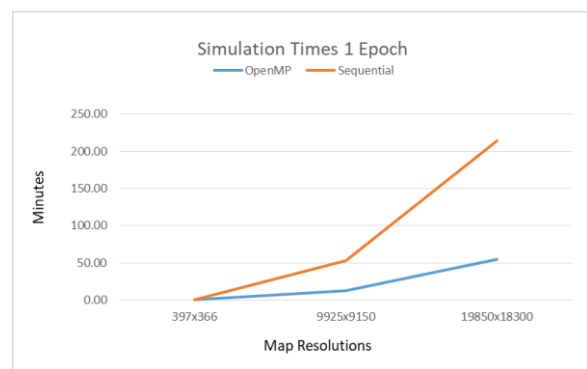


Fig. 13. Simulation times with different resolutions, 1 Epoch each.



TABLE IV. RUNNING TIMES FOR EACH RESOLUTION

pixel resolution		pixel size	Execution times (minutes)	
cols	rows	(meters)	Sequential	OpenMP
397	366	50	0.46	0.11
9,925	9,150	25	230.23	58.91
19,850	18,300	1	929.18	245.42

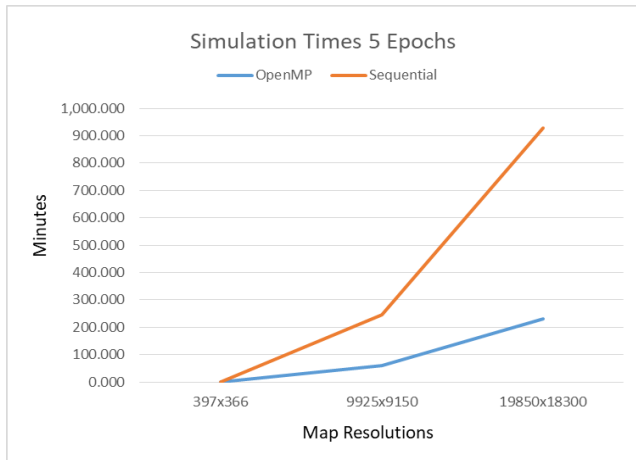


Fig. 14. Simulation times with different resolutions, 5 epochs each.

Table III shows the first results for 1 epoch, times are shown in minutes.

It is evident the correlation in the maps sizes and the running times, as shown in Fig. 13 and 14, simulation times grow as we incremented the map sizes.

Table IV shows the first results for 5 epoch, times are shown in minutes.

### B. Discussion

Execution times using sequential and parallel algorithms increase along with the maps size, but not linearly. As we expected, sequential method is the one that takes the most time to complete the calculations. OpenMP helped to reduce in almost 4 times the execution times.

At lower resolutions, there is no big difference due to the minimum execution times. Resolutions around  $500^2$  pixels should not represent a big challenge when working with a small number of urban uses, for our experiments we use 3.

Since most simulations require from 3 to 8 urban uses and maps sizes from  $1000^2$  to  $5000^2$  pixels, it is necessary to implement a strategy to enhance calculations in the development of this kind of tools.

The number of epochs is critical in this, depending on the kind and configuration, every simulation needs several epochs. As shown in Table IV, for map sizes lower than  $1000^2$  pixels, we expect times under 59 minutes in our future simulations.

## IV. CONCLUSION AND FUTURE WORK

OpenMP provides mechanisms that help to reduce execution times when implemented in a simulation model based on a cellular automaton obtaining improvements of up to 4 times.

Since numerous simulations must be performed in order to achieve different tasks such as calibrating the simulation model, perform sensitivity analysis or tests with different urban uses, OpenMP must be considered as a very interesting option when implementing algorithms for this area.

Future work: First, our CA based simulation tool became faster after the implementation of the parallelization strategy to calculate transition potentials, now we need to continue testing more resolutions and urban uses. Second, we are considering implementing CUDA along with artificial neural networks in order to improve the forecast of urban growth.

### REFERENCES

- [1] W. Plata-Rocha, M. Gómez-Delgado, y J. Bosque-Sendra, "Simulating urban growth scenarios using GIS and multicriteria analysis techniques: A case study of the Madrid region, Spain", *Environ. Plan. B Plan. Des.*, vol. 38, pp. 1012–1031, 2011.
- [2] C. G. Ralha, C. G. Abreu, C. G. C. Coelho, A. Zaghetto, B. Macchiavello, y R. B. Machado, "A multi-agent model system for land-use change simulation", *Environ. Model. Softw.*, vol. 42, pp. 30–46, 2013.
- [3] E. F. Lambin, B. L. Turner, H. J. Geist, S. B. Agbola, A. Angelsen, J. W. Bruce, O. T. Coomes, R. Dirzo, G. Fischer, C. Folke, P. S. George, K. Homewood, J. Imbernon, R. Leemans, X. Li, E. F. Moran, M. Mortimore, P. S. Ramakrishnan, J. F. Richards, H. Skånes, W. Steffen, G. D. Stone, U. Svedin, T. a. Veldkamp, C. Vogel, y J. Xu, "The causes of land-use and land-cover change: Moving beyond the myths", *Glob. Environ. Chang.*, vol. 11, pp. 261–269, 2001.
- [4] F. Aguilera Benavente, W. Plata Rocha, y J. Bosque Sendra, "Diseño y simulación de escenarios de demanda de suelo urbano en ámbitos metropolitanos", *Rev. Int. sostenibilidad, Tecnol. y humanismo*, pp. 57–80, 2009.
- [5] F. Aguilera Benavente, L. M. Valenzuela Montes, J. A. Soria Lara, M. Gómez Delgado, y W. Plata Rocha, "Escenarios Y Modelos De Simulación Como Instrumento En La Planificación Territorial Y Metropolitana", *Ser. Geográfica*, vol. 17, pp. 11–28, 2011.
- [6] R. White y G. Engelen, "High-resolution integrated modelling of the spatial dynamics of urban and regional systems", *Comput. Environ. Urban Syst.*, vol. 24, pp. 383–400, 2000.
- [7] R. Chandra, *Parallel Programming in OpenMP*. 2001.
- [8] J. A. Inzunza, "La Planeación Urbana en el Municipio Mexicano: Culiacán, Un Caso de Estudio", 2003.
- [9] J. I. Barredo y M. Gómez Delgado, "TOWARDS A SET OF IPCC SRES URBAN LAND-USE SCENARIOS: MODELLING URBAN LAND-USE IN THE MADRID REGION European Commission – DG Joint Research Centre Institute for Environment and Sustainability Department of Geography, University of Alcalá", pp. 1–16, 2000.
- [10] R. White, G. Engelen, y I. Ujje, "The use of constrained cellular automata for high-resolution modelling of urban land-use dynamics", *Environ. Plan. B Plan. Des.*, vol. 24, núm. 3, pp. 323–343, 1997.
- [11] B. Chapman, G. Jost, y R. Van Der Pas, *Using OpenMP*. The MIT Press, 2008.
- [12] T. Mattson y L. Meadows, "Introduction to OpenMP". [On Line]. Disponible en: <http://www.openmp.org/wp-content/uploads/omp-hands-on-SC08.pdf>.