

**NINTH INTERNATIONAL CONFERENCE ON  
ADVANCED COMPUTING (ICoAC 2017)**

**14-16 December 2017**



**Organized by  
DEPARTMENT OF COMPUTER TECHNOLOGY  
ANNA UNIVERSITY, CHENNAI**

**ABOUT ICoAC 2017**

**TOPICS OF INTEREST**

**ABOUT DEPARTMENT**

**PANELS**

**PROGRAMME SCHEDULE**

**KEYNOTE SPEECHES**

**SESSIONS DETAILS**

**PAPERS**

**SPONSORS**



# ABOUT ICoAC 2017



ICoAC 2017 is an international conference in the field of Computer Science and Communication, focusing to address issues and developments in advanced computing. This conference seeks to bring together international researchers to present papers and generate discussions in recent trends and developments of computing. The conference will feature a range of presentations on latest research activities as well as stimulating talks and keynote addresses. It is organized by the Department of Computer Technology from 14<sup>th</sup> -16<sup>th</sup> December 2017.

The conference features, stimulating keynote talks, peer-reviewed technical paper presentation with short papers and posters, student paper presentation.





# TOPICS OF INTEREST

---



## SOCIAL COMPUTING

- **Virtualization and Visualization**
- **Green IT**
- **Cluster Computing**
- **E-Commerce and E-Governance**
- **Natural Language Processing**
- **Sentiment Analysis**
- **Social Network Mining**
- **Semantic Web**

## HIGH PERFORMANCE COMPUTING

- **Pattern Recognition**
- **Image Processing**
- **Intelligent Agents**
- **Machine Learning**
- **Soft Computing**
- **Multicore Programming**
- **Compiler Optimization**
- **Cloud and Fog Computing**
- **Data Mining and Data Analytics**





# TOPICS OF INTEREST

---



## NETWORK SCIENCE

- Ad hoc & Sensor Networks
- Body Area Network
- Network Security and Management
- Neural Networks
- Vehicular Communication
- Next Generation Networks
- Mobile Computing

## SMART COMPUTING

- Internet of Things
- Ambient Intelligence
- Agent Based Systems
- Virtual-Learning
- Access and Home Networks
- Ubiquitous Computing
- Evolutionary Computing





# ABOUT DEPARTMENT



The Department of Computer Technology at the MIT campus was formed by bifurcating the Department of Information Technology in the year of 2010 and offers Computer Science and Engineering programmes since 2001. This department promotes the pursuit of excellence in the field of Computer Science and Engineering. Ministry of Science and Technology sanctioned FIST program for this department in 2014. The CARE centre in the department is sponsored by Ministry of Communication and Information Technology (MCIT), Government of India. Next Generation Networks Lab (NGNLabs) was established with DST-SERB sponsor. The department has established research laboratories in the areas of Big Data Analysis, Data Science, Grid Computing, Internet of Things, Media as a Service (MaaS), and Vision Intelligence.





# ICoAC 2017 CHAIRS

---



Patron

**Dr. Geetha T.V.**, *Convener Committee, Anna University, Chennai, India.*

Chairman

**Dr. Ganesan S.**, *Registrar, Anna University, Chennai, India.*

Co Chairman

**Dr. Rajadurai A.**, *Dean, MIT Campus, Anna University, Chennai, India.*

General Chair

**Dr. Thamarai Selvi S.**, *Director, CTD, Anna University, Chennai, India.*

Head of the Department

**Dr. Anandhakumar P.**, *Department of Computer Technology, Anna University, Chennai, India.*

Convener

**Dr. Gunasekaran R.**, *Department of Computer Technology, Anna University, Chennai, India.*

---





# ADVISORY COMMITTEE

---

Ali Kashif Bashir, University of Faroe Islands, Denmark  
Babak Moazzez, Kennesaw State University, USA  
Dipak Ghosal, University of California, USA  
Geoffrey Charles Fox, Indiana University, USA  
Gopal A, CSIR, CEERI, USA  
Jeny Rajan, NIT, Surathkal  
John Ebenezer Augustine, IIT Madras, Chennai  
Kanmani S, Pondicherry Engineering College  
Kannan G, Indiana University, USA  
Kinshuk, University of North Texas, USA  
Kishore Kumar, NIT, Warangal  
Leela Velusamy, NIT, Trichy  
Mateen M. Rizki, Wright state university, USA  
Narayanasamy P, PSG, Coimbatore  
Nicola Marchetti, Trinity College Dublin, Ireland  
Nilavalan Rajagopal, Brunel University, UK  
Nirmala Shenoy, Rochester Institute of Technology, USA  
Palanisamy P, NIT, Trichy

Purusothaman T, GCT, Coimbatore  
Ramanujam R, IMSc, Chennai  
Rajkumar Buyya, The University of Melbourne, Australia  
Raj K Bhatnagar, University of Cincinnati, USA  
Ravi Sankar, University of South Florida, USA  
Ruppa K Thulasiram, University of Manitoba, Canada  
Sasikumar M, CDAC, Pune  
Srinivasan Narayanamurthy, NetApp India Pvt. Ltd, Bangalore  
Suvra Sekhar Das, IIT, Kharagpur  
Sricharan M S, Wipro Technologies, Chennai  
Sridhar Radhakrishnan, University of Oklahoma, USA  
Thavasi Raja, NIT, Trichy  
Vaidehi Vijayakumar, VIT University, Chennai  
Vincenzo Piuri, University of Milan, Italy  
Vivekanandan Kumar, Athabasca University, Canada  
Waleed Ejaz, Ryerson University, Canada  
Xavier Fernando, Ryerson University, Canada



# KEYNOTE SPEECHES

---

## ❖ New Frontiers in Cloud Computing for Big Data and IoT Applications

Dr. Rajkumar Buyya, University of Melbourne, Australia

## ❖ Challenges in Stochastic Modeling for Telecommunication Systems

Dr. S. Dharmaraja, IIT-Delhi

## ❖ Machine Vision System for Industrial Applications

Dr. A. Gopal, CSIR-CEERI, Chennai

## ❖ Applications of Graph Theory in Image Processing

Dr. V. Masilamani, IIITDM, Chennai

---







# KEYNOTE SPEECHES

---

## ❖ Waveforms for Next Generation Wireless Communication Systems

Dr. Suvra Sekhar Das, IIT, Khargpur

## ❖ Think Like a Vertex

Dr. John Ebenezer, IIT, Madras

---





# SESSION DETAILS

---

- SESSION 3A      NETWORK SECURITY / MANET
  - SESSION 4A      IMAGE PROCESSING / NETWORKS
  - SESSION 5A      CLOUD COMPUTING / GRID COMPUTING / WSN-I
  - SESSION 6A      COMPUTER ARCHITECTURE / NETWORK SECURITY
- 





## SESSION 3A – NETWORK SECURITY / MANET

**[57] Application of NLTK for Voice Based Sentiment Analysis**

*Pankaj Keserwani, Ashish Mishra and Shefalika Ghosh Samaddar*

**[69] Extracting the features of Medicinal Plants and improving the accuracy of classification using Fuzzy Local Binary Pattern Approach**

*Sriniva M*

**[107] Comparison of Machine Learning Algorithms- Opinion Mining for Social Network Data**

*Parkavi R*

**[149] An Empirical Study of Deep CNN Models towards Semantic Segmentation**

*Arunekumar Balasubramanian and Suresh Joseph*

**[289] A Review of Hierarchical Fuzzy Text Clustering**

*Seema Wazarkar, Bettahally Keshavamurthy and Amrita Manjrekar*

**[319] Fuzzy-knowledge-inferred Edge Directed Image Enhancement Technique**

*Reshmalakshmi Chandrasekharan and Sasikumar Madhavan Nair*

**[362] Design of Automated Data Extraction System for Medical Web Forums using Semantic Analysis**

*Umamageswari Kumaresan and Kalpana Ramanujam*





# SESSION 5A – CLOUD COMPUTING / GRID COMPUTING / WSN-I

**[106] Energy Efficient MAC Protocol for Body Centric NANO-Networks (BANNET)**  
*Sivapriya S and Sridharan D*

**[250] Anomaly Based Intrusion Detection System using Classifiers**  
*Aarathi M and NarasimaMallikarjunan K*

**[282] Optimized Channel Awareness Routing For Congestion Avoidance by Dynamic Queue Space Management in MANET**  
*R.Sundar Dass, A.Kathirvel Ayyaswamy and S.Narayanan Sakthi*

**[321] IoT-Based Charging Station for Electric Vehicles**  
*Chellaswamy C, Archana V, Arunraj J and Bhagirathi S*

**[358] Automated Trust Evaluation Mechanism for Future Generation Internet**  
*Umamaheswari S and Arun Fera M*

**[77] A new Heuristic Similarity Model to improve the accuracy for data sparsity issues in Collaborative Filtering Recommendation Algorithms**  
*Mohana H and Suriakala M*

**[279] Reliability Aware Self-Test Schedule for Aging Systems**  
*Harini Sriraman and Pattabiraman Venkatasubbu*





# SESSION 6A – COMPUTER ARCHITECTURE / NETWORK SECURITY

- [5] **Monitoring as a Service (MaaS) and its usage in Containers across Multicloud**  
*Baskaran Jambunathan and Kalpana Yoganathan*
- [6] **Big Biomedical Data Engineering**  
*Ripon Patgiri, Sabuzima Nayak and Samir Kumar Borgohain*
- [84] **Convergent Scheduling with EDF and BackFill Algorithm to improve the Performance of the Grid Scheduler**  
*Kalyani Vaidyanathan*
- [324] **Impact of performance analysis of varied subjects on overall result: An empirical discourse of educational data mining**  
*Mudasir Bhat, Majid Baba and Muheet Butt*
- [225] **Optimized Feature Selection in Dimensionality Reduction for Big data with Random Forest Algorithm**  
*Vinayagasundaram B, Swathy R, Roshini C, Sarshini D and Senthil kumar G*
- [80] **Design and Development of a System for The Impact Analysis of Internet usage among Engineering Students**  
*Rajalaxmi R.R., Natesan P, Krishnamoorthy N and Ponni S*



**[205] Dynamic Remote Data Auditing using Privacy Preserving Auditing Protocol in cloud Environment**

*Raja, Ramakrishnan, Hariharan S, Ramprasath M and Arunkumar G*



## Programme Schedule ICoAC 2017

Time	Thursday, 14 <sup>th</sup> December, 2017		
08.30am - 09.00am	Conference Registration, CB Block, Department of Computer Technology		
09.00am - 10.30am	Keynote Address on “New Frontiers in Cloud Computing for Big Data and Internet-of-Things (IoT) Applications” <b>Dr. Rajkumar Buyya</b> , <i>Professor, Computing and Information Systems, University of Melbourne, Australia</i>		
10.30am - 11.00am	TEA BREAK		
<b>Session 1</b> Conference Hall CB Block 11.00am - 01.00pm	Oral Paper Presentations on <b>Cloud Computing and Social Engineering</b>		
	Session Chairs <b>Dr. Dhananjay Kumar, Dr. Y. Nancy Jane</b>		
	Session Coordinator <b>Mr. K. Murugan</b>		
	Hall Management <b>Mr. S. Gowrishankar</b>		
	<b>S.No</b>	<b>ID</b>	<b>TITLE</b>
	1	290	<b>Enhanced Holter Monitoring in Adaptive Environment with HIVEing</b> <i>Chandra Priya J and Perinba Jothi T</i>
	2	197	<b>Load Balancing in Cloud Environment using Stackelberg’s approach</b> <i>Vinayagasundaram B and Swathy R</i>
	3	121	<b>#BigBoss - Twitter Driven Temporal Peak Recognition in Long Run Events</b> <i>Karthika S, Parvathi A. V and Bose S</i>
	4	242	<b>A Social Network Analysis of Football - Evaluating Player and Team Performance</b> <i>Balasubramaniam Srinivasan</i>
	5	172	<b>Federated Cloud Services using Virtual API Proxy Layer in a Distributed Cloud Environment</b> <i>Shreyas M M</i>
	6	163	<b>An Efficient Ciphertext Policy-Attribute Based Encryption for Big Data Access Control in Cloud Computing</b> <i>P. Praveen Kumar, P. Syam Kumar and P.J.A. Alphonse</i>
	7	263	<b>An Comparative Study on Attribute Based Encryption Schemes for Secure Cloud Data Outsourcing</b> <i>Thangavel M, Varalakshmi P and Abinaya C</i>
	8	175	<b>Gender Classification of Blog Authors: With Feature Engineering and Deep Learning using LSTM Networks</b> <i>Vijay Prakash Dwivedi, Saurav Jha, Deepak Kumar Singh and Ranvijay</i>
	9	159	<b>Finding Instantaneous Community of Ideologically Similar Users in Social Forums</b> <i>Jitendra Kumar, Kumari Roshni VS</i>
10	179	<b>Task Scheduling in Big Data - Review, Research Challenges, and Prospects</b> <i>Kannan Govindarajan, Supun Kamburugamuve, Pulasthi Wickramasinghe, Vibhatha Abeykoon and Geoffrey Fox</i>	
11	115	<b>A Combinatorial Optimization Algorithm for Load Balancing in Cloud Infrastructure</b> <i>Kannan Govindarajan and Thamarai Selvi Somasundaram</i>	
12	83	<b>LVC MOS based Green Data Flip Flop Design on FPGA</b> <i>Amanpreet Kaur, Gunjan Gupta and Bishwajeet Pandey</i>	
01.00pm - 02.00pm	LUNCH BREAK		
02.00pm - 03.15pm	Keynote Address on “Challenges in Stochastic Modeling for Telecommunication Systems” <b>Dr. S. Dharmaraja</b> , <i>Professor, Department of Mathematics, IIT Delhi</i>		
03.15pm - 03.30pm	TEA BREAK		
<b>Session 2</b> Conference Hall CB Block 03.30pm - 05.15pm	Oral Paper Presentations on <b>Network Engineering</b>		
	Session Chairs <b>Dr. R. Kathirolu, Dr. D. Sangeetha</b>		
	Session Coordinator <b>Dr. Fouzul Hidayat</b>		
	Hall Management <b>Ms. M. Kavitha</b>		
	<b>S.No</b>	<b>ID</b>	<b>TITLE</b>
	1	271	<b>An Enhanced Ant Colony Optimization Algorithm for Vehicle Routing Problem with Time Windows</b> <i>Ashima Gupta and Sanjay Saini</i>
	2	237	<b>Top-N Recommendation using Bi-Level Collaborative Filtering</b> <i>Suman Banerjee, Pratik Banjare, Mamata Jenamani and Dilip Kumar Pratihari</i>
	3	334	<b>Wavelength assignment and adaptive shortest path algorithm in cognitive radio networks using ant colony optimization</b> <i>D Arivudainambi and S Mangairkarasi</i>
	4	292	<b>Detection of Control Layer DDoS Attack using Entropy Metrics in SDN: An Empirical Investigation</b> <i>Kshira Sagar Sahoo, Manikanta Vankayala, Bibhudatta Sahoo, and Ratnakar Dash</i>
	5	64	<b>Track Health Care using Location Based QoS Web Service Recommendation System</b> <i>Justin Dhas Y and Jeyanthi P</i>
	6	368	<b>Link Failure Detection and Alternate Path Tracing in OpenFlow Based Ethernet Networks</b> <i>Muthumanikandan Vanamoorthy, Valliyammai Chinnaiah and Harish Sekar</i>
	7	294	<b>An Anomaly Behavior based Detection and Prevention of DoS Attack in IoT Environment</b> <i>Santhosh Kumar S and Kulothungan K</i>
8	89	<b>Blind Interference Alignment Simulation over Rayleigh fading channel</b> <i>Vepuri Prasanna Vani</i>	
9	375	<b>Congestion Control in 6LoWPAN Networks using Fuzzy Logic (FLCC)</b> <i>Rajesh G, Swetha Chandrasekar, Priyanka Radhakrishnan and Vaishnavi Ramamurthy</i>	
10	326	<b>ARRDA: Adaptive Reliable Routing for QoS based Data Aggregation in Wireless Sensor Networks</b> <i>Gomathy Prathima E, Venugopal K R, Sundaraja Sitharama Iyengar and Lalit Mohan Patnaik</i>	



Time		Friday, 15 <sup>th</sup> December, 2017	
08.30am - 09.00am	Conference Registration, CB Block, Department of Computer Technology		
09.00am - 10.30am	Keynote Address on “Machine Vision System for Industrial Applications” <b>Dr. A. Gopal</b> , Chief Scientist & Scientist-in-charge, CSIR-CEERI, Chennai		
10.30am - 11.00am	TEA BREAK		
<b>Session 3</b> <b>Conference Hall</b> <b>CB Block</b> <b>11.00am - 01.00pm</b>	Oral Paper Presentations on <b>Learning Analytics and Databases</b> Session Chairs <b>Dr. M.R. Sumalatha , Dr. S. Muthurajkumar</b> Session Coordinator <b>Dr. O. S. Gnanaprakasi</b> Hall Management <b>Ms. R. Vanaja</b>		
	S.No	ID	TITLE
	1	328	<b>A Multi-objective approach for DG and Capacitor placement using Harmony Search Algorithm</b> <i>Venkatesh Kona and Dr. Ravindra Kollu</i>
	2	181	<b>Blood pressure prediction based on Pattern Classification</b> <i>Yue Zhang and Feng Zhimeng</i>
	3	152	<b>Multi-Agent based Artificial War</b> <i>Ajitha Santhakumari, Ananya Datta and Suresh Kumar T.V</i>
	4	355	<b>Performance Analysis of Classification Approaches for the Prediction of Type II Diabetes</b> <i>Durgadevi Mullaivanan, R.Kalpna</i>
	5	145	<b>Computer Assisted System for Predicting Human Behavior using Time Delay Neural Networks</b> <i>Nancy Jane Y</i>
	6	354	<b>Distributed and Parallel Processing of Location based spatial query with Approximate Transformation</b> <i>Priya M and.Kalpna R</i>
	7	366	<b>Feature Constrained Parallel Data Processing Approach for Spatiotemporal Event Detection</b> <i>Bhuvaneshwari A, Valliyammai C and Devakunchari R</i>
	8	35	<b>A Comparative Study of Machine Learning Methods for Generation of Digital Forensic Validated Data</b> <i>Nandan Kumar, Pankaj Keserwani and Shefalika Samaddar</i>
9	9	<b>A Framework for Quality Enhancement of Multispectral Remote Sensing Images</b> <i>Shilpa Suresh, Devikalyan Das and Shyam Lal</i>	
<b>Session 3A</b> <b>Vision Intelligence</b> <b>Lab</b> <b>CB Block</b> <b>11.00am - 01.00pm</b>	Oral Paper Presentations on <b>Network Security / MANET</b> Session Chairs <b>Dr. B. Lydia Elizabeth, Dr. S. Umamaheswari</b> Session Coordinator <b>Mr. T. Sudhakar</b> Hall Management <b>Ms. K. S. Ezhilaisai</b>		
	S.No	ID	TITLE
	1	57	<b>Application of NLTK for Voice Based Sentiment Analysis</b> <i>Pankaj Keserwani, Ashish Mishra and Shefalika Ghosh Samaddar</i>
	2	69	<b>Extracting the features of Medicinal Plants and improving the accuracy of classification using Fuzzy Local Binary Pattern Approach</b> <i>M Srinivas</i>
	3	107	<b>Comparison of Machine Learning Algorithms- Opinion Mining for Social Network Data</b> <i>Parkavi R</i>
	4	149	<b>An Empirical Study of Deep CNN Models towards Semantic Segmentation</b> <i>Arumekumar Balasubramanian and Suresh Joseph</i>
	5	289	<b>A Review of Hierarchical Fuzzy Text Clustering</b> <i>Seema Wazarkar, Bettahally Keshavamurthy and Amrita Manjrekar</i>
	6	319	<b>Fuzzy-knowledge-inferred Edge Directed Image Enhancement Technique</b> <i>Reshmalakshmi Chandrasekharan and Sasikumar Madhavan Nair</i>
7	362	<b>Design of Automated Data Extraction System for Medical Web Forums using Semantic Analysis</b> <i>Umamageswari Kumaresan and Kalpana Ramanujam</i>	
01.00pm - 02.00pm	LUNCH BREAK		
02.00pm - 03.15pm	Keynote Address on “Applications of Graph Theory in Image Processing “ <b>Dr. V. Masilamani</b> , IITDM, Chennai		
03.15pm - 03.30pm	TEA BREAK		
<b>Session 4</b> <b>Conference Hall</b> <b>CB Block</b> <b>03.30pm - 05.15pm</b>	Oral Paper Presentations on <b>Smart Computing</b> Session Chairs <b>Dr. P.T.V. Bhuvaneshwari, Dr. Ponsy R. K. Sathiabham</b> Session Coordinator <b>Ms. J. Chandra Priya</b> Hall Management <b>Ms. S. Arpana</b>		
	S.No	ID	TITLE
	1	184	<b>Fixed head Short-term Hydrothermal Scheduling using Whale Optimization algorithm considering the Uncertainty of Solar Power</b> <i>Sujoy Das, Aniruddha Bhattacharya and Ajoy Kumar Chakraborty, Vibhav Pandey</i>
	2	4	<b>Building and Tree Detection by Fusing LiDAR and Aerial Images for Urban Development Planning</b> <i>Pavan Malbhage and Suchitra Khoje</i>
	3	162	<b>An SLA Design with Digital Forensic Capabilities</b> <i>Pankaj Keserwani and Shefalika Ghosh Samaddar</i>
4	254	<b>Intuitionistic Fuzzy Id3: An Approach to E-Transactional Fraud Detection</b> <i>Sikdar Md Sultan Askari and Md Anwar Hussain</i>	

	5	223	<b>Behaviour Profiling of reactions in Facebook posts for Anomaly Detection</b> <i>Savyan P V and Mary Saira Bhanu S</i>
	6	217	<b>Adaptive Neuro-Fuzzy Inference System for Assessing the Maintainability of the Software</b> <i>Therasa P.R and Vivekanandan P</i>
	7	51	<b>Top-k Similar Access Behaviour based on Structural Equivalence</b> <i>T Ramalingeswara Rao, Pabitra Mitra and A Goswami</i>
	8	374	<b>Performance Evaluation of Classifiers for Analysis of Road Accidents</b> <i>Sugetha C, Karunya L, Prabhavathi E and Kola Sujatha P</i>
	9	372	<b>Detection of Vulnerabilities Caused by WebView Exploitation in Smartphone</b> <i>Fouzul Hidhaya S. and Angelina Geetha</i>
<b>Session 4A</b> <b>Vision Intelligence</b> <b>Lab</b> <b>CB Block</b> <b>03.30pm - 05.15pm</b>	Oral Paper Presentations on <b>Image Processing / Networks</b> Session Chairs <b>Dr. P. Varalakshmi, Dr. J. Dhalia Sweetlin</b> Session Coordinator <b>Ms. Cinu C. Kiliroor</b> Hall Management <b>Mr. S. Gowrishankar</b>		
	<b>S.No</b>	<b>ID</b>	<b>TITLE</b>
	1	148	<b>Sensitive Data Protection in Cloud-Based on Modified Elliptic Curve Cryptographic Technique</b> <i>Sumathi Muruganandam and Sangeetha S</i>
	2	171	<b>An Analysis of Multifarious Character Recognition</b> <i>M Sornam and M Poornima Devi</i>
	3	211	<b>A Novel Approach to Optimized Hybrid Item-based Collaborative Filtering Recommendation Model using R</b> <i>Abhaya Kumar Sahoo and Chittaranjan Pradhan</i>
	4	204	<b>LangTool: Identification of Indian Languages for Short Text</b> <i>Sreebha Bhaskaran, Geetika Paul, Deepa Gupta and Amudha J</i>
	5	214	<b>Urban Slum Extraction using GLCM Based Statistical Approach from Very High Resolution Satellite Data</b> <i>Prabhu R, Umasree S, Alagu Raja R.A. and Avudaiammal R</i>
	6	243	<b>Analysis and Estimation of Brain Tissue Atrophy using Magnetic Resonance Images</b> <i>N Ahana Priyanka and G Kavitha</i>
	7	246	<b>A Multifamily Android Malware Detection using Deep Autoencoder Based Feature Extraction</b> <i>Teenu S John, Tony Thomas and Md. Meraj Uddin</i>
<b>Time</b>	<b>Saturday, 16<sup>th</sup> December, 2017</b>		
<b>08.30am - 09.00am</b>	<b>Conference Registration, CB Block, Department of Computer Technology</b>		
<b>09.00am - 10.30am</b>	Keynote Address on <b>“Waveforms for Next Generation Wireless Communication Systems”</b> <b>Dr. Suvra Sekhar Das, Associate Professor, G S Sanyal School of Telecommunications, IIT Kharagpur</b>		
<b>10.30am - 11.00am</b>	<b>TEA BREAK</b>		
<b>Session 5</b> <b>Conference Hall</b> <b>CB Block</b> <b>11.00am - 01.00pm</b>	Oral Paper Presentations on <b>Future Generation Networks</b> Session Chairs <b>Dr. B. Vinayagasundaram, Dr. P. Pabitha</b> Session Coordinator <b>Mr. R. Nandhakumar</b> Hall Management <b>Ms. M. Kavitha</b>		
	<b>S.No</b>	<b>ID</b>	<b>TITLE</b>
	1	178	<b>A Novel Gamification Approach to Recommendation Based Mobile Applications</b> <i>Neeraj S, Oswald C and Sivaselvan B</i>
	2	306	<b>Enhanced Negative Selection Algorithm for Malicious Node Detection in MANET</b> <i>Kathiroli Raja and Indira Natarajan</i>
	3	318	<b>A Lightweight Incognito Key Exchange Mechanism for LTE-A Assisted D2D Communication</b> <i>Sheeba Backia Mary Baskaran and Gunasekaran Raja</i>
	4	142	<b>Path Planning for Mobile Sink in Wireless Sensor Networks</b> <i>Mayur Shrirame and S Mini</i>
	5	228	<b>Intrusion Resilient Concealed Data Aggregation in Wireless Sensor Networks</b> <i>Maivizhi Radhakrishnan and Yogesh Palanichamy</i>
	6	151	<b>NLMS/F Based Adaptive Beam former for Indoor Wireless Channel</b> <i>Basabadatta Mohanty, Harish Kumar Sahoo and Bijayananda Patnaik</i>
	7	177	<b>Pyramidal-Based Connected Dominating Set for Tactical and Energy Harvest Networks</b> <i>Ceronmani Sharmila and George Amalanathan</i>
	8	244	<b>Adaptive Joint Call Admission Control Scheme for Heterogeneous Wireless Networks</b> <i>Anupam Gautam and Selvamuthu Dharmaraja</i>
	9	377	<b>Access Log Anomaly Detection</b> <i>Ma. Tharshini, M. Ragavinodini and Radha Senthil Kumar</i>
<b>Session 5A</b> <b>Vision Intelligence</b> <b>Lab</b> <b>CB Block</b> <b>11.00am - 01.00pm</b>	Oral Paper Presentations on <b>Cloud Computing / Grid Computing / WSN-1</b> Session Chairs <b>Dr. C. Valliyammai, Dr. P. Kola Sujatha</b> Session Coordinator <b>Ms. A. Bhuvaneswari</b> Hall Management <b>Ms. R. Vanaja</b>		
	<b>S.No</b>	<b>ID</b>	<b>TITLE</b>
	1	106	<b>Energy Efficient MAC Protocol for Body Centric NANO-Networks (BANNET)</b> <i>Sivapriya S and Sridharan D</i>
	2	250	<b>Anomaly Based Intrusion Detection System using Classifiers</b> <i>Aarthi M and Narasimamallikarjunan K</i>

	3	282	<b>Optimized Channel Awareness Routing for Congestion Avoidance by Dynamic Queue Space Management in MANET</b> <i>R.Sundar Dass, A.Kathirvel Ayyaswamy and S.Narayanan Sakthi</i>
	4	321	<b>IoT-Based Charging Station for Electric Vehicles</b> <i>Chellaswamy C, Archana V, Arunraj J and Bhagirathi S</i>
	5	358	<b>Automated Trust Evaluation Mechanism for Future Generation Internet</b> <i>Umamaheswari S and Arun Fera M</i>
	6	77	<b>A new Heuristic Similarity Model to improve the accuracy for data sparsity issues in Collaborative Filtering Recommendation Algorithms</b> <i>Mohana H and Suriakala M</i>
	7	279	<b>Reliability Aware Self-Test Schedule for Aging Systems</b> <i>Harini Sriraman and Pattabiraman Venkatasubbu</i>
<b>01.00pm - 02.00pm</b>	<b>LUNCH BREAK</b>		
<b>02.00pm - 03.15pm</b>	Keynote Address on “Think Like a Vertex” <b>Dr. John Ebenezer</b> , Associate Professor, Department of Computer Science, IIT Madras		
<b>Session 6 Conference Hall CB Block 03.15pm - 05.15pm</b>	Oral Paper Presentations on <b>Image processing and Applications</b> Session Chairs <b>Dr. P. Jayashree, Dr. G. Rajesh</b> Session Coordinator <b>Mr. V. Muthumanikandan</b> Hall Management <b>Ms. K. S. Ezhilaisai</b>		
	S.No	ID	TITLE
	1	109	<b>Accelerated Digitally Reconstructed Radiograph generation scheme for 2D to 3D Image Registration of Vertebrae based on Sparse Sampling and Multi-Resolution</b> <i>Vidya Bhat, Shyamasunder Bhat N and Anitha H</i>
	2	195	<b>Colon Cancer Detection in Biopsy Images for Indian Population at Different Magnification Factors using Texture Features</b> <i>Tina Babu, Tripty Singh, Deepa Gupta and Shahin Hameed</i>
	3	221	<b>Recursively seeded Image Encryption Algorithm</b> <i>Vidhya Ramamoorthy and Brindha Murugan</i>
	4	155	<b>Simulation Study on Tokamak Relevant Visual Servoing System</b> <i>Pramit Dutta, Amit Kumar Srivastava, Naveen Rastogi and Krishan Kumar Gotewal</i>
	5	185	<b>Modified Level Cut Liver Segmentation from CT</b> <i>Rekha R Nair, Tripty Singh, Ravi Nayar and Shiv Kumar</i>
	6	327	<b>Double Density Wavelet with Fast Bilateral Filter Based Image Denoising for WMSN</b> <i>Rekha Haridoss and Samundiswary Punniakodi</i>
	7	58	<b>Machine Intelligence Prospective for Large Scale Video based Visual Activities Analysis</b> <i>Naresh Kumar</i>
	8	169	<b>Suitability Analysis of Fractal Compression Technique for Medical Images</b> <i>Panjavarnam B and P.T.V. Bhuvaneshwari Mohan</i>
	9	146	<b>An Amalgam Approach to Detect Edges using Ultrametric Contour Map in Natural Scene Images</b> <i>Soma Debnath and Suvamoy Changder</i>
	10	167	<b>A Survey on Image Classification and Activity Recognition using Deep Convolutional Neural Network Architecture</b> <i>M Sornam, Muthusubash Kavitha and V Vanitha</i>
	<b>Session 6A Vision Intelligence Lab CB Block 03.15pm - 05.15pm</b>	Oral Paper Presentations on <b>Computer Architecture / Network Security</b> Session Chairs <b>Dr. S. Neelavathy Pari, Dr. V. P. Jayachitra</b> Session Coordinator <b>Mr. P. Ramesh</b> Hall Management <b>Ms. S. Arpana</b>	
S.No		ID	TITLE
1		5	<b>Monitoring as a Service (MaaS) and its usage in Containers across Multicloud</b> <i>Baskaran Jambunathan and Kalpana Yoganathan</i>
2		6	<b>Big Biomedical Data Engineering</b> <i>Ripon Patgiri, Sabuzima Nayak and Samir Kumar Borgohain</i>
3		84	<b>Convergent Scheduling with EDF and BackFill Algorithm to improve the Performance of the Grid Scheduler</b> <i>Kalyani Vaidyanathan</i>
4		324	<b>Impact of performance analysis of varied subjects on overall result: An empirical discourse of educational data mining</b> <i>Mudasir Bhat, Majid Baba and Muheet Butt</i>
5		225	<b>Optimized Feature Selection in Dimensionality Reduction for Big data with Random Forest Algorithm</b> <i>Vinayagasundaram B, Swathy R, Roshini C, Sarshini D and Senthil kumar G</i>
6		80	<b>Design and Development of a System for The Impact Analysis of Internet usage among Engineering Students</b> <i>Rajalaxmi R.R., Natesan P, Krishnamoorthy N and Ponni S</i>
7	205	<b>Dynamic Remote Data Auditing using Privacy Preserving Auditing Protocol in cloud Environment</b> <i>Raja, Ramakrishnan, Hariharan S, Ramprasath M and Arunkumar G</i>	
<b>05.30pm</b>	<b>High Tea</b>		

Convener, ICoAC 2017

Home Page

S.No	TITLE	PAGE NO
1	<b>Monitoring as a Service (MaaS) and its usage in Containers across Multicloud</b> <i>Baskaran Jambunathan and Kalpana Yoganathan</i>	382
2	<b>Big Biomedical Data Engineering</b> <i>Ripon Patgiri, Sabuzima Nayak and Samir Kumar Borgohain</i>	389
3	<b>Application of NLTK for Voice Based Sentiment Analysis</b> <i>Pankaj Kumar Keserwani, Ashish Kumar Mishra and Shefalika Ghosh Samaddar</i>	396
4	<b>Extracting the features of Medicinal Plants and Improving the Accuracy of Classification using Fuzzy Local Binary Pattern Approach</b> <i>Srinivas M</i>	402
5	<b>A new Heuristic Similarity Model to Improve the Accuracy for Data Sparsity Issues in Collaborative Filtering Recommendation Algorithms</b> <i>Mohana H and Suriakala M</i>	406
6	<b>Design and Development of a System for the Impact Analysis of Internet usage among Engineering Students</b> <i>Rajalaxmi R.R., Natesan P, Krishnamoorthy N and Ponni S</i>	411
7	<b>Convergent Scheduling with EDF and BackFill Algorithm to improve the Performance of the Grid Scheduler</b> <i>Kalyani V</i>	417
8	<b>Energy Efficient MAC Protocol for Body Centric NANO-Networks (BANNET)</b> <i>Sivapriya S and Sridharan D</i>	422
9	<b>Comparison of Machine Learning Algorithms - Opinion Mining for Social Network Data</b> <i>Sriprasath G.P, ShanmugaSundaram S and Parkavi R</i>	427
10	<b>Sensitive Data Protection in Cloud-Based on Modified Elliptic Curve Cryptographic Technique</b> <i>Sumathi M and Sangeetha S</i>	433
11	<b>An Empirical Study Of Deep CNN Models Towards Semantic Segmentation</b> <i>Arunekumar N.B and Suresh Joseph K</i>	439
12	<b>An Analysis of Multifarious Character Recognition</b> <i>Sornam M and Poornima Devi M</i>	447
13	<b>LangTool: Identification of Indian Languages for Short Text</b> <i>SreebhaBhaskaran, Geetika Paul, Deepa Gupta and Amudha J</i>	455
14	<b>Dynamic Remote Data Auditing using Privacy Preserving Auditing Protocol in cloud Environment</b> <i>Raja, Ramakrishnan, Hariharan S, Ramprasath M and Arunkumar G</i>	461
15	<b>A Novel Approach to Optimized Hybrid Item-based Collaborative Filtering Recommendation Model using R</b> <i>Abhaya Kumar Sahoo and ChittaranjanPradhan</i>	468
16	<b>Urban Slum Extraction using GLCM Based Statistical Approach from Very High Resolution Satellite Data</b> <i>Prabhu R, Umasree S, Alagu Raja R.A. and Avudaiammal R</i>	473
17	<b>Optimized Feature Selection in Dimensionality Reduction for Big data with Random Forest Algorithm</b> <i>Vinayagasundaram B, Swathy R, Roshini C, Sarshini D and Senthilkumar G</i>	479
18	<b>Analysis and Estimation of Brain Tissue Atrophy using Magnetic Resonance Images</b> <i>AhanaPriyanka N and Kavitha G</i>	485
19	<b>A Multifamily Android Malware Detection using Deep Auto Encoder Based Feature Extraction</b> <i>Teenu S John, Tony Thomas and MerajUddinMd</i>	492
20	<b>Anomaly Based Intrusion Detection System using Classifiers</b> <i>Aarathi M and NarasimaMallikarjunan K</i>	500
21	<b>Reliability Aware Self-Test Schedule for Aging Systems</b> <i>HariniSriraman and PattabiramanVenkatasubbu</i>	506
22	<b>Optimized Channel Awareness Routing For Congestion Avoidance by Dynamic Queue Space Management in MANET</b> <i>Sundar R, Kathirvel A and Narayanan Si</i>	510
23	<b>A Review of Hierarchical Fuzzy Text Clustering</b> <i>Seema Wazarkar, Bettahally Keshavamurthy N and Amrita Manjrekar</i>	516
24	<b>Fuzzy-knowledge-inferred Edge Directed Image Enhancement Technique</b> <i>Reshmalakshi C and Sasikumar M</i>	520
25	<b>IoT-Based Charging Station for Electric Vehicles</b> <i>Chellaswamy C, Archana V, Arunraj J and Bhagirathi S</i>	526
26	<b>Impact of performance analysis of varied subjects on overall result: An empirical discourse of educational data mining</b> <i>Mudasi Ashraf, Majid Zaman and Muheet Ahmed</i>	534
27	<b>Automated Trust Evaluation Mechanism for Future Generation Internet</b> <i>Umamaheswari S and Arun Fera M</i>	542
28	<b>Design of Automated Data Extraction System for Medical Web Forums using Semantic Analysis</b> <i>Umamageswari Kumaresan and Kalpana Ramanujam</i>	549

# Monitoring as a Service (MaaS) and its usage in Containers across Multicloud

Baskaran Jambunathan  
Research Scholar, Vels University, Chennai.  
shriyabaskaran@yahoo.com

Kalpana Yoganathan  
Asst. Professor, Vels University, Chennai  
ykalpanaravi@gmail.com

**Abstract** — Software development has turned into a new paradigm shift in cloud based hosting and deployment through the rapid growth in Container based application development. Platform as a Service (PaaS) provides means of application development in the form of micro service and deployed through containers over the infrastructure layer. The shift from VM based deployment to container based deployment helps the developers to focus on development alone and help is agile based development and help in focussing more on innovations and less of operational task. Monitoring, managing and tracking these application containers are the key industry focus areas and enable companies to make decisions in moving from one cloud service providers to another as containers are more light weight and portable across platforms. In this article, the authors would like to analyse various tools and techniques to analyse container performance and arrive at the logical decision making in moving containers across cloud platforms.

**Keywords:** Containers, Dockers, Kubernetes, Pod, multicloud, cluster management, Predictive analytics

## I. INTRODUCTION

In recent years cloud based application development becomes very prominent as companies started building cloud native application development beyond doing lift and shift of their existing applications. Application developed on premise are re-hosted and migrated on to cloud and deployed in Virtual Machines in earlier days. These applications are difficult to be ported across different cloud platforms and manageability is a big challenge as the VMs are heavy weight. Container based application development are catching up in the industry and enable companies to build more agile and innovative application though the flexibility and porting on to different platform seamlessly. Containers are more lightweight as the Operating system and under laying infrastructure are shared across different containers. It supports all languages and is highly secure, scalable and portable across platforms. Applications can be decomposed into small services called micro services and each service of group of services can be hosted in containers and managed more efficiently through PaaS platform. There are different containers are available, but Docker is the most popular and well accepted in the industry.

Docker[7] is an open platform for developers and operations teams to build, ship, and run distributed applications. A “Dockerized” application is portable and can be run anywhere, locally or in the cloud. It addresses the complications of large, distributed applications running across complex environments. Docker is finding favour as a solution for supporting and incorporating microservices into application architectures.

## II. CONTAINER AS A SERVICE

Container [7] management is the emerging technology area and Container management as a service [6] ( CaaS ) is the new model of application development and deployment[4] in cloud platform. CaaS stands in between Infrastructure as a Service ( IaaS ) and Platform as a Service ( PaaS ). There are many tools and products available in the market for building container based application and this platform provides lots of out of box features for managing complete life cycle of this containers. Some of the products available are like Pivotal cloud foundry, RedhatOpenshift, IBM Bluemix, Apprenda etc. These PaaS platform helps developers to build Container based application development and host the micro services in the containers and deploy them. It provides services like security, scalability, availability and many other Services as out of box. Developers need to only configure and consume these services and focus only on application development and deployment. These platforms also helps in automating deployment[4] across different environment and support complete DevOps[2][7] process, enable developer to focus only on application development.

Containers are flexible, adaptive, and portable across many platforms due to their lightweight nature. These containers can host application as a whole or a service or a group of services based on the way applications can be packaged and deployed. Containers needs to be constantly measured and monitored and hence various metrics has be collected both at infra level and at the application level in order to manage them more efficiently. There are many container based monitoring tools are available like dynatrace, cAdvisor etc which helps in capturing various key parameters and provide a self-service dash boards to monitor and track though their dashboards. Authors in this article would like to analyse different tools available in the market and evaluate their performance and suitability in the context of monitoring containers and collect metrics to analyse.

## III. MONITORING AS A SERVICE

Monitoring as a Service (MaaS)[2][3] in the Cloud is a concept that combines the benefits of cloud computing technology and traditional on-premise IT infrastructure monitoring solutions. MaaS is a new delivery model that is suited for organizations looking to adopt a monitoring framework quickly with minimal investments. Monitoring-as-a-service (MaaS) is one of many cloud delivery models under anything as a service (XaaS)[3]. It is a framework that facilitates the deployment of monitoring functionalities for various other services and applications within the cloud. The most common application for MaaS is online state monitoring,

which continuously tracks certain states of applications, networks, systems, instances or any element that may be deployable within the cloud. MaaS offerings consist of multiple tools and applications meant to monitor a certain aspect of an application, server, system or any other IT component. There is a need for proper data collection, especially of the performance and real-time statistics of IT components, in order to make proper and informed management possible.

#### A. When to use MaaS?

Monitoring as a service (MaaS) are gaining popularity in cloud ecosystem where companies are looking for fully automated and sensible approach for monitoring their various cloud resources deployed in their environment. Following are the key reasons for companies to use MaaS.

##### 1) Price Sensitive Customers:

For small and medium enterprises, MaaS provides cost effective pay per use pricing model. Customers don't need to make any heavy investments neither in capital expenditures (capex) nor in operating expenditures (opex).

##### 2) Cloud Based SaaS and PaaS offering Add-On:

MaaS provides a better technology fit for monitoring cloud based SaaS and PaaS offerings. MaaS can be provided as an add-on product offering along with SaaS and PaaS.

##### 3) Distributed Infrastructure Assets:

In scenarios where the IT infrastructure assets are distributed across different locations and branch offices, MaaS is a good option since the monitoring infrastructure is centralized in the cloud and can easily monitor all distributed infrastructure assets.

##### 4) Mixture of Cloud and On-Premise Infrastructure:

MaaS is already in the cloud. Hence in deployments where customer has a mix of on-premise and cloud infrastructure, MaaS provides good monitoring options for the hybrid environment.

##### 5) Multitenant Monitoring Requirements:

For vendors offering multi-tenant functionality on their hosted services, MaaS provides a strong backend framework for monitoring the multi-tenant services and their availability.

#### B. Assets that can be monitored using MaaS

MaaS is capable of monitoring all aspects of IT infrastructure assets. Following are the few monitoring aspects

##### 1) Servers and Systems Monitoring:

Server Monitoring provides insights into the reliability of the server hardware such as Uptime, CPU, Memory and Storage and is an essential tool in determining functional and performance failures in the infrastructure assets.

##### 2) Database Monitoring:

Database monitoring is to provide performance analysis and trends which in turn can be used for fine tuning the database architecture and queries, thereby optimizing the database for our business requirements.

##### 3) Network Monitoring:

Network monitoring provides pro-active information about network performance bottlenecks and source of network disruption.

##### 4) Storage Monitoring:

Storage monitoring for SAN, NAS and RAID storage devices ensures that your storage solution are performing at the highest levels and ensures anytime availability of business critical data by reducing the downtime of storage devices and hence improves availability.

##### 5) Applications Monitoring:

Applications monitoring provides insight into resource usage, application availability and critical process usage for different types of applications. It is essential for mission critical applications that cannot afford to have even a few minutes of downtime and by application we can prevent application failures before they occur and ensure smooth operations.

##### 6) Cloud Monitoring:

Cloud Monitoring for any cloud infrastructure such as Amazon or Azure gives information about resource utilization and performance in the cloud. Cloud monitoring provides insight into exact resource usage and performance metrics that can be used for optimizing the cloud infrastructure.

Virtual Infrastructure monitoring: Monitoring virtual machines and related infrastructure gives information around resource usage such as memory, processor and storage and reliability of any such infrastructure failures.

#### C. Challenges in container monitoring

In the world of containers – monitoring infrastructure alone or application alone may not be able to provide the complete picture. Since the applications and the underlying infrastructure are highly dependent and their corresponding software components are inseparable, we need to monitor together to get the complete picture.

Complete Monitoring = (App + software defined components/devices + Infra)

There are multiple tools are available for monitoring and the challenges with the monitoring tools are

- There are vast set of monitoring tools to collect various statistics available
- Each tool gives different set of attributes in different format, which needs to be carefully analyzed
- Many times data collection tools may tend to overload the container itself, making the statistics inaccurate.
- At times, differentiating metrics for containers that are related and share resources
- Lot of computation is required to come up with meaningful inferences from all the data that is collected.
- Categorizing container utilization and statistics for multitenant applications is complex
- Different applications provide different format of logs.
- Analyzing the interconnectivity between applications in different containers, hosts or regions.

#### IV. OUR APPROACH IN MONITORING

In our approach, we are analysing the containers performance across multiple cloud environment against all the above parameters, gather metrics and use this metrics for analytics and decision making for moving or migrating the containers along with application and data from one cloud platform to another. We have used different monitoring tools to assess the performance of containers and use this metric to monitor the changes and update the containers during the deployment. We use this metrics insight at different levels like containers, clusters and nodes. These metrics collected are analysed in terms of changes, usage, errors and configurations and are correlated as part of monitoring the performance and arrived at a decision on moving the containers, either single or a group (clusters) to different platform or regions within the same platform. We use all types of analytics like proactive, predictive and adaptive analytics in analysing the current performance and do the decision making.

Our approach to container monitoring and analysis has the following considerations

- Apps are embedded within the containers which are in turn within a VM or physical host
- Containerization requires monitoring at these different levels in order to collect complete statistics.
- Containers can be linked – ability to monitor and make sense of statistics from linked containers becomes critical.
- Ability to intelligently correlate collected data in the context of App à Container à Host relation.
- Abstraction of monitoring methods and data in order to enable integration with any monitoring tool of choice.
- Ability to do proactive, reactive and adaptive monitoring.

Our objective and key focus area is that based on the metrics collected, structured analysis are performed and based on the analysis, migration of containers across cloud platform are performed for optimization. Our intension is to make the monitoring as modern using modern tools and focus on the business outcome. Traditional approach focus on fault detection and our approach focus more on metrics and automation based, which leads to migration and optimization.

Our key considerations are

- How do we know what to monitor and finding it all together which are useful for analysis.
- Although data collection is cheap, but not having it when we really need it can be expensive.
- Monitoring constantly and alerting the stakeholders proactively is what matters the most.
- Do the analysis recursively until we find the root cause of the problem and resolve.



Fig 1: Our process for analysis

We will discuss further in this article on the different environment created and used for analysis, different tools used

for monitoring and metrics collection, Analysis and co-relation of these metrics and decision making in migrating the containers further.

#### V. MONITORING AS A SERVICE – OUR DESIGN

We design our system as a self- service portal which helped in monitoring multiple resources like network, servers (containers), applications etc., and defined Monitoring parameters as service. As part of the services, we defined the components like event monitoring, logging, reporting, alert notification etc.

User can monitor the different components of container ecosystem. User account management, configuration management, alert management, event management etc can be done using the self-service portal and email will be triggered based on the configuration. All data (historical) are stored in the database for analysis and study of the system and take corrective action as per the decision made.

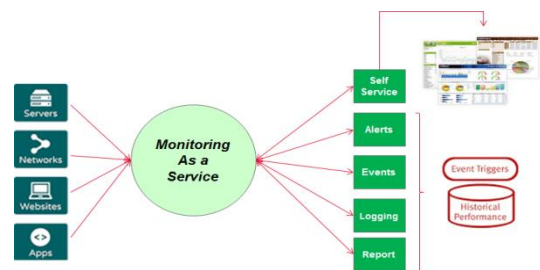


Fig 2: Our MaaS Monitoring system

#### VI. MONITORING TOOLS USED FOR OUR ANALYSIS

For our analysis, we have considered three cloud platforms viz., Amazon web services, Microsoft Azure and RedhatOpenstack which are interconnected through networking. Our intension is to deploy containers in each of the cloud platform and monitor the performance through various metrics analysis using some of the tools like sematext, weavescope, cAdvisor and dynatrace.

##### 1) *Sematext*: Sematext

Docker Agent is a modern, open-source, Docker-native monitoring and log collection agent. It runs as a tiny container on every Docker host and provides automatic collection and processing of Docker Metrics, Events and Logs for all cluster nodes and all auto-discovered containers. Easy to use and Deployable to any Docker-enabled platform within a few minutes: Docker Cloud, AWS ECS, etc. it does Automatic discover of any newly launched containers without manual intervention and Collect all Docker operational bits – metrics, logs, and events, both from Containers and their Hosts.

##### 2) *CAdvisor*:

(Container Advisor) provides container users an understanding of the resource usage and performance characteristics of their running containers. It is a running daemon that collects aggregates, processes, and exports information about running containers. Specifically, for each container it keeps resource isolation parameters, historical resource usage, histograms of complete historical resource usage and network statistics. This data is exported by

container and machine-wide. cAdvisor has native support for Docker containers and should support just about any other container type out of the box

3) *Weavescope:*

Weave Scope automatically detects and monitors every host, Docker container and process in your infrastructure, builds a map showing their inter-communications and then presents an up-to-date view of your infrastructure in a web interface. We can visualize, monitor and control your distributed applications and troubleshoot bottlenecks, memory leaks or any other issues. It does this without requiring changes to our code or configuration, and without having to make declarations about your infrastructure that become out-of-date and stale. Weave Scope can be deployed to any infrastructure, and works well in all cloud and bare-metal environments

4) *Dynatrace:*

It is designed to deal with highly dynamic infrastructure where containers come and go frequently. By monitoring your containers with Dynatrace you're all set for the monitoring of microservices and the associated micro-deployments, which are commonly delivered via Docker containers. Monitoring with Dynatrace is easy, as it should be. Dynatrace automatically scales with our environment by discovering and monitoring new containers. Track deployments of your Dockerized micro services and monitor distributed applications across your network of hosts or cloud instances.

We considered these four tools after lots of study and analysis of multiple monitoring tools. The choice of parameters considered for analysis is based on our observation and considerations of various tool analysis and scenarios. The table below shows the comparative study of various container parameters and its availability in these four tools considered for analysis. Following table explain the feature wise mapping of the tools.

TABLE 1: Comparison of four Container monitoring tools from business perspective

Metrics	cAdvisor +ELK	Weave scope	Dyna trace	Sematext
Commercial / Opensource	Open source	Open source	Commercial	Commercial
Ease of Deployment	Easy	Easy	Easy	Easy
Resource overhead	High (upto 10%)	Medium (upto 5%)	Very Low (<1%)	Very Low (<1%)
On-premise / Cloud	On-premise	On-premise	Both	Both
Trigger Alerts	No	No	Yes	Yes
Kubernetes monitoring	Yes with Heapster	No	Yes	Yes

Following table explain the performance metrics comparison with the tools.

Based on this study we found that dyanatrace provides comprehensive features and metrics both infrastructure and application level and it provides a good snapshot of all these features and very good reports.

The key objective is to analyse the metrics collected and make decision in moving the containers to better cloud platform for performance optimization.

TABLE 2: Comparison of four Container monitoring tools from features perspective

Metrics	cAdvisor +ELK	Weave scope	Dyna trace	Sematext
Docker Host CPU	Yes	Yes	Yes	Yes
Host Memory	Yes	Yes	Yes	Yes
Host Disk Space	Yes	Yes	Yes	Yes
Total Containers Running	Yes	Yes	Yes	Yes
Container CPU	Yes	Yes	Yes	Yes
Container Memory	Yes	Yes	Yes	Yes
Container Memory Usage	Yes	Yes	Yes	Yes
Container Swap	No	No	Yes	Yes
Container Disk I/O	Yes	No	Yes	Yes
Container Network Metrics	Yes	Yes	Yes	Yes
Application Metrics	No	No	Yes	Yes
Inter-Container traffic	No	Yes	Yes	Not sure

VII. OUR ANALYSIS AND OBSERVATIONS

We wanted to analyse the performance of the containers inside the VM and see how it is impacting the VM and its memory consumption against different cloud environment. To carry out our analysis we created a VM of size Large with 2 CPU, memory 4GB and hard disk 50GB. Our idea is to look at the container performance in this VM in each of the cloud platform like AWS, Azure and Openstack by

- a) Adding containers more and more and evaluate the performance
- b) Create a pod of four containers, replicate the pod[7] and evaluate the performance

We considered dynatrace as explained above, as a monitoring tool to look at the key performance metrics like CPU percentage, memory percentage, network Utilization and disk latency and conducted the lab exercise to study the performance.

Following table shows the performance of containers in AWS. We created the 4GB Virtual machine in AWS and added containers started from 4, added 2 containers more and more and at each time analysed the metrics and see at what point of time AWS reaches close to 100% utilization. In our analysis, when we reached 14 containers and above, CPU was hitting close to 100 %.

TABLE 3: Performance metrics in AWS for container deployment

No.of Container	CPU %	Mem %	Disk latency 9mSec	N/W Util
4	1%	34%	4.8	47.3
6	1.33	45	2.37	61
8	1.36	56	3.37	74
10	1.62	66	4.28	84



14	59	90	18.2	147
16	100	92	33.4	32.8

3	10	77	65.7	125
4	15	81	822	136
5	20	84	59.8	178
6	50	85	304	427
7	70	90	350	460
8	95	99	400	480
9	100	100	475	560

Following table shows the performance of containers in Azure. When we conducted the same exercise in Azure, starting with 4 containers and keep adding additional two containers every time and studied the performance, Azure was able to withstand even after 18 containers and CPU was reaching 90% and not 100% as against AWS. Following observations depicts the behaviour.

TABLE 4: Performance metrics in AWS for Pod deployment

No. of PODs	CPU %	Mem %	Disk latency (ms)	N/W Util (Kbps)
1	5	31%	3.93	70.2
2	6.35	52	2.83	87.2
3	6.73	73	2.53	121
4	15	86	50.7	96.7
5	99	88	147	140
6	100	36	50.7	96.7

From these experiments, we feel that the way Azure manages the Containers and VM is different from AWS obviously. Our study is not to analyse the reason for this difference in behaviour, rather to study the container behaviour at that particular instance and decide whether to move it to different cloud platform as it is reaching performance saturation in AWS after 14 containers and the same is working better in Azure. Since containers are portable, we wanted to make use of it and migrate to different cloud platform, where it can perform better at that particular instance.

TABLE 5: Performance metrics in Azure for container deployment

No. of Container	CPU %	Mem %	Disk latency (mSec)	N/W Util
4	1.62	38%	58.8	47.3
6	1.97	52	30.8	58
8	2.36	62	3.34	70
10	2.91	72	138	84
14	5.02	88	26.4	147
16	17	87	283	196
18	28	90	309	207

Our Observations are as follows

- AWS and Azure show similar strength during initial raise in containers.
- Beyond 14 containers in case of AWS, CPU gets saturated. Unable to add more than 14 containers in AWS due to cpu saturation and existing containers were crashing in AWS after saturation point.
- In Azure, even after 19 containers, CPU did not get saturated.

TABLE 6: Performance metrics in Azure for pod deployment

No. of PODs	CPU %	Mem %	Disk latency (ms)	N/W Util (Kbps)
1	7.04	34	45.4	62.8
2	8.88	56	85.5	91.3

As part of our further analysis, since container uses microservice[7] and group of containers has to be used in conjunction with each other in the form of a cluster and has to be managed within a pod, which is the common scenario in many container based development, we considered a group of four containers in the form of pod. We conducted similar exercise, by increasing the number of pod within the VM and did the cluster management with kubernetes[7].

Following are our observations in both AWS and Azure. From our analysis and observations, it is found that in the case of AWS, it reaches the saturation when then the number of pod goes beyond 5 (each has four containers) and in the case of Azure, it can with stand even the number exceeds 8 (each pod has four containers). Below graph depicts the performance and our observation

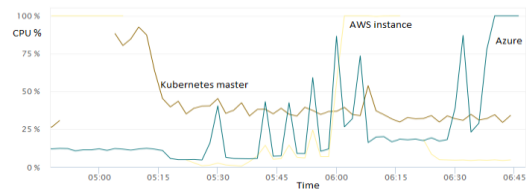


Fig 3: CPU Performance AWS vs Azure

Following fig shows AWS reaching saturation when the pod hits 5 (four each leading to 20 containers in total) whereas Azure reaches only 15% when the pod is 5.

Name *	CPU usage	Memory usage	Disk latency	Network traffic	Number of containers	State
mars	37%	55 % of 3.36 GB	6.48 ms	54.4 kbit/s	30	Running
ip-172-31-24-212.ap-south-1.compute.internal AWS	100%	90 % of 3.45 GB	53.4 ms	55.7 kbit/s	20	Running
asanki Azure cloud	15%	84 % of 3.36 GB	38.3 ms	146 kbit/s	24	Running

Fig 4: CPU performance reaching saturation for AWS

Following table shows Azure reaching saturation when the pod hits 8 (four each leading to 32 containers in total)

Name *	CPU usage	Memory usage	Disk latency	Network traffic	Number of containers	State
mars	35%	27 % of 3.36 GB	41.2 ms	49 kbit/s	19	Running
ip-172-31-24-212.ap-south-1.compute.internal AWS	4.54%	12 % of 3.45 GB	2.87 ms	33.3 kbit/s	21	Running
asanki Azure	100%	90 % of 3.36 GB	247 ms	140 kbit/s	33	Running

Fig 5: CPU performance reaching saturation for Azure

In both of our analysis (containers getting added more and more & Number of pods increased more and more), it is clearly evident from our analysis that Azure performs better at a given situation where AWS is reaching saturation. As

mentioned earlier, we did not analyse the reason behind it as our intention is to monitor the container behaviour and based on that, chose the right platform for a given situation and make decision in moving from one platform to another. Here both simple container management and kubernetes based cluster management are used for analysis in both the platform.

We conducted the same exercise on OpenstackMitaka version, by creating the VM of same size and performed the similar exercise of both addition containers directly on to VM and also added the pod of four containers gradually. Following are the results obtained from similar exercise on VM running in Openstack single node instance.

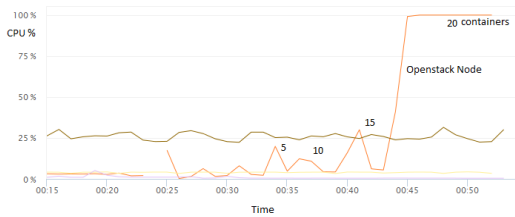


Fig 6: CPU performance reaching saturation for Openstack with containers

In the first case, when we add the containers directly on to VM, we see the CPU getting saturated when the number of containers hitting 18, which is similar to AWS and in case of Azure it could go beyond as well.

Secondly in the case of number of pod, with each having four containers, Openstack could withstand up to 5 Pods and when it hits 6th one it is reaching saturation, which is similar to AWS and in case of Azure it is reaching up to 8.

TABLE 7: Performance metrics in Open stack for Container deployment

No. of Containers	CPU %	Mem %	Disk latency	N/W Util
4	2.4	15%	128	72
6	11	20	98	278
10	12	47	184	266
14	23	66	165	343
16	30	75	290	360
18	100	91	65	54

Following table describes the performance parameters for the pods getting added gradually in Openstack

TABLE 8: Performance metrics in Openstack for Pod deployment

No. of PODs	CPU %	Mem %	Disk latency (ms)	N/W Util (Kbps)
1	19	14	115	1024
2	11	29	114	259
3	9	45	144	151
4	29	66	156	337
5	60	85	148	313
6	100	90	160	330

Diagram below describes the CPU performance in Openstack and its behaviour.

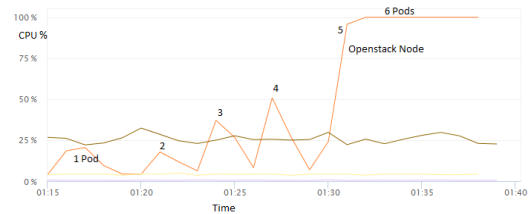


Fig 7: CPU performance reaching saturation for Openstack with pods

## IX. MIGRATION ON TO MULTICLOUD

In above analysis, we are laying a strong foundation in studying container performances in different cloud platform. Our intention is to integrate different cloud platform like AWS, Azure, Google, Openstack and VMWare and create a multicloud[7] environment and manage it through a single integrated overcloud platform with tool like Redhat - cloudforms or Docker Datacenter. Keep monitoring the containers running within the VM and monitor the metrics both at VM level and container level. Based this analysis, monitor the performance at regular intervals and see when the performance gets impacted in terms of CPU, memory etc. At any point of time, if we find that there is a performance issue in one cloud platform, then we recommend to migrate the containers along with its dependencies to another cloud. But in this exercise, there are lots of challenges if the containers need to be moved live from one platform to another and also if it has lots of dependencies across the platform, we need to evaluate these dependencies and make decisions. Here the intention is to see how we can monitor the performance of resources in each cloud at any given time and make decision by means of statistical analysis and migrate to the best cloud suitable for the given resource at a given point of time. Too much of movement across cloud is also not a recommended approach and at the same time if we let the resource struggle to run due to various reasons in a cloud is also not a good practice. Hence constant monitoring of these resources are critical when we manage multicloud or a single cloud across different regions, and take proactive decision in migration to different cloud without impacting the business and to the notice of the users.

## VIII. OUR RECOMMENDATION AND FUTURE WORK

Our future research is to focus on the movement of containers across platform, based on the predictive, proactive and adaptive analysis and alert the stakeholders accordingly. We would like to bring more of predictive analytics and decision making into our analysis and improve the analysis and make a decision in migrating and optimizing cloud resources. We want to build a single dash board which can bring multiple metrics across different cloud and monitor periodically. These metrics will be analysed, co-related and evaluated constantly over a period of time and knowledge database will be created. Adaptive statistical analysis will be performed on the data collected and some level of business intelligence will be added to study the current performance and make the decision in moving to the right cloud with reasons. In this process, the other legalities in terms of process, governance, polices & procedures are to be considered and

need to develop the best practices and standards into consideration for this optimization procedure. We are working towards achieving this objective which will be a potential platform for many industries in future.

#### REFERENCES

- [1] "MONITORING-AS-A-SERVICE IN THE CLOUD", A Thesis Presented to The Academic Faculty by ShicongMeng, School of Computer Science, Georgia Institute of Technology May 2012.
- [2] "Enhanced Monitoring-as-a-Service for Effective Cloud Management", ShicongMeng, Student Member, IEEE, and Ling Liu, Senior Member, IEEE.
- [3] "Unified Monitoring and Analytics in the Cloud", Ricardo Koller, CanturkIsci IBM T.J. Watson Research, SahilSuneja, Eyal de Lara, University of Toronto.
- [4] "A SURVEY ON AUTOMATED DEPLOYMENT OF CLOUDERA DISTRIBUTION ON DOCKER CONTAINERS"; SHRIKANT S. RAUT, SWATI SALEM, RUPALI KALOKAR, SUNIL NAGARGOJE.
- [6] Model-Driven Management of Docker Containers; FawazParaiso, StéphanieChallita, Yahya Al-Dhuraibi, Philippe Merle; HAL Id: hal-01314827 <https://hal.inria.fr/hal-01314827>, Submitted on 1 Jun 2016.
- [7] MyroslavRys' Containers as a Service: Comparing Providers and Evaluating the State of the Market <http://sandhill.com/article/containers-as-a-service-comparing-providers-and-evaluating-the-state-of-the-market/>.
- [8] BaskaranJambunathan and DrKalpanaYoganathan "Multi Cloud Deployment with Containers", International Journal of Engineering and Technology (IJET) - e-ISSN : 0975-4024.

# Big Biomedical Data Engineering

Na Ripon Patgiri, Sabuzima Nayak, and Samir Kumar Borgohain

Department of Computer Science & Engineering

National Institute of Technology Silchar, India-788010 ripun@cse.nits.ac.in,

sabuzimanayak@gmail.com, and samir@nits.ac.in

**Abstract**—The Big Data, a massive amount of data, is the most popular buzzword and popular paradigm to change a game of any data-intensive field. The engagement of Big Data technology provides a new direction to an organization. The Big Data gives a vision to biomedical data engineering. There are numerous data-intensive fields to engage Big Data technology to achieve their vision. Interestingly, the Big Data play a crucial role in Big Biomedical Data Engineering (BBDE). The massive amount of biomedical data becomes a dilemma in analysis, diagnosis and prediction. Besides, the large-scale medical data cannot be stored, and processed without employing Big Data technology. The deploying Big Data technology can change the game of biomedical engineering. This paper exploits the role of Big Data in biomedical data engineering and its storage dilemma.

**Keywords**—Big Data, Big Data Analytics, Big Data storage, Big Biomedical Data Analytics, Biomedical, Big Biomedical Data Engineering, Healthcare, Cancer, Genome, Neurology

## I. INTRODUCTION

The Big Data is exploding every data intensive field around the world since its inception, especially, in the IT industries. The Big Data is able to show its existence and dominate the market within a few years. The Big Data is high volume consisting of a variety of data. These data are increasing with a very high velocity. The highly increasing massive dataset is the perplex dilemma to store, process and manage. The conventional system does not work in large-scale data. Therefore, there are numerous way to deal with the dilemma of BBDE using Big Data. NoSQL, for instance. The Big Data seems more theoretical, but Big Data engineer can feel the pain of maintaining the mammoth sizes of data. Besides, the scientist deals with tremendous biomedical data daily basis. Thanks to the Big Data paradigm that makes our life easy in Biomedical Data Engineering (BDE). Moreover, an organization can enhance their performance by employing Big Data technology. Therefore, the boundary of Big Data extends to every field nowadays. For example, science, engineering, economy, politics, and business.

The biomedical research is known as useful research for the well being of human being. Engaging Big Data in biomedical can generate numerous possibilities. It is a common practice nowadays. Every year enormous amount of health data is created [1]. Interestingly, every individual generates 0.4 TB data, 6 TB of genomic information and 1100 other health data by 2020[2]. The clinical data also be doubled every 73 days from 2020 onwards [2]. The collective healthcare data are tremendous in size and increasing its size at a very high pace on a daily basis. The huge data visualization enhances the abilities to find trends, identify

outliers and perform quality checks [3]. The visualization process stitches together the variety of data types for a common goal. There is a research opportunity in the technological singularity [4] that can exist in Big Biomedical Data Engineering (BBDE) which is to be exposed. However, this singularity can only be discovered after excessive experimentation of the technology or after deployment of the technology in the field of healthcare.

The Internet of Things (IoT) is a disruptive technology predicted by [5]. Things are becoming smart and connected to the Internet. The IoT is booming research area and there are billions of IoT devices have already been shipped in the market. The IoT is integral part of smart hospital nowadays [6]. The IoT devices deployed in hospital to upgrade the conventional hospital to smart hospital. Therefore, the IoT devices are the key sources of generating massive amount of data in healthcare system.

Ta et al. [7] classifies the healthcare data into seven major categories, namely, Electronic Healthcare Records (ERRs), Social media, clinical text, Genomic data, Biomedical Signals, Biomedical images, and sensing data. Various data are grouped into a single category based on the nature of the data. For example, brain signal and heartbeat signal are grouped into Biomedical Signal. The biomedical data is fed into Big Data technology for processing and stored in the Big Data environment. The advancement of Big Data technology makes healthcare system easier than earlier (conventional system). However, there are many barriers in modern biomedical systems [8]. Interestingly, some data are personal and extremely private, for example genome. These data should not be exposed to the outer world and to be maintained their privacy [9], [10]. Privacy becomes limited by deploying Big Data in genome [11]. In addition, the most of the health data require privacy. Aziz et al. [12] emphasizes on the secure and efficient genome data computation. Moreover, the large set of such kind of data transmission is also a key issue. The Big Data Smart Socket (BDSS) is an example to transfer large-scale biomedical data, provides an alternative mirror for data seamlessly [13]. On the contrary, the data discovery is also a prominent research area [14]. The metadata enhances the process of the data discovery. However, there are plenty of data Indexing techniques available nowadays. With the technological advent, the biomedical data scientists apply computational and statistical approaches for data mining and machine learning to gain insight of the huge dataset. Precisely, the tools of Big Data are extremely helpful in the field of the biomedical research, namely, Hadoop.

The paper is worded as follows- section II describes Big Data in the context of biomedical engineering. Section III exposes the role of Big Data analytics in biomedical data engineering. Section IV-A discusses challenges and opportunity in cancer research. Section III-F and IV discusses about neurology and genome research using Big Data. Section V discusses possible health center set up by Big Data. Section VIII discusses future of Big Biomedical Data Engineering. Finally, section IX concludes the paper.

## II. BIG DATA

First, confirm t The Big Data is defined by Doug Laney using three Vs, namely, volume, velocity, and variety [15]. However, there are many V's to define the Big Data characteristics, namely,  $V_3^{11} + C$  [16]. Simply, the Big Data is enormous data handling which is impossible for a conventional system. The Big Data paradigm is proven as a game changer paradigm in many data-intensive fields, for example, environment, science, engineering, business etc. [16]. The evolving of Big Data paradigm creates many hope and possibilities in the various data-intensive fields. The real-life application and system of Big Data are representing a large set of data in the hetero-geneous environment, modelling and processing the plethora of data, querying and mining the massive amount of data in databases and data warehouse [17]. The Big Data is employed in the real-life system due to the gigantic size of data is stored, processed, monitored, analyzed, and visualized. For instance, Square Kilometer Array telescope data are exabytes [18]. However, the data-intensive fields use Big Data technology to improve and organizational performance and revenue, like BDE. Undoubtedly, the Big Data is a good choice for BDE due to a massive amount of data to be analyzed [17]. The Big Biomedical Data Engineering (BBDE) requires huge storage spaces, processing capacities, visualization, and analysis [14]. The article [19] ask a question- "why do we write?". The assumed environment may differ, but the answer is similar. However, the BBDE requires the data to write so that someone will use in future to study the diseases in a curing and diagnostic process. The answer converges with article [19]. The Muade et al. [19] indicates that writing is the first phase of Big Data.

The figure 1 exposes a complete solution for a healthcare system using Big Data paradigm. The solutions are the Hadoop-based which are more common to other Big Data based solution. The Hadoop is two tiered, namely, Hadoop Distributed File System (HDFS) - a storage engine and MapReduce - a programming engine. The HDFS is used to store the biomedical data and MapReduce is used to define what to do with these stored data. The Hadoop stack is used to represent the Big Data technology. The selection of Hadoop tools depends on the purposes of designing the complete healthcare system. For example, HBase can be used for large-scale tabular data. As figure 1 depicts, the remote doctors/researchers play a vital role. The remote doctors are located in geographically different place as shown in the figure 1. The Apache Spark is also another emerging tool for BBDE. In addition, it is easy to develop a framework for BBDE using Hadoop by collaborating between Engineer and Doctors. The most of the current solutions are Hadoop-based for clinical data in large-scale space, for instance, BigBWA

[20], GMQL [21]. Moreover, Dive is another tool to visualize biological data [22]. The technology is ready to solve the very complex dilemma of BBDE. The Big Data technology is deployed to make a decision very correctly. Most of the time, we rely on the technology rather manual decision. However, there may not be an existing solution for a particular scenario of biomedical large-scale data. In this case, the problem is converted into the key/value to define the required action by the MapReduce program. The power of MapReduce lies in self-defining solution. In addition, the Big Data tools are ready to provide a solution to BBDE dilemmas. The open source tools are HBase, Hive, Spark [23], Mahout etc. [24]. These tools leverage heterogeneous medical data from heterogeneous sources [25]. The tools provide import, export, compare, combine and understand data [26]. Bourne et al. [27] urges better way of storing, processing and managing the EHR data. Distributed Application Runtime Environment (DARE) is a gateway for scalable scientific data-intensive research [28]. Moreover, Scientific Workflow Management Systems (SWfMSs) are modern workflow systems for Big Data in distributed and parallel environment [29]. Another cloud-based workflow engine is developed by Szabo et al [30].

## III. BIG DATA ANALYTICS

Big Data Analytics (BDA) is a merger of Big Data and Ana-lytics [4]. The analytics means the logical method of analysis. BDA provides a platform for discovery of hidden jewels from data. The machine learning algorithms are used to implement the BDA. The BDA is categorized into five main subcategories, namely, descriptive analytics, predictive analytics, prescriptive analytics, decision analytics, risk analytics and security analyt-ics. These analytics are very useful in medical data analysis. The BDA helps in decision making, analyze the large dataset properly and report the analysis results. The past decades the healthcare data are stored in digital form. The analytics can help in discovering hidden patterns on digitally stored medical data. For instance, the predictive analytics helps in better disease forecast in Biomedical system [31], [32]. Bates et al. [32] emphasizes on Big Data analytics to improve the healthcare system [33]. A description-driven system enhance the biomedical data analytics [14], for instance, CRISTAL[34]. The description-driven system is a metadata system to re-use the stored data again and again in future. Therefore, the metadata is stored alongside the health data. The metadata ensures the availability of the stored data. We may or may not know what kind of data is available in the databases. In this scenario, a descriptive query must be fired to get a similar case to retrieve these data from a very large set of data. Moreover, most of the cases we want to know analysis report from healthcare data analytics about the possibilities of making

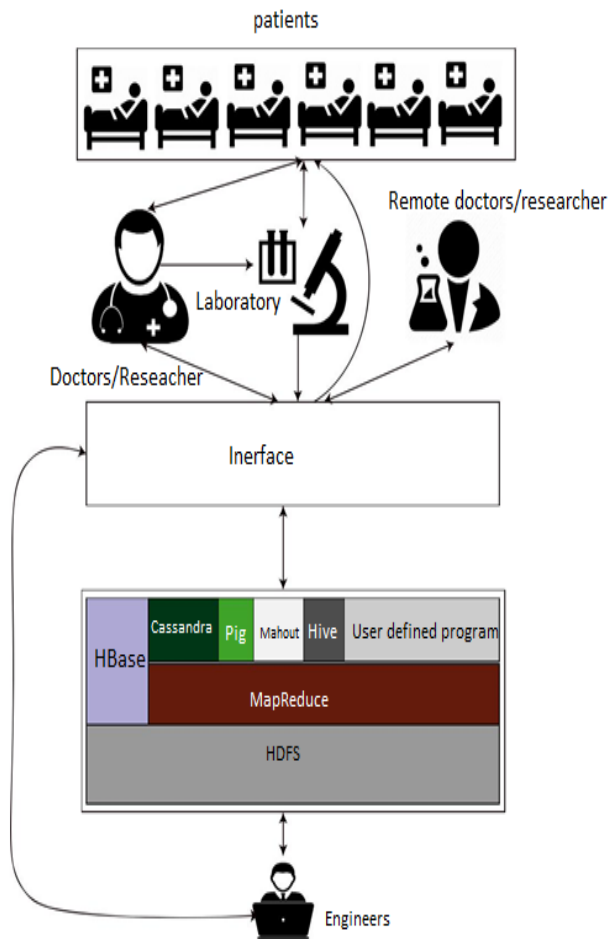


Fig. 1. The Landscape of the Big Data solution of healthcare system.

a decision. The analytics are great tools to make a decision. Without BDA, the decision making is impossible in a large-scale environment. The large-scale medical data are applied not only in the decision making process, but also forecasting, prescription, risk and possibility analysis.

The purpose of Big Data analytics is logical analysis on data. The data are enormous in size. It cannot be performed in a conventional way. Therefore, the Big Data analytics is used to analysis on a plethora of data. The Big Data analytics dubs many problems to solve, for example, Business, Economy, Science etc. Shahand et al. [35] disclose the biomedical data analysis. GVSS performs analysis to detect dengue and flu[36].

Moreover, the most modern biomedical data analysis is cloud-based data analysis [37]. However, let us discuss the Big Data analytics by taking an example of cancer to understand, that is, Big Biomedical Data Analysis (BBDA). The BBDA is a merger of Big Data, Big Data Analytics and Biomedical Data. This BBDA creates enormous solution in Biomedical data engineering. The BBDA is yet to develop. The taxonomy of Big Data Analytics is discussed with respect to BBDA.

### A) Descriptive analytics

The descriptive analytics is a logical analysis of large datasets. A massive dataset has been explored to discover and gain insight. The descriptive analytics is categorized into two, namely, description (or exploration) and discovery analytics. The descriptive or explorative analytics is performed to gain in-depth insight on large-scale medical data and which outputs an informative output. These data are explored to study, diagnose and treat a patient. Discovery analytics discovers some hidden truth on very large scale storage system which contains medical data of many years. However, this is very helpful in biomedical data analytics analysis reasons are unknown. For example, we do not know the patterns of a DNA. We would like to sequence and discover something new patterns.

### B) Predictive analytics

The predictive analytics is logical analysis on the massive amount of data where the future depends on those massive data. The predictive analytics is further sub-categorized into three, namely, prediction or forecast, possibility and indicative

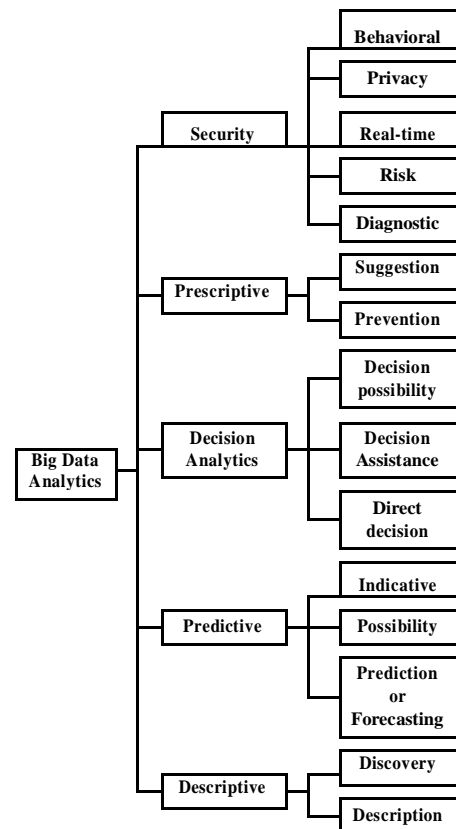


Fig. 2. Taxonomy of Big Data Analytics

analytics. Now, when we perform DNA analysis, the DNA has the code of Adenine (A), Cytosine (C), Guanine (G) and

Thymine (T). Prediction or Forecast: what happen if we alter some information? Possibilities: what is its future and what are the possible effect? And indication: where will it be beneficial?

### C. *Decision analytics*

The decision analytics is a logical analysis of the mammoth size of data to make a good decision. It is classified into three subcategories, namely, direct decision analytics, decision assistance and decision possibility analytics. The Big Data analytics perform analysis and reports the results. Direct decision: is it cancer? Decision assistance: comparison among the non-cancerous DNA cell and cancerous DNA cell. What are the patterns to declare it as a cancer? Decision possibilities: If it is a cancer, then what is the possible stage of the cancer? And it is not a cancer, then what symptoms can it contain (it is decided that there is cancer)?

### D. *Prescriptive analytics*

The prescriptive analytics is a logical analysis of a huge dataset to fetch a suggestion. The prescriptive analytics also further classified into two subcategories, namely, suggestion (recommendation) and preventive analytics. Suggestion or recommendation: List of curing such kind of cancer. How has the curing process been done? Prevention: What is the reason for cancer?

### E. *Security analytics*

The security analytics is a logical analysis of huge data for security purposes. The analysis of the very large dataset is not possible in a conventional way. That's why the Big Data analytics is deployed to ensure security. The security analytics is classified into five subclasses, namely, diagnostic analytics, risk analytics, real-time analytics, privacy analytics and behavioral analytics. Diagnostic: What is the type of cancer detected? Risk: Is it false detection? Real-time: What is the current stage of cancer after performing X treatment? Privacy: Does the DNA made public or keep under privacy policy? Behavioral: What is the abnormality seen after performing X treatment?

### F. *Neurology Research*

The size of a brain is enormous to store and process. It requires very large-scale computing capability. The brain consists of neurons. The neurons of the human brain are 86 billion to compute which 17 million years in a conventional system[38]. Therefore, Big Data is a solution to such mammoth sized data. The data can be downloaded to analyze locally, but the data sizes are too big to do the task locally and that's why, there is no point to move out from a cloud-based solution[39]. A neuroscientist manually tracks each axon from a very large dataset [40]. The neuron networks and nervous systems are very complex to analyze. Deploying Big Data able to analyze, and visualize those perplex neurons. For example, the Big Graph can be deployed to analyze the relation among the neuron network. Furthermore, it is also very helpful in analysis, such kind of gigantic database by deploying Big Data analytics.

## IV. GENOME RESEARCH

The Genome contains all requirements of maintaining the structure of an organism. Genomes are made up of DNA. The DNA is unique information to keep an organism's growth, health, and development. The genome-wide association studies (GWAS) analyze billions of single-nucleotide polymorphisms (SNPs), along with disease status [41], [11]. The analysis can be used to predict disease status. The Big Data creates an opportunity of data-driven prediction of diseases [42]. The key research challenge in a genome is the sequencing and alignment of genome data in large-scale space. A DNA contains ACGT code. Thus, human DNA contains more than 3billions of such code. This code specifies every detail of a person and slide variation of this code can produce different species. Therefore, the information about detail characteristics of any living thins is encoded in the form of ACGT. A few such kind of DNA can form petabytes. It is reported that size of genome data are petabytes [43], [44]. The DNA database sizes become enormous to store. It is impossible to manage the Genome database without Big Data technology. The Big Data analytics can play a vital role in the analysis of such mammoth size of a database.

### A. *Cancer Research*

The temThe Big Data and Cancer research are very hot topic nowadays. The Cancer research requires computational skills [45], [46], [41]. The data of Cancer are very large to analyze, and therefore, the Big Data play a vital role in cancer treat-ment, cancer prevention, and screening. A doctor can match available cancer trials with the patient sample rapidly even if the data size is huge [47]. The International Cancer Genome Consortium (ICGC) consist of two petabytes of cancer data in a five year period [43]. The genome sequences of tumors from patients with breast cancer are compared with the numer-ous available breast-cancer genomes and searched for similar patterns in diverse cancer kinds which is a time-consuming process [39], [48]. In addition, Arend Sidow from Standford university quote as "If I could, I would routinely look at all sequenced cancer genomes. With the current infrastructure, that's impossible" [39]. Because a genome size is 100GB that contributes tremendous size of a database. Most of the clinical data are unstructured [49]. It requires Big Data to store, pro-cess and analyze this plethora of data where 1,00,000 genomes of 75,000 patients have been studied by [50]. Moreover, the conventional database cannot cope up with the exponentially increasing size of the number of patients. The future lies within Big Data technology, the sample data can easily be sequenced for better study and analyze a huge set of a sample of cancer patients. Chen et al. [51] develop a framework, called BUFAM, for breast cancer in the environment of highly heterogeneous biomedical big data. Moreover, biomarker, a new idea is emerging for complex biomedical feature analysis based on Big Data paradigm, like cancer [52], [53]. It is believed that engaging Big Data improves the screening, diagnostic, and treatment of tumor [54]. Fortunately, the National Institute of Health (NIH) reported that the cancer patient death rate decline nowadays.

## V. HEALTHCARE

Another aspect of BBDE is biocuration [55]. The Big Data enhance the biocuration process through a well establish a database to make a good decision. Howe et al. [55] state that the biocuration still lags behind the data generation in the funding, development, and recognition. It requires Big Data technology in a large-scale environment. The Big Data assists in providing good health by exposing in-depth insights on causes and outcomes of a disease, precision medicine, and disease prediction and prevention [56]. The Big Data analytics are very useful in the prevention of epidemic disease. Mooney et al. [57] predict that the Big Data will be practicing in the future as today in an epidemic. The healthcare system must ensure the eco-friendly diagnostic, screening, treatment, and prevention. Philip Bourne quotes

“Our mission is to used data science to foster an open digital ecosystem that will accelerate efficient and cost-effective biomedical research to enhance health, lengthen life and reduce illness and disability” [58].

The Big Data helps in generating new knowledge in biomed-ical and keep practitioner up-to-date [8].

## VI. ISSUES AND CHALLENGES

### A. DATA GATHERING

The first challenge is gathering the biomedical data. There is a diverse source of gathering biomedical data. However, the gathering biomedical data requires many years. Thus, collecting real biomedical data concerns not only money, but also a time. The challenges are data collection of DNA of every citizen, all data from every hospital, pathology and research center. In addition, the biomedical data gathering also requires human resources.

### B. DATA STORAGE

Let us assume, DNA database of the most populous country; India or China. Now, the government wishes to collect the DNA information of every person for the various purposes, for instance, security. Now, collecting those data will takes many years by the government. Imagine, what will be the database size? There is a requirement of a mammoth sized data storage engine. Moreover, it also requires a creation of a different logical database which contains breast cancer, brain cancer etc. separately depending on the disease. Now, these plethora of DNA data is used to study for diagnosing and treatment purposes. It makes the medical system easy if a system can analyze the gigantic database in a few seconds. However, deploying the Big Data Analytics can make it possible to analyze all those data. But still, it takes a time to analyze all those data. It will take huge manpower and processing to understand entire DNA structure. Even today, the Biomedical data engineering has a negligible amount of progress in research and yet to do much more. For instance, TRENCADIS is an effort on medical database creation [59].

Moreover, the medical information requires the metadata server to ensure the future accessibility of the data efficiently and effectively. The metadata defines- how to store, where to

store and how to retrieve [60], [61]. Li et al. [62] implements metadata management systems for High Content Screening. The data management requires the scalable metadata management system to cope up with ever growing dataset in Biomedical system. However, the metadata server is decoupled from data server for high manageability. Therefore, the data server does not affect the metadata server and vice-versa. Moreover, the scalability is the biggest issue of healthcare system. The scalability can easily be achieved by decoupling metadata server from data server [61]. Besides, it also provides location independent metadata management system.

### C. DATA PROCESSING AND VISUALIZATION

There is a call for a Big Data processing engine to process a jumbo sized data. The biomedical data processing and visualization engine can be developed using MapReduce directly or using some famous framework, for instance, Spark. The data size is gigantic to process and visualize. The processing takes many days to many years in the conventional system and visualization cannot be done manually or using conventional system due to bulky in size. The data are collected from diverse source and stitched together to present the data in a meaningful way. Therefore, there is always a call for engagement of Big Data technology to handle Biomedical data.

### D. VALUE

The collected data are enormous in size. Why are the data collected? Can we able to assign a worth to the collected data or simply dumping it? The Big Data always concerns about extractions of worthy information from the collected data. The collected data is worthy only when if we use those data for a betterment of the Biomedical system. Thus, the involvement of Big Data in Biomedical data always gives a worthy sense. The challenge lies in assigning value to the collected data and extracting value from the same data.

### E. ACADEMIC RESEARCH

There are many countries which have a huge gap between doctors and engineers in the biomedical academic research fields. Introducing the BBDE course in both medical institute and engineering institute bridges the gap between the doctors and engineers. Thus, Biomedical Engineering can be enhanced. The research collaboration among Engineer, Doctors (Medical) and Industry people make advancement the Biomed-ical data engineering. However, this is the most prominent issues in all countries. The collaborative research leads better results. e-Infrastructure is an example of collaborative research [63].

### F. DATA PRIVACY

Most of the Biomedical data are private data, and thus, the privacy is the prime factor to be maintained in biomedical database. This privacy requirement creates an issue in creating open database worldwide. This is the reason the worldwide database creation requires privacy protection, and thus, the database cannot grow. Eric Schadt [11] emphasizes on privacy of personal data and quote as



“Genomic information has been the main focus of past debates on the protection of privacy and is subject to more legal regulations than other forms of high-dimensional molecular data such as RNA levels.”

Moreover, Schadt also indicates that the Internet breaches the privacy of individuals. Earlier, the protection of privacy was easier than today’s landscape of BBDE.

#### G. REAL-TIME PROCESSING

The modern healthcare system requires real-time profiling of patient which poses a big challenge to achieve. It is required to monitor continuously every parameters of a patient automatically and remotely to enhance healthcare system [64],[7]. For example, heartbeat. However, there are modern stream processing engines to implement the real-time healthcare system requirements. For example, Apache Spark. However, it is still called for a large-scale stream processing for healthcare system.

#### VII. OPPORTUNITY

There are abundant of the database available to perform research work. However, the privacy requirement becomes a barrier in creating open databases. There is numerous opportunity in biomedical research. Moreover, there are numerous biomedical database available, namely, International Cancer Genome Consortium (ICGC) [43], National Institute of Health (NIH), The Cancer Genome Atlas (TGCA), National Cancer Institute (NCI), the National Human Genome Research Institute (NHGRI) [65], [66], CanSAR, NCI genomic data commons, International cancer genomics consortium, Cancer Genomics Hub, Pennsylvania Cancer Alliance Bioinformatics Consortium (PCABC) Biorepository, genome-wide association studies (GWAS) [41], Pancreatic Expression Database (PED), and METABRIC Repository etc [67], [50], [48]. There are also research opportunities in biocuration process, personalize medicine, cardiovascular care [68], MOOM [69], BigBWA [20], ENS@T-CANCER [70] and SUPER-FOCUS [71] etc. Moreover, there are endless opportunity to work on Biomedical Engineering as an engineer. Orthology research is also another important field in data-intensive computation [72]. There are numerous opportunity in the field of genome research. The massive, open, online, and medicine (MOOM) is used to detect the genetic cause of disorder [69]. Topol [69] target to collect nearly five millions individual genomes for sequencing, which will contribute petabytes of data. However, the collecting, storing and processing the genome data are the very time-consuming process [12]. The BigBWA is a Hadoop-based solution for Genome sequencing and alignment[20]. Moreover, the metagenomic shotgun sequencing is also another key research focus. The SUPER-FOCUS is an unannotated shotgun metagenome data sequencer [71]. Moreover, the GMQL is Hadoop-based query language for abstractions of genomic region data and associated experimental, biological and clinical metadata and interoperability between many data formats [21]. The GeneGrid provides a seamless integration of diverse data source to make strides of Bioinformatics research work [73].

The opportunity for cancer research is provided by The Cancer Genome Atlas (TGCA) for better research, treatment, and prevention. The TGCA is the most famous framework for cancer research which is a result of a joint collaboration of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) [65], [66]. In addition, more cancer-related data are found, namely, CanSAR, NCI genomic data commons, International cancer genomics consortium, Cancer Genomics Hub, Pennsylvania Cancer Alliance Bioinformatics Consortium (PCABC) Biorepository, genome-wide association studies (GWAS) [41], Pancreatic Expression Database (PED), METABRIC Repository etc [67], [50], [48].

#### VIII. BIOMEDICAL FUTURE

The future of biomedical imaging will be 3D or beyond. To visualize these types of images requires huge computational power on a massive scale. Moreover, most of the clinical data will be recorded in the form of digital data. Therefore, the database size will continue to increase. However, there is also Moore’s law. The Big Data will make an easier prediction, analysis, prevention, and curation. The cloud model also helps in reducing the cost of diagnosis, treatment, research and analysis, but vendor lock-in is a problem too. However, the future will be a very cost-efficient treatment of a disease, because the Biomedical will depend on data hosted on Big Data paradigm. In addition, the future of epidemiologist will require technical skills [57]. The knowledge of computer programming will be the sole criteria for the epidemiologist. There is a rare human resource of having a combination of science and computing knowledge for BBDE [74]. A new data scientist will be required for the BBDE. Most of the health care system will be transferred to Big Data for research, prediction, prevention, etc. Those systems will be a computer program to assist the doctors automatically. Furthermore, the cancer database will continue to grow. Therefore, the death due to cancer will continue to decline. Moreover, the prior prediction of cancer disease will be possible in near future. In the current scenario, the implication is data availability. However, this will not be an implication in the near future, but the technological singularities will arise.

#### IX. CONCLUSION

The Big Data and Biomedical data have a close relationship, and the merger forms a new paradigm, called BBDE. This is an opportunity for the research community to work on this useful research. The future of the BBDE depends on Big Data paradigm. The treatment, prevention, and curing process are done using Big Data technology. Moreover, there are many frameworks yet to be developed. The biomedical data are stored in digital form. Thus, the research on biomedical data become easier for Engineers and Doctors. BBDE has many barriers to overcome. Moreover, the advancement of Big Data technology creates an easier way to perform research work on BBDE. In addition, the Big Data enables Big Biomedical Data Analytics and BBDE. BBDE suffers from storage issues, and thus, deploying Big Data technology can solve the issue. However, the BBDE is emerging area for the research community and the WELFARE of human being.

## REFERENCES

- [1] A. K. Rider and N. V. Chawla, "An ensemble topic model for sharing healthcare data and predicting disease risk," in Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. New York, NY, USA: ACM, 2013, pp. 333:333–333:340.
- [2] Editorial, "The power of big data must be harnessed for medical progress," *Nature*, vol. 539, no. 7630, p. 467468, 2016.
- [3] W. Dunn, A. Burgun, M.-O. Krebs, and B. Rance, "Exploring and visualizing multidimensional data in translational research platforms," *Brief Bioinformatics*, vol. bbw080, 2016.
- [4] B. C. Desai, "Technological singularities," in Proceedings of the 19th International Database Engineering & Applications Symposium. New York, NY, USA: ACM, 2014, pp. 10–22.
- [5] N. I. Council, "Disruptive technologies global trends 2025. six technologies with potential impacts on us interests out to 2025. 2008." Accessed on 25 november 2017 from <https://fas.org/irp/nic/disruptive.pdf>.
- [6] S. Baker, W. Xiang, and I. Atkinson, "Internet of things for smart healthcare: Technologies, challenges, and opportunities," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2017.
- [7] V.-D. Ta, C.-M. Liu, and G. W. Nkabinde, "Big data stream computing in healthcare real-time analytics," in 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2016, pp. 37–42.
- [8] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.
- [9] B. C. Desai, "The state of data," in Proceedings of the 18th International Database Engineering & Applications Symposium. New York, NY, USA: ACM, 2014, pp. 77–86.
- [10] Z. Huang, E. Ayday, H. Lin, R. S. Aiyar, A. Molyneaux, Z. Xu, J. Fellay, L. M. Steinmetz, and J. P. Hubaux, "A privacy-preserving solution for compressed storage and selective retrieval of genomic data," *Genome Research*, pp. 1687–1696, 2016.
- [11] E. E. Schadt, "The changing privacy landscape in the era of big data," *Molecular Systems Biology*, vol. 8, no. 612, pp. 1–3, 2012.
- [12] M. M. Al Aziz, M. Z. Hasan, N. Mohammed, and D. Alhadidi, "Secure and efficient multiparty computation on genomic data," in Proceedings of the 20th International Database Engineering & Applications Symposium. New York, NY, USA: ACM, 2016, pp. 278–283.
- [13] N. A. Watts and F. A. Feltus, "Big data smart socket (bdss): a system that abstracts data transfer habits from end users," *Bioinformatics*, vol. 33, no. 4, pp. 627–628, 2017.
- [14] R. McClatchey, A. Branson, and J. Shamdasani, "Provenance support for biomedical big data analytics," in Proceedings of the 20th International Database Engineering & Applications Symposium. New York, NY, USA: ACM, 2016, pp. 386–391.
- [15] D. Laney, "Gartner predicts three big data trends for business intelligence," *Gartner*, 12, February, 2015, Retrieved on 10, December, 2016 from <http://www.forbes.com/sites/gartnergroup/2015/02/12/gartner-predicts-three-big-data-trends-for-business-intelligence/>.
- [16] R. Patgiri and A. Ahmed, "Big data: The v's of the game changer paradigm," in 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS). Sydney, NSW, Australia: IEEE, 2016, pp. 17–24.
- [17] A. Cuzzocrea, D. Sacca, and J. D. Ullman, "Big data: A research agenda," in Proceedings of the 17th International Database Engineering & Applications Symposium. New York, NY, USA: ACM, 2013, pp. 198–203.
- [18] C. Seife, "Big data: The revolution is digitized," *Nature*, vol. 518, no. 7540, pp. 480–481, 2015.
- [19] M. Bonenfant, B. C. Desai, D. Desai, B. C. M. Fung, M. T. Ozsu, and J. D. Ullman, "Panel: The state of data: Invited paper from panelists," in Proceedings of the 20th International Database Engineering & Applications Symposium. New York, NY, USA: ACM, 2016, pp. 2–11.
- [20] J. M. Abuin, J. C. Pichel, T. F. Pena, and J. Amigo, "Bigbwa: approaching the burrowswheeler aligner to big data technologies," *Bioinformatics*, vol. 31, no. 24, pp. 4003–4005, 2015.
- [21] M. Masseroli, P. Pinoli, F. Venco, A. Kaitoua, V. Jalili, F. Palluzzi, H. Muller, and S. Ceri, "Genometric query language: a novel approach to large-scale genomic data management," *Bioinformatics*, vol. 31, no. 12, pp. 1881–1888, 2015.
- [22] D. Bromley, S. J. Rysavy, R. Su, R. D. Toofanny, T. Schmidlin, and V. Daggett, "Dive: a data intensive visualization engine," *Bioinformatics*, vol. 30, no. 4, pp. 593–595, 2014.
- [23] C. B. Nielsen, H. Younesy, H. O'Geen, X. Xu, A. R. Jackson, A. Milosavljevic, T. Wang, J. F. Costello, M. Hirst, P. J. Farnham, and S. J. M. Jones, "Spark: A navigational paradigm for genomic data exploration," *Genome Research*, vol. 22, no. 11, pp. 2262–2269, 2012.
- [24] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.
- [25] Y. Karasneh, H. Ibrahim, M. Othman, and R. Yaakob, "A model for matching and integrating heterogeneous relational biomedical databases schemas," in Proceedings of the 2009 International Database Engineering & Applications Symposium. New York, NY, USA: ACM, 2009, pp. 242–250.
- [26] C. Lynch, "Big data: How do your data grow?" *Nature*, vol. 455, no.7209, pp. 28–29, 2008.
- [27] P. E. Bourne, J. R. Lorsch, and E. D. Green, "Perspective: Sustaining the big-data ecosystem," *Nature*, vol. 527, no. 7576, pp. S16–S17, 2015.
- [28] S. Maddineni, J. Kim, Y. El-Khamra, and S. Jha, "Distributed application runtime environment (dare): A standards-based middleware framework.

# Application of NLTK for Voice Based Sentiment Analysis

Pankaj Kumar Keserwani

Department of Computer Science  
and Engineering  
National Institute of Technology,  
Sikkim-737139, India  
pankaj.keserwani@gmail.com

Ashish Kumar Mishra

Checktronix India Private Limited  
Software Company Dharmapuri,  
Tamil Nadu, India.  
am67663@mail.com

Shetalika Ghosh Samaddar

Department of Computer  
Science and Engineering  
National Institute of Technology,  
Sikkim-737139, India  
shetalika99@yahoo.com

**Abstract**—Customer Relationship Management (CRM) is to gather feedbacks from customers for offline/online purchases they made or for services experienced by them. Information gathering plays a crucial role in efficiency analysis for deciding upon marketing and financial strategy. The sudden change of activity due to sentiment analysis is the computational treatment of sentiments and determination of subjectivity of text processed for determination of efficient system inputs that builds up efficiency analysis of a system. The paper contributes towards sentiment analysis of audio pieces that are purely voice based. The paper provides a treatment of sentiment analysis making use of tool; Natural Language Toolkit (NLTK). NLTK is a suite of libraries and programs for symbolic and statistical Natural Language Processing (NLP) for English, written in the Python programming language. NLTK is intended to support research and teaching in NLP or closely related areas, including empirical linguistics, cognitive science, artificial intelligence, information retrieval etc. The results obtained though NLTK are used in Decision Support System (DSS) for deriving in efficiency building strategy in finance, marketing, HR etc.

**Keywords**—Data Mining, NLTK, NLP, Bag of Words model, Neural Networks.

## I. INTRODUCTION

Consumers' perspective has always been important for client service mechanism. Decision making process [1] gets rightly influenced by such mechanisms. Reviews of users are collected time to time as in the format of text or in voice samples. For example, for various products, sales and related feedback systems of customer relationship management. These opinions are not personal as they are aggregated and open to all for making powerful and crucial decisions.

Following surveys conducted on more than 2000 American adults [2] [3]. The study reveals the following pertinent points:

- Online research on a product is a normalcy as far as Internet users are concerned. The percentage of users is as high as 81%. Moreover, 60% off them happen to be American making online research survey a cult for customer relationship management.

- The readers of online reviews of various services such as travel agency, medical services claim that their service

selection had a significant influence of these services and the percentage of such review users are between 73 to 87% for their selection of services worldwide.

- The services rating has considerable commercial impact on purchase or access to services. 22 to 99% of the users go for 5 star rated item rather than a lower star rated item irrespective of the price of the item.

- Due to such rating consideration a product/service/person is always associated with an online rating system. More than 30% of the items is associated with an online review or comments and the rating gives a comparison between the products of the same class reflecting the sentiments of the customer e.g trivago.com for hotel room rate comparison.

- The motivation behind considering rating and other related information are mainly for consumption of goods and services. However, strong sentiments are reflected for political review, lifestyle arrangement, attitudinal discourses, spiritual perception etc. The sentiment analysis in these cases goes beyond simple ratings/liking/acceptance/yes-no type of binary sentiments. The fuzziness of acceptance with appropriate preposition has also been depicted to the usage of the Likert scale [4] such as a agreeing strongly, general agreement, agreement on an average, not agreeing weakly, disagree completely etc. 78% of the Internet users are in the upper two categories on the preposition of convenience for shopping online. Similarly, 68% of the Internet users agree with the notion that online shopping is time effective.

- Every commercial activity on Internet is backed up by research about the product or service and 81% of Internet users have used Internet to generate such supporting information. The purchase of a product online is also called cost saving mechanism and 66% of online users are using such services. Similarly, 64% of Internet users used goods and services for travel and tourism.

- Even a commercial activity like online auction enhances its participation by deriving information from users' reflection. 26% of Internet users participated in an online auction backed up with research or analyzed information of sentiment of other users.

- In order to download digital content paid and unpaid services are rated by users' sentiments. 17% of the Internet users access or download digital content as a paid service. It may be conjectured that this lesser percentage of Internet users are giving importance to quality content rather than just getting the service as a freebie.

- Internet user's reliance on online advice/recommendation/rating by other unknown user is one of the pertinent point in case of sentiment analysis. The system of sentiment analysis use this identified sentiments as an object of usage. The objects are processed/studied from predefined objective with envisioned possibilities.

The following excerpt from a whitepaper is illustrative of the envisioned possibilities:

Web 2.0 platform encourages creation of objects such as blogs, discussion forum, peer to peer network, social engineering sites etc. The unprecedented reach and power by the users of such social media has given the opportunity for the Internet users to share their brand experiences, opinion poll, attitude reflection, agreement/review etc. on services, products or goods, opinion, issue (local as well as global) and folk news. Interaction through the social media has enormous influence in shaping the opinion of the other non- introduced consumers and the consumers knowingly or unknowingly build up their brand loyalties. The other users may change their purchase decision and in turn provides their own brand advocacy using sentiment objects. Consumer insights get percolated through social media monitoring their sentiments providing marketing messages, CRM feedbacks, brand positioning product activities etc. [5].

Any system of sentiment analysis needs to recognize the sentiment information for use and detection of pertinent sentiment. The object of sentiment may involve certain action for generation of sentiment information such as:

- Aggregation of opinion poll that are required to be registered on familiar scale (star system, Likert scale, letter grades, Fuzzy interpretation technique)
- Highlighting provision of some opinion based on predefined selection mechanism of reflected sentiment through an opinion.
- Identification of point of disagreement and points of consensus building
- Identification of relevant communities built up on the basis of sentiment of opinion holders
- Level of authorization among opinion holders for direction control and accounting

After the introduction of the field of problem, section II divulges different parameters of sentiment analysis system (SAS) in voice data mining and their prediction capability. Section III proposes the actual SAS based on various techniques already available. It is an intelligent assembly of the techniques that has been proposed. Experimentation to obtain maximum accuracy has been conducted. The proposed system also contains block diagram for the software used and

total experimental setup. Section IV is for concluding remarks and future direction of work of the present paper.

## II. BACKGROUND STUDY OF SENTIMENT ANALYSIS SYSTEM (SAS)

### A. (SAS) IN VOICE BASED SENTIMENT MINING

The introduction of sentiment analysis using opinion mining started in early 2001 and quickly spread its wings to incorporate research activity through various projects on identification of faith and believes, influence of opinion, prospective generation. It was used for taking conclusive action on the basis of opinion poll and processing through opinion mining. Even analysis of sentiments started to take its toll on share market, making stocks to climb up or down the ladder. Sentiment analysis require not only techniques for processing of sentiment objects but also requires linguistic actions on interpretation of metaphor, narratives, extraction of point of view, influence spread of a sentiment object in text etc.

Like any other system, sentiment analysis system starts with the definition of various objects used in the system and its processing due to diversity of application. A general agreement on terminology is missing all together. Moreover, a sentiment object include both structured as well as semistructured/unstructured data. According to the industrial estimates only 21% of the available data used in any sentiment analysis system is in structured form. Other data forms extracted from textual or audio/video data are at most semi structured if not fully unstructured.

Sentiment objects can be of anyone of the following:

- Pieces of information: tweets/posts on social networking portals
- Chat room conversation
- News reviews and opinion
- Blogs and visionary articles
- Product and services reviews and ratings
- Users/patient records in Healthcare sector

All these data are identified by various common and uncommon features giving the sentiment objects, a category of high dimension data. Any sentiment object with lack of commonality cannot deliver ready information unless it is processed manually or analyzed by an automated system. Involvement of textual voice and video data make the analysis system. An automated analysis system of linguistic based sentiment gets help from natural language processing (NLP).

Data preprocessing involves identification of objects, its format, size and other physical features contributing to analysis of static or dynamic data. Preprocessing contains predominantly following steps (Fig. 1.):

- Noise Removal
- Lexicon Normalization
- Object Standardization

Any piece of text which is not relevant to the context of the data can be specified as the noise. A general approach for noise removal is to prepare a dictionary of noisy entities, and iterate the text object by tokens (or by words), eliminating those tokens which are present in the noise dictionary.

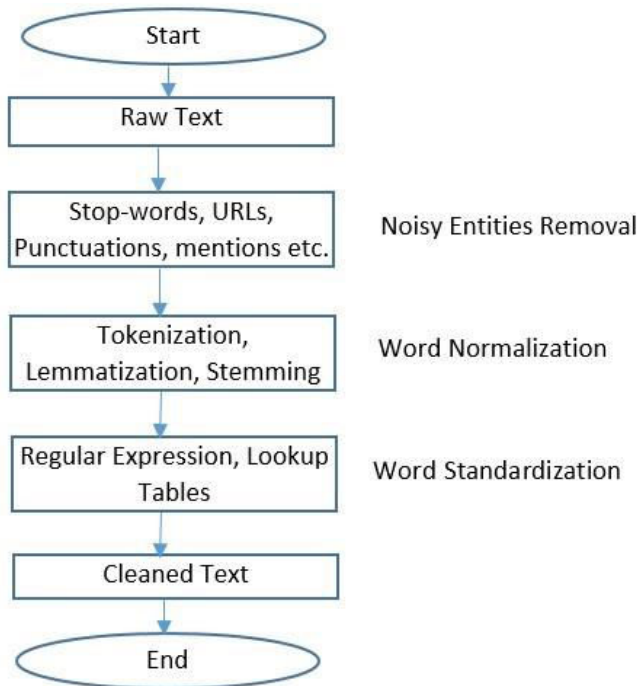


Fig. 1. Text cleaning Process

Another type of textual noise is about the multiple representations exhibited by single word. For example "play", "player", "played", "plays" and "playing" are the different variations of the word "play". The most common lexicon normalization practices are:

- **Stemming:** Stemming is a rudimentary rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word.
- **Lemmatization:** Lemmatization, on the other hand, is an organized and step by step procedure of obtaining the root form of the word. It makes use of vocabulary (dictionary importance of words) and morphological analysis (word structure and grammar relations) for processing.

Text data often contains words or phrases which are not present in any standard lexical dictionaries. These pieces are not recognized by search engines and search models or meta search engines.

Some of the examples are acronyms, hashtags with attached words, and colloquial slangs. With the help of regular expressions and manually prepared data dictionaries, this type of noise can be fixed or identified.

Social media comprises of a majority of slang words having different contextual meanings. These words should be transformed into standard words to make free text with single meaning considering the effect of context. The words like 'luv' will be converted to 'love', 'Helo' to 'Hello' etc. The

similar approach of apostrophe look up table can be used to convert slangs to standard words. A number of sources are available on the web, which provides lists of all possible slangs with or without context. Such sources can be used as lookup dictionaries for conversion purposes.

Grammar checking is primarily learning based, huge amount of proper text data is learned and models are created for the purpose of grammar correction. There are many online tools that are available for grammar correction purposes. In fact, all the document processing and document preparation system come ready with syntactic grammar checking and suggest various correction.

In natural language, misspelled errors are encountered. Companies such as Google, Microsoft have achieved a decent accuracy level in automated spell correction. One can use algorithms like the Levenshtein Distances, Dictionary Lookup etc. [6] or other modules and packages to fix these errors and get automated correction.

The variations of experiments in sentiment analysis makes the literature survey a difficult job. A main purpose of such analysis when applied to programming is software quality improvement of the proposed assembly of systems. The application of sentiment analysis on requirement analysis phase of software development has also been looked into. There are atoms to introduce static analysis framework for writing good quality Java programs [7]. A marking process has also been devised. Software engineering metrics are used for providing feedback to the students. Sentiment analysis in these prospective has been divided into two: synthetic and semantic analysis. The two has been used for automated testing against SQL injection. The technique has been named, Sania [8] for detecting SQL injection vulnerabilities in web applications devising the development and debugging phases. Sania is capable of intercepting the SQL queries with web application and database. Sania also compares parse trees of the intended SQL query. Sania has been tested in real life applications. Deep learning mechanism is used for textual analysis performing deep semantic analysis of text [9]. A graphical logical form as a semantic representation for understanding of the text has been described [10]. The system is able to capture rich semantic content. The semantic text processing has a parser at the core and augmented with the statistical pre-processing techniques and online lexical lookup as usual.

Semantic content has been extracted and analysed of existing software systems. The semantic information can be derived from comments, documentation and identifier names associated with the source code using information retrieval methods [11]. The paper happens to be a synopsis of doctoral dissertation of the author. Application of latent semantic analysis (LSA) to program source code and associated documentation has been presented [12] which is a corpus based statistical method for inducing and representing meaning of the words and passages. Applying LSA to the domain of source code and other internal documentation for the purpose of modular software reuse is a new dimension of sentiment analysis. Similarly, LSA has been applied for object recognition and clustering [13]. The method proposed has

been successfully overcome the problem of synonymy and polysemy. The method has been applied on indigenous data bases of objects from cases of commercial product having financial and marketing significance.

### III. PROPOSED SENTIMENT ANALYSIS SYSTEM

The sentiment analytics framework can be applied in many observatory systems, especially, in case of online shopping, digital marketing and opinion based decision systems. The data regulated and generated are again a feedback to the system producing required decisions and analytic support in realizing current trends of any or every context better. Basic structure can be understood from the figure below (Fig. 2.):

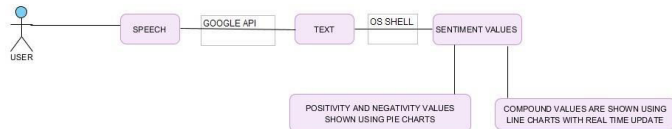


Fig. 2. Activity Structure

Sentiment indicates a view or opinion that is held or expressed to be conveyed or understood. Words such as happy are considered positive while hate is considered as negative. Sentiment analysis is an approach by which a machine tries to assess the sentiment conveyed in a source text. The number of machine learning methods in natural language processing and information retrieval have been in use from a long time. Cheaper and much more efficient.

Computing resources to process such data together are available and they convey meaningful sentiment on the basis of data aggregation and processing [7]. There are various tools and techniques that are applicable to formulate proposals in application of sentiment analysis. Some of the techniques, though not exhaustive, have been presented below:

i.NLP (Natural Language Processing): Natural Language Processing enables computers to derive meaning from humans or natural language input. It uses a variety of methodologies to decipher the ambiguities in human language. It answers questions like what are the keywords or parts of speech, or category of the document, given the categories or classes.

ii.Text Mining: Text mining derives relevant, novel and interesting information from the text that are not obvious from the given dataset. It provides a metadata (describes data for further work on). The goal of text mining is to discover relevant information in text by transforming text into data that can be used for further analysis. Text Mining answers questions such as frequency of words, length of sentences, or presence or absence of certain words in a sentence and/or a paragraph.

Opinion Mining uses NLP, text mining and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. It can be used for evaluating subjective opinions.

iii.Models for evaluation: There are two popular approaches for sentiment analysis:

a) Bag of words model: [14] Bag-of-words model focuses completely on the words, or sometimes a string of words, but usually pays no attention to the "context" so-to-speak. The bag of words model usually has a large list, probably better thought of as a sort of "dictionary," which are considered to be words that carry sentiment. These words each have their own "value" when found in text. The values are typically all added up and the result is a sentiment valuation. The equation to add and derive a number can vary, but this model mainly focuses on the words, and makes no attempt to actually understand language fundamentals. It can be further understood by figure below (Fig. 3.):

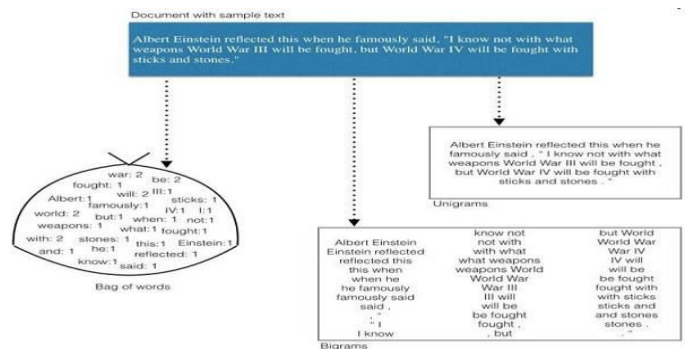


Fig. 3. Bag of words model

b) Natural language processing model: This model attempts to have the machine actually understand the sentences structures, context, and is more focused on the succession of a string of words. Usually, this structure requires the machine to have understanding of grammar principles. To do this, Natural Language Processing (NLP) techniques are used to tag parts of speech, named entities, and more, in order to actually understand the "language" of the text, and not just look for target words. It also uses WordNet [15], where similar words used alternatively for each other (not necessarily synonyms), are grouped together (Fig. 4.).

#### A. Experimental Set-up of the Proposed SAS (sentiment analysis system)

The system SAS has been created by assembly of tools and techniques running consecutively so that a result can be obtained for use and analysis. The system has been represented through following steps:

a) Key files used in the project: Python language has been used for the development of snippets of programs that has been used intermittently with the tools already

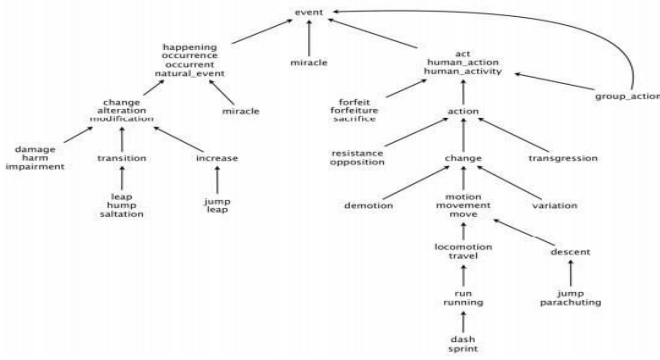


Fig. 4. WordNet

available such as NLTK. Django framework is a free and open source integrated development environment for python, a programming language for computing and programming.

- `settings.py`: Contains the project settings including directories containing templates, static files (like HTML, CSS and images) and database options.
- `models.py`: This file contains the database models which are used in the project.
- `urls.py`: This is responsible for URL routing. It uses regular expressions to match URLs to the page which contain the relevant information.
- `views.py`: `urls.py` decides the function to call on getting a particular URL, `views.py` contains the actual function which needs to be called. Functions in Django can display a web page or be used as an API endpoints which can receive certain arguments using GET or POST and process those arguments and return a result. In the present case, since the data which are processing is not assumed to be sensitive, GET method is used but POST can be used just as well. The predictive algorithm which analyses the sentiment is presented here as well in the form of an API endpoint.
- `tests.py`: Unit tests are written here or style sheets.

1) `static` (directory): This contains the static files which are not changed and are used repeatedly, such as images which are used as icons.

2) `templates` (directory): This contains the webpages which are displayed along with their styling.

The web application takes speech from the user as input. Third party API, Google WebSpeech translates it to text. The text is sent using an AJAX query to a Django API which has the algorithm to extract the sentiment from it using Natural Language and Toolkit (NLTK) [16]. The result is sent as a JavaScript Object back to the query which initiated the process. Finally, the results are displayed in real time using Chart.JS and the result is updated every few seconds without navigating away from the page. Following technologies were required to build this product:

- **Python**: This is the core of the project. This has been used as a backend using Django framework as well as for text mining program and NLTK library.
- **JavaScript**: This is responsible for animations as well as to render charts and converting speech to text using Google WebSpeech API.
- **HTML and CSS** For designing and styling parts of the output.

### B. IMPLEMENTATION OF SENTIMENT ANALYSIS SYSTEM

The application developed is to showcase the usage of sentiment analysis. The application is being run on the said platform and the reaction of the web application is captured using screenshots that have been presented in Fig. 5. and Fig.6.

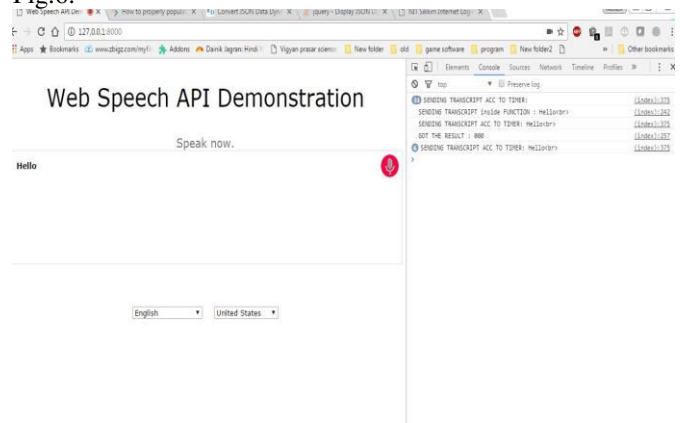


Fig. 5. Waiting for User Input



Fig. 6. Implementation Results

Sequence Diagram of proposed system is depicted as below(Fig. 7.):

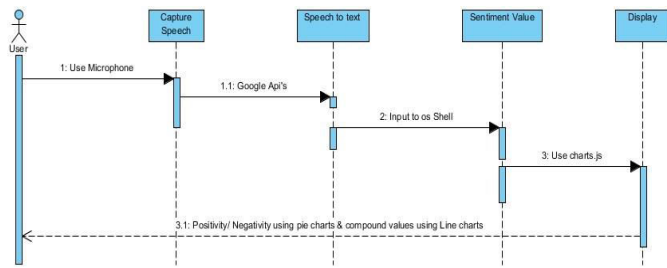


Fig. 7. System Sequence Diagram

The link of the working scenario is provided in the youtube link as: [https://youtu.be/Oz\\_S8hd4x4](https://youtu.be/Oz_S8hd4x4)

#### IV. PERFORMANCE AND COMPARATIVE ANALYSIS

An article about illegal poaching in Africa is read aloud. The accuracy in translation of speech to text is limited by clarity of the microphones being used. The result is mostly accurate up to a certain level. The pie chart shows relative positive and negative sentiments expressed in the text. The line chart labeled “Net Sentiment” is the overall sentiment expressed in the text. It ranges from -1 to 1. If it is above 0, it indicates a positive sentiment and if it is below 0, it indicates a negative sentiment. A number close to 0 indicates a neutral sentiment. A number close to +1 or -1 indicates an extreme sentiment, either very positive or very negative.

Both the charts are updated in real-time with changes in text. The result obtained shows the prospect of the application after further modification on limitation.

#### V. CONCLUDING REMARKS AND FUTURE DIRECTION OF WORK

##### A. LIMITATION

The algorithm for text analysis fails to recognize sarcasm. Although, it is accurate at most of the times, a lot more can be done before it can classify the sentiment inferred from the text as efficiently as human beings can. It requires a fast internet connection for translating speech to text using Google WebSpeech API. Hence, it cannot work offline on a local server. This can be remedied with an offline speech to text converter. Since, the backend uses Django Web Framework, it is quite heavy in a bit on the slower side. While it makes the application robust and secure, it is not as fast as one would like to have for outline use.

##### B. CONCLUDING REMARKS

The work carried out shows the prospect of sentiment analysis in web application. The percentage of accuracy maintained or better yet, improved in such case. Future direction of work will go towards a definitive project of

handling a situation that involves sentiment and expression of a multi-cultural, multi religious crowd. The work can further be taken up for patient easing and monitoring based on the stakeholders input.

#### REFERENCES

- [1] B. Pang, L. Lee *et al.*, “Opinion mining and sentiment analysis,” *Foundations and Trends R in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [2] “Online consumer-generated reviews have significant impact on offline purchase behavior - comscore, inc,” <https://www.comscore.com/Insights/Press-Releases/2007/11/OnlineConsumer-Reviews-Impact-Offline-Purchasing-Behavior>, (Accessed on 07/31/2017).
- [3] J. A. Horrigan, “Online shopping,” *Pew Internet & American Life Project Report*, vol. 36, pp. 1–24, 2008.
- [4] “Likert scale — simply psychology,” <https://www.simplypsychology.org/likert-scale.html>, (Accessed on 07/31/2017).
- [5] J. Zabin and A. Jefferies, “Social media monitoring and analysis: Generating consumer insights from online conversation,” *Aberdeen Group Benchmark Report*, vol. 37, no. 9, 2008.
- [6] R. Haldar and D. Mukhopadhyay, “Levenshtein distance technique in dictionary lookup methods: An improved approach,” *arXiv preprint arXiv:1101.1232*, 2011.
- [7] N. Truong, P. Roe, and P. Bancroft, “Static analysis of students’ java programs,” in *Proceedings of the Sixth Australasian Conference on Computing Education-Volume 30*. Australian Computer Society, Inc., 2004, pp. 317–325.
- [8] Y. Kosuga, K. Kono, M. Hanaoka, M. Hishiyama, and Y. Takahama, “Sania: Syntactic and semantic analysis for automated testing against sql injection,” in *Computer Security Applications Conference, 2007. ACSAC 2007. Twenty-Third Annual*. IEEE, 2007, pp. 107–117.
- [9] J. F. Allen, M. Swift, and W. De Beaumont, “Deep semantic analysis of text,” in *Proceedings of the 2008 Conference on Semantics in Text Processing*. Association for Computational Linguistics, 2008, pp. 343–354.
- [10] C. Cattuto, D. Benz, A. Hotho, and G. Stumme, “Semantic analysis of tag similarity measures in collaborative tagging systems,” *arXiv preprint arXiv:0805.2045*, 2008.
- [11] A. Marcus, “Semantic driven program analysis,” in *Software Maintenance, 2004. Proceedings. 20th IEEE International Conference on*. IEEE, 2004, pp. 469–473.
- [12] J. I. Maletic and N. Valluri, “Automatic software clustering via latent semantic analysis,” in *Automated Software Engineering, 1999. 14th IEEE International Conference On*. IEEE, 1999, pp. 251–254.
- [13] V. Hebballi and V. Rojit, “Latent semantic analysis (lsa) based object recognition and clustering,” in *Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on*. IEEE, 2015, pp. 416–421.
- [14] “Part 1: For beginners - bag of words - bag of words meets bags of popcorn — kaggle,” <https://www.kaggle.com/c/word2vecnlp-tutorial/details/part-1-for-beginners-bag-of-words>, (Accessed on 07/31/2017).
- [15] “About wordnet — wordnet - about wordnet,” <https://wordnet.princeton.edu/>, (Accessed on 07/31/2017).
- [16] “Natural language toolkit nltk 3.2.4 documentation,” <http://www.nltk.org/>, (Accessed on 07/31/2017).



# Extracting the features of Medicinal Plants and improving the accuracy of classification using Fuzzy Local Binary Pattern Approach

M.Srinivas

Assistant Professor, Dept. Of IT, CVR College of Engineering,  
Ibrahimpatan (M), R.R Dist-501510,  
Telangana  
msrinivasremote1975@gmail.com

**Abstract**—In this paper, a method to extract the features of leaf texture of medicinal plant leaves with Fuzzy Local binary patterns (LBP\_F) approach is proposed. The extracted feature data set is reduced with numerical approximation approach. To evaluate the approach, Flavia data set is chosen and experiments are conducted on this data set to classify the medicinal plant species. The classification results obtained are more encouraging and the results obtained with the proposed approach are better than the conventional LBP approach.

**Keywords**—Texture, Leaf Classification, Local Binary Pattern, Flavia data set

## I. INTRODUCTION

Ayurveda is the science of medicinal plants originated from India. It is described from the ancient ages that Ayurveda is the transmission of knowledge from Gods to sages and then to human physicians. Researchers considered the Ayurveda as protoscience and it will balance both the physical and mental conditions. Machine learning techniques will be applied to identify ayurvedic medicinal plants and classification of medicinal plants will play a major role to design an expert system.

Herbarium in the digital form is available online to provide the information about plant species. Scientists possessing explicit knowledge on plant taxonomy finding difficult in identifying plant species. For a given text based query, information from virtual herbarium databases can be obtained but they are not accurate. Hence there is a need for computer vision experiments to obtain the information based on image for a given text query. Plant species can be identified from the properties like colour, shape and texture. Texture plays an important role in classifying plant species, apart from shape and colour.

## II. LITERATURE REVIEW

First, Several authors have proposed data mining techniques to classify the plant leaves based on their shapes and surface textures, but texture plays an important role in identifying the properties of plant leaves for similar shapes.

Wang et.al. has identified the plant leaves based on the properties like eccentricity form factor, perimeter ratio of convex hull, circularity, rectangularity and aspect ratio to describe the local features of a plant leaf for classification and

recognition [2]. Stephen et.al. has evaluated physiological width and physiological length of a plant leaf based on the geometry of the shape and classified the plant species with morphological features and made it online with a huge data set of plant leaf images [3]. Zeng et.al. has classified the plant leaves based on shape descriptors using the Histogram of gradients [10].

Several mathematical models have been developed for object recognition and classification of artificial textures [1]. Feature distribution models are popular as compared with texture models in describing the texture pattern because of its theoretical simplicity. To determine the uniform patterns in texture a generalized gray-scale rotation invariant Local Binary Pattern operator(LBP) of any quantization of the angular space has been proposed[6]. Further, Local Binary Pattern operator is enhanced by combing the operators with varying spatial resolutions for multi resolution analysis of different surface textures. A new biometric system has been proposed to extend the concept of LBP with varying radius to determine the finger print images[7]. Finger print images are decomposed into overlapping sub images and each sub image is convoluted with Gabor filter to identify the local features for classifying the LBP features of convoluted sub images. Alexander et.al. has proposed a pixel pattern texture feature vector (PPBTF) method for gender recognition[8]. To obtain the local features, a pattern map is constructed with the geometry of shapes of edges and lines. Principal Component Analysis technique is applied to evaluate the eigen values of the feature data set and developed an expert system for gender classification.

It is observed from the literature review that several methods are exposed to classify the plant species and tried to develop an expert system for recognizing the plant species. Most of the models are explained based on the shape features for classification of plant leaves to describe plant taxonomy. The classification models described based on shape features will vary from one plant species to another and these models are not able to explain for similar shapes of plant leaves for two different species. To differentiate plant species with similar shapes, texture models to capture leaf texture is important for classification. With this backdrop, a Fuzzy threshold based LBP numerical approximation approach is proposed for the classification of plant leaves based on texture features.

In this paper, a method of extracting the features of different plant species using leaf texture properties is explained. Fuzzy threshold local binary pattern algorithm is proposed to extract the texture features of plant leaves. Further numerical approximation using Simpsons one-third rule is exploited to refine the feature representation and interval valued symbolic approach is applied for classification [9].

In section III, basics of LBP operator are illustrated to understand the proposed Fuzzy threshold based LBP algorithm. To refine the Fuzzy LBP histogram features, numerical approximation using Simpson's one-third rule is applied to the extracted features data followed by interval valued symbolic approach representation technique for classification. In section IV, classification results obtained by the Fuzzy threshold based LBP algorithm are discussed and compared with the conventional LBP method. In Section V, results are concluded with further directions in improving the classification.

### III. METHODOLOGY

The proposed methodology derives the Fuzzy threshold based LBP algorithm to capture leaf texture properties and these extracted features are refined through numerical approximation technique to reduce the number of features to provide effective description for plant leaf texture are explained. Also, the proposed methodology exploits the concept of interval valued symbolic approach for effective representation of texture properties. The following subsections provide a detailed description of proposed methodology.

#### A. Basics of LBP Operator

In [6], non-parametric LBP operator based on local neighbourhood around a pixel to represent the texture was proposed. Further, LBP operator is combined with different spatial resolutions to form a new operator Local Binary Pattern Variance (LBPV) which is rotation invariant to explain uniform local binary patterns for multi-resolution gray-scale texture classification [6].

Given an image to describe texture properties, Local Binary Pattern operator with P equally spaced pixels and with uniform radius R from the centre pixels which acts as threshold is computed for all the pixels in the image as follows:

$$LBP_{P,R} = \sum_{i=0}^{P-1} S(x) 2^i \dots\dots\dots (1)$$

$$x = G_P - G_C \dots\dots\dots (2)$$

In (2)  $G_P$  and  $G_C$  are the gray values of the neighbourhood pixel and centre pixel which act as threshold respectively. The LBP code is evaluated for all the pixels of the image and LBP codes are obtained with possible values of  $2^P$ . Finally, similar values of LBP codes are binned together to form a LBP histogram which is used to describe the texture properties.

#### B. Threshold based LBP

The procedure followed to evaluate Fuzzy threshold based LBP (LBP\_F) is similar to Local Binary Pattern approach described in section 3.1. In LBP\_F, the binary value of each pixel is represented based on fuzzy threshold which is computed from sigmoid function. Membership value is assigned to each pixel with either 0 or 1 by comparing the neighbouring pixel with centre pixel. The algorithm designed for LBP\_F was implemented to extract the texture features of leaf image. The inputs I, n, and R refer to an input image, neighbourhood size, and radius respectively.  $Diff_j$  is the difference between  $j^{th}$  neighbour  $p_j$  and centre pixel  $p_c$ . Norm is a mean of differences,  $Diff_j$  and  $\mu_j$  is the membership value obtained with the norm of differences. Fuzzy threshold(FT) is evaluated using sigmoid function which is obtained from the ratio of  $Diff_j$  and  $\mu_j$ .

#### Algorithm: Fuzzy Threshold LBP (LBP\_F)

1. While not end in image I
  2. Choose Window  $W_i$  from image I with radius R.
  3. Compute  $Diff_j = p_j - p_c$
  4. Compute the sum of differences, call it as  $Diff_j$  for 'n' iterations
  5. Calculate the Norm of the differences
  6. Obtain the ratio  $\mu_j = \frac{Diff_j}{Norm}$
  7. Obtain the fuzzy threshold with the sigmoid function
- $$FT = \frac{1}{1 + \exp(-\mu_j)}$$
- given by
8. Calculate the LBP\_F similar to the LBP procedure defined in equation (1)
  9. End while
  10. Explore with different values of R.

#### C. Numerical Approximation of Fuzzy LBP histogram

The LBP histogram obtained through Fuzzy LBP operator to capture leaf texture can be further refined to reduce the feature set through numerical approximation. The Fuzzy LBP histogram features is divided into equal sized intervals and a polynomial of suitable order is chosen for each interval to fit the data. In the proposed work, the Fuzzy LBP histogram data bins with equal sized intervals are derived and for each bin the data is fitted with the corresponding polynomial. A polynomial is a function expressed as follows:

$$Y = P(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n \dots\dots\dots (3)$$

for some coefficients  $a_0, a_1, a_2, \dots, a_n$ . If  $a_n = 0$  the polynomial is an equation of order n. Based on the values of n, the equation represents a straight line (n = 1), quadratic (n=2) simulating parabola, cubic (n=3) and so on with n + 1 unknowns, being x as an independent variable and y as a dependent variable and constant coefficients  $c_i$ . Constructing a polynomial arises from one (x, y) data point and the hypothesis about the order of the polynomial shall govern the process. Further the values for those coefficients  $c_i$  are computed. Fig. 1 shows the best polynomial for a sample of 10

histogram data bins representing histogram features which is chosen from the first ten histogram data bins.

From the sub plots, it is observed that 5<sup>th</sup> order polynomial suits for the data presented above and this polynomial is chosen for our experiments to obtain better performance than the conventional LBP features data. After constructing the 5<sup>th</sup> order polynomial, the area is computed using Simpson's one-third rule, because it is accurate when compared to other numerical methods like Trapezoidal rule, Midpoint rule, and left or right approximation using Riemann sums on the basis of error bounds of these approximations.

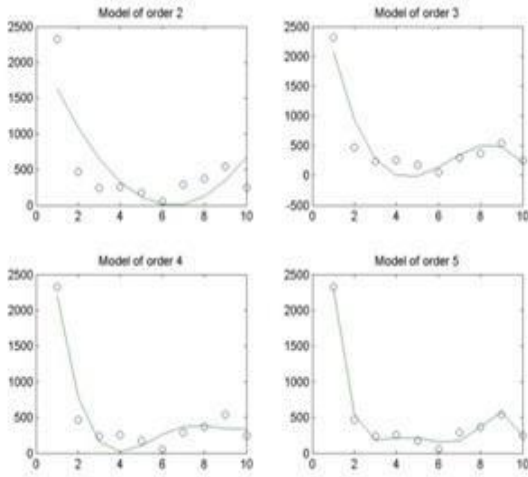


Fig. .1. Polynomial curves fitting points generated for a sample of 10 data points

To refine the data, the area is computed with Simpson's one-third rule by dividing the interval in to sub intervals of same length.

Fig. . 2 shows the LBP histogram for the first leaf and Fig. . 4 shows samples of the model plant leaves provided by[3]. Fig. . 3 shows the refined histogram data with numerical approximation to compare with the LBP histogram shown in Fig. . 2.

#### D. Feature Representation

The intra-class variations for different surface textures of the leaf will vary due to different maturity levels. To capture these variations a method is chosen to select the multiple representatives for a class by combining similar leaves into one cluster and choose a interval representative for the particular group within a class.

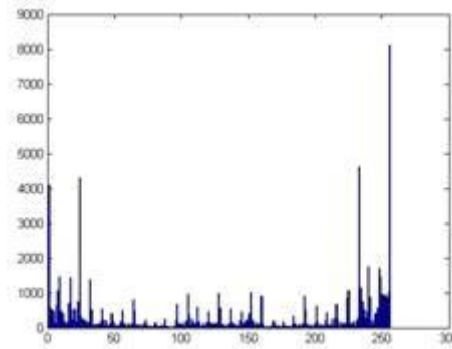


Fig. 2. LBP histogram with 256 bins

Further the data has been reduced to with a considerable size in order to classify the system properly. With the reduced data, an interval valued feature representation is applied so that the classification accuracy can be improved in near future. Because in the reduced data the features will not be lost as compared with the original data as the data is under control of the user in developing the system. It is observed that the reduced data without loss of information will provide better classification results with an improved accuracy for any system of importance in nature. The reduced data histogram where the area is calculated using Simpson's one-third rule is provided in Fig. . 3 and it is observed that the gaps are filled and better histogram features are obtained.

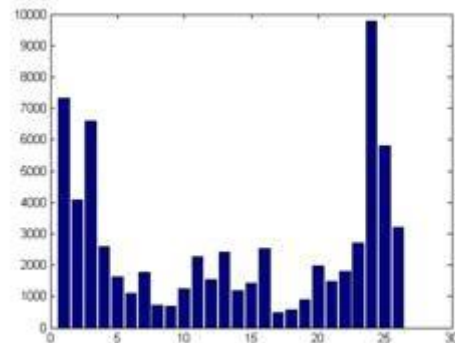


Fig. . 3. Area under the polynomial for every 10 bins

#### IV. EXPERIMENTAL RESULTS

To validate the proposed algorithm, the Flavia data set is chosen which is available online[3]. All the plant species in the data set are recognized as medicinal plant species which contains medicinal values. The database consists of 32 plant species. The features extracted data set histogram is shown in Fig. . 2 and the refined data histogram using numerical approximation is shown in Fig. . 3.

To conduct experiments, 60 samples from each species by varying the ratios of training and testing samples is chosen randomly to design an expert system as discussed. In Table. 2 results are shown with two different sets of experiments with varying training and testing ratio of samples for each species of the proposed algorithm.

Table. 2 shows the average accuracy obtained for varying training and testing ratios with Fuzzy LBP symbolic approach. The code provided by [6] is utilized for extracting the basic LBP features for classification with nearest neighbourhood classifier as chi square distance measure.

TABLE 1. CLASSIFICATION ACCURACY WITH NN CLASSIFIER OF LBP FEATURES

Ratio of Training and Testing	Average Accuracy
60:40	82.6563
50:50	77.8123

TABLE 2. CLASSIFICATION ACCURACY WITH NN CLASSIFIER OF FUZZY THRESHOLD BASED LBP FEATURES

Training and Testing Ratio	Mean Accuracy
60:40	97.7163
50:50	95.1123



Fig. .4. Samples of the model plant leaves

It is clear that the results obtained with the Fuzzy LBP representation symbolic approach shows better performance when compared with the LBP features extracted with NN classifier shown in Table. 1. For experiments, 5<sup>th</sup> order polynomial is considered for approximation to compute the area under the polynomial and normalized them so as to make them invariant to the size of the image.

## V. CONCLUSION

In this paper, a Fuzzy Threshold Local Binary Pattern approach to extract the features and classifying medicinal plants is proposed. A Numerical approximation technique using Simpsons one-third rule is utilized for refining the Fuzzy LBP histogram features for better representation. For, experiments Flavia data set is chosen which is available online[3] and the results obtained are encouraging and comparable with different versions of LBP techniques. There is a wide scope for exploring the suitability of soft computing techniques for texture description and classification.

## REFERENCES

- [1] Jianguo Zhang and Tieniu Tan, "Brief review of invariant texture analysis methods", *Pattern Recognition*, 35:47, 2002.
- [2] Xiao-Feng Wang Ji-Xiang Du and Guo-Jun Zhang, "Leaf shape based plant species recognition," *Applied Mathematics and Computation*, 185:883(2007).
- [3] Stephen Gang Wu, Forrest Sheng Bao, Eric You Xu, Yu-Xuan Wang, Yi-Fan Chang and Chiao-Liang Shiang, "A Leaf Recognition Algorithm for Plant classification Using Probabilistic Neural Network," *International Symposium on Signal Processing and Information Technology*, ,IEEE, 2007, Cario, Egypt.
- [4] Xiao, X., Hu, R., Zhang, S., Wang, X.: "HOG-Based Approach for Leaf Classification," In *ICIC*, 149-155(2010)
- [5] R. Makwana, V. Thakar, and N. Chauhan, "Fuzzy Threshold based Local Binary Pattern for Illumination Invariant Face Recognition", *International Conference on Signal System and Automation*, GCET, V.V. Nagar, January, 2011.
- [6] T. Ojala and Matti Pietikainen, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 24:87, 2002.
- [7] Nanni L A Lumini, "Local binary patterns for a hybrid fingerprint matcher". *Pattern Recognition*, " 41(2008) 3461:3466.
- [8] Alexander Souto Martinez, Andre Ricardo Backes, Wesley Nunes Goncalves and Odemir Martinez Bruno, "Texture analysis and classification use deterministic tourist walk." in *Pattern Recognition*, 43:94, 2010.
- [9] Bock, H.H., Diday, E.: "Analysis of Symbolic Data," Springer-Verlag.(2000).
- [10] Xiang-Yan Zeng, Yen-Wei Chen, Zensho Nakao, Hanqing Lu., "Texture representation based on pattern map," *Signal Processing*, 84(2004), 589:599.
- [11] Nagendraswamy H.S., Naresh Y.G, M.Srinivas Classification of Medicinal Plants using LBP: A Numerical Approximation approach, *ICSIP-2013*.

# A New Heuristic Similarity Model to improve the Accuracy for Data Sparsity issues in CF Recommendation Algorithms

N H Mohana

Ph.D Research Scholar, Department of Computer Science,  
Dr Ambedkar Government Arts College (Autonomous),  
Vyasarpadi, Chennai – 39  
Vkmohana85@gmail.com

Dr M Suriakala

Assistant Professor, Department of Computer Science,  
Dr Ambedkar Government Arts College (Autonomous),  
Vyasarpadi, Chennai – 39  
suryasubash@gmail.com

**Abstract**— Ecommerce is a growing proliferation in today's online trading industry. One of the greatest success behind the ecommerce world is said to be predicting the users preferences in advance by examining the users past records called 'recommender system'. Collaborative Filtering (CF) is the most prominent approach in generating recommendations. CF matches the "like – minder" people those who have similar taste in buying an item. The main objective of this paper is to make accurate recommendations to the like-minder people by user – item rating matrix factorization method. So that the system can perform a reliable recommendation to the target user. This paper address the problem of data sparsity, accuracy by framing a new similarity model called Tuned cosine (TCOS) using cross product reference which enhances the recommendation performance when only few ratings are available. To illustrate the TCOS similarity model a sample training dataset for user – item rating table was generated along with missing data. There are several predefined user similarity algorithms such as Pearson correlation coefficient (PCC), Cosine, Jaccard, MSD etc. Unfortunately, these algorithms are not much effective in several circumstances especially data sparsity and cold start problem. A traditional predefined similarity model and the newly proposed similarity model were applied in to that rating table and thus the results are compared to show that, the resultant proposed TCOS similarity model works well in data sparsity limitations while comparing with the predefined one. Its ensured that the proposed new similarity model makes the reliable recommendation to the target users or like - minder users.

**Keywords**— E-Commerce, CF, user – item rating matrix, PCC, Cosine, Jaccard, TCOS.

## I. INTRODUCTION

In an internet world, E- commerce industry plays a vital role in the WWW. E-commerce websites becomes an unbeatable one and it shows a rapid development in it. Enterprises mainly focuses on rebuild the relation with their old customer and simultaneously focussing on new customer [1] by recommending a new and trendy product to the customer those who have similar taste in the past using recommender system. Recommender system plays a vital role in today's online E-Commerce industry. CF is the most widely used and successful method in recommending a product or an item to the end user. The key feature of this technique is to

find a similar user by applying several similar measures on the user - item rating matrix [2]. CF recommendation algorithm is the backbone to the ecommerce world. The CF includes memory based and model based methods [3]. In memory based method CF recommendation is possible for user based (UB) and item based (IB) called UBCF and IBCF method. For UBCF, the algorithm searches for a similar user with the active user, for IBCF, the algorithm searches for a similar item with the items which are rated by the active user. The memory based method first calculates the similarities among users and then selects the nearest neighbour of the target user. Finally, it recommends a product or an item accordingly to the nearest neighbours. However, the model based method first construct a model to describe the behaviour of users and therefore, to predict the ratings of an item [3]. This paper focuses on memory based CF recommendation algorithms. Pitfalls occurs in CF are data sparsity, poor scalability, cold start problem, user similarity is hard to distinguish. The main objective is to calculate the similarities among users or an item when the data's are sparse. Comparing the predefined similarity algorithms with my tuned cosine similarity algorithm. The predefined similarity algorithms are pearson correlation coefficient, cosine, jaccard similary measures, MSD mean square difference etc. besides these predefined similarity algorithms an improved similarity algorithms were proposed by several authors like weighted pearson coefficient[5], constrained pearson coefficient (CPCC)[6], sigmoid person coefficient (SPCC)[7], Adjusted Cosine (ACOS)[8] etc. The pre defined similarity algorithms have some drawbacks in calculating the user similarity index (USI). The following section will briefly explain the limitations of the predefined similarity algorithms. Therefore to overcome the limitations a new heuristic tuned similarity formula was proposed and it gives a better results comparing to the predefined one. In order to test and verify a sample training dataset for user – item missing rating table was generated and applied in to the predefined and tuned similarity formula. Finally the result was compared and analysed. The resulting matrix factorization records that the new trending formula gives a better recommendation performance than the predefined one. This paper focuses on how recommendation takes place? Using what algorithms, what are all the existing algorithms are there and their pitfalls, how to overcome the pitfalls using what

technique. These are all the things which we are going to concentrate in the upcoming sections. Though everything is come to an online era, all trading industries were widely connected to the E-Commerce world. Once if they become a registered users. Fortunately, the huge volume of user's information's were tracked and saved in to the database like users click stream, browsing behaviour, scale of rating etc. Here forth, the recommender system works out well to recommend a product or an item and to provide a personalised service to the target user

The need of improving the data sparsity issues in CF recommendation algorithm is to upgrade the precision of predicting the users missing ratings. So, higher the precision leads to well grounded Top N recommendation to the target user. Thus the online commercial ecommerce industry acquire a territorial net profit, it enhance the marketing strategy in a booming way, improves the sales frequency range, well equipped customer support system, targets the company's bottom-line and increases the profitability.

## II. RELATED WORK

Saranya k[2] et.al shows an improve modified heuristic similarity measure for personalization using CF technique. She analysed a new similarity model for data sparsity problem. Her main goal is to improve the prediction of the nearest neighbour by combining the local context as well as global preferences of the user's behaviour. Haifeng liu[4] et.al first analyse the drawbacks of the predefined similarity models. Then he proposed a new heuristic similarity model(NHSM) in order to improve the accuracy of CF. Finally he made a comparative study between his NHSM model with the predefined one. Experiments were carried out in to two real datasets and his result demonstrate the effectiveness of the NHSM model. M.Jamali[7] et.al analysed one of the common limitations occurs in PCC was if both the users have rated more common items then the similitaty will be more credible. So that a sigmoid function based PCC (SPCC) was proposed.

H.J.Ahn [8] et.al focuses the limitations exist in the traditional cosine similarity formula will not consider the user preference rating. For considering the preference of users ratings he introduced a adjusted cosine(ACOS) similarity formula .Leily sheugh[9] et.al focuses on the PCC as a metric of similarity in recommender system. The limitations taken by him is variance of rating is needed to calculate the PCC in denominator fraction but the measure becomes zero , so therefore a new modified version of PCC is proposed and he proved that the volume of useful information has been increased potentially and it will create a suitable platform for rising the recommender system. J.Bobadila[10] et.al proposed a combination of jaccard as well as mean squared difference whose metric will complement each other for a new user cold start problem. Chenyang Li [11] et.al proposed a model based CF recommendation algorithms for an item which is integrated with empirical analysis. He evaluates the similarities between two items and his similarity measure helps to predict the ratings for the user's scalability problem were addressed and shows better result by optimizing the map reduce for IB collaborative algorithms. Mengsi Liu [12] et.al developed a novel recommendation algorithm called pairwise

factored mixed similarity model (P-FMSM) focusing on recommending a product or an item using implicit feedback rating. His model captures the locality of user – item interactions which correlates in to the global context.

Manu M.N [13] et al implemented CF recommendation algorithms using apache mahout a big data platform results in a better scaling capability in handling the huge volume of datasets compared to the other data mining tools. Asma sattar [14] et al introduced a novel hybrid recommendation framework which integrates the content based CF algorithms with CF to overcome the data sparsity and cold start problem. His proposed hybrid approach performs well than the individual content based approaches and naïve hybrid approaches. Jaime rai goza [15] et.al implemented a predefined CF similarity algorithms in cloud based environment called software as a service (SAAS). The author produces a list of recommended movies depends on the users browsing behaviour. CF in cloud based environment outputs a more reliable and personalizes movie recommendation to the target userirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the file "MSW\_USltr\_format".

## III. A COMMON FRAMEWORK FOR CF RECOMMENDER SYSTEM[16]

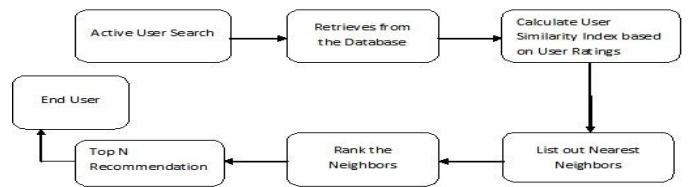


Fig (1). CF Framework [1]

### a) Predefined of Similarity methods

The traditional predefined similarity methods results in evaluating that how far the two users were correlated with each other. In this section we first analyse the existing similarity methods and its drawbacks. In CF pearson correlation coefficient (PCC), Cosine and Jaccard similarity measures are the most widely used methods in recommender systems. Whereas, CPCC [6] and ACOS [8] similarity measures were proposed by several authors and derived from the predefined one. Thus the defined formula evaluates that how x and y are correlates each other. The formulas are defined as follows.

$$sim(x, y)^{PCC} = \frac{\sum_{p \in I} (r_{x,p} - \bar{r}_x)(r_{y,p} - \bar{r}_y)}{\sqrt{\sum_{p \in I} (r_{x,p} - \bar{r}_x)^2} \cdot \sqrt{\sum_{p \in I} (r_{y,p} - \bar{r}_y)^2}} \quad (1)$$

$$sim(x, y)^{COS} = \frac{\vec{r}_x \cdot \vec{r}_y}{\|\vec{r}_x\| \cdot \|\vec{r}_y\|} \quad (2)$$

$$sim(x, y)^{Jaccard} = \frac{|I_x \cap I_y|}{|I_x \cup I_y|} \quad (3)$$

$$sim(x, y)^{CPCC} = \frac{\sum_{p \in I} (r_{x,p} - r_{med})(r_{y,p} - r_{med})}{\sqrt{\sum_{p \in I} (r_{x,p} - r_{med})^2} \cdot \sqrt{\sum_{p \in I} (r_{y,p} - r_{med})^2}} \quad (4)$$

$$sim(x, y)^{ACOS} = \frac{\sum_{p \in P} (r_{x,p} - \bar{r}_x)(r_{y,p} - \bar{r}_y)}{\sqrt{\sum_{p \in P} (r_{x,p} - \bar{r}_x)^2} \cdot \sqrt{\sum_{p \in P} (r_{y,p} - \bar{r}_y)^2}} \quad (5)$$

The Assume,  $X = \{X_1, X_2, X_3 \dots X_N\}$ ,  $Y = \{Y_1, Y_2, Y_3 \dots Y_N\}$  are set of users. Let  $P = \{P_1, P_2, P_3 \dots P_M\}$  are set of items. Therefore, the user – item rating matrix is abbreviated as  $R = (r_{i,j})_{N \times M}$ ,  $i = 1, 2, 3, \dots, N$ ,  $j = 1, 2, 3, \dots, M$ . Where,  $X, Y$  are set of users i.e. user  $x$ , user  $y$ .  $I$  represents a set of common ratings for an item  $p$  given by user  $x$  and  $y$ .  $r_{x,p}, r_{y,p}$  indicates the rating value of item  $p$  by user  $x$  and  $y$  respectively.  $\bar{r}_x, \bar{r}_y$  denotes the mean average rating value of user  $x$  and  $y$ .  $\vec{r}_x$  and  $\vec{r}_y$  Denotes the rating Vector value of user  $x$  and  $y$ .  $\|\vec{r}_x\|, \|\vec{r}_y\|$  indicates the magnitude representation of  $\vec{r}_x$  and  $\vec{r}_y$ . Where,  $r_{med}$  is the median value in the rating scale.  $p \in P$ ,  $P$  is set of all items.  $r_{u,p}$  is set to be zero for user missing ratings.

#### Drawbacks:

Some limitations exists in PCC, Cos and Jaccard. Since the scale of rating is absolute the system can know which one is positive rating and which one negative rating. The Constrained pearson correlation coefficient (CPCC) was proposed for considering the impact of +ve and –ve rating [6]. Similarly, if two users rated more common items then the similarity will be more credible [5] so that weighted pearson correlation coefficient (WPCC) was proposed. Results in poor recommendation when the scale of rating is absolute(i.e.) results in full of 1's and 0's. If two users rated more common items then the similarity will be more credible. Scalability.

#### b) The formalization of new similarity method

The predefined cosine similarity method does not support the preferences of the user's ratings. Predefined similarity measures like PCC, Cosine and Jaccard are not much effective when user – item rating matrix is sparse [1]. For considering the preference of user's ratings a new heuristic tuned cosine (TCOS) similarity formula using cross product is proposed. The TCOS (cross product) is defined as follows.

$$sim(x, y)^{TCOS} = \frac{\vec{r}_x \times \vec{r}_y}{\|\vec{r}_x\| \times \|\vec{r}_y\|} \quad (6)$$

Where,  $\vec{r}_x$  and  $\vec{r}_y$  Denotes the vector product of user  $x$  and  $y$  respectively.  $\|\vec{r}_x\|, \|\vec{r}_y\|$  Denotes the magnitude vector representation of user  $x$  and  $y$ .

#### c) TCOS Algorithm

Input : Movilense / Epinion dataset

Output: Top – N Recommendation

Step 1: create a user – item rating matrix for the experimental dataset.

Step 2: compute the similarity  $Sim(x, y)^{TCOS}$  using equation (4).

Step 3: Predict the rating using equation (4).

Step 4: Find the K-Nearest neighbour according to the similarity resultant matrix.

Step 5: Analyse the resultant matrix.

Step 6: user similarity index ranges from +1 to -1.

Step 7: 1 indicates high similarity index.

0 indicates low similarity index.

Step 6: Produce recommendation.

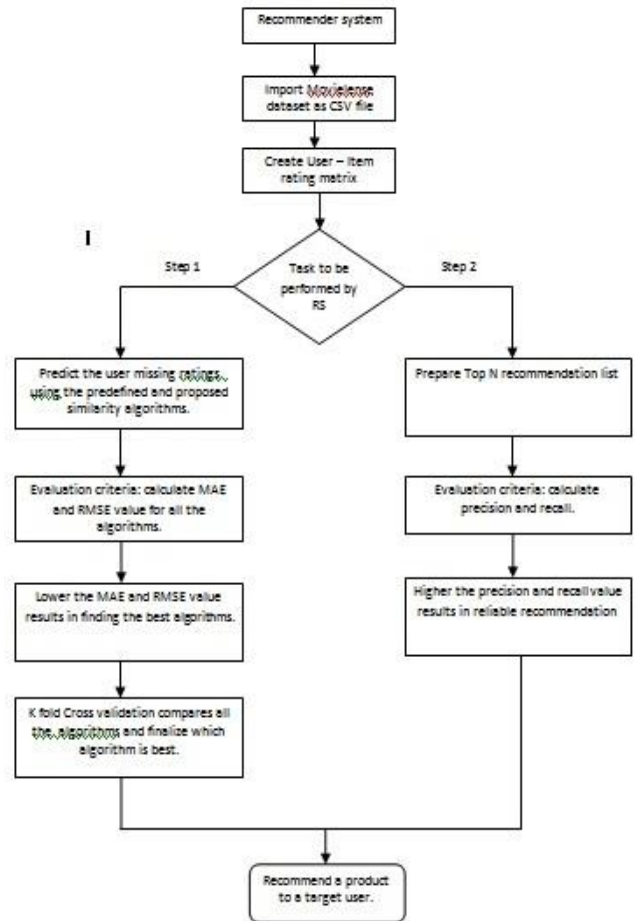


Fig (2). System architecture and work flow diagram.

d) Sample user-item rating matrix

TABLE 1. USER – ITEM RATING MATRIX EXAMPLE.

User/Item	I1	I2	I3	I4
U1	4	3	5	4
U2	5	3	-	-
U3	4	3	3	4
U4	2	1	-	-
U5	4	2	-	-

For this table we are going to find the relationship between two users. For e.g. (u1, u2), (u1, u3), (u1, u4), (u1, u5) etc. So that we can analyse which pair is closest to each other. The above example is applied to the predefined formula as well as the Tuned cosine formula. The resulting matrix is as follows.

Evaluation metrics

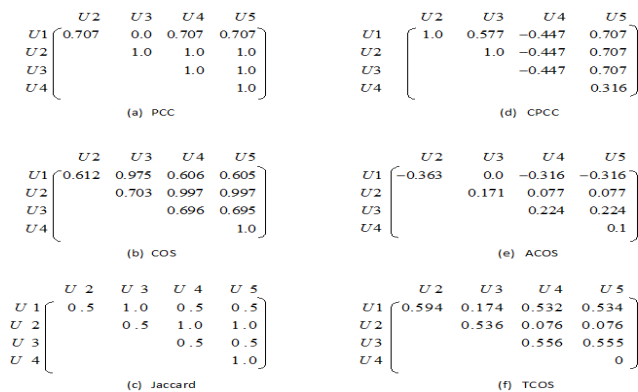


Fig (3) UB similarity matrix for Table 1 example.

e) Figures and Tables

From fig (3) we can see the PCC, cosine, jaccard, CPCC and ACOS user similarity resultant matrix for table 1. Since the scale of rating is absolute in PCC, CPCC and jaccard method. There is no absolute value in TCOS.

Results in low similarity index instead of high similarity index. For example we can see that user 1 and user 3 have very similar rating vectors called (4,3,5,4) and (4,3,3,4). Therefore, user 1 and user 3 were highly correlated each other but the result shows that zero in PCC and ACOS. i.e no correlation. This leads to bad recommendation performance. Thus, there is a slight improvement in TCOS similarity measures.

Results in high similarity index instead of low similarity index. For example we can see that user 2 and user 4 have dissimilar rating vectors called (5, 3,-,-) and (2, 1,-,-). Therefore, user 2 and user 4 are the least pair who does not correlate with each other, but the result shows that 1.0 in PCC,

cosine and jaccard. As we can see that user4 and user5 rating vector is (2,1,-,-) and (4, 2,-,-) will be the least pair but it shows high similarity index. Overall (u2, u4), (u2,u5) and (u4, u5) should show a low similarity index because it does not correlate with each other but it is high in PCC, COS, Jaccard, ACOS. This limitation is overcome by TCOS method.

Each user has different similarities in TCOS method but this is not the case in PCC, Cosine and jaccard method. i.e PCC results in all 1.0, jaccard there are only two kinds of similarities 1.0 or 0.5.

Misleading still exist in PCC, cosine and jaccard that rating vector for user 3 and user 5 is (4,3,3,4) and (4,2,-,-) got low similarity index than user 4 and user 5 is (2,1,-,-) and (4,2,-,-). TCOS overcomes the misleading capability by showing user 3 and user 5 got high similarity index than user 4 and user5. In order to compare the accuracy of the proposed method with the other methods, we use of MAE (Mean Absolute Error) and RMSE (Root Mean Square Error) which are two most common measures of predictive accuracy [17].

IV. CONCLUSION

The paper first analyse the CF recommendation algorithms for E-Commerce websites. How the product is recommended to the end user, on what criteria recommendation is done, what are all the existing predefined similarity algorithms are there and their limitations, how to overcome the limitations were discussed. The new TCOS similarity algorithm using cross product was defined and proposed. It overcomes the limitations exist in the predefined PCC, COS similarity measure whose recommendation is not reliable. Thus, TCOS results in a reliable recommendation to the target user. The proposed methodology is applied in to the sample user-item rating matrix and can obtain a rapid improvement than the predefined one. The resultant TCOS matrix demonstrates the effectiveness of the similarity measure and thus results in a credible recommendation. In future, TCOS similarity algorithm will be implemented with several experimental datasets like movielense, epinions and jester5k in any one of the implementation area Apache mahout a java library, LensKit a java library, LibRec, Python, R environment, PHP/MySQL.

REFERENCES

- [1] H.Mohana and Dr. M Suriakala. An overview study on web mining in e-commerce, international journal of scientific research (IJSR), VOLUME-6 | ISSUE-8 | AUGUST - 2017 | ISSN No 2277 - 8179 | IF : 4.176 | IC Value : 78.46.
- [2] SARANYA K. G, G. SUDHA SADASIVAM. Modified Heuristic Similarity Measure for Personalization using CF Technique. IN Applied Mathematics & Information Sciences An International Journal. Appl. Math. Inf. Sci. 11, No. 1, 317-325 (2017).
- [3] F. Cacheda, V. Carneiro, D. Fernández, V. Formoso, Comparison of CF algorithms: limitations of current techniques and proposals for scalable, high-performance recommender system, ACM Trans. Web 5 (1) (2011) 1–33.
- [4] Haifeng Liu, Zheng Hu, Ahmad Mian, Hui Tian, Xuzhen Zhu. A new user similarity model to improve the accuracy of CF. Knowledge-Based



- Systems, Elsevier, Volume 56, January 2014, Pages 156–166. doi.org/10.1016/j.knosys.2013.11.006.
- [5] J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing CF, in: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1999, pp. 230–237.
- [6] U. Shardanand, P. Maes, Social information filtering: algorithms for automating word of mouth, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1994, pp. 210–217.
- [7] M. Jamali, M. Ester, TrustWalker: a random walk model for combining trustbased and item-based recommendation, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 397–406.
- [8] H.J. Ahn, A new similarity measure for CF to alleviate the new user cold-starting problem, *Inform. Sci.* 178 (1) (2008) 37–51.
- [9] Leily Sheugh, Sasan H. Alizadeh. A note on Pearson Correlation Coefficient as a metric of similarity in recommender system. in: *AI & Robotics (IRANOPEN)*, DOI: 10.1109/RIOS.2015.7270736, Sep 2015 IEEE.
- [10] J. Bobadilla, F. Ortega, A. Hernando, J. Bernal. A CF approach to mitigate the new user cold start problem. *Knowledge-Based Syst.*, 26 (2011), pp. 225–23.
- [11] Chenyang Li and Kejing He. CBMR: An optimized MapReduce for item-based CF recommendation algorithm with empirical analysis. *Concurrency and Computation: Practice and Experience*. DOI: 10.1002/cpe.4092, Feb 2017.
- [12] Mengsi Liu, Weike Pan, Miao Liu, Yaofeng Chen, Xiaogang Peng, Zhong Ming. Mixed Similarity Learning for Recommendation with Implicit Feedback, *Knowledge-Based Systems* (2016), doi:10.1016/j.knosys.2016.12.010.
- [13] Manu M N , Ramesh B. Single-criteria Collaborative Filter Implementation using Apache Mahout in Big data. *International Journal of Computer Sciences and Engineering (IJCSE)*. Volume-5, Issue-1. E-ISSN: 2347-2693. Jan 2017.
- [14] Asma Sattar, Mustansar Ali Ghazanfar, Misbah Iqba. Building Accurate and Practical Recommender System Algorithms Using Machine Learning Classifier and CF. *COMPUTER ENGINEERING AND COMPUTER SCIENCE*. Arab J Sci Eng . DOI 10.1007/s13369-016-2410-1. Dec 2016.
- [15] Jaime Raigoza and Vikrantsinh Karande . A Study and Implementation of a Movie Recommendation System in a Cloud-based Environment. *International Journal of Grid and High Performance Computing (IJGHPC)*. DOI: 10.4018/IJGHPC.2017010103. 2017.

# Design and Development of a system for the Impact analysis of Internet usage among Engineering students

Dr.R.R.Rajalaxmi,  
Professor & Head,  
Department of Computer  
Science &Engineering,  
Kongu Engineering College,  
TamilNadu, India.

Dr.P.Natesan,  
Associate Professor,  
Department of Computer  
Science &Engineering,  
Kongu Engineering College,  
TamilNadu, India.

Dr.N.Krishnamoorthy,  
Assistant Professor(Sr.Gr),  
Department of Computer  
Science &Engineering,  
Kongu Engineering College,  
TamilNadu, India.

S.Ponni,  
Research Assistant,  
Department of Computer  
Science &Engineering,  
Kongu Engineering College,  
TamilNadu, India.  
ponnis94@gmail.com

**Abstract**—Internet is one of the most important information and communication technology which cause a ground breaking change in the computing world. It has a wide variety of users. One such wide user is the college students who use the internet for various purposes. The aim of the study is to investigate the pattern of internet usage among the undergraduate engineering college students. Also, the work depicts the view of the student on the usage of internet. A questionnaire is designed for collecting the data from undergraduate students in Kongu Engineering College. The data is collected from 682 students of three programme. The survey revealed almost 97% of the students use the internet daily. Results shows that the number of respondents have several characteristics. On testing, the results shows that the questionnaire is valid and reliable. Statistically, two hypothesis is tested. The results shows that there is a strong relationship between the gender and usage of internet for social media sites than the education related sites.

**Keywords**—Internet usage, Engineering stream, Undergraduate students.

## I. INTRODUCTION

The introduction of various technology like mobile phones and highly intellectual embedded systems makes a revolutionary change all over the world. The invention of smart phones makes the internet to play a major part of the human society. Nowadays it's difficult to find a person without the smart phone and the internet. It became a routine habitual in our day to day life. The Internet is a large global network which consists of interconnected networks that provide various services such as email, file transfer protocol, accessing the database and many others. Nowadays the internet has most of the information which can be retrieved and accessed easily at anywhere at any time. There are millions of information available on the internet which can be used by anyone. It provides communication between the users through various social media. Social media is an online technological platform used for the purpose of communication and as a source for sharing the information. The social media networks play a vital role nowadays. These networks are mainly used for the smart phones and computers by most of the people. The major users of the internet and the social

media network seems to be the student's community. The usage of the internet became an important part of the student life. It is also used by the teachers to refer various learning materials and for research. The influence of the internet is also examined by various researchers. It helps the student to gain their academic knowledge and increase the performance on the academic side.

## II. RELATED WORK

Harlina(2015) proposed a study to determine the association between the usage of Internet and the performance of the students in a public university. They conducted a survey in public university where the survey consists of Internet Addiction Diagnostic Questionnaire with Yes/No type Questions. Majority of the users are dependent and some users are said to be likely dependent on the internet based on the analysis of the data collected from the survey. This shows that the usage of internet is highly associated with improvement in academics.

A case study is suggested on the behavior analysis of students while using the social media. Facebook is the social media they considered for the case study. The analysis mainly focuses on the relationship between the confidence and participation in social media and also relating to the behavior of the students. The general scanning model was used to observe the behavior and the attitude of the students for the purpose of collecting the data. The analysis shows that Facebook is mainly used for communication and entertainment purpose as per the data collected. It also shows that the students were aware of good and bad impact of social media which are said to be a good sign of behavior analysis which was proposed by Kaya(2016).

An investigation is carried out on the usage of Facebook mediating the personality of the students and their performance in academics by Muzamil(2017). A survey was prepared based on the different types of personalities and usage of Facebook which can predict the academic performance of the students. It shows positive impact on academics of some types of personalities while some other personalities exhibit the negative impact on academics. The

study shows that the personality factor is a main component that has a major impact on their performance in academics. While using Facebook, there is an improvement in their academic results which sometimes shows as a relaxing parameter among the students.

The relationship between the use of social media among the students and their academic performance is explored by Esam(2015). The data was gathered through an online survey to gather information about the usage and effects of social media and their grades points from the students. The results show that there is a nonlinear relationship between the usage of social media and academic performance of the students. It also conveys that the time factor affects the education of the students. It also tells a way to balance the time factor for improving the academic performance effectively.

Chou(2005) proposed a research study which identifies the adaptive patterns of internet use that constitute behavioral addiction. The work mainly explores the research on the social effects of Internet addiction. It contains four major sections. The first section explores the field and introduces definitions, terminology, and assessments. The second section describes research findings and focuses on several key factors related to Internet addiction, including Internet use and time, identifiable problems, gender differences, psychosocial variables and computer attitudes. The third section reviews the addictive potential of the Internet in terms of the Internet, its users, and the interaction of the two. The fourth section ends with the current and projected treatments of Internet addiction, suggests future research agendas, and provides implications for educational psychologists.

Emily(2008) worked on how internet plays a vital role in student's life. It is important to monitor students' attitudes and usage to ensure students use technology effectively by recognizing credible sources and utilizing the correct technology for each situation. The study utilizes a descriptive survey to explore the current usage and attitudes toward the Internet by students enrolled in college of agriculture courses at the University of Florida. The results indicate that these students are substantial users of the Internet and programs like Facebook, Myspace, and search engines. Students believe that Internet is easy to understand, important, beneficial, believable, and accurate. Recommendations are offered as to what these findings mean for instructors in the classroom.

The connection between the individual's personality and the behavior while using online is shown in a study by Yair(2010). They produce more objective criteria and the information uploaded on Facebook of users and measured the personality traits. They concluded that there is a strong connection between the personality and the usage of social media.

A study is to measure the degree of Internet addiction among a particular domain students is given by Pramanik (2012). A survey questionnaire was prepared which can able to assess the mild, moderate and severe addiction. Results show that most of the students were in the moderate category. Late night internet surfing leads to sleeplessness and rest of the students tried to reduce the time spent in online but it got failed and some experienced the restlessness due to internet

access. Mostly students from mild addiction slowly move to moderate addiction which is a bad sign for the students. A prevention measure should be done to prevent the internet addiction of the students.

Peter M.Ogedebe(2012) researched on a paper that examines the extent of usage of Internet among Nigerian University undergraduate and how this has affected their performance. The impact of the Internet on academic performance is discussed in literatures and that it could be positive and/or negative. A questionnaire was designed. The paper was therefore of the opinion that if Internet Services are fully exploited, the performance of students in institutions of higher learning in Nigeria will get improved.

The main aim of the work is to examine the satisfaction and usability of internet usage on students' assignment completion tasks and their performance. The study was conducted and the survey was conducted for post -graduate students of Management Studies. The results shows that the usage of internet significantly describes the difference in student's performance. Internet usage is the interpreter of Technology satisfaction and student's performance. Since these factors are found to have significant relationship with students' performance, the management and decision makers in universities and institutes need to give higher importance as the students could use the internet efficiently and effectively. The work is proposed by Manju Bhaga (2011).

In all these works, the authors concentrated on the students of different stream but this work mainly focus on undergraduate students of Engineering stream who use the internet for various purposes. Moreover the statistical methodology is defined for validating the questionnaire. The rest of the paper is unfolded as follows, Section 3 discusses the proposed methodology followed by Section 4 which contains the results analysis and the hypothesis is examined in Section 5. Finally it ends with the conclusion along with the future work of the paper.

### III. METHODOLOGY

The proposed study is to analyze and predict the behavior and learning experiences of the students based on their usage of web and social media in their day-to-day life. The problem is divided into three categories such as usage of internet, social media and their learning methodology among the undergraduate college students. Hence a questionnaire is prepared accordingly with three sections along with demographic information. Each section contains 15 questions related to the topic under consideration. The senior faculty members in Engineering department in the institution tested the validity of the questions. Also, an expert from our institution evaluated the questionnaire validity and suggested some points that were revised in the questionnaire. Then the pilot survey is conducted with a sample of 10 students. The participants were asked to answer all the questions in order to carry out the analysis from the given data. Based on the inputs given by the students, some modifications in the questionnaire were made. The corrected survey questionnaire is given to 682 Engineering students of the different branch in the institution.

The data is analyzed to identify the correlation between the variables. These are defined in the Fig.1 as shown in below,

TABLE I. NUMBER OF COLLEGE STUDENTS ON DIFFERENT CHARACTERISTICS

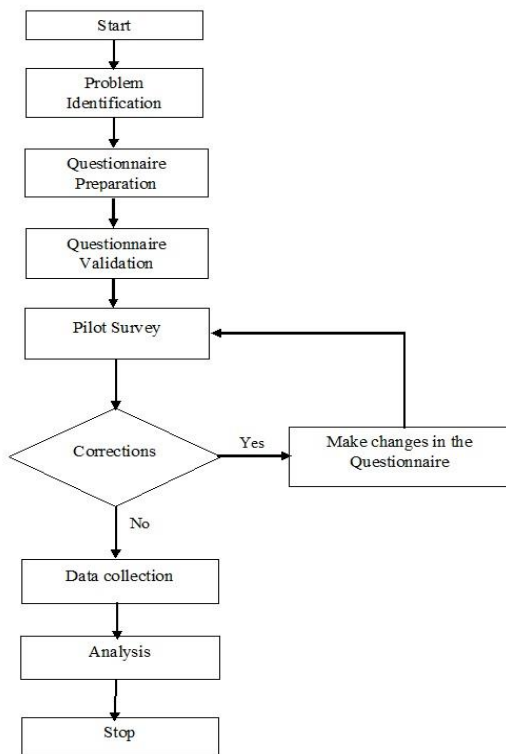


Fig 1. Flowchart of the proposed work

The questionnaire was given to the undergraduate students of a particular department and the strength is about 682 out of which 323 were male and remaining were female undergraduate students answered the survey conducted in the college. The levels of academics range from: second year (222), third year (238) and final year (222). The survey also revealed the background of the students: the students of about 225 came from the city area, 264 from the town and 193 from the village side. The survey also revealed that the students of about 608 are from English medium and 74 are from Tamil medium. Out of 682 students, 326 are of day scholars and remaining 356 are of hostellers. The students of about 476 are using internet in Mobiles/Smartphones, and 203 are using internet in Desktops/Laptops and remaining 3 are using through the device called Tablets. These are all shown in Table.I . From all these, it shows that the students feels that the usage of internet is mandatory nowadays.

Characteristics	N	%
<b>Gender(N=682)</b>		
Male	323	47.4
Female	359	52.6
<b>Native(N=682)</b>		
City	225	32.99
Town	264	38.7
Village	193	28.29
<b>Medium of instruction</b>		
English	608	89.2
Tamil	74	10.8
<b>Mode of Admission</b>		
Management	283	41.5
Counselling	399	58.5
<b>Devices frequently used by the student for the usage of internet</b>		
Mobiles/Smartphones	476	69.8
Desktop/Laptops	203	29.8
Tablets	3	0.04
<b>Primary usage of internet</b>		
E-mail	99	14.5
Education	224	32.8
Entertainment	212	31.1
E-shopping	14	2.05
Chatting	80	11.7
Gaming	34	4.98
Others	19	2.78

#### IV. RESULT ANALYSIS

In order to gain a confidence in the upcoming results of the study, the questionnaire must assure the consistency. In short, the questionnaire must be a valid and reliable. So the internal consistency was checked and thereby ensuring that the various items measuring the different questions delivers the consistent scores or not. The internal consistency consists of three methods. They are Split Halves test, Kuder-Richardson test and Cronbach's alpha test.

The split halves test involves dividing a test into two halves. It is simple and speed. But, nowadays the machines can take over the laborious number crunching and it needs a powerful test. The Kuder-Richardson test is more advanced and it is more complex. It works out the average correlation for all the possible split half combinations in a test and gives more accurate result. The main disadvantage of this approach is that the answer for each question must be a simple right or wrong answer that is either zero or one. To overcome the disadvantages in the above two methods, Cronbach's alpha method is introduced.

The Cronbach's Alpha test averages the correlation between every possible combination of split halves and also it allows multi-level response. It will also tell if the

questionnaire designed is accurately measuring the variable of interest. The calculation carried to be simpler one and it seems to be best statistical measure for calculating the internal consistency. The Equation 1 depicts how to compute the value of Cronbach's alpha is:

$$\alpha = \frac{N \cdot c'}{v' + (N-1) \cdot c'} \quad (1)$$

where N is the number of items, c' is the average covariance between item-pairs and v' is the average variance. The value of alpha decides the reliability of the content. To test the content reliability, the SPSS tool was used.

A commonly accepted rule for checking the internal consistency using Cronbach's alpha (13) is given as  $\alpha \geq 0.7$  is said to be acceptable and  $0.6 < \alpha < 0.7$  is said to be questionable and  $\alpha < 0.6$  is unacceptable.

To check the reliability, the data should be in numeric form. So the data collected is converted based on their priority of the question. The 5 point Likert scale used to collect the opinion from the students was converted into numeric form. For example, let's consider a question from social media usage section. The question contains three options: Agree, Neutral and Disagree. The encoding of the options is Agree as 3, Neutral as 2 and Disagree as 1. The following Table.II shows the data encoding,

TABLE II. DATA ENCODING

Does your content sharing in social media network based on your mood?	changes
Agree	3
Neutral	2
Disagree	1

Likewise all the attributes are encoded that contains different scaling of the options.

In this work, the questionnaire is divided into three section based on the requirement of the project. Each section is tested for its reliability.

The first section is internet usage where the questions related to the usage of internet in different sectors is asked. When this section undergoes the reliability test, the result shows the alpha value as 0.707 which is said to be acceptable as per the common rule. The Table.III shows the reliability measure of the internet usage section,

TABLE III. RELIABILITY ON INTERNET USAGE

**Reliability Statistics**

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.707	.681	13

The second section is mainly concentrated on the usage of social networking sites among the college students. It also underwent the reliability test, the result shows the alpha value as 0.712 which is said to be acceptable as per the common rule. The Table.IV is generated to identify the Cronbach's alpha value,

TABLE IV. RELIABILITY ON SOCIAL NETWORKING SITES

**Reliability Statistics**

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.712	.852	6

The last section focuses on the learning patterns of the college students. This section carries the reliability test, the result produces as the alpha value 0.697 which is said to be questionable as per the common rule. The Table.V is calculated to measure the reliability,

TABLE V. RELIABILITY ON LEARNING EXPERIENCES

**Reliability Statistics**

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.697	.775	7

The above reliability shows that the first two section is acceptable but the last one is questionable. While considering as overall, the reliability result shows that the questionnaire is acceptable. Thus the content reliability shows the positive result.

**V. EXAMINING THE HYPOTHESIS**

Analysis of Variance (ANOVA) is a test to measure the differences between two means. It is a way to find out if a survey is significant. It helps to figure out whether the null hypothesis should be rejected or accepted. The significance level tells whether to accept or to reject the hypothesis.

**Hypothesis one:**

H<sub>0</sub>: There is a significant relationship between gender and the use of internet for education purpose.

H<sub>a</sub>: There is no significant relationship between gender and the use of internet for education purpose.

In Table.VI & Table.VII, model summary and ANOVA illustrates the weak relationship between gender and the use of internet for education purpose but it was not significant with (0.619). This shows that the rejection of the null hypothesis. So the alternate hypothesis is accepted. The results shows that there is no significant relationship between gender and the use of internet for education purpose.

TABLE VI. ANOVA TEST FOR HYPOTHESIS ONE

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.661	4	.165	.661	.619 <sup>a</sup>
	Residual	156.768	627	.250		
	Total	157.429	631			

a. Predictors: (Constant), [OTHERS(LIKE ACADEMIC FORUMS,BLOGS)], [PLACEMENT], [CO-CURRICULAR ACTIVITIES], [STUDIES]

b. Dependent Variable: GENDER

TABLE VII. MODEL SUMMARY FOR HYPOTHESIS ONE

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.065 <sup>a</sup>	.004	-.002	.500

a. Predictors: (Constant), [OTHERS(LIKE ACADEMIC FORUMS,BLOGS)], [PLACEMENT], [CO-CURRICULAR ACTIVITIES], [STUDIES]

**Hypothesis two:**

H<sub>0</sub>: There is a significant relationship between gender and the use of internet for social media sites.

H<sub>a</sub>: There is no significant relationship between gender and the use of internet for social media sites.

In Table.VIII & Table.IX, model summary and ANOVA shows that the weak relationship between gender and the use of internet for social media sites but it is significant with (0.048) so the null hypothesis is accepted in this case. This shows that the importance of using internet for online social networking sites.

TABLE VIII. MODEL SUMMARY FOR HYPOTHESIS TWO

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.079 <sup>a</sup>	.006	.005	.498

a. Predictors: (Constant), [SOCIAL MEDIA(LIKE FB, INSTAGRAM, TWITTER)]

TABLE IX. ANOVA TEST FOR HYPOTHESIS TWO

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.971	1	.971	3.908	.048 <sup>a</sup>
	Residual	156.458	630	.248		
	Total	157.429	631			

a. Predictors: (Constant), [SOCIAL MEDIA(LIKE FB, INSTAGRAM,TWITTER,INSTAGRAM)]

b. Dependent Variable: GENDER

Moreover from the analysis of data, some patterns were generated which might be useful for further study. If a student spent more hours on the usage of internet for academic activities on daily basis and the usage pattern of internet before the semester examination shows that there is an improvement in their grades. On gender wise evaluation, the female students tend to have more interest towards the academic activities than male students. Also if the student have more number of social media accounts then they tend to spend more number of hours in social media sites and there is a disturbance in their physical and mental health. When considering the number of hours spent in E-shopping sites shows that there is a slight change in their buying habits.

VI. CONCLUSION

The use of internet seems to be important as everything in this world becomes online. So the usage of internet is also increased among the college students. In this paper, the opinion on the internet was collected from the college students and formed a dataset. The study is conducted and the reliability is tested. The Cronbach's alpha measure is considered and overall the value shows that the questionnaire is valid and it is reliable. The hypothesis also reveals that there is a significant relationship between the gender and online social networking sites. Also some patterns on the usage of internet is studied based on the analysis of data that were listed which reveals some traits towards the internet usage by the undergraduate students.

## VII. FUTURE WORK

Furthermore, the survey will be conducted in different branches of engineering stream in different colleges. The analysis will be carried out to explore the inferences, so a model can be created. The data mining algorithms will be helpful in creating a model and it will be done in near future.

## ACKNOWLEDGMENT

This part of work is supported by the Indian Council of Social Science Research (ICSSR) under Research project scheme.

## REFERENCES

- [1] Harlina Halizah Siraj, Abdus Salam, Nurul Ashiqin Hasan, Tan Hiang Jin, Raihanah Binti Roslan, Muhammad Nazam Bin Othman, "Internet Usage and Academic Performance: A Study in a Malaysian Public University", *International Medical Journal* Vol. 22, No. 2, pp. 83 – 86, April 2015.
- [2] Tugberk Kaya, Huseyin Bicen, "The effects of social media on student's behaviors; Facebook as a case study", *Computers in human behavior* Vol. 59, pp. 374 – 379, February 2016.
- [3] M. Muzamil Naqshbandi, Sulaiman Ainin, Noor Ismawati Jaafar, Nor Liyana Mohd Shuib, "To Facebook or to Face Book? An investigation of how academic performance of different personalities is affected through the intervention of Facebook usage", *Computers in Human Behavior*, Vol. 75, pp. 167-176, May 2017.
- [4] Esam Alwagait, Basit Shahzad, Sophia Alim, "Impact of social media usage on students academic performance in Saudi Arabia", *Computers in Human Behavior*, Vol. 51, pp. 1092-1097, 2015.
- [5] Chien Chou, Linda Condron and John C. Belland, "A Review of the Research on Internet Addiction", *Educational Psychology Review*, Vol. 17, pp.363-388, December 2005.
- [6] Emily B. Rhoades, Tracy Irani, Brian E. Myers, "Internet as an Information Source: Attitudes and Usage of Students Enrolled in a College of Agriculture Course", *Journal of Agricultural Education* Volume 49, pp. 108 - 117, June 2008.
- [7] Yair Amichai-Hamburger, Gideon Vinitzky, "Social network use and personality", *Computers in Human Behavior*, Volume 26, pp. 1289-1295, May 2010.
- [8] T.Pramanik, MT Sherpa and R. Shrestha, "Internet addiction in a group of medical students: a cross sectional study", *Nepal Med College*, Vol. 14, pp.46-48, March 2012.
- [9] Peter M. Ogedebe, "Internet Usage and Students' Academic Performance in Nigeria Tertiary Institutions: A Case Study of University of Maiduguri", *Academic Research International*, Vol. 2, pp. 334-343, May 2012.
- [10] Ela Goyal, Seema Purohit, Manju Bhaga, "Study of satisfaction and usability of the Internet on student's performance", *International Journal of Education and Development using Information and Communication Technology (IJEDICT)*, Vol. 7, pp. 110-119, 2011.
- [11] <https://www.youtube.com/user/WekaMOOC>
- [12] <http://www.cs.ccsu.edu/~markov/weka-tutorial.pdf>
- [13] <http://www.statisticshowto.com/cronbachs-alpha-spss/>

# Convergent Scheduling with EDF and BackFill Algorithm to improve the Performance of the Grid Scheduler

Kalyani V  
ME CSE  
MIT, Anna University  
Chennai  
vkalyani29@gmail.com

**Abstract**—Grid computing is a high performance computing environment to solve larger grid environment scale computational demands. Grid computing has resource management, task scheduling, security problems, and information management and so on for increasing the performance. Task scheduling is a basic issue in achieving high performance in grid computing systems. In this paper we studied the performance of the scheduler while using the Earliest Deadline First Algorithm with a new approach Convergent Scheduling. When this algorithm is combined with the Backfill algorithm the performance of Earliest DeadLine Algorithm is improved. The resources which are available are used efficiently. The overall performance of the scheduler is improved.

**Keywords**—Grid computing; Scheduling; Resource manager; Quality-of-Service; Time sharing; Space sharing.

## I. INTRODUCTION

Grid computing is mainly focused on the synchronized problem solving and resource sharing in dynamic, multi institutional virtual organizations. Grid does not enforce absolute control over these resources and resource management is subject to multiple and dissimilar organizational administrative policies. From user's view, a Grid is a collective problem solving environment in which many user jobs can be submitted without knowing where the resources are or even who own the resources. A Grid must assure the quality of service of a job's execution. In order to construct a Grid computing environment, it is difficult to have a common Grid infrastructure software system.

Globus has been successful in developing open and standard protocols and interfaces, and is the de facto standard in Grid world. Its major contribution is a PKI based Grid certificate solution to the Grid security problem, which enables cross organizational resource access control. It also provides the toolkits and mechanisms for VO scope job submission (GRAM) and resource discovery (GRIS) which we refer to a cluster computing system as a "local scheduler", as compared to a global Grid scheduler. Grid scheduling is the process of scheduling applications over Grid resources.

Scheduling and resource management are salient in optimizing multiprocessor Grid resource allocation and

determining its ability to deliver the negotiated Quality of Service needs. This need has been confirmed by the Global Grid Forum in the special working group dealing with the area of scheduling and resource management for grid computing. The resource manager gets information about the job characteristics and determines when and on which processor each job will execute.

Now, the major issue concerns the effective system integration by correctly utilizing the existing tool with all research results that will build Grid computing applicable to many commercial schemas play an important role. When the desired users' requirement can't be achieved, the degradation should be graceful and fair to every user. As the tasks requirements, we refer to the ready to be executed on a processor. This naturally tends to the need of congestion control and the associated notion of fairness problems that we address in this paper.

## II. RELATED WORK

[19] paper proposes a optimized schedule based approach for scheduling a continuous stream of batch jobs on the machines of a computational Grid. Our new answers represented by dispatching the rule Earliest Gap - Earliest Deadline First (EG-EDF) and Tabu search are based on the idea of filling gaps in the existing schedule. EG-EDF rule is able to construct the schedule using this rule for all jobs gradually by applying technique which fills earliest existing gaps in the schedule with latest jobs. If no gap for a coming job is available EG-EDF rule uses Earliest Deadline First (EDF) strategy and Tabu search for including new job into the existing schedule. Such schedule is then utilized using the Tabu search algorithm proceeding tasks into earliest gaps again. Scheduling choices are made to meet the Quality of Service requested by the submitted tasks, and to optimize the usage of hardware processes.

In [16] this paper, propose a new method, called Convergent Scheduling, for scheduling a continuous stream of batch jobs on the machines of large-scale computing farms. This method utilizes a set of heuristics that train the scheduler in making decisions. Each heuristics controls a specific problem constraint, and provides to carry out a Fig. that calculates the degree of matching between a job and a machine. Scheduling choices are taken to meet the Quality of



Service requested by the submitted jobs, and optimizing the usage of hardware and software processes. We have compared it with some of the most used job scheduling algorithms that is Backfilling, and Earliest Deadline First. Convergent Scheduling is able to compute good assignments, while being a simple and modular algorithm.

Dynamic priority scheduling algorithm is used to find the virtual priority value in the scheduled tasks. In [14] paper they consider a virtual computing environment that provides computational resources on demand to users with multi attribute task descriptions that include a valuation, process (CPU) requires and a completion deadline. Achieving a high QOS in this environment depends on calculating a balance between processing high priority tasks before their deadlines expire, while maximizing resource utilization and its needs. The problem becomes more challenging in an economic setting, where the task valuation is made local.

### III. PROPOSED ARCHITECTURE AND ALGORITHM

Grid scheduler defined as the process of making scheduling decision involving resources and processes over multiple administrative domains. The mandatory jobs are provided to the scheduler and the scheduler then gets the necessary information from the hardware, software monitors and the policy descriptor. The scheduler process executes the jobs one by one till all the jobs are processed.

For each job, the list of clusters that can satisfy the software requirements are used and then the scheduler finds the hardware resources and policies of each of those clusters. From this information, it decides which cluster can give the best completion time, which is the sum of wait time and

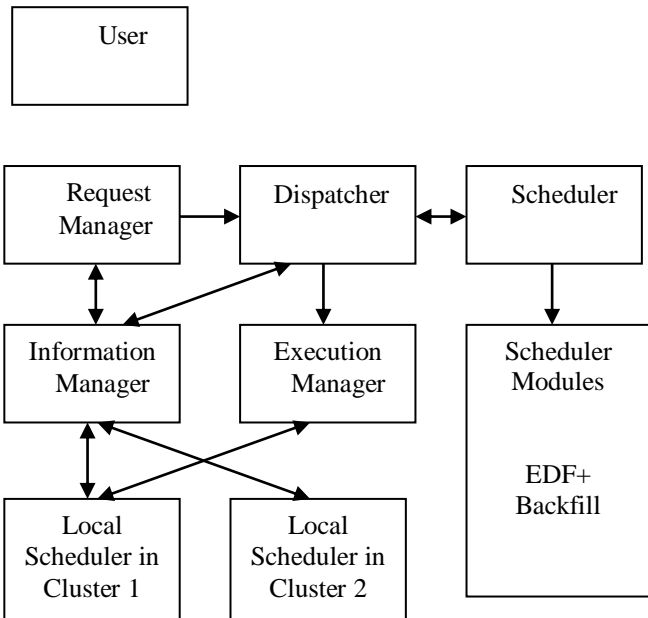


Fig. 1: Scheduler Architecture

execution times. It will select the cluster and assign the job to the cluster.

Since the clusters are selected dynamically during the execution of the scheduler, load balancing is also addressed if any cluster is heavily loaded then the scheduler will automatically not pick that cluster, since it will give us a longer wait time and the heavy load can be seen from the hardware resource monitor as well as from the cluster policy description language.

#### A. Workflow of scheduler

Fig. 1 shows the architecture of Scheduler. In grid computing environment there will be number of users who run their job in the grid environment. The users submit the job to the Request Manager. Dispatcher is component, which dispatches the job to various components in the scheduler. Information manager is the one, which collects the resource information in the grid environment from different local Scheduler. Local Scheduler will maintain their own information such as Operating system, CPU speed, Free memory, software present in the system etc. Dispatcher gets their input from the Request manager and Information manager. Incoming jobs and Resource information's are the inputs given to the scheduler for scheduling the incoming jobs. In Scheduler, it first schedules the incoming jobs based on the Earliest DeadLine First Algorithm and it matches the scheduled job with the resource information's. Then based on the matched information, the backfill algorithm is applied to the scheduled job. After scheduling the result is given to the dispatcher. Dispatcher receives the scheduled job and it submits to the Executer, which executes the job. The Executer gives the job to the corresponding resources and executes the job.

#### B. Modified scheduling algorithm

In this approach, jobs are scheduled using Earliest DeadLine Algorithm then Match making process is done and then Backfill algorithm is applied.

##### 1) Earliest Deadline First rules:

The most widely used urgency based scheduling scheme is EDF method, also known as the deadline-driven rule. This method dictates that, at any point, the system must assign the highest priority to the task with the most imminent deadline. The most urgent tasks with the earliest deadline are served first, followed by the remaining tasks according to their urgency.

Earliest deadline first (EDF) scheduling is a dynamic scheduling algorithm. It places job in a priority queue. Whenever a scheduling event occurs the queue will be searched for the process closest to its deadline. EDF has lead to high levels of SLO fulfillment in both small and big jobs, as it considers deadlines when scheduling. The incoming jobs will be arrived in the Central Scheduler Queue. High Priority is give to the job which has Earliest Deadline. Other parameter such as Total CPU utilization, RAM size, Total execution time, Operating system, number of node required are to be consider while allocating the job to the resources. Earliest

Deadline First is an optimal scheduling algorithm on pre-emptive in the following sense. The Schedulability test for EDF is defined below.

$$U = \sum_{i=0}^n \frac{C_i}{T_i} \leq 1$$

Where the  $\{C_i\}$  are the worst case computation times of the  $n$  processes and the  $\{T_i\}$  are their respective inter-arrival periods (assumed to be equal to the relative deadlines).

If a collection of independent jobs, each characterized by an arrival time, an execution requirement, and a deadline, can be scheduled (by any algorithm) such that all the jobs complete by their deadlines, the EDF will schedule this collection of jobs such that they all complete by their deadlines.

### 2) Match Making:

In Match Making approach the user will submit the Job Details and the Information Manager will provide Host Detail. Based on the user job requirement the job detail will be available. Host Detail will be available based on the resources availability. Based on the Resources available, the match is done between the incoming jobs and the available resource. There might be a peculiar case for example; a single job may get many hosts available which matches its requirements. In this case, the match is done with the Best host available. In other case; the Resources available might not match/fulfill the requirements necessary for the incoming jobs. In this case, the job will not be submitted. Finally thus the matches are made (between the resources available and job requirements) and the output is produced.

### 3) The backfill algorithm:

In Backfill algorithm, when a job reaches at the top of the queue, no job below in the queue will be allowed to start that will cause it to be delayed any longer than it is by jobs that were running when it reached the top of the queue. This is particularly important in massively parallel systems where jobs requiring large numbers of nodes are common. Working flow of Backfill Algorithm,

BackFill following steps to be apply:

- 1) Feasible backfill jobs are filtered selecting those that actually fit the current backfill window.
- 2) Base degree-of-fit on scheduling criteria (i.e. processors, seconds etc.)
- 3) Job with best fit is started
- 4) While backfill jobs and idle resources remain repeat 1

## IV. IMPLEMENTATION DETAILS

Fig. 2 show job submission to the recourses using Earliest Deadline First Algorithm. Fig. 3 show the job submission to the resources using Earliest Deadline First + Backfill Algorithm. In this example we have consider eight users User A, User B, User C, User D, User E, User F, User G, User H submit their job to the gird environment. While submitting the

job we consider the following parameters: User Name, Earliest Deadline First, and Numbers of node required by the user, Time take for the executions, Status and resources Availability. For scheduling periodic real time tasks, Deadline Monotonic algorithm (DMA) is used. Context switching overhead is possible in RMA, so DMA is preferred. For example, User A wants to complete the job before 1<sup>th</sup> Feb, the nodes required is 32 and the execution time is 120 minutes. Based on Deadline, the job gets scheduled. Thus the resource gets matched with the host available.

### Earliest Deadline First Algorithm Workflow:

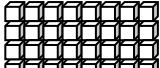

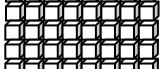
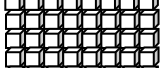



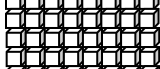
User Name	EDF	Nodes	Minutes	Status	Resources
User A	01/2	32	120	R	
User B	03/2	64	60	R	
User C	08/2	24	180	R	
User D	12/2	32	120	W	
User E	16/2	16	120	W	
User F	24/2	10	480	W	
User G	25/2	4	30	W	
User H	26/2	4	120	W	

Fig. 2: EDF Algorithm before Applying Backfill Algorithm


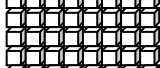

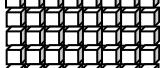


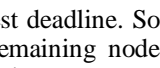
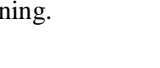
User Name	EDF	Nodes	Minutes	Status	Resources
User A	01/2	32	120	B	
User B	03/2	64	60	B	
User C	08/2	24	180	B	
User G	25/2	4	30	B	
User H	26/2	4	120	B	
User D	12/2	32	120	W	
User E	16/2	16	120	W	
User F	24/2	10	480	W	

Fig. 3: EDF Algorithm after Applying Backfill Algorithm

1. User A needs 32 nodes and it has earliest deadline. So job A is submitted to resources and remaining node available are 96. User A job status is running.

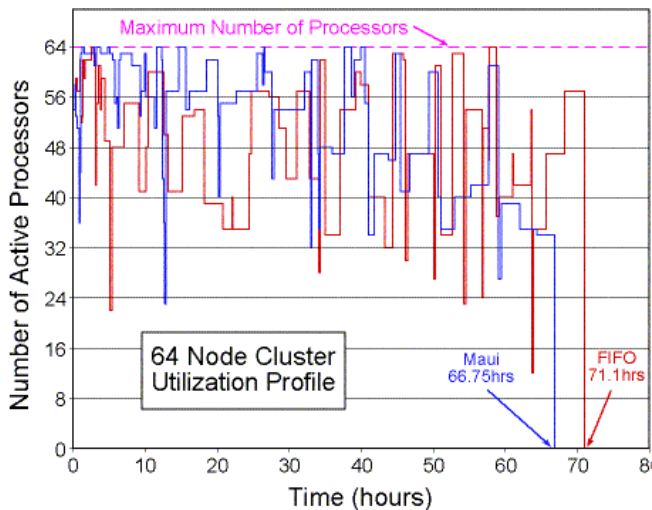


Fig. 4: Number of active Processors utilized in a time interval for Maui and FIFO

2. User B needs 64 nodes and it has next earliest deadline for the job submission. So User B job is submitted to the resources. Then remaining available node is 32.
3. User C need 24 nodes and submitted to available resources and remaining available node is 8.
4. Then User D job has next earliest deadline. This job need 24 nodes but available node is 8 only. So this job and other remaining jobs should wait for the resources availability.

#### Earliest Deadline First and Backfilling Algorithm Work Flow:

1. User A needs 32 nodes.  
Total nodes available are 128. So the job gets submitted to 32 nodes. Hence after submitting the remaining nodes available are 96.
2. Now the User B needs 64 nodes and is submitted. Remaining nodes 32.
3. User C needs 24 node and is also submitted. Remaining nodes 8
4. User D, E, F needs nodes which are not sufficient with the remaining nodes available.
5. Thus User G which requires only 4 nodes gets submitted.
6. User H the same, which requires only 4 nodes also gets submitted.
7. Initially the jobs were in waiting status.
8. Now the User A, B, C, G, H's jobs are in Running status.
9. User D, E, F's jobs are in Waiting status.
10. After the completion of the Running status jobs, Waiting status job gets submitted accordingly on the above basis.

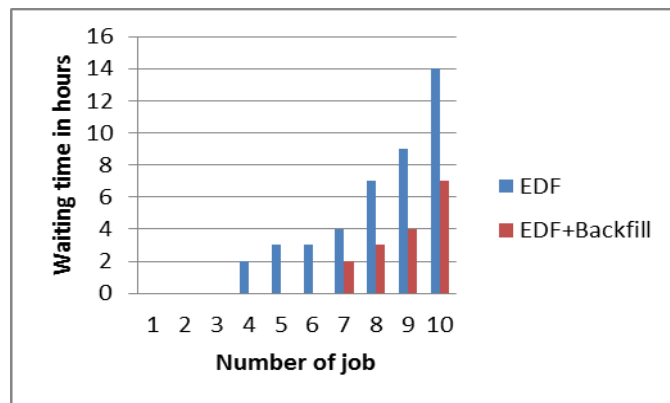


Fig. 5: Number of job Vs Waiting time for EarliestDeadline First Algorithm and Earliest Deadline First + Backfill Algorithm.

#### V. SIMULATION RESULT

By using the EDF algorithm, Service Level objective is satisfied. Thus Quality of Service (QoS) is improved. By combining BF algorithm to this approach, Job Completion Ratio is increased and the Resource Utilization is done efficiently.

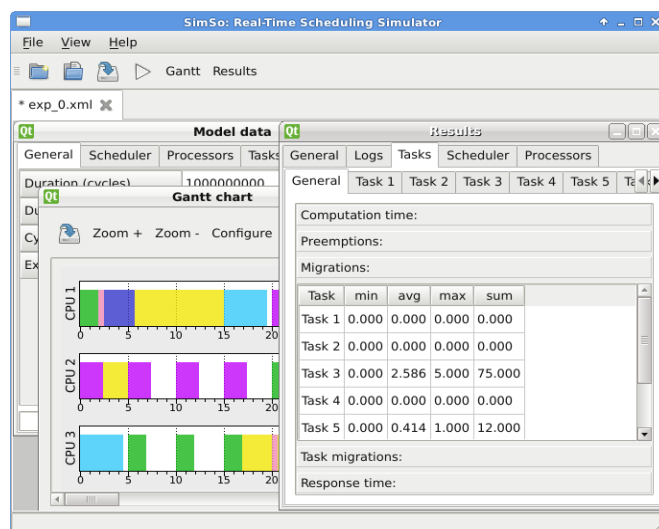


Fig. 6: Scheduling results of all the tasks and their gantt charts

The tasks assigned in the model data is used for Gantt charts and the corresponding min, max and sum values are calculated. For the running tasks, the CPU utilization falls due to the availability of resources. Fig. 6 show the performance analysis of Earliest Deadline First Algorithm and Earliest Deadline First with Backfill Algorithm. When considering the Earliest Deadline First Algorithm the waiting time increases. When we combine the Earliest Deadline First algorithm with Backfill the waiting time of the job decreases. In Fig. 5 show while submitting the tenth job the waiting time for that job is fourteen hours. When we use Earliest Deadline First with Backfill Algorithm, the waiting time is reduced to half that is

only seven hours. While using this approach overall waiting time will reduce.

## VI. CONCLUSIONS

In this work, we proposed a new approach for scheduling and its main advantage is to be maximizes of Job Completion ratio and provide Quality of service. Based on the user requirement the incoming jobs are executed. We compared the performance of Earliest Deadline First and Earliest Deadline First with Backfill. By combining the Earliest Deadline First with Backfill the overall job completion time decreases and the performance increases compared to Earliest Deadline First algorithm.

## REFERENCES

- [1] I. Foster, S. Tuecke and C. Kesselman, "The Anatomy of the Grid Enabling Scalable Virtual Organizations," For Supercomputer Applications, vol.15, no.3, 2001.
- [2] W. Leinberger and V. Kumar, "Information Power Grid: The New Frontier in Parallel Computing," IEEE Concurrent , vol.7, no.4, pp.75-84, Oct-Dec.1999.
- [3] "Scheduling Working Group of the Grid Forum, " published: 10.5, Sept.2001.
- [4] Zdzislaw Pawlak, "Rough set theory and its applications," published in Journal of Telecommunications and Information Technology, 2002.
- [5] D. Lifka, "The ANL/IBM SP scheduling system," Proc. 1st Workshop on JSSPP, 1995.
- [6] B. Nitzberg and J.P. Jones, "Scheduling for Parallel Supercomputing: A Historical Perspective of Achievable Utilization," Proceedings in the 5th Workshop on JSSPP, 1999.
- [7] S. Zhou, X. Zheng, J. Wang, and P. Delisle, "Utopia: a load sharing facility for large, heterogeneous distributed computer systems," Software-Pract. & Exp. 23(12), pp.1305-1336, Dec 1993.
- [8] R. L. Henderson, "Job scheduling under the portable batch system," Proc. 1st Workshop on JSSPP, 1995.
- [9] S. Kannan, M. Roberts, P. Mayes, D. Brelsford, and J.F. Skovira, Workload Management with LoadLeveler," first ed. IBM, Nov. 2001.
- [10] D. Jackson, Q. Snell and M. Clement named "Core Algorithms of the Maui Scheduler on multiprocessor system," Proc. 7th Workshop on JSSPP in 2001.
- [11] D. Jackson, "Maui/Moab Default Configuration" personal/general communication with CTO of Cluster Resources, Jan. 2006.
- [12] D. Talby and D. G. Feitelson, "Supporting priorities and improving utilization of the IBM SP2 scheduler using slackbased backfilling," 13th Intl. Parallel Processing Symp., pp.513-517, Apr 1999.
- [13] S. Srinivasan, R. Kettimuthu, V. Subramani, and P. Sadayappan, "Selective reservation strategies for backfill job scheduling," Proc. 8th Workshop on JSSPP, 2002.
- [14] W. A. Ward, Jr. Carrie L. Mahood, and J. E. West, "Scheduling jobs on parallel systems using a relaxed backfill strategy for distributed systems," Proc. 8th Workshop on JSSPP, 2002.
- [15] B.G. Lawson and E. Smirni, "Multiple-queue Backfilling Scheduling with Priorities and Reservations for Parallel Systems," ACM SIGMETRICS Performance Evaluation Review, vol. 29, no. 4, pp:40-47, 2002.
- [16] E. Shmueli and D. G. Feitelson, "Backfilling with lookahead to optimize the performance of parallel job scheduling," Proc. 9th Workshop on JSSPP, 2003.
- [17] D. G. Feitelson and D. Talby, "Improving and stabilizing parallel computer performance using adaptive backfilling," Proceedings 19th IEEE IPDPS, April 2005.
- [18] Avi Nissimov and Dror G. Feitelson, "Probabilistic backfilling," Proc. 13th Workshop on JSSPP, 2007.
- [19] " Comparison Of Multi-Criteria Scheduling Techniques", Grid Computing *Achievements and Prospects*, 10.1007/978-0-387-09457-1\_15, Sergei Gortlatch, Paraskevi Fragopoulou and Thierry Priol.
- [20] Gabriele Capannini "A job scheduling framework for large computing farms " Conference on High Performance Networking and Computing ,Proceedings of the 2007 ,ACM/IEEE conference on ,Supercomputing ,Article No. 54 ,Year of publication: 2007 ISBN:978-1-59593-764-3
- [21] Inbal Yahav "Bid based scheduler with backfilling for a multiprocessor system" ACM International Conference Proceeding Series; Vol. 258, Proceedings of the ninth international conference on Electronic ,Pages: 459 – 468, Year of Publication: 2007 ,ISBN:978-1-59593-700-1

# Energy Efficient MAC Protocol for Body Centric Nano-Networks (BANNET)

S.Sivapriya

Department of Electronics & Communication  
College of Engineering, Chennai  
sivapriya.bsn@gmail.com

Dr.D.Sridharan, Professor

Department of Electronics & Communication  
College of Engineering, Chennai  
srid.cegece@gmail.com

**Abstract**—The micro-scaled biosensor networks have some limitations in cellular level precision and do not provide accurate medical care. This leads to the development of nano biosensor networks which will streamline the diagnostic process aid in the treatment of patients through accurate and localized drug delivery and tumor detection in healthcare system. In this paper, a comparative analysis between an energy efficient MAC protocol for body centric nano networks (EEMAC-BANNET) and an energy efficient MAC protocol for Terahertz band communication (EEWNSN-MAC) have been done. The channel model for human tissue environment have been developed and used as a channel for simulation purposes. The main objective of this work is to study and analyze the efficiency of nano nodes under different channel condition (free space & human tissue environment). This evaluation is done by using Nano-Sim module in Network Simulator 3 (NS-3) for critical metrics in WNSN such as packet loss ratio and energy consumption/packet with different density of nano-nodes.

**Keywords**—EEMAC-BANNET, Medium Access Control (MAC), Nanosensors, Wireless Nano-Sensor Networks (WNSN)

## I. INTRODUCTION

Although micro-scaled body sensor nodes have significantly advanced health care options, these WBANs have restricted applicability under certain constrained situations such as invasive cellular level treatment. The component sensors of a microscaled WBAN are unable to provide absolute invasive medical care and also fail to provide cellular-level precision in measurement and actuation. The latest advancements in nanotechnology have provided a suitable solution to the aforementioned limitations of the micro scale WBAN based health care.

Nano medicine is a field with continuous progress, introducing novel applications in many health care areas. Some of the most promising areas are: nano diagnostics, nano pharmaceuticals, reconstructive surgery, nano robotics, nano surgery and ultrafast DNA sequencing. Nano carriers have great potential in the field of drug delivery, medical diagnostics, therapeutics and molecular targeting. The specific targeting of nano carriers to cells or organs and the controlled release of drug improves efficacy at the target and reduces toxicity in non-target organs.

The recent boost in nanotechnology fosters extension of control and networking to nano scale by deployment of nanosensors having a size of one to few hundred nanometers.

Being equipped with a nano-antenna, a memory, a CPU, and a power supply, such nanosensors are enabled to perform simple operations and wireless communication in short distances. Such wireless communication among nanosensors bolsters emergence of a new paradigm called wireless nanosensor networks. In the field of medicine and health care, nanomachines can monitor inaccessible parts of the human body, such as the aortic heart valve and collect sensory data of the valve. A nano network, in turn, can carry the sensory data to an external device such as a Smartphone or an Internet gateway enabling nanodevices to wirelessly communicate with powerful external processing devices. A nanonetwork connected to internet gateways enables a new network paradigm called the IoNT.

As discussed in [16], the two main novel communication techniques used for intrabody application are: molecular communication and electromagnetic communication. The electromagnetic communication techniques proved to be better than molecular communication techniques for intrabody applications. In this paper, electromagnetic communication based on the modulation and demodulation of electromagnetic waves using components that are made based on novel nano materials such as graphene and carbon nano-tubes.

In general with reference to complete healthcare systems with IoT & IoNT architecture as shown in Fig.1, monitoring devices communicate with the nanointerface and a network coordinator (which provides the connectivity with a remote health-care server through a wireless/wired broadband technology) by using IEEE 802.15.4285 radios.

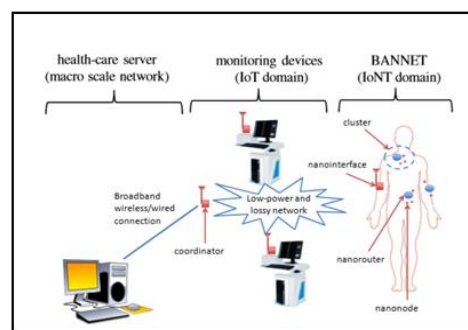


Fig 1: Complete healthcare system with IoT & IoNT paradigms [5]

In fact, it is assumed that each monitoring device is configured for tracking some biological functionalities of a

given health-care server (macro scale network). To this end, it sends specific request messages to the nanointerface of the BANNET through the IoT network infrastructure. The nanointerface will deliver the received request to all nanorouters, thus allowing them to retrieve an answer from their corresponding clusters. Then, the requests generated by a sub-set of nanonodes are sent back to the monitoring device in the opposite direction.

## II. RELATED WORKS

Due to special characteristics of nanonetworks, traditional wireless MAC protocols (e.g. TDMA, CDMA, CSMA /CA) or micro scaled sensor network protocols are not applicable in the domain of nanonetworks.

Jornet et al. proposed and analyzed a MAC protocol, PHLAME [10] which makes use of joint estimation of transmitter and receiver channel and modulation characteristics so that the interference and packet loss would be reduced. However, the transmitter and the receiver may not have sufficient computational resources to find the suitable computation parameter and the handshake operation may reduce the terahertz band real transmission speed and the achieved throughput which also makes synchronization as an issue. Later, Mohrehkesh [11] proposed a receiver initiated MAC protocol for nano-sensor networks which describes both centralized and distributed scenario of nanosensor network by using coordinated energy consumption scheduling algorithms. If there is no synchronization used among nano-nodes, several ready to receive (RTR) packets will be sent with no DATA replay. Also the transmitters (nanonodes) may listen for RTR packets, but may not receive anyone. In both cases, energy would be wasted. In [13], Wang et. al. developed energy and spectrum aware MAC protocol for centralized nano-sensor network which uses symbol compression scheduling and packet level timeline scheduling algorithm. In downlink situation when the nano controller sends data to the nano-node, as the nano-nodes energy is limited, this data packet could be sent several times until all the nano-nodes in the field received the data. This will waste the limited nano-nodes energy. TAB-MAC protocol by Xin-Wei-Yao and Josep Miquel Jornet [9] uses beamforming techniques of antennas with simple handshaking process for channel accessing.

In EEWNSN-MAC by Negar Rikhtegar [3], mobile multihop communication among nano-nodes is discussed. The author tries to discuss about the energy efficient MAC protocol for nano-sensors via multihop communication in centralized approach. The main issue in the related work is the technological limitations of nano-scale devices are not considered and proper channel model for the defined protocol is not defined.

This paper is organized as follows: Section III will provide the system model for body centric nano-networks. Section IV discusses about channel model for human tissue at nano scale in intrabody applications. In Section V, simulation proceedings of the MAC protocol under two different channel conditions have been simulated using NS3 and their performance is analyzed. The future scope of the work has

been discussed in Section VI and concludes the paper in Section VII.

## III. SYSTEM MODEL

In this section, we present a system model for EEMAC-BANNET. The system model for EEMAC-BANNET is same as EEWNSN-MAC as shown in Fig.2 and the described MAC protocol in [3] is used for comparison under two different channel conditions (general & human tissue environment).

The mobile nano-nodes which are around the cluster but far off from nano-router reach the cluster head (nano-router) via multi hop communication through cluster members (nano-nodes connected to nano-router). The nano-node which is nearer to the nano-router will be allotted different time slots for transferring the collected information via TDMA scheduling. The procedure for MAC protocol can be divided into two phases: Route selection by mobile nanonodes & scheduling algorithm by nano router. The mobile nano node which has a data packet to transmit to nano-router communicates via. Multi hop communication of fixed nano nodes.

The mobile node which has a packet to transmit sends REQUEST MESSAGE to neighboring nano nodes. The nanonodes which are in the transmission range of nano-router replies back with reply message containing the available energy and node type (nano-router or nano-node).The mobile nano-node on receiving the REPLY MESSAGE based on available energy selects the next hop. If the neighboring node is a nano-router, then it directly sends the data to nano-router instead of multi hop communication. This has been explained in Fig 3.

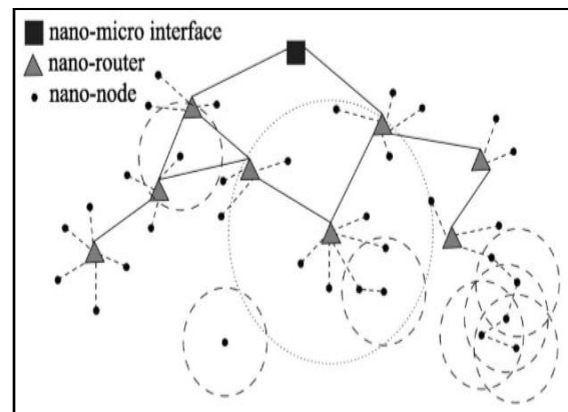


Fig. 2. Mobile Multihop Communication in EE-WNSN [3]

The nano-router periodically broadcasts NEIGHBOR DISCOVERY MESSAGE to know whether any cluster members have packets to deliver. The nanonodes which have packet to transmit replies with ACK MESSAGE. Based on this information from different nano nodes, the nano-router assigns different time slots to different nano-nodes as explained in Fig 4.

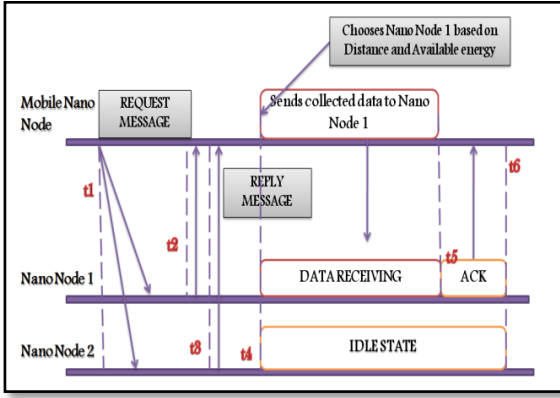


Fig. 3. Timing Diagram for Route Selection of Mobile Nano Nodes [3]

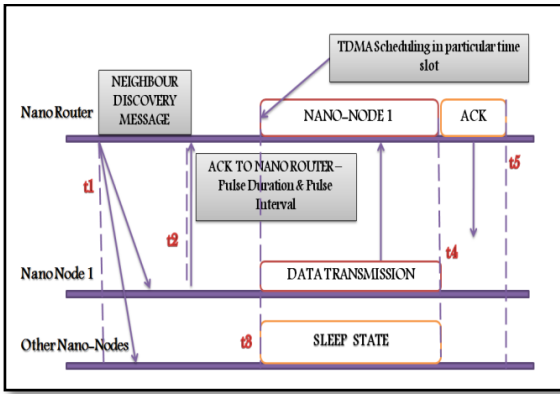


Fig. 4. Timing Diagram for Intra-cluster Communication [3]

Consequently, the idle periods and the data transmission to the closest nano-router can highly decrease the energy consumption of nano-nodes and prolong the network lifetime.

#### IV. CHANNEL MODEL UNDER HUMAN TISSUES AT NANOSCALE FOR INTRABODY APPLICATION

EM communications in the terahertz band (i.e., 0.1–10.0THz) can be supported by graphene-based nano-antennas. As terahertz band communication for intra body application do not support high frequency carrier signals for transmission, a low-weight simple modulation technique has to be used for body centric nano-networks. The dielectric properties of the human body in [17] are tabulated in Table 1. As the nano-nodes are placed inside muscles, their dielectric properties are taken into account for designing a channel model for body centric nano-networks.

TABLE 1: DIELECTRIC PROPERTIES OF HUMAN BODY

Dielectric Properties	SKIN	FAT	MUSCLE
Permittivity	47.6	12.1	57.6
Conductivity	0.71	0.01	0.85

Generally, terahertz communications use impulse radio-ultra wide band (IR-UWB) communication, but it is too complex for a nano-node to support this kind of communication due to its energy scarcity problems. So, time spread on-off shift keying (TS-OOK) is used where logic 1 represents a single bit and logic 0 represents silence which reduces energy requirement of nano-node to a desired level.

The signal at the receiver can be expressed based on the Friis equation (1):

$$P_{RX} = P_{TX} + G_{TX} + G_{RX} - P_{PL} - A \text{ (dB)} \quad (1)$$

where  $P_{RX}$ ,  $P_{TX}$ ,  $G_{TX}$ ,  $G_{RX}$  and  $P_{PL}$  are the received power, transmitted power, gain of the transmitting antenna, gain of the receiving antenna and path loss incurred due to different channel conditions. Moreover, the path loss,  $A$ , i.e., the power attenuation contributed by the human tissues can be of two forms namely the spreading path loss,  $A_s$ , due to the expansion of the waves in tissues, and the absorption path loss,  $A_a$ , due to the absorption of tissues. According to [18], it can be expressed as equation (2):

$$A = A_s + A_a = 10 \cdot \log(4\pi d/\lambda_g) + 10 \cdot \log(\alpha d) \quad (2)$$

where  $d$ ,  $\lambda_g$ , and  $\alpha$ , are the propagation distance of the wave, the wavelength in the considered medium, and the absorption co-efficient measuring the amount of absorption loss of the EM wave in the medium, respectively.

$$P_{PL} = 10 \cdot n \cdot \log(d/d_0) \text{ (dB)} \quad (3)$$

where  $d_0$  is the reference distance 0.5mm, and  $n$  is the path loss exponent. Since the analysis is for two different channel conditions the path loss exponent for human tissue environment is 4.65-6.26 as given in [18]. The value of fitting coefficient  $n$  is based on the communication channel. The on-body path loss characteristic was formulated based on the path loss as a function of distance.

#### V. SIMULATION PROCEEDINGS

The performance evaluation of the existing system in free space and under human tissue implant have been evaluated using NANO-SIM module in NS3(Network Simulator 3). The simulation parameters considered have been listed in Table 1. The evaluation parameters are total energy consumption sent (or) received/packet and packet loss ratio(PLR).

**Energy Consumption Parameter:** One of the challenges in WNSN is the limited energy resource that is stored in a nano battery. As described, for pulse-based communication, it can be observed that the energy is consumed only during transmission of “1”. The energy is needed to send and receive a packet is given by equations (4) & (5):

$$E_{\text{packet-tx}} = N_{\text{bits}} * W * E_{\text{pulse-tx}} \quad (4)$$

$$E_{\text{packet-rx}} = N_{\text{bits}} * W * E_{\text{pulse-rx}} \quad (5)$$

where  $E_{\text{pulse-tx}}$  and  $E_{\text{pulse-rx}}$  are the energies that are needed in order to transmit and receive a pulse, respectively and  $N_{\text{bits}}$  represents the number of bits per packet. The code weight  $W$  is assumed to be 0.5, considering equal number of 0’s and 1’s in a packet.

**Packet Loss Ratio (PLR):** The PLR parameter refers to the ratio of number of packets lost due to nano-node incapability and different channel condition to total number of packets sent. The packet loss ratio can be estimated from equation (6):

$$\text{PLR} = (\text{packet}_{\text{sent}} - \text{packet}_{\text{received}}) / \text{packet}_{\text{sent}} \quad (6)$$

where  $\text{packet}_{\text{sent}}$  is the number of packets sent by sources (nano-nodes) and the  $\text{packet}_{\text{received}}$  is the number of successfully received packets by nano-micro interface.

TABLE 2: SIMULATION PARAMETERS

PARAMETER	VALUE
Available energy of Nano-node	800pJ
Range of nano-nodes	0.001m
Transmitted energy/pulse	1pj
Received energy/pulse	0.1pj
Pulse duration	100fs
Pulse interval time	10ps
Available energy of Nano-router	15uJ
Range of nano-routers	0.02m
Packet Size	1024 bits

In simulation scenario, we considered EEWNSN-MAC [3] as our criteria and evaluated the results under human tissue environment. The results are compared with the existing work [3] which uses general channel model and EEMAC-BANNET which uses human tissue channel model for the above mentioned performance metrics which have been tabulated in Table 3.

TABLE 3: SIMULATED RESULTS

No. of nano-nodes	Packet Loss Ratio (%)		Energy cons/packet (nJ)	
	EEWNSN-MAC	EEMAC-BANNET	EEWNSN-MAC	EEMAC-BANNET
10	25	60	0.32	1.01
20	18	46	0.40	1.04
30	15	26	0.50	1.08
40	8	10	0.60	1.11

Fig. 5. Packet Loss Ratio for EEWNSN-MAC and EEMAC-BANNET

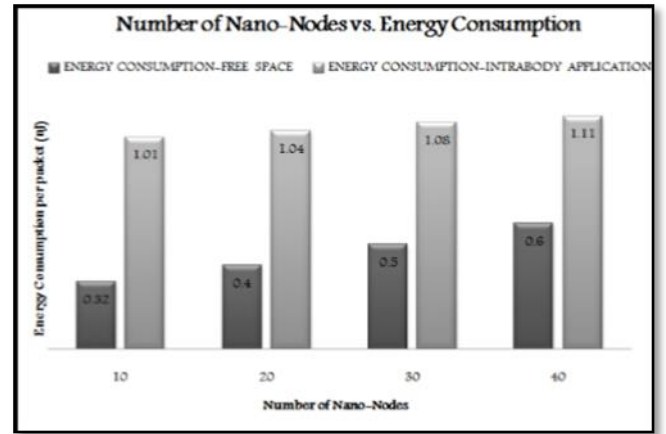


Fig. 6. Energy Consumption Parameter of EEWNSN-MAC and EEMAC-BANNET

From the above results as shown in Fig 5 & Fig 6, the average PLR for EEWNSN-MAC and EEMAC-BANNET is 10.49% & 35.15% respectively and the average Energy consumption/packet for EEWNSN-MAC and EEMAC-BANNET is 0.5nJ and 1.05nJ respectively.

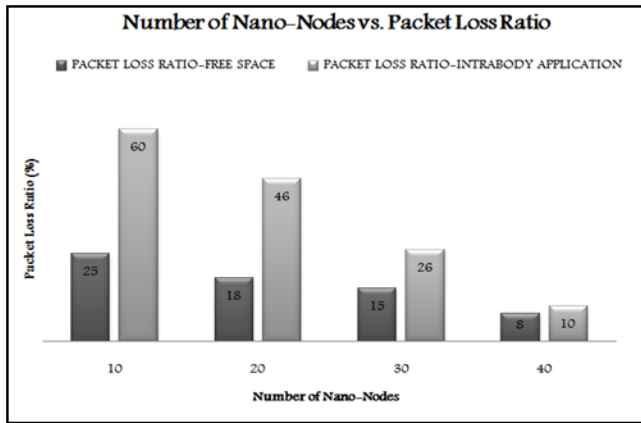
The system model and defined MAC protocol [3] is for general Terahertz band communication and does not describe it for specific application purposes. In specific for healthcare applications, the MAC protocol should be based on an energy control mechanism with all technological limitation considerations which would yield less delay with maximum throughput for the different payloads. Hence, the system model in existing work [3] is NOT SUITABLE for BODY CENTRIC NANO-NETWORKS and the PLR & energy consumption parameter is very high in human tissue environment which can be reduced by efficient transmission algorithms.



Thus, a novel MAC protocol have to be implemented which should overcome aforementioned problems which yields low complex and high energy efficient MAC for Terahertz band intrabody applications.

## VI. FUTURE SCOPE

The later proceedings of this particular aspect of study



would be analysis of transmission parameters under human tissue at nanoscale for Terahertz band communication. As the nano sensors are limited in energy resources, they have to be efficiently used with proper harvesting mechanisms.

Conventional energy harvesting mechanisms cannot be used for nano-scale technologies because of technological limitations. The energy harvesting mechanisms can be implemented by means of mechanical vibrations such as air-conditioning, heart beat and so on. But the energy harvesting mechanism in BANNET is quite different and different analysis and studies have to be made for energy efficient MAC in Nano sensor networks.

Hence a different state of art technique need to be implemented for body centric nano networks which takes into account the effects of energy harvesting mechanisms. This in turn reduces PLR and increases energy efficiency parameter.

## VII. CONCLUSION

WNSNs will boost the applications of nanotechnology in many fields of our society, ranging from healthcare to homeland security and environmental protection. In this paper, we analyzed EE-WNSN MAC protocol for two different environments and finally made an conclusion that the traditionally developed MAC protocol is not suitable for Body centric Nano-NETworks (BANNET). In order to increase the energy efficiency and to enhance the proper use of limited nano-nodes capability, proper scheduling mechanism with allotted time slots have to be implemented. The algorithm should distribute transmission and sleeping time slots between the nanosensors according to the available energy and channel conditions.

## REFERENCES

- [1] I.F. Akyildiz, F. Brunetti, C. Bl'azquez, "Nanonetworks: A New Communication Paradigm", Elsevier Computer Networks Journal, vol. 52, no. 12, pp. 2260-2279, May 2008.
- [2] M. Gregori, I.F. Akyildiz, "A New NanoNetwork Architecture Using Flagellated Bacteria and Catalytic Nanomotors", IEEE Journal on Selected Areas in Communications, vol. 28, no. 4, pp. 612-619, May 2010.
- [3] Negar Rikhtegar, Manijeh Keshtgari, Zahra Ronaghi, "EEWNSN: Energy Efficient Wireless Nano Sensor Network MAC Protocol for Communications in the Terahertz Band", Springer Journal, June 2017.
- [4] N. Agoulmine, K. Kim, S. Kim, T. Rim, J.-S. Lee and M. Meyyappan, "Enabling communication in bio-nanosensor networks: toward innovative healthcare solutions," IEEE Wireless Communications, vol. 19, no. 5, pp. 42-51, 2012.
- [5] Giuseppe Piro, Gennaro Boggia, and Luigi Alfredo Grieco, "On the design of an energy-harvesting protocol stack for Body Area Nano-NETworks", Nano Communication Networks Journal (Elsevier), November 2014.
- [6] J.M. Jornet, I. F. Akyildiz, "Femtosecond-long pulse-based modulation for terahertz band communication in nanonetworks", IEEE Trans. on Commun.62 pp.1742-1754, 2014.
- [7] A. Dohr, R. Modre-Oprian, M. Drobics, D. Hayn, G. Schreier, "The Internet of Things for Ambient Assisted Living", in: IEEE Int. Conf. on Information Technology: New Generations, ITNG, pp. 804-809,2010.
- [8] Balasubramaniam, Sasitharan, and Jussi Kangasharju. "Realizing the internet of nano things: challenges, solutions, and applications." ,2008.
- [9] Josep Miquel Jornet , Joan Capdevila Pujol and Josep Solé Pareta , " PHLAME: A Physical Layer Aware MAC protocol for Electromagnetic nanonetworks in the Terahertz Band" in Nano Communication Networks (Elsevier),January 2013.
- [10] S. Mohrehkesh and M. C. Weigle. "RIH-MAC: receiver initiated harvesting-aware MAC for nanonetworks." Proceedings of ACM, pp. 3-9, 2014.
- [11] Rawan Alsheikh', Nadine Akkari and Etimad Fadel , "MAC Protocols for Wireless Nano-sensor Networks: Performance Analysis and Design Guidelines" in IEEE Trans. On Communication , 2016.
- [12] Pu Wan, Ian F. Akyildiz , " Energy and spectrum-aware MAC protocol for perpetual wireless nanosensor networks in the Terahertz Band", in Nano Communication Networks (Elsevier),June 2013.
- [13] J.M. Jornet, I.F. Akyildiz, "Graphene-based nano-antennas for electromagnetic nanocommunications in the terahertz band", in: Proc. of 4th European Conference on Antennas and Propagation, EUCAP, April 2010.
- [14] N. Agoulmine, K. Kim, S. Kim, T. Rim, J.-S. Lee and M. Meyyappan, "Enabling communication and cooperation in bio-nanosensor networks: toward innovative healthcare solutions," IEEE Wireless Communications, vol. 19, no. 5, pp. 42-51, 2012.
- [15] Qammer Abbasi, Nishant Chopra, "Nano-Communication for Biomedical Applications: A Review on the State-of-the-Art From Physical Layers to Novel Networking Concepts", IEEE access letters, July 2016.
- [16] Kamran Sayrafian-Pour ,Kamya Yekeh Yazdandoost "A Statistical Path Loss Model for Medical Implant Communication Channels", IEEE Transactions,2015.
- [17] Zekeriyya Esat Ankarali ,Qammer Abbasi "In Vivo Communications: Steps toward the Next Generation of Implantable Devices", IEEE Vehicular Technology, June 2016.
- [18] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

# Comparison of Machine Learning Algorithms - Opinion Mining for Social Network Data

G.P.Sriprasath, S.Shanmuga Sundaram, R.Parkavi

Information & Technology Department

Thiagarajar College of Engineering

Madurai, India

Sriprasath101197@gmail.com

**Abstract**—This paper deals with comparison of 3 data mining algorithms which come under classification based on predictive modeling. In this work of machine Learning, the current scenario of conflicts between North Korea and USA is analyzed by extracting opinions from online social media twitter by using three algorithms, namely, Decision Tree (J48), Naive Bayes and K-Nearest Neighbor (IBK) in WEKA tool. By comparing the results of three algorithms it has been found that the K-Nearest Neighbor (IBK) algorithm is more efficient than the Decision Tree algorithm and Naive Bayes algorithm. It is concluded that K-Nearest Neighbor (IBK) algorithm is more preferable to Decision Tree and Naive Bayes algorithms with respect to accuracy and speed, not only to small datasets but also for large datasets, to form opinion for such current scenario from online social media.(Abstract)

**Keywords**—Machine Learning, Naive Bayes algorithm, Decision Tree (J48) algorithm, K-Nearest Neighbor (IBK) algorithm, WEKA tool, online social media.

## I. INTRODUCTION

There are many situations in the present world where predictions have to be made in current scenario. These predictions are generally made using sample surveys. These sample surveys consume large volume of man-power as well as cost and time. But heaps of information can be extracted from the opinions of the users of social media. This information is very helpful to make major predictions/decisions without using much time and energy. This information can be compiled into datasets which can be analysed using various data mining algorithms to make major prediction/decision in a political scenario.[7] There are plenty of data mining algorithms to analyse the datasets from online social media.

Data mining can be classified into two models Descriptive and Predictive Models.[2] In Descriptive model patterns in a sample data can be determined and subdivided into clustering, summarization and association rules. Predictive model is a process that forecasts results/outcomes by adopting probability methods. They are classified into three methods, namely, classification, regression and time series.

Classification is a data analysis technique that extracts models which predict categorical class labels such as discrete or unordered.[1] The opinions from online social media on the

prevailing situation of conflict between North Korea and US is taken as the dataset. This paper compares three algorithms namely, Naive Bayes algorithm, Decision Tree algorithm and K-Nearest Neighbor algorithm, coming under classification methods of algorithm in Predictive model using the sample datasets. It does not predict the results/decisions of the online social media opinions on the conflicts between North Korea and US. WEKA is the tool used to analyze these algorithms. The sample data set, used for comparing the three data mining algorithms, has been prepared by filtering the set of opinions on the current prevailing scenario between North Korea and USA. The various opinions of the online social media users about the scenario are taken as the input for the algorithm.[3] Outputs of these three algorithms are compared on the basis of parameters such as accuracy, TP Rate, Time taken, Kappa statistics, FP Rate, ROC area etc. Based on the analysis, K-Nearest Neighbor algorithm has proved to be more efficient than Decision Tree and Naive Bayes algorithms with respect to speed and accuracy.

## II. LITERATURE SURVEY

### A. Data Classification: Algorithms and Applications

This book written by Charu C. Aggarwal explains various algorithms and applications of classification addressing various problematic areas such as multimedia, biological data, text, social network. The book initially deals with common techniques of classification, including instance-based methods, rule-based methods and support vector machine methods. The book then explains particular methods used for data domains like time-series, network, uncertain data and discrete sequence,. It also covers data sets of various sizes. The book concludes with analysis of differences of the classification process. It discusses transfer learning, ensembles, rare-class learning, active learning, rare-class learning , and semi-supervised learning as well as evaluation aspects of classifiers.

### B. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples and Case Studies

This book written by John D. Kelleher and Brian Mac Namee introduces machine learning methods of predictive data analysis including theoretical hypothesis and practical implementations. This introductory textbook offers a detailed

and focused treatment of the most important machine learning approaches used in predictive data analytics, covering both theoretical concepts and practical applications. Technical and mathematical material is augmented with explanatory worked examples, and case studies illustrate the application of these models in the broader business context. It is added with detailed worked examples of technical and mathematical materials. It also includes many practical instances explaining the applications of these models. This book explains the following 4 methods of machine learning: similarity based learning, error based learning, information based learning and probability based learning. Each of these methods is dealt by a nontechnical description of basic concept, followed by mathematical examples and algorithms explained by detailed worked examples.

### C. Machine Learning: An Algorithmic Perspective

This book written by Stephen Marsland assists the students in grasping the concepts of machine learning algorithms. It includes the practical implementation and programming of the various concepts in machine learning. It contains experimentation of the Support Vector Machine that is used for its implementation, and also introduces various python functions which helps in making the code simpler at many instances of programming. It also includes the statistical interpretations of the machine learning algorithms.

### D. Machine Learning Algorithms: A reference guide to popular algorithms for data science and machine learning:

This book written by Giuseppe Bonaccorso focuses on the feature selection and Feature engineering process of machine learning algorithms. After building a data model it assists us to learn the behaviour of various algorithms on that model. It allows us to build a ML architecture from its initial stage. It also allows us to build clusters and learn Natural Language Processing. It allows to learn the concepts of supervised, unsupervised and semi-supervised learning and also provides us the practical implementations of various famous algorithms.

### E. Word Sense Disambiguation: Aspects Concerning Feature Selection:

This book written by Florentina T. Hristea provides us the recent progress of the naïve bayes model in unsupervised word sense disambiguation(WSD). It discovers the use of Naïve Bayes algorithm in unsupervised WSD which leads to more disambiguation results which remains to be insufficiently explored. It explores different sources of knowledge for feature selection such as WordNet, dependency relations and web N-grams using Naïve Bayes algorithms. It highlights the uses of knowledge feeding to a knowledge-lean algorithm which uses the Naïve Bayes algorithm for unsupervised WSD.

## III. WEKA TOOL

Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms and tool for data mining tasks. The algorithms can either be applied directly to a data set or called from written java code. It contains a set of

tools for data preprocessing, classification, clustering, association, attribute selection and visualization through Pilot matrix. There are plenty of parsing attributes under supervised and unsupervised filters which can be utilized for preprocessing of the dataset. There are varieties of classifiers under the headings of bayes, lazy, trees, functions and Meta which can be used for analyzing the data. It is fully built by using Java programming language which makes it very easy to run on any modern computing platforms. It supports GUI interface to easy access of the functions.

## IV. ALGORITHMS

The three algorithms compared in this work are 1) Naive Bayes algorithm which comes under Bayes type classifier, 2) Decision Tree (J48) algorithm which comes under Tree type classifier and 3) k-nearest neighbor algorithm which comes under Lazy type classifier. Brief descriptions of those three algorithms are given below:

### A. Naive Bayes Algorithm

It is a machine learning algorithm for classification problems. It comes under bayes type of classifiers. It is primarily used for text classification, which involves high dimensional training data sets.[8] A few examples are spam filtration, sentimental analysis and classifying news articles. It is represented by the equation shown in Fig. 1.

$$P(c | x) = \frac{P(x | c) P(c)}{P(x)}$$

$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$

Fig. 1. Equation for Naive Bayes Algorithm

Where  $P(X|C)$  denotes the conditional probability of occurrence of event X given the event C is true,  $P(X)$  and  $P(C)$  denote the probability of the occurrences of event X and event C respectively and  $P(C|X)$  denotes the probability of occurrence of event C given event X is true.

### B. Decision Tree Algorithm

A decision tree algorithm is a predictive modeling approach used in data mining. It uses a tree structure to specify sequences of decisions and consequences.[9] A decision tree employs a structure of nodes and branches. The depth of a node is the minimum number of steps required to reach the node from the root and eventually a final point is reached and a prediction is made. It implicitly performs variable screening or feature selection. Decision Tree (J48) algorithm in decision tree is chosen to analyze the topic.

### C. K-Nearest Neighbor Algorithm

It is a simple classification and regression algorithm. It is a supervised learning algorithm. The default method that is used in this algorithm is Euclidean distance.[10] It can be defined as a type of instance Bayes learning. It is the simplest

algorithm among all the other data mining algorithms. A peculiarity of this algorithm is that it is sensitive to the local structure of the data.[10]

Each algorithm is being compared by its performance values of each factor. The significance of each performance factor is explained below:

- The number of correctly and incorrectly classified instances shows the percentage of sample instances that were correctly and incorrectly classified.
- A true positive test result (TP) is one that matches the condition when that condition is present.
- A true negative test result (TN) is one that does not match the condition when that condition is absent.
- A false positive test result (FP) is one that matches the condition when that condition is absent.
- A false negative test result (FN) is one that does not match the condition when that condition is present.
- Sensitivity measures the ability of a test to match the condition when that condition is present. So, Sensitivity = TP/ (TP+FN).
- Specificity measures the ability of a test to correctly exclude the condition (not match the condition) when that condition is absent. So, Specificity = TN/ (TN+FP).
- Predictive value positive is the percentage of positives that correspond to the presence of that condition. Thus, Predictive value positive = TP/ (TP+FP).
- Predictive value negative is the percentage of negatives that correspond to the absence of that condition. So, Predictive value negative = TN/ (TN+FN).

The percentage of correctly classified instances is mostly called accuracy or sample accuracy. For more accuracy some of the other measures, ROC Area or area under the ROC curve, are the preferred one.

Kappa value is a metric that measures the agreement between the classifications and the true classes.[5] It is calculated by taking the expected agreement from the observed agreement and dividing the maximum possible agreement. A value greater than zero indicates that the classifier is doing better than chance.

ROC area measurement is one of the most important values output by WEKA. An Optimal classifier will have ROC area values approaching to 1, with 0.5 being comparable to random guessing (similar to a Kappa statistic of 0).

### V. DATASET

The current prevailing situation between North Korea and USA is used as the subject. Different Opinions of the subject, gathered from online social media twitter, is taken as dataset to analyze the three algorithms of WEKA tool.[3] Out of a range of many tweets from all over the world, 5 sample datasets, each containing 27, 62, 123, 169, 193 instances respectively, are utilized as the input for the analysis. Each

dataset contains instances of class yes or no. The characteristics of those 5 datasets are given in Table 1.

TABLE 1. CHARACTERISTICS OF SAMPLE DATASETS

SLNO	DATASET	INSTANCES	ATTRIBUTES
1)	Dataset27	21-6	316
2)	Dataset62	56-6	544
3)	Dataset123	117-6	1026
4)	Dataset169	138-31	1265
5)	Dataset200	151-49	1265

### VI. EXPERIMENT

First of all, the unsupervised filter String To Word Vector was applied on each sample data to filter the strings into words.[3] This String To Word Vector filter changes String attributes into a collection of attributes indicating the word occurrence data from the text present in the strings. The collection of attributes is decided by the first batch filtered data. This filter generated distinct attributes according to the yes or no conditions given in the sample data. Next the algorithms of Naive Bayes, Decision Tree (J48) and K-Nearest Neighbor (IBK) were applied on each of the 5 sample datasets individually. The WEKA tool generated the models and the models were tested by using the training sets for each dataset and for each algorithm. Due to shortage of space, Performance Report of each algorithm for all datasets was not able to be produced, but the screenshot of output of each algorithm for the dataset containing 193 instances only is given in Fig. 2, Fig. 3 and Fig. 4.

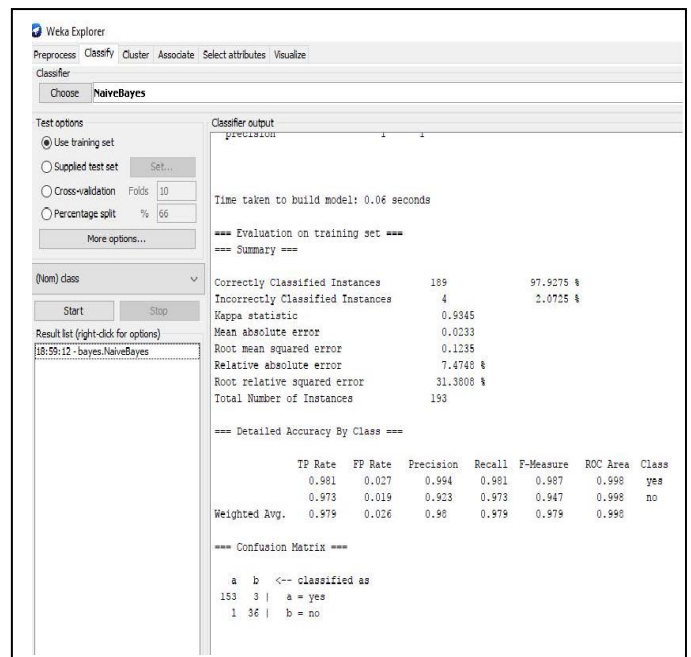


Fig. 2. Screenshot of Output for Naive Bayes Algorithm

## VII. RESULT ANALYSIS

The performance values of each algorithms on all the five sample datasets has been given in Fig. 5, 6 and 7 for each algorithm. Comparative study of three algorithms shows that:

1) The percentage of correctly classified instances is more for K-Nearest Neighbor (IBK) as compared to Decision Tree (J48) and Naive Bayes as given in Fig. 8 and the percentage of incorrectly classified instances is very meager as compared to Decision Tree (J48) and Naive Bayes as given in Fig. 9. Hence K-Nearest Neighbor (IBK) algorithm performs more accurate than Decision Tree (J48) and Naive Bayes algorithms, with respect to precision.

2) When the number of instances in dataset increases, it does not affect much to the number of incorrectly classified instances for K-Nearest Neighbor (IBK) algorithm whereas the number of incorrectly classified instances for Decision Tree (J48) and Naive Bayes algorithms increases significantly as given in Fig. 9.

3) With respect to confusion matrix, the number of cross classification of instances is very meagre for K-Nearest Neighbor (IBK) algorithm whereas it is considerably high in case of Decision Tree (J48) and Naive Bayes algorithms as given in Fig. 5, 6 and 7.

4) The time taken for building the model is almost zero for K-Nearest Neighbor (IBK) algorithm, but it is significantly high for Decision Tree (J48) and Naive Bayes algorithms as given in Fig. 5, 6 and 7.

5) The time taken for testing model with training data is negligible for Decision Tree(J48) algorithm whereas it is considerably less for K-Nearest Neighbor(IBK) algorithm as compared to Naive Bayes algorithm as given in Fig. 5, 6 and 7.

### A. Performance Report for each Algorithm

SLNO	DATASET	INSTANCES (YES/NO)	ATTRIBUTES	NAÏVE BASE ALGORITHM				TIME FOR BUILDING MODEL (sec)	TIME FOR TESTING MODEL WITH TRAINING DATA (sec)
				CORRECT PERCENT (INSTANCES)	INCORRECT PERCENT (INSTANCES)	CONFUSION MATRIX			
1)	Dataset27	21-6	316	100% (27)	0% (0)	21(a-a) 0(a-b)	0(b-a) 6(b-b)	0.08	0.11
2)	Dataset62	56-6	544	98.3871% (61)	1.6129% (1)	55(a-a) 0(a-b)	1(b-a) 6(b-b)	0.08	0.16
3)	Dataset123	117-6	1026	98.374% (121)	1.626% (2)	115(a-a) 0(a-b)	2(b-a) 6(b-b)	0.09	0.34
4)	Dataset169	138-31	1265	96.4497% (163)	3.5503% (6)	150(a-a) 0(a-b)	6(b-a) 13(b-b)	0.48	1.04
5)	Dataset193	156-37	1423	97.9275% (189)	2.0725% (4)	153(a-a) 1(a-b)	3(b-a) 36(b-b)	0.06	0.82

Fig. 5. Screenshot of Parameter Values for Naive Bayes Algorithm

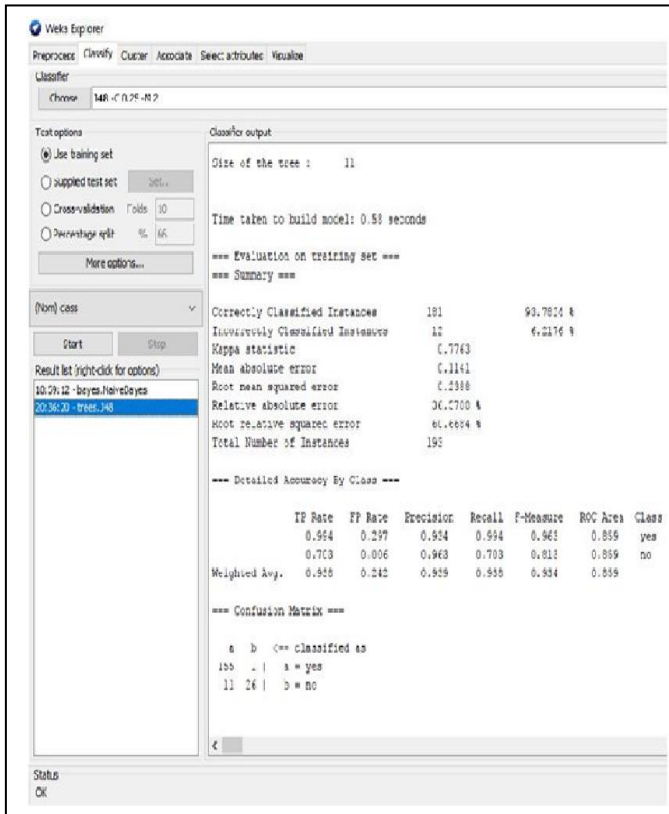


Fig. 3. Screenshot of Output for Decision Tree(J48) Algorithm

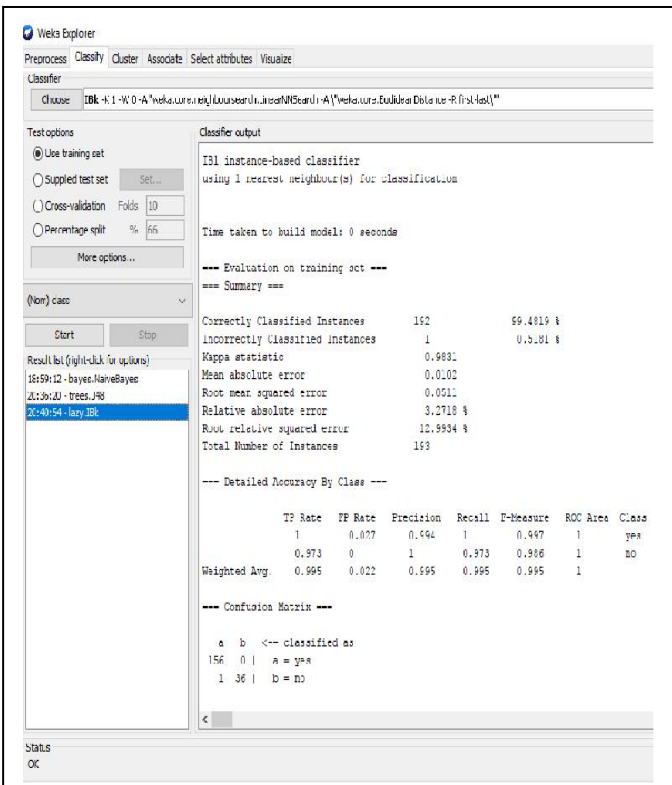


Fig. 4. Screenshot of Output for K-Nearest Neighbor(IBK) Algorithm

J48 ALGORITHM									
SLNO	DATASET	INSTANCES (YES/NO)	ATTRIBUTES	CORRECT PERCENT (INSTANCES)	INCORRECT PERCENT (INSTANCES)	CONFUSION MATRIX		TIME FOR BUILDING MODEL (sec)	TIME FOR TESTING MODEL WITH TRAINING DATA (sec)
						21(a-a)	0(b-a)		
1)	Dataset27	21-6	316	96.2963% (26)	3.7037% (1)	1(a-b)	5(b-b)	0.09	0
2)	Dataset62	56-6	544	98.3871% (61)	1.6129% (1)	56(a-a)	0(b-a)	0.14	0
3)	Dataset123	117-6	1026	99.1870% (122)	0.8130% (1)	117(a-a)	0(b-a)	0.14	0.01
4)	Dataset169	138-31	1265	95.8580% (162)	4.1420% (7)	155(a-a)	1(b-a)	1.16	0.09
5)	Dataset193	156-37	1423	93.7824% (181)	6.2176% (12)	155(a-a)	1(b-a)	0.58	0.03

Fig. 6. Screenshot of Parameter Values for Decision Tree (J48) Algorithm

IBK ALGORITHM									
SLNO	DATASET	INSTANCES (YES/NO)	ATTRIBUTES	CORRECT PERCENT (INSTANCES)	INCORRECT PERCENT (INSTANCES)	CONFUSION MATRIX		TIME FOR BUILDING MODEL (sec)	TIME FOR TESTING MODEL WITH TRAINING DATA (sec)
						21(a-a)	0(b-a)		
1)	Dataset27	21-6	316	100% (27)	0% (0)	21(a-a)	0(b-a)	0.01	0.06
2)	Dataset62	56-6	544	100% (62)	0% (0)	56(a-a)	0(b-a)	0	0.11
3)	Dataset123	117-6	1026	100% (123)	0% (0)	117(a-a)	0(b-a)	0	0.23
4)	Dataset169	138-31	1265	99.4083% (168)	0.5917% (1)	156(a-a)	0(b-a)	0.01	0.54
5)	Dataset193	156-37	1423	99.4819% (192)	0.5181% (1)	156(a-a)	0(b-a)	0	0.19

Fig. 7. Screenshot of Parameter Values for K-Nearest Neighbor Algorithm

B. Comparative Analysis of Algorithms based on each parameter using Bar Graph

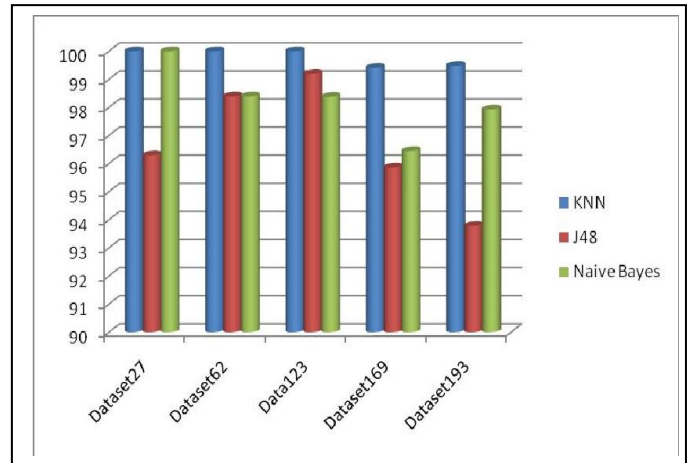


Fig. 8. Comparison of algorithms based on Correct Instances:

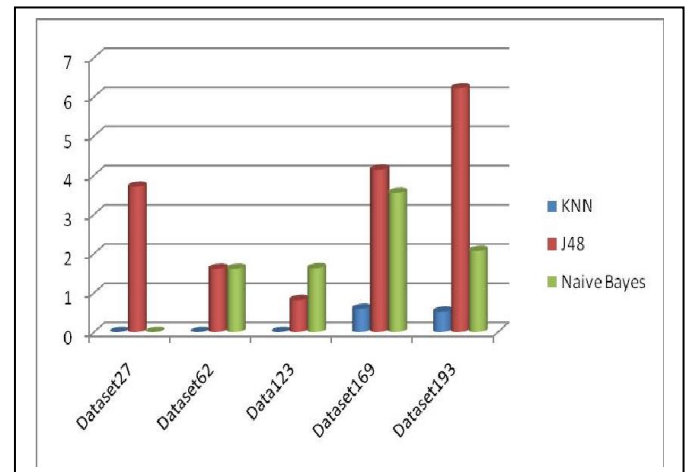


Fig. 9. Comparison of algorithms based on Incorrect Instances:

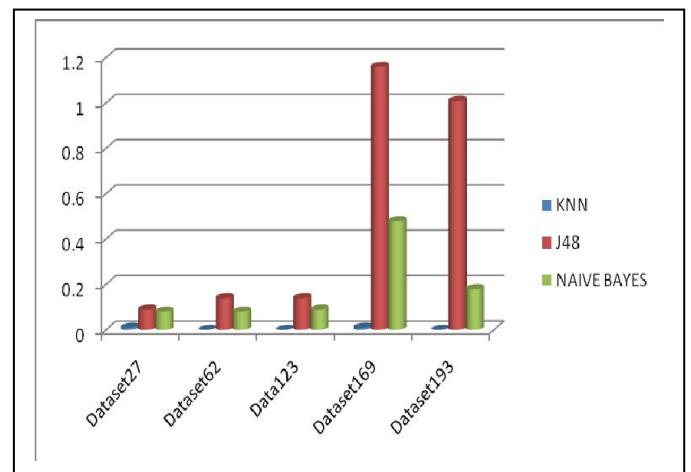


Fig. 10. Comparison of algorithms based on Time for Building the Model:

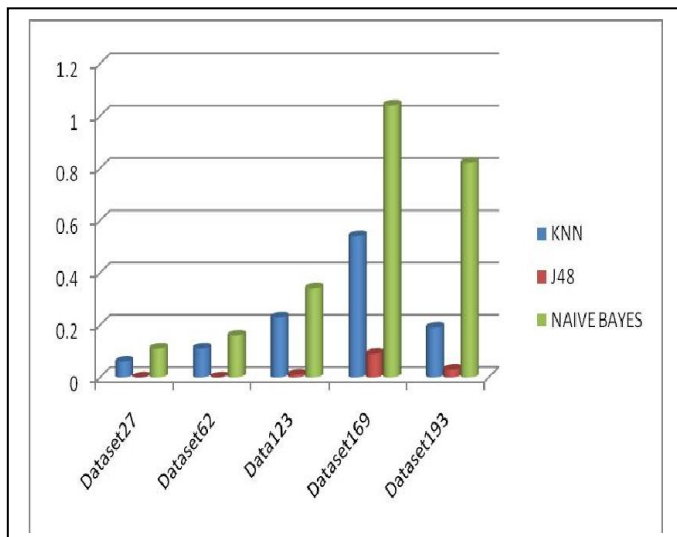


Fig. 11. Comparison of algorithms based on Time for Testing Model

### VIII. CONCLUSION

From the comparative study of performance of K-Nearest Neighbor (IBK), Decision Tree (J48) and Naive Bayes algorithms on distinctive sample sets of dataset collected from the online social media twitter on current topic of conflicts between North Korea and U.S.A. it is being concluded that K-Nearest Neighbor (IBK) algorithm performs opinion mining more accurately without much cross classification of instances as seen from Fig. 8 and 9 and also that it performs the analysis within least insignificant time by speedily building the model

and testing the model, as compared with the Decision Tree (J48) algorithm and Naive Bayes algorithm as seen from Fig. 10 and 11.

In general, to analyze and arrive at a decision/prediction for plenty of situations prevailing in the current fast growing world it is recommended to use K-Nearest Neighbor algorithm for better, accurate results within short period as compared to other algorithms.

### REFERENCES

- [1] Pang-Ning Tan, Micheal Steinbach, Vipin kumar, "Introduction to Data Mining", Pearson Publications, March 25, 2006.
- [2] Jiawei Han, Micheline Kamber and Jian Pei, "Data Mining: Concepts and Techniques" Third Edition, Morgan Kaufmann Series, Third Edition.
- [3] Mamdouh Refaat, "Data Preparation for Data Mining using SAS", Morgan Kaufmann Series, First Edition, September 29, 2006, ISBN: 9780123735775.
- [4] Paulraj Ponniah, "Data Warehousing: Fundamentals for IT Professionals", Wiley Student Second Edition, ISBN: 978-0-470-46207-2.
- [5] Tom M. Mitchell, "Machine Learning", Mc Graw Hill International Editions Computer Science Series, First Edition.
- [6] M. Young, "The Technical Writer's Handbook. Mill Valley", CA: University Science, 1989.
- [7] Tan, Steinbach & Kumar, "Introduction to data mining", Second Edition.
- [8] Xindong Wu, Vipin Kumar, "The Top Ten Algorithms in Data Mining", 1st Edition, Chapman & Hall.
- [9] Brian Steele, John Chandler, Swarna Reddy, "Algorithms for Data Science", 2016 Edition.
- [10] Bezdek JC, Chuah SK, Leep D, "Generalised k-nearest neighbor rules", 1986.

# Sensitive Data Protection in Cloud – Based on Modified Elliptic Curve Cryptographic Technique

Sumathi M

Research Scholar, Department of Computer Applications  
National Institute of Technology  
Tiruchirappalli, Tamil Nadu, India  
sumathishanjai.nitt@gmail.com

Dr.Sangeetha S

Assistant Professor, Department of Computer Applications  
National Institute of Technology  
Tiruchirappalli, Tamil Nadu, India  
sangeetha@nitt.edu

**Abstract**—Cloud computing is a current technology used to store large volume of data with minimal cost. Security is a challenging task of cloud computing. Nowadays data owners are interested to know about the security level of their data and also transferred their data to other organization for performing different task. An existing security technique offered security to complete data with high encryption / decryption time, storage cost and also provides less security to sensitive data without the knowledge of data owner and other organization admins. To provide trade-off between high security and low storage cost, the sensitive data are identified from entire data and security is applied to sensitive data instead of complete data. In the proposed method, the security mechanisms are applied to sensitive data with the knowledge of data owner and inter-organization admin. Hence, security of sensitive data is to be increased with minimal storage cost. High security is demanded by Modified Elliptic Curve Cryptography (MECC) algorithm. Elliptic Curve Cryptography yield better security through random keys, but true random key generation is a challenging task. MECC algorithm merged the pseudo random key with data owner private key and specific organization admin key. Compared to ECC algorithm MECC offers better security with the knowledge of customer and other organization admin.

**Keywords**—Sensitive Data, Modified ECC, Attribute Grouping, Inter-Organization, Storage Size.

## I. INTRODUCTION

Cloud Computing is the fastest growing technology, which is used to store and retrieve a large volume of data in an efficient manner. The Cloud Services are Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [1]. Cloud computing benefits are high reliability, scalability, flexibility and reduced maintenance cost. Two major issues of cloud storage system are focused in this paper. **In the first issue**, unauthorized internal/external members steal valuable and financial information through denial-of-services, side channel attacks, data leakage through Cloud Service Provider (CSP) or Third Party Authority (TPA). To resolve this problem strongest security and privacy techniques are required to provide data confidentiality, integrity and data access controllability. These requirements are achieved by confidentiality assured cloud data services, owner controlled cloud data sharing and

integrity guaranteed cloud data storage [2]. Existing symmetric key techniques such as Data Encryption Standard (DES), Advanced Data Encryption Standard (AES) and Homomorphic Encryption schemes are applied to both entire and selective data encryptions. The symmetric key encryption techniques strength depends on secure key exchange channel, secrecy of symmetric key and number of permutations. The drawback of existing selective data with homomorphic encryption technique is, the encrypted selective data has to be fragmented into number of parts and stored into separate places. Combining these fragmented parts into original results by authorized user is a complex task [3]. The asymmetric key techniques such as Ron Rivest, Adi Shamir, Leonard Adleman (RSA) and Message Digest (MD) are applied to entire data with larger key size. Hence, the encryption and decryption time increases [4]. To overcome these issues, strongest security mechanisms is required in selective data encryption. In the second issue, existing complete data encryption increases the storage space utilization in the cloud. Due to pay per use characteristic of cloud, the complete data encryption process increases storage cost for large volume of encrypted data [5, 6]. To resolve this issue, the sensitive data are identified from entire data and strongest security mechanisms are applied to it instead of entire data. This process will increase security to sensitive / confidential data reduces storage cost and illegal access in the cloud storage system.

The rest of the paper is organized as follows: Section II focused on related works based on Elliptic Curve Cryptography and Attribute Based Encryption techniques. Section III gives the details about Modified ECC encryption and decryption process. Section IV discusses the results of the proposed system. Finally, the paper is concluded with current work and future enhancement.

## II. RELATED WORKS

This section focuses the overview of existing concepts related to Elliptic Curve Cryptography (ECC) system and various Attribute based Encryption (ABE) systems. The existing security techniques like symmetric key techniques (DES, AES) and asymmetric key techniques (RSA, MD5) are depends on a specific key with higher key size. In comparison with RSA and ECC mechanisms, RSA requires higher key size (2048 bit) than ECC (224 bit) algorithm.



The 160bit ECC provides same security as 1024 bit RSA [7]. Rajdeep et al, analysed various encryption techniques and provided the comparison results, as ECC offered higher security level with smaller key size in a single round than other encryption techniques [9]. Compared to these algorithms ECC provides stronger security with smaller key size via a pseudorandom key. The pseudorandom numbers are generated by solving the Elliptic Curve equations. For example, consider the equation  $y^2 = x^3 + x + 4$  over  $F_{11}$ . Let  $G=(2,5)$  be a point on E and choose  $U_0 = (0,2)$  as the initial value, the EC points are  $U_i(x,y) = (9,4)$ ,  $U_i(x,y)_2 = (1001, 0100)$ ,  $B_i(U_i)_{2 \times 2} = (01,00)$ , and  $B_i(U_i)_{3 \times 3} = (001,100)$  like that the binary sequence will be generated for each pair [9]. When a data is encrypted by pseudo random key sequences, the unauthorized access will be a challenging task (identification of pseudo random number). ABE based on selective attributes encryption instead of entire data encryption provides better access control mechanisms. Table 1. show the existing works related to ECC, pseudo random number generation, ABE and selective data encryption with its benefits and drawbacks.

#### A. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Based on this discussion, the pros of existing system are:

- ECC and ABE encryption provides better security, access control and key management.
- ABE allows selective attribute encryptions.
- ECC provide better security with minimal key size and stronger security through discrete logarithms.

The cons of existing system are:

- Difficult to generate true random numbers for key generation and manual identification of random number.
- Higher cipher-text size.
- Difficult to upgrade polices and revocation in ABE.

To resolve these issues ABE technique combined with ECC technique and applied security to selected data (sensitive data) instead of whole data. In the proposed work the pseudorandom numbers dynamically generated by random number generation function, selective data encryption reduces cipher-text size and key revocation is required to it.

TABLE I. WORKS FOR ECC AND ABE

Techniques in ECC & ABE	Benefits and Issues in ECC & ABE		
	Benefits	Issues	References
Pseudo-random generators using ECC	Random key usage with the combination of x and y coordinates	Hard to generate true random numbers	10
Elliptic Curve Cryptosystem for Hierarchical Access control	Hierarchical relationship based access control and changing of secret keys	Any user can inherit any of his inheritors key straightly	11
Secure Online Bank System	Using bankers identity ID to generate public key for encryption	Double encrypted client bank record	12
Cipher Text Policy ABE	Data sharing with Collaborative Key Management Protocol. Key escrow problem is rectified. More Secure.	High Ciphertext size, high encryption and decryption cost.	13
Weighted Attribute Based Encryption	Fine grained access control.	Difficult to upgrade each modification.	14
Decentralized access controls	Anonymous authentication. Key distribution was decentralized and robust access control.	Access policies need to store in cloud. Access policies are not protected.	15
Dynamic groups for secure anti-collision data sharing	Anti-collision data sharing, low maintenance and key management.	Privacy preservation.	16
Role Based Encryption	The role hierarchy and user membership are used for encryption and decryption.	Revocation problem	17
Selective data Encryption	Time constraint based data size selection and AES is used for encryption	Security depends on AES Key size and data packet size	18

### III. PROBLEM STATEMENT AND OBJECTIVE

Nowadays data owners are interested to protect their sensitive data from others in the online process. In addition to that, inter-organizations are highly involved in cloud data transfer. Based on above analysis, "Traditional data security mechanisms are applied to entire data, without knowledge of data owner and involved inter-organization admin and also there is no trade-off between data security and usability of cost effective sensitive data storage".

To overcome these problems, the following steps are proposed in this work:

- I. To identify sensitive data from the data set based on data owner preference.

- II. To involve inter-organization members for providing better security to inter-organization transferred data.
- III. To provide high-end security to sensitive data and usability to non-sensitive data with minimal storage cost and processing time.

#### IV. PROPOSED SYSTEM

In the proposed work the whole data is categorized into two groups such as sensitive and non-sensitive data with the data owner preference. Based on Likert scale and Dichotomous scale value for an individual data owner is calculated it and stored into Response Matrix. Sensitivity and visibility value is estimated for each data owner from the response matrix. Using this sensitivity and visibility value, individual data owner privacy score is identified for each attributes and also average privacy score is estimated from it. The privacy score of individual attribute is compared to average privacy score value. The less valued privacy score attribute is identified as sensitive data and other attributes are identified as non-sensitive data [19]. Now the sensitive data is further categorized into a number of groups (5 groups) based on number of organization involved in the process – E.g Data owner private attributes, Insurance attribute, Marketing attribute, Loan attributes and attributes that are common for all.

##### A. Modified Elliptic Curve Cryptography

Elliptic Curve Cryptography (ECC) is a public key cryptosystem, used to generate keys by the key generation function. Compared to other cryptographic technique ECC provides better security through random numbers. The difficulty of existing ECC algorithm is, to generate the true random numbers. Hence pseudo-random numbers are used for key generation. Pseudo-random numbers offered less security than true random numbers. To overcome this problem, the Modified ECC (MECC) algorithm is proposed here. The dataset used for the proposed work is banking dataset. Hence information are included in the system. Fig 1. shows the system architecture for the proposed work.

The attribute groups are:

- a. **Sensitive Attribute (S)** – This information is unknown to others.
- b. **Sensitive Attributes with Loan (SL)** – Information is viewed by Load application process.
- c. **Sensitive Attributes with Marketing (SM)** – Information is viewed by Marketeing application process.
- d. **Sensitive Attributes with Insurance and Marketing (SIM)** – Information is viewed by Insurance and Marketing process.
- e. **Common Attributes** – Information is viewed by Insurance, Marketing and Loan application process.

Now each group attributes are encrypted by different keys with MECC. The inter-organization admin's are involved for providing key for encryption and encryption process.

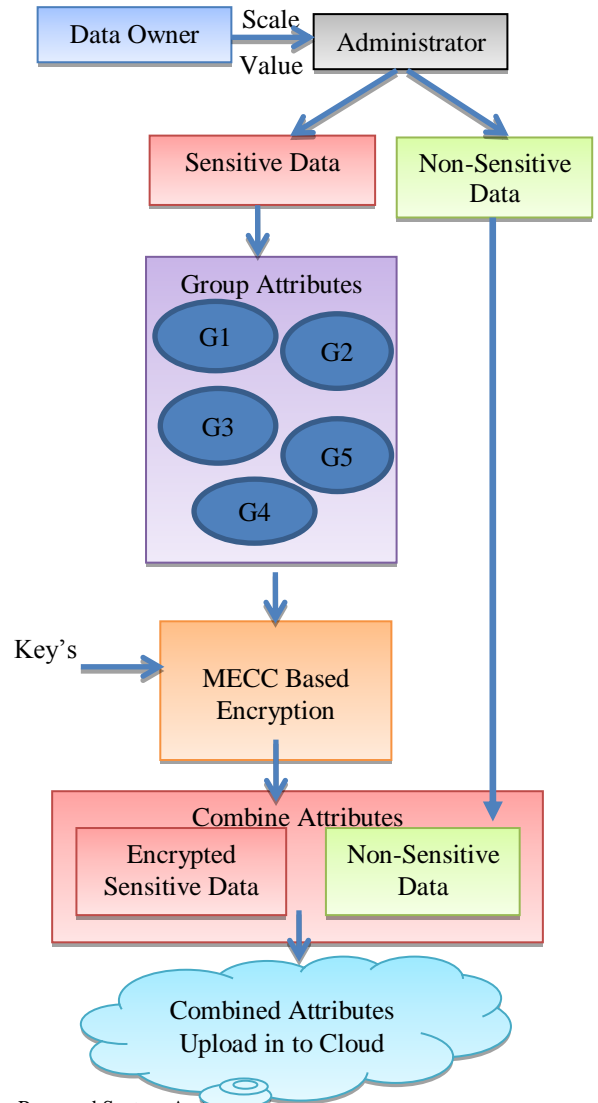


Fig. 1. Proposed System Architecture

##### B. Key Generation

The MECC uses four different keys for encryption and decryption process. The keys are

- a. **Data owner key** – Generated by bank admin and maintained by data owner.
- b. **Pseudo Random Number** – Generated by solving elliptic curve equations
- c. **Partial account Number** – Taken from data owner account number

**Group admin key** – Generated by bank admin and maintained by inter-organization admin.

When a data is sent for encryption process, the admin obtained keys from data owner, group admin and adds partial account number, pseudo random number. Now the sensitive data is encrypted by admin with four keys. The public key is generated by applying scalar multiplication between data owner's private key and partial account number. The public key is multiplied (scalar multiplication)

with pseudo random number  $k$  and added with group admin key. Now the final key is used encryption process of particular attribute group.

$P$  ← Part of Account Number  
 $P_r$  ← Private Key of Data owner  
 $P_u$  ← Public Key of Data owner  
 $k$  ← Pseudo Random Number  
 $P_k$  ← Scalar product pseudo random key and public key  
 $P_u = P_r * P$   
 $P_k = k * P_u$

The key management issue is rectified by maintaining the key by individual data owner and group admin's not by the bank admin.

### C. Modified ECC Encryption

In the encryption process the sensitive data are encrypted by bank admin with the preference given by data owner. The encryption process is given in pseudo code 1. After performing encryption process the encrypted group attributes are combined with non-encrypted attributes and the complete data are uploaded into cloud storage.

#### Pseudo Code 1- Modified ECC Encryption

**Input :**

$S_d$  ← Load sensitive dataset  
 $N_{sd}$  ← Load Non-sensitive dataset  
 $N_{ou}$  ← Number of users  
 $S_{da}$  ← Sensitive data attributes  
 $E_d$  ← Encrypted data  
 $P_r$  ← Private key  
 $P_u$  ← Public key  
 $I_k$  ← Insurance key  
 $L_k$  ← Loan key  
 $M_k$  ← Marketing key

**Output:**

$E_s, E_{ILM}, E_{IL}, E_L, E_M$  ← Encrypted Attribute Groups

**Pseudo code:**

for all users ( $1: N_{ou}$ ) do  
 $P$  ← part of account number of  $N_{ou}.i$   
 $P_u = P_r * P$   
 for all sensitive data ( $1$  to  $s_{da}$ )do  
 if  $Attr == Password$  //  $Attr == ifsc\_code$  //  
 $Attr == micr\_code$  //  $Attr == cif\_number$  //  
 $Attr == pin\_number$  then  
 $E_s = S_{da}.data + (k * P_u)$

else if  $Attr == id$  //  $Attr == email$  //  $Attr == phone\_no$   
 then  
 $E_{ILM} = S_{da}.data + (k * P_u) + I_k + L_k + M_k$   
 else if  $Attr == salary$  //  $Attr == acc\_no$  //  
 $Attr == credit\_card\_number$  then  
 $E_{IL} = S_{da}.data + (k * P_u) + I_k + L_k$   
 else if  $Attr == age$  //  $Attr == Nominee\_reg\_no$  then  
 $E_L = S_{da}.data + (k * P_u) + L_k$   
 else if  $Attr == aadhar\_no$  then  
 $E_M = S_{da}.data + (k * P_u) + M_k$   
 end if  
 end for

end for

#### D. Modified ECC Decryption

To decrypt the individual customer key, the scalar multiplication is performed between the private key, pseudo random number, part of account number. Then the product is subtracted from inter-organization key and specific encrypted group attributes instead of entire attributes. The decryption process requires specific key from data owner and inter-organization admin's. Hence, the decryption details are known by data owner and inter-organization member. The decryption details are given in pseudo code 2.

#### Pseudo code 2- Modified ECC Decryption

**Input:**

$D_d$  ← Decrypted data  
 $R_c$  ← Record for Specific customer  
 $D_k$  ← Decryption Key

**Output:**

$D_s, D_{ILM}, D_{IL}, D_L, D_M$  ← Decrypted Attribute

**Pseudo code:**

for specific customer record ( $R_c$ ) do  
 $P \leftarrow D_{k.i}$   
 if  $Attr == Password$  //  $Attr == ifsc\_code$  //  
 $Attr == micr\_code$  //  $Attr == cif\_number$  //  
 $Attr == pin\_number$  then  
 $D_s = E_s.data - (P_r * (k * p))$   
 else if  $Attr == id$  //  $Attr == email$  //  $Attr == phone\_no$   
 then  
 $D_{ILM} = E_{ILM}.data - (P_r * (k * p)) - I_k - L_k - M_k$   
 else if  $Attr == salary$  //  $Attr == acc\_no$  //  
 $Attr == credit\_card\_number$  then  
 $D_{IL} = E_{IL}.data - (P_r * (k * p)) - I_k - L_k$   
 else if  $Attr == age$  //  $Attr == Nominee\_reg\_no$  then  
 $D_L = E_L.data - (P_r * (k * p)) - L_k$   
 else if  $Attr == aadhar\_no$  then  
 $D_M = E_M.data - (P_r * (k * p)) - M_k$   
 end if  
 end for

The decryption details are maintained in the cloud storage log records and sent to customer with accessed information.

### V. EXPERIMENTAL RESULTS

The dataset for sensitive data analysis is a false banking dataset created by ourselves. The dataset consist of 1000 records with 25 attributes taken from different Indian Banking systems. Through the data owner preference, the sensitive attributes are classified from a non-sensitive attribute by the privacy score calculation. The sensitive attributes are categorized into 5 groups and applied different keys for each category instead of a single key and the keys are managed by individuals instead of bank admin. This process increases security and resolves the key management problem. The encrypted sensitive and non-sensitive attributes are combined together and stored into the CloudMe – cloud storage. CloudMe is a private cloud and it provides Infrastructures as a Service to the user. Hence, high security is provided by our data with minimal encryption / decryption time, storage cost and key management.

#### A. Encryption Time

The following table ii and figure 2 shows the encryption time taken by RSA, MD5, Combination of MD5+RSA+AEC, ECC and MECC algorithm. Based on this result, MECC works well for large volume of data size [4, 5]. The encryption time is measured in Millisecond's (ms).

TABLE II. ENCRYPTION TIME ANALYSIS

Data Size	Asymmetric Algorithms Encryption time (msec)				
	RSA	MD5	MD5 + RSA+AES	ECC	MECC
75KB	0.1	0.167	0.492	0.56	0.356
100 KB	0.089	0.231	0.098	0.75	0.082
125 KB	0.078	0.269	0.192	0.83	0.175
150 KB	0.324	0.359	0.212	1.74	0.195

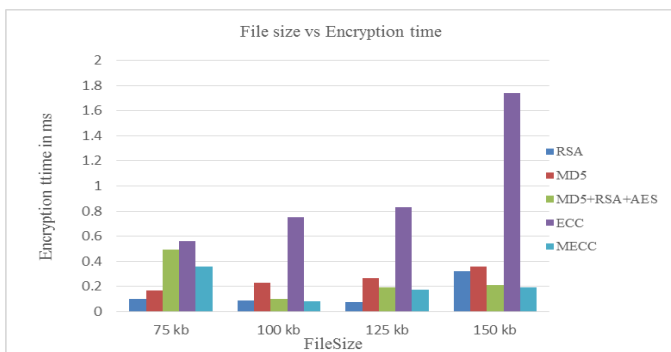


Fig. 2. Encryption Time Analysis

#### B. Decryption Time

The decryption time details are given in table iii and figure 3 shows the decryption time taken by RSA, MD5, and Combination of RSA + MD5+ AES, ECC and MECC

algorithm. When compared to other algorithm MECC takes lesser time.

TABLE III. DECRYPTION TIME ANALYSIS

Data Size	Asymmetric Algorithms				
	RSA	MD5	MD5 + RSA+AES	ECC	MECC
75KB	7.701	0.427	6.353	<b>0.68</b>	0.175
100 KB	8.237	0.609	6.72	<b>0.85</b>	0.181
125 KB	4.847	0.758	5.018	<b>0.96</b>	0.186
150 KB	6.383	0.852	4.755	<b>1.13</b>	0.195

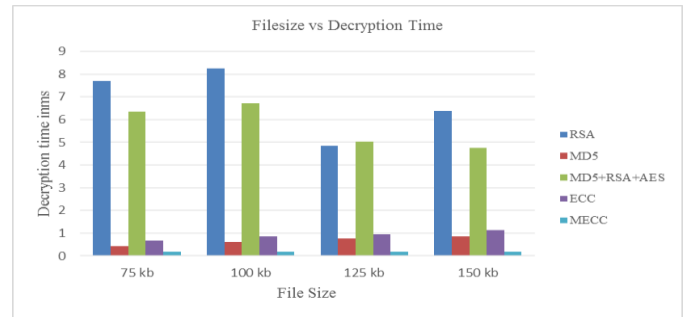


Fig. 3. Decryption Time Analysis

#### C. Storage Cost

The storage space between the entire data and sensitive data is shown in figure 4. The sensitive data encryption requires less storage size compared to entire data encryption storage size.

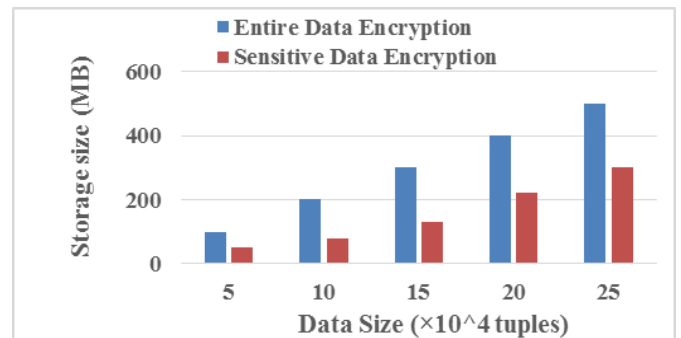


Fig. 4. Storage Size Comparison

The proposed method avoids various security attacks like man-in-the-middle attack, eavesdropping, brute force attack, offline dictionary attack and insider attacks as given below.

- Man-in-the-middle attack* – The attacker may collect some group attributes from genuine user. Other attributes cannot be identified using these attributes as each attribute group is encrypted by different combination of keys which makes the system more secure.
- Eavesdropping* – By knowing either single key or partial keys, the attackers are unable to compute the

other keys which makes the system more secure against eavesdropping attack.

- c. *Brute force attack* – The attacker is unable identify the private keys exactly in polynomial time because of discrete logarithmic approach. As, partial account number and group keys also added to private keys, the attacker cannot compute the other keys.
- d. *Offline dictionary attack* – Even if the attacker captures communication message between user and cloud, he will not be able decrypt the message, as it is difficult to identify four different types of keys.
- e. *Insider attacks* –The admin, group admins and Cloud service providers are unable to access the data without the knowledge of data owner as part of the key is available with the user.

## VI. CONCLUSION AND FUTURE WORK

With the rapid growth of data size, the organizations are moving from traditional storage system to cloud storage system to store huge volume of data with minimal storage and maintenance cost. The biggest challenge of cloud storage is security breaches. To overcome this challenge with minimal storage cost, the sensitive data are identified and applied for Modified Elliptic Curve Cryptography technique. MECC provides strong security to sensitive data with minimal cost and security is provided with the customer and other organization admin knowledge. This MECC provides better security when compared to ECC, RSA and MD algorithms. The key management problem is fully rectified by this system because the key is managed by the individual users and admins. The selective data encryption reduces the storage size, transmission cost and increase processing speed when compared to whole data encryption without compromising the security. In the future work, the confidentiality and integrity is verified by dynamic auditing system.

## REFERENCES

- [1] S. Nagaraju and L.Parthiban, "SecAuth: Provably secure multi-factor authentication for the cloud computing system", Indian Journal of Science and Technology, Vol.9, 2016.
- [2] Jun Tang, Young Cui and Qi Li, Kui Ren, Jiangchuan Liu, Rajkumar Buyya, "Ensuring Security and Privacy Preservation for Cloud Data Services", ACM Computing Surveys, Vol.49, No.1, Article 13, 2016.
- [3] Sabrina De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati, "Selective and Fine grained access to data in the cloud", Secure Cloud Computing, Springer 2014.
- [4] Dindyal Mahto, Dilip Kumar Yadav, "RSA and ECC: A Comparative Analysis", International journal of applied engineering research, Vol.12, 2017, PP 9053-9061.
- [5] Cloud Security Framework for Indian Banking Sector, Institute for Development and Research in Banking Technology.
- [6] W.Janse, T.Grace, "Guidelines on Security and Privacy in Public Cloud Computing", NIST SP 800-144.
- [7] Yongjun Ren, Jian Shen, Jin Wong, Jin Han, Sung young Lee, "Mutual verifiable provable data auditing in Public Cloud", Journal of Internet Technology, Vol.16, No.2, 2015.
- [8] Rajeev Bhanot and Rahul Hans, "A reiew and comparative analysis of various encryption algorithms", International Journal of Security and Applications, Vol.9, No.4, 2015, PP 289-306.
- [9] Omar Reyad , Zbigniew Kotulski, "On Pseudo Random Number Generators Using Elliptic Curves and Chaotic Systems", An International Journal Applied Mathematics & Information Sciences, No.1, 2015, PP 31-38.
- [10] Toughi Shahriyar, Mohammad H.Fathi, Yoones A.Sekhvat, "An Image encryption scheme based on Elliptic curve pseudo random and advanced encryption system" , An Intenational Journal on Signal Processing, 2017.
- [11] Morteza Nikooghadam, Ali Zakerolhosseini, Mohsen Ebrahimi Moghaddam, "Efficient utilization of elliptic curve cryptosystem for hierarchical access control", The journal of systems and software, 83, 2010, PP 1917-1929.
- [12] Khaled Ahmed Nagaty, "A framework for secure online bank system based on Hybrid Cloud Architecture", Journal of Electronic Banking Systems, Vol.2015, ID 614386, 13 Pages.
- [13] M.Y. Shabir, A.Iqbal, Z.Mahmood, and A.Ghafoor, "Analysis of Classical encryption techniques in Cloud Computing", Tsinghua Science and Technology, vol.21, 2016, PP 102-113.
- [14] A.Alrawais, A.Alhothaily, C.Hu, X.Xing, and X.Cheng, "An attribute based encryption scheme to secure Fog communication", IEEE Access, 2017.
- [15] S.Ruj, M.Stojmenovic, and A.Nayak, "Decentralized access control with anonymous authentication of data stored in clouds", IEEE Transactions on parallel and distributed systems, vol.25, 2014, PP 384-394.
- [16] Z.Zhu and R.Jiang, "A secure anti-collusion data sharing scheme for dynamic groups in the cloud", IEEE Transactions on parallel and distributed systems, Vol.27, 2016, PP 40-50.
- [17] L.Zhou, V.Varadarajan, and M.Hitchens, "Achieving secure role based access control on encrypted data in cloud storage", IEEE Transactions on Information forensics and security, vol.8, 2013, PP 1947- 1960.
- [18] Keke Gai, Meikang Qiu, Hui Zhao, "Privacy Preserving data encryption strategy for big data in mobile cloud computing", IEEE Transaction on big data, vol. xx, No.xx, 2017.
- [19] M.Sumathi, Dr.S.Sangeetha, "Sensitive data identification in cloud based online banking system", Proceedings of International Conference on Communication and Security (ICCS 2017) at SASTRA University, Thanjavur, Tamilnadu.
- [20] Sujithra M, Padmavathi G, Sathya Narayanan, "Mobile Device Data Security: A Cryptographic approach by outsourcing mobile data to cloud", Procedia Computer Science, 2015, PP 480-485.

# An Empirical Study of Deep CNN Models Towards Semantic Segmentation

N.B.Arunekumar

Department of Computer Science, Pondicherry  
University Puducherry, India.  
arunekumarbala@gmail.com

Dr. K.Suresh Joseph

Department of Computer Science, Pondicherry University  
Puducherry, India.  
ksjoseph.csc@gmail.com

**Abstract**—The image classification and segmentation has been a challenging task for decades. Even though several attempts were made, nearing the human accuracy was a humongous task until the advent of deep learning. This paper discusses the prime classification architectures and several segmentation architectures using deep learning. For each model, its novelty, architecture along with the performance is discussed.

**Keywords**—image classification, image segmentation, deep learning

## I. INTRODUCTION

The deep learning has been gaining huge popularity in the recent years due to its capability of fitting any complex function. Deep learning models have been used in computer vision tasks such as the image classification and segmentation. The basic component of deep learning is the perceptron which is a mathematical model that depicts the biological neuron. This perceptron works as a good binary classifier but when it comes to complex function perceptron could not handle it. Stacking these neurons into layers the feed forward neural network has been formed. These feed forward networks are capable of approximating any complex functions. These deep networks are trained to approximate the function after which they are put into test for giving the results. The training is done using the back propagation which relies on the gradient decent algorithm to alter the weight and the bias parameters of the neurons in the deep network. The network uses the negative log likelihood to compute the error and then back propagates the error inside the network to adjust the weights of the network. Propagating the error for each element in the training set would be a hectic task thus training set is divided into chunks whose errors are accumulated and then propagated into the network. Based on these errors the weights are adjusted. The rest of this paper is arranged in the manner such that section II covers the basics of the convolution neural network, Section III lists the popular datasets along with its properties, Section IV gives a brief description about the segmentation, Section V covers the metric used by most of the CNN models, Section VI discusses the most vital models of classification which have been the base model for the several other models that came later. Section VII the review tables for several tasks like detection and semantic segmentation, Section VII conclusion.

## II. CONVOLUTION NEURAL NETWORK

Even though capable of classifying images, the resolution of images became a huge computation factor for the ordinary neural network. The input layer's parameters increase with respect to the number of pixels. An ordinary color image may contain three channels R,G,B. Due to the huge input parameters the subsequent layer's parameters also increases, which exponentially explodes the computation need of the feed forward network. Convolution neural networks easily averted these restrictions as the weights are being shared across the images using the filters and the stride. These convolution neural networks were modeled based on the visual cortex. Early research in the CNN was infamous but after the alexnet proposed by Krizhevsky et al., 2012 [7], the use of cnn came into limelight.

The main components of the CNN are the convolution layers which are the filters that convolve with the image and produces feature maps corresponding to the image. In the deeper layers the filters are convolved over the feature maps of the previous layers. The pooling layers are used to reduce the size of the feature maps as the layers go deep. The depth of the filters increases in the deeper layers. The non-linearity layers like Relu are being inserted at different places as required by the architecture. All the convolutional filter's weights are learnt using the back-propagation algorithm. The final n-way classifier is used for classification problem which could also be a softmax classifier.

## III. DATASETS

There are several standard datasets classification and the segmentation tasks. We have listed some of the datasets which have been used by the prime architectures for their benchmark test. For the classification task, each training image in the dataset would be encoded with its respective class. For semantic segmentation, masks would be present on the object according to the respective classes.

### A. Image Net

- The imagenet [29] dataset is one among the prime datasets used for the classification and localization benchmark.
- The latest version of the image net dataset used for the LSVRC challenge 2017 challenge

- The dataset had 1000 different categories for the object localization.
- 200 categories which are labeled for object detection in images.

#### B. Pascal Voc

- The Pascal VOC challenge which was held till 2012 used this dataset.
- Till 2012 LSVRC was conducted along with the VOC challenge.
- The VOC 2012 [30] data set which is the final release comprises of 11,530 image in total in which 27,450 ROI object annotation.
- 6,929 segmentations are present

#### C. Ms Coco

- The 2017 stuff segmentation challenge was hosted based on this COCO dataset.
- This challenge's motto is to segment the background stuff which occupies of about 66% pixels in this dataset
- The data set has 91 different categories for the stuff in images.
- The dataset is also well suitable for captioning
- The MS COCO dataset's latest release has 330K images in total
- 80 categories each image contains about 5 captions.

#### D. Synthia

- The Synthia dataset is specific for driving scenarios.
- It comprises of 13 classes such as building, sky, car, pedestrian and so on which is related to driving.
- This dataset has more than 20,000 images
- All images are photo-realistic images

#### E. Cityscape

- The cityscape dataset[31] comprises of 20,000 and 5000 images which are annotated coarse and fine respectively
- Likewise, this dataset is based on the urban street scenario with diverse situations.
- This dataset is annotated with 30 classes for semantic and instance-wise segmentations

#### F. ADE20K Dataset

- The dataset comprise a total of 150 categories all are being derived from diverse scene ensuring non – uniform distribution instances of objects in the images.
- This bench mark was utilized for the Scene parsing challenge 2016.

- The training set of the dataset has a total of 20k images.
- The validation set consists of 2 k images.

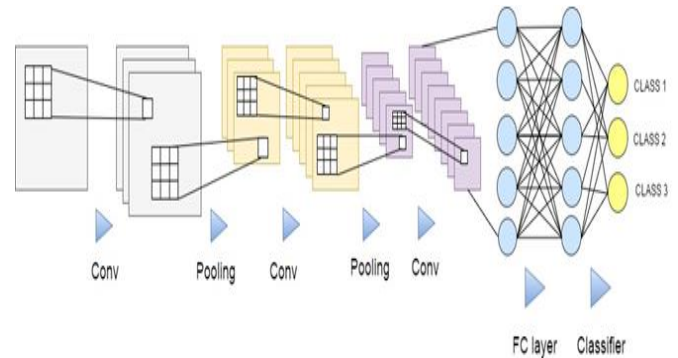


Fig. 1. Convolution Neural Networks

#### IV. SEGMENTATION

Classification networks outputs the probability of an image among the given classes. Instead, the segmentation network identifies multiple objects present in an image. The bounding box segmentation networks draws boxes around each object in the image and labels it. In bounding box detection method the boundary is not tight and may overlap with other objects.

The other type is the per-pixel segmentation or semantic segmentation where each pixel is labeled based on their respective classes. Here the boundaries are very tight thus segmentation would be perfect. The semantic segmentation is further intensified as instance segmentation where each instance of a same object is segmented separately.

#### V. METRICS

The most commonly used metric is the intersection over union. Each pixels class is categorized as TP, FP and FN which is used for the calculation of the metric.

$$IOU = \frac{\text{true positive}}{\text{true positive} + \text{false positive} + \text{false negative}} \quad (1)$$

#### VI. PRIME CLASSIFICATION MODELS

The models which are explained below have a vital role in fabricating the other architectures. These architectures are being the base architecture for the latest models which are performing above than the human accuracy levels in classification. Many of segmentation models are also being built over these models. As models became deeper and deeper the accuracy rate also increased. These classification architectures get the images as input and provide the probability distribution for the given image. The class with the highest probability is decided to be the correct class.

##### A. Alex Net

The alex net by Krizhevsky et al., 2012 [7] uses 2 GpUs because of the computation celling on a single gpu. The reLu non linearity  $f(x) = \max(0; x)$  was introduced . Even though the relu does not saturate due to the lack of normalization,

Local Response Normalization is used here to help in the improvement of generalization and decrease in the test error rate.

Unlike normal pooling where the size of the stride is equal to the filter size, here overlapping pooling is utilized. The size of the stride is lesser than the size of the filter by which filter may overlap on the pooling operation. The prime architecture of the alexnet comprises of 8 layers of which, the last three layers are fully connected while the initial 5 layers are convolutional layers.

Only the third convolution layer has interconnection between both the GPUs sharing the weights of its previous layers. Second fourth and fifth are connected only to the previous layers of the same GPU. Response normalization

The main components of this deconvnet network proposed by Zeiler et al., 2011 are the deconvolution and the unpooling layers, the unpooling layer is a simple inverse of the max pooling layer. It records example specific structures i.e. it records the max pixel during the pooling operation and uses the same spot during the unpooling for placing the pixel. The activations are again traced back to the original locations, thus the detailed object structure is reconstructed even at low resolution. The deconvolution layer operates by convolving the filters with the coarse output of the unpooled layers. Low level filters capture the object's class specific coarse configurations such as the region location and shape, while high level filters give the tuning on the class specificity i.e. complex patterns. This enhances the pixels of target class and suppresses the activations that are noisy. For visualizing the feature maps of the higher layers, an image is given as the input and the feed forward operation is done till the layer whose feature maps is to be visualized.

For visualizing a single feature map, rest of the feature maps are made to zero. Passing that feature map through the max activation we could find which part of the image activated this feature. Likewise, giving this activation map as the input to the deconvolution network the visualization could be done easily.

The Max pooling layers are placed after the First and second convolutional layers. Max pooling by overlapping pooling method is placed after each response normalization layer and the 5th convolution layer. Relu Nonlinearity is used after each layer. Overfitting is avoided by methodologies such as training over the patches of original image, its horizontal reflection, varying the RGB channel intensity and by introducing dropout Neurons in the last two fully connected layers.

### B. ZF-Net

Without knowing how actually convolution neural network works it would take a humongous time to improve it by trial and error method. Thus only after visualizing a network it would be easy to modify a network to get a reduced error rate. The ZF-NET by Zeiler et al., 2012 [11] uses de-convolution network for visualization and to get the idea about the feature maps in the network. This paper adapts 2012 alexnet with a little modification of using a single GPU with dense connections.

The features of the higher layers are more Complex than the simple edges and their blobs in the initial layers. The drawbacks identified by these visualizations are the first layer had least coverage over the information of medium frequencies and the large stride of 4 in the 1st conv layer results in the aliasing artifact caused in second layer visualizations. Thus small modifications are made where 11x11 filter size of 1st layer is reduced to 7x7 and the stride is reduced to 2. The complete architecture comprises of 8 layers of which 5 convolutional, 2 fully connected and lastly a C-way softmax classifier.

### C. Google Net

The GoogLeNet by Szegedy et al., 2015 [10] considers two subordinate challenges of reduced computation cost and managing the scarcity of data even after data augmentation techniques used for deeper networks along with the classification of images.

This network has parameters 12 times lesser than Alex net, Likewise it also utilized only half of the computational cost used by the Alex net. The GoogLeNet introduced a novel module called the Inception module.

This module works by the intuition of clustering the neurons based on the statistical correlation of the data set. When applied layer by layer neurons are clustered according to the correlation of the previous layer.

The same methodology can be used for the images, the size of the filter for the next layer is decided according to the statistical correlation of pixels in the local patches of the previous layer. For the pixels with high correlation in the neighborhood patch or dense clusters, 1x1 convolutions can be used. For the spread out clusters the filters of size 3x3 and 5x5 are used according to the spread. All the filters could also be used at once, the filters are applied and the results are concatenated in the naive approach.

Max pooling can also be done in Parallel. The 1x1 proposed by Lin et al. [12] along with 1x1 relu, is used before the 3x3 and 5x5 for the reduction of dimensionality and computation. This gives the final shape of inception module and pipelines are also used for the parallel performance. Global average pooling layers averts overfitting. The fully connected layers are also avoided for the same reason of overfitting. The factors like dead Relu facilitate the gradient vanishing in the network, due to which the softmax classifiers are introduced in the middle layers and the respective loss are added to the final loss with a reduced factor of 0.3.



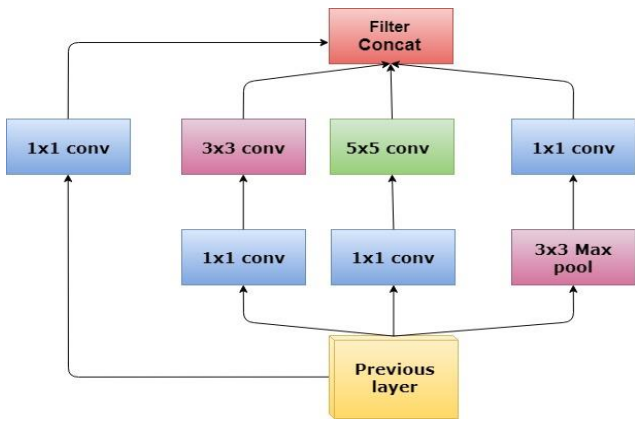


Fig. 2. Inception Module

The architecture considers an inception module as a layer. The initial 4 layers are convolution and max pooling which are interleaved alternatively after which the inception layers are stacked on top of each other with the max pooling layers amid of these inception layers. After the inception module stacking, the global average pooling layer with drop outs of 40% and linear layer for accommodating the user specific label sets and final softmax loss function are used.

#### D. VGG-Net

This This architecture by Karen Simonyan al., 2015 [25] comprises of different configuration where each configuration differs by depth and the number of convolutional layers.

The main architectural components followed among these configurations are the small receptive field of 3 x 3 along with the stride of 1 for the convolution layers. The 3x3 convolution layers stacked on each other uses a small amount of Parameters but provides the equivalent performance of the layers with large receptive field. mBy stacking 2 3x3 conv layers we get an equivalent of a 5x5 filter likewise stacking 3 3x3 conv layers is equivalent to a 7x7 filter. All the configurations have Max pooling layers of the size of 2 x 2 and a stride of 2 , the usage of relu nonlinearity between the layers have also been increased. The configuration with 16 weighted layers of which 13 convolution layers with max pooling amid of the groups, 3 fully connected layer and finally a softmax is considered for the classification task which provides the least error rate of 7.5 % with the imagenet dataset.

Similar to Krizhevsky et al. (2012), training was done with mini-batch gradient descent and for Regularization, dropout methodology was used. With ILSVRC-2012 dataset, Top 1 val

error has reduced Up to 23.7 percentage, top 5 val error to 6.8 Percentage and top 5 test error to 6.8 percentage. Apart from the classification this architecture was also used with localization where Euclidean loss was used instead of the previous loss function. This architecture also performed well with localization.

#### E. Resnet

Increase in the depth of the network, will result in the capturing complex features, but on the other hand, plain increase of depth also increases the error rate thus deteriorating the learning on datasets like cifar-10 and imagenet. Thus, resent by He, Kaiming et al., 2015 [6] was formulated as a network where the depth must be high but training error must be low, a new method was chosen where a shallow architecture VGG 16 was adapted. Then identity layers were introduced in between to increase the depth with another consideration where, if the size of the feature map is reduced by half using the stride of 2 on convolution layer then the filter's count is doubled by increasing the depth. But, again the optimization became a hindrance. Thus, a novel residual framework was considered to be inserted in the plain network above.

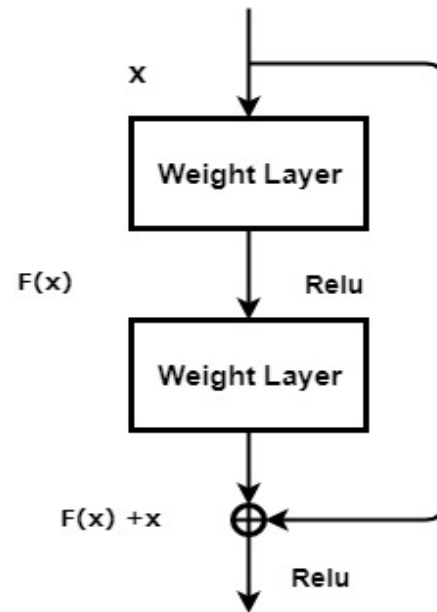


Fig. 3. Residual Block

The residual learning framework consists of an identity mapping done over 2 convolutional layers with a Relu in between. Mathematically, we consider the  $H(x)$  as the output of convolution- relu- convolution layers with input  $x$ . Now the identity mapping is introduced where the input must be added with the existing  $H(x)$  to get the new  $H(x)$  as output.

Thus the old  $H(x)$  without adding the input is considered as  $F(x)$  giving the residual mapping as in (2) which would be easier to optimize than the unreferenced mapping and these identity mapping does not take any additional computatio cost.

$$H(x)=F(x)+x \quad (2)$$

Now these residual blocks are stacked to get deep layers The overall architecture consists of initial 7x7 conv layer , final average pooling, 1000 fully connected and softmax classifier with fully stacked residual layers in-between totaling of upto 152 layers. It has an error rate of 3.57 % on the imagenet classification challenge. it also has best results with the localization and segmentation datasets

### F. Squeeze Net

The main objective is to reduce the size of the convolutional network to accommodate the micro devices and to send the model on air. Novel fire module of the squeeze net by Iandola et al., 2017 [5] comprises of a squeeze [1x1] convolution - relu - expand [1x1] convolution - expand [3x3] convolution - relu microarchitecture. The squeeze layer reduces the number of the input channels restricting the parameters and elevates the accuracy.

While the expand layers [1x1] and [2x2] carries out convolution with least parameters preserving the accuracy. The prime architecture comprises of an initial convolution layer, followed by 8 fire modules , 10th convolution layer and a global average pooling at last. After which a softmax classifier is used. The max pooling layers are used in between below the 1st conv, fire 4 and the fire modules. The max pooling is used amid of the fire modules.

The ordinary architecture has 50x parameters lesser than the alexnet with size 4.8 MB, while deep compression of 8 bit and 6 bit data type applied on the squeeze architecture, the parameters reduce to about 363x with size: 0.66 MB and 510x with size 0.47 MB but has the same accuracy rate of the alexnet To increase the accuracy rate of the squeeze net the skip connection are introduced amid of the fire models inspired by RESNET proposed by He et al., 2015. Two variants with simple connection and complex skip connections were used where the simple connection module has a higher accuracy rate of 82.5% with the size of 4.8 MB.

### VII. TABLE

In the below given review some of the classification models are also being discussed.

#### A. Bounding box detection:

The classification networks just reveal the object's class based on the complete image, it's not capable of locating the objects inside the image. To specify the objects individually inside the image we go for the object detection. In these networks, the input image is given and the pre-trained network would draw bounding boxed around the objects in the image either by labeling the box or by giving a specific class based color to the boxes The problems of these boxes are overlapping of boxes with other objects and these networks give un-tight boundaries.

#### B. Per-pixel segmentation:

Here the architectures are being trained such that the tight boundaries are made for the object. The tight boundaries are given by the masks which are drawn over the objects. Each mask color corresponds to a class. By this form of segmentation we can have more accurate boundaries. In the below given survey some of the architectures which are used for the medical image segmentation are also described. In the below given survey problem specific architectures such as architectures for medical imaging and so on. In some cases there are multiple architectures being coupled to form a complete one.



Fig. 3. Object Detection

TABLE I. CLASSIFICATION NETWORKS

Author And Year	Properties					
	<i>Novelty</i>	<i>Architecture</i>	<i>Accuracy</i>	<i>Pros</i>	<i>Cons</i>	<i>Base Architecture</i>
Jaderberg et al., 2016[28]	Spatial Transformer module	1) localisation network- gives transformation parameters 2) grid generator –generates sampling grid 3) sampler – output map	84.1% CUB-200-2011	Tolerates spatial variance	Huge computation cost	N/A
Xu, Yan et al., 2017 [18]	patch based architecture with SVM	image patches are resized and fed into alexnet, resulting 4096 feature vector is fed into SVM for classification	91.1% - MICCAI brain	pre-trained network- high accuracy	non end to end, patches have fixed size	Alexnet [7]
He, Kaiming et al., 2015 [19]	spatial pyramid pooling layer	conv layer- spatial pyramid pooling layer - fully convolutional layer	ILSVRC 2014 8.06%, caltech 101 - 93.42% ,VOC 2007 - 82.44%	Tolerates scale variance	slower	zf-net [11]

TABLE II. BOUNDING BOX OBJECT DETECTION NETWORKS

Author And Year	Properties					
	<i>Novelty</i>	<i>Architecture</i>	<i>Accuracy</i>	<i>Pros</i>	<i>Cons</i>	<i>Base ARC</i>
Girshick et al., 2012 [17]	pipelined architecture	region proposals - alexnet - svm -	31.4% - ILSVRC 2013	localization of objects image	huge computation because of pipeline	Alexnet [7]
Aubreville et al., 2017 [8]	localizer, classifier and STN	3 part architecture : localizer, Spatial Transformer Networks to crop the image and classifiers to give a bounding box	92.8% - Aperio ScanScope	Faster execution on segmentation	Less number of images in dataset	STN [28]
Girshick et al., 2015 [8]	ROI layer, parallel use of classifier and regressor	convnet - ROI pooling layer- parallel softmax and bb regressor	voc 2007 - 66.9%	25x faster than the Rcnm	a bit slower	SPP net [19]
Ren, Shaoqing et al., 2016 [12]	region proposal network- anchors	convnet- region proposal network	69.9% - voc 2007	Region proposal of different scale and aspect ratio	Gives bounding boxes no per pixel segmentation	VGG-16 [25]

TABLE III. SEMANTIC SEGMENTATION NETWORKS

Author And Year	Properties					
	Novelty	Architecture	Accuracy	Pros	Cons	Base ARC
Shelhamer et al., 2017 [9]	fully convolutional network without fc layers	convolution network - bilinear upsampling with skip architecture	62.7% - voc 2011 and 62.2% - voc 2012	dense per pixel prediction	can't handle the scale variance of images	VGG-16 [25]
Badrinarayanan et al., 2016 [15]	batch norm and relu in encoder and decoder network, pooling indices	encoder: conv+BN+Relu-pooling. decoder: up sampling 2x(conv+BN+Relu)	CamVid 11-60.10%.	perfect upsampling	Memory and computation consumption is high	encoder-decoder
Jiabin Ma et al., 2017 [14]	irregular kernels	positional parameters of the filters are added along with bilinear interpolation	VOC- 43.1%, cityscape - 72.9%	Takes geometrically irregular input features	huge computation because more parameters	Resnet [6]
Fisher Yu et al., 2016 [13]	dilated convolutions	Context network: 7 layers of 3x3 each of which has dilation of 1,16,8,4,2,1 and 1 for 7 <sup>th</sup> to 1 <sup>st</sup> layer and truncation of pointwise max(·; 0) - final 1x1	71.3% - VOC-2012	more receptive field with less parameters	More parameters for computation	front end VGG-16 [25]
Rabinovich et al., 2015 [27]	global context in solving the local ambiguity	Global average pooling – early fusion-late fusion – L2 normalization before merging	69.8% - voc 2012	Robust learning	Less accurate	N/A
Ronneberger et al., 2015 [22]	upsampling operators instead of pooling layers, overlap tile methodology	Contraction: conv - maxpool. Expansion: conv - upconvolution. Copy and crop between each corresponding network layer	IOU of 92%, on PhC-U373, 77.5% on DIC-HeLa	easy segmentation of high Res. images	huge training time	FCN [9]
Xu, Yan et al., 2016 [1]	3channel channel fusion	multi-channel 1) Background classification by FCN 2)object detection by faster Rnn 3)edge detection done by HED - then all channels are fused	4.5 - MICCAI 2015	easily segment instances	The fcnn by default is not tolerant to scale variance	FCN [9] faster rcnn
Li, et al., 2017 [4]	uses superpixels and crf	super pixels from image- clustering using crf	average per-class accuracy, our approach achieves 47.7%,	annotation over image level is enough	Intolerant to strong illumination	N/A
Noh et al., 2015 [21]	Multiple Adjacency Trees and Multiscale Features	vgg-16 convolution net - fully connected networks 2 layers - deconvolution network	72.5% - VOC 2012	Tolerant to variance of scale, no fixed receptive field constraint	Can be weak on small datasets	De-conv net [24]
Hong et al., 2015 [23]	Coupling of two networks	Classification network: gets score labels-bridge layer: transfers details – segmentation layer: produces prediction	66.6 - VOC 2012	Best for small datasets	Not advisable for heterogeneous datasets	VGG -16 [25] and De-conv net [24]
Chen et al., 2017 [26]	contextual features of multi-level	upsampling layers attached to the final 3 layers and fused with a softmax classifier	score- 0.812 in 2015 MICCAI nuclei / glands	tolerates scale variance and vanishing gradient	More parameters	FCN [9]
Linhui L 2017[32]	Batch norm after each conv in encoder	Encoder :5 lyers of conv +batchnorm+relu and max -pooling. Decoder: corresponding up sampling and conv layers	73.1% - RGB-D	depth maps to enhance accuracy	Increase number of images for generalization	Alexnet [7], encoder decoder

## VIII. CONCLUSION

In the above review the deep learning architectures that are capable of performing semantic segmentation are being reviewed. Initially the important classification architectures were discussed. Then based on them the bounding box object detection and semantic segmentation architectures were being surveyed. All the above architectures are vitally using the convolution neural networks while the recurrent neural networks are also being coupled along with the CNN's for image annotation tasks.

## REFERENCES

- [1] Y. Xu *et al.*, "Gland Instance Segmentation by Deep Multichannel Neural Networks," pp. 1–10, 2016
- [2] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive Deconvolutional Networks for Mid and High Level Feature Learning.",2011
- [3] A. Giusti, D. C. Cire??an, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks," *2013 IEEE Int. Conf. Image Process. ICIP 2013 - Proc.*, pp. 4034–4038, 2013.
- [4] Y. Li, Y. Guo, Y. Kao, and R. He, "Image Piece Learning for Weakly Supervised Semantic Segmentation," vol. 47, no. 4, pp. 648–659, 2017.
- [5] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "50 X FEWER PARAMETERS AND < 0 . 5MB MODEL SIZE," pp. 1–13, 2017.
- [6] K. He, "Deep Residual Learning for Image Recognition.", 2015
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Nips*, pp. 1–9, 2012.
- [8] M. Aubreville, M. Krappmann, C. Bertram, R. Klopfeisch, and A. Maier, "A Guided Spatial Transformer Network for Histology Cell Differentiation," pp. 1–5, 2017.
- [9] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," vol. 39, no. 4, pp. 640–651, 2017.
- [10] C. Szegedy *et al.*, "Going deeper with convolutions," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07–12–June, pp. 1–9, 2015.
- [11] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," 2012.
- [12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks," pp. 1–14, 2015.
- [13] F. Yu and V. Koltun, "Multi-Scale Context Aggregation By Dilated Convolutions," *Iclr 2016 Multi-Scale*, 2016..
- [14] J. Ma, W. Wang, and L. Wang, "Irregular Convolutional Neural Networks."- 2017
- [15] V. Badrinarayanan, A. Kendall, R. Cipolla, and S. Member, "SegNet : A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation," vol. 8828, no. c, pp. 1–14, 2016.
- [16] S. Lazebnik and C. Schmid, "Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories." 2006
- [17] R. Girshick, J. Donahue, T. Darrell, J. Malik, and U. C. Berkeley, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2012.
- [18] Y. Xu *et al.*, "Large scale tissue histopathology image classification , segmentation , and visualization via deep convolutional activation features," pp. 1–17, 2017.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," pp. 1–14.,2014
- [20] R. Girshick, "Fast R-CNN.",2015
- [21] H. Noh, S. Hong, and B. Han, "Learning Deconvolution Network for Semantic Segmentation," vol. 1. ,2015
- [22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," pp. 1–8.2015
- [23] S. Hong, H. Noh, and B. Han, "Decoupled Deep Neural Network for Semi-supervised Semantic Segmentation," pp. 1–9.2015
- [24] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional Networks."2010
- [25] F. O. R. L. Arge and C. I. Mage, "V d c n l - s i r," pp. 1–14, 2015.
- [26] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P. Heng, "DCAN : Deep contour-aware networks for object instance segmentation from histology images," *Med. Image Anal.*, vol. 36, pp. 135–146, 2017.
- [27] A. Rabinovich and A. C. Berg, "ParseNet: Looking Wider to See Better."2015
- [28] M. Jaderberg and G. Deepmind, "Spatial Transformer Networks," pp. 1–15.2015
- [29] Olga Russakovsky\*, Jia Deng\*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg and Li Fei-Fei. (\* = equal contribution) ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.
- [30] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.
- [31] Cityscapes Cvpr2016 M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [32] L. Li, B. Qian, J. Lian, W. Zheng and Y. Zhou, "Traffic Scene Segmentation Based on RGB-D Image and Deep Learning", *IEEE Transactions on Intelligent Transportation Systems*, pp. 1-6, 2017.

# An Analysis of Multifarious Character Recognition

Dr. M. Sornam  
Department of Computer Science  
University of Madras  
Chennai, India

Poornima Devi. M  
Department of Computer Science  
University of Madras  
Chennai, India  
poornima160492@gmail.com

**Abstract**—The study has been done to identify the research work for Tamil Character recognition and to review the methodologies used so far. In this paper, presented an analysis for ancient stone inscription characters, handwritten characters, epigraphical and palm leaf characters, historical printed documents and text from videos for various languages. From the existing research, character recognition for Tamil characters was very less compared to other languages. Most of the systems has three phases; first is the preprocessing phase, second is the classification and recognition phase and the third phase is postprocessing. From this analysis, it has been observed that the research for Tamil character recognition has to be improved.

**Keywords**—CNN; SVM; Image Zoning; SURF; Segmentation; MLP.

## I. INTRODUCTION

Character recognition plays a vital role in today's research work. Characters can be recognized from historical documents, ancient stone inscription, ancient epigraphical script, Handwritten Documents, ancient palm leaf document, video frames, natural scene images etc., for the purpose of knowing ancient cultures and secrets, for the verification of handwriting, for the verification of signature, for identifying the centuries, for identifying the country by referring the language script type, for identifying the place for investigations etc. Some of them were implemented using Image processing and some were implemented using Artificial Neural Network. Most commonly used architecture includes preprocessing, feature extraction, Segmentation, Classification and Recognition in fig.1.

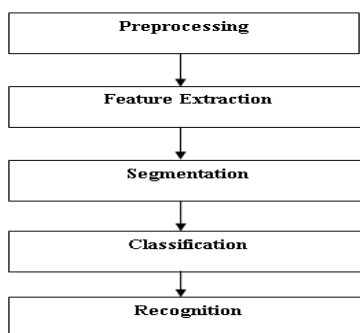


Fig. 1. Flowchart of overall process for character recognition.

In preprocessing stage, input images was resized and then converted into grayscale images and removed noises from the

image using some noise removal techniques. The various preprocessing used in this study was binarization, normalization, smoothing, noise removal, image thinning and image acquisition.

Binarization [28] [8] [18] [6] [23] [3] [30] [21] [9] [12] is the process which converts the images to binary images. Normalization [6] [21] [12] is the process in image processing where it changes the intensity values of pixels. Smoothing [4] [23] [29] [21] is the process which is used to reduce the noises from the image. Most commonly used technique for smoothing is low-pass filter. Noise removal [6] [3] [25] [29] [30] [9] is the process of removing noise from the signal; some of the commonly used noise removal techniques are bilateral filter, Gaussian filter, local pixel grouping, median filter, salt and pepper noise, wiener filter. Image thinning [6] [9] [12] is the morphological process which is used to separate the foreground image and applied only in binary image and produces the output as binary image. Image acquisition [6] [25] [31] [9] [20] is the first process in image processing which is used to retrieve the images.

The various feature extraction used in this study was zoning [3] [31], SURF [10] [12], chain code histogram of characters [29] [19], structural regional [6], moment based feature [19], GSLRE [32], Geometric feature [24], HOG [5], BOW and tf-idf [11].

Line segmentation [3], character segmentation [28] [24] [8] [25] [30] [31] [9], temporal over segmentation [27] and word segmentation [7] [8] were used for segmentation.

Various classifiers used in this study was fuzzy [29], k-mean [29], c-mean [29], SVM [29] [5] [1], TSVM [30], BPN [21], PSO [19], KNN [5], MLP [1] [28], GFF [1], DNN [11], CNN [15]. Overview of ancient character recognition with publication, Area of research, script language used and the year of publication were tabulated in Table I.

TABLE I. OVERVIEW OF ANCIENT CHARACTER RECOGNITION

Authors	Area	Script	Year
G. Bhuvanewari et al.[6]	Image Processing	Tamil Stone Inscription	2015
G. Janani et al.[9]	Image Processing	Tamil Inscription	2016
N. Jayanthi et al.[10]	Image Processing	English, Tamil, Other Language inscription	2017
Karunaratne et al.[12]	Image Processing and NN	Sinhala inscription	2017
Rajakumar et al.[21]	Image Processing and NN	Tamil inscription	2012
Rajan et al.[23]	Image Processing and NN	Tamil inscription	2009
Suriya Kala et al.[29]	Image Processing and NN	Tamil inscription	2016
Ajimore [1]	Image Processing and NN	Handwritten devanagari Characters	2017
Antony et al.[3]	Image Processing and NN	Handwritten Tulu script	2011
Apoorva et al.[5]	NN	Handwritten Hindi Numeric characters	2017
Hafemann et al.[15]	Image Processing and NN	Handwritten English Signature	2017
Nair et al.[17]	Image Processing and NN	Handwritten Malayalam characters	2017
Prathima et al.[18]	Image Processing	Handwritten Nandinagari characters	2017
Zarro et al.[24]	Image Processing and NN	Handwritten Kurdish characters	2017
Saikat et al.[26]	NN	Handwritten Bangla characters	2017
Obaidullah et al.[28]	Image Processing and NN	Handwritten Numeral Script	2015
Xuefeng et al.[32]	Image Processing and NN	Handwritten Kurdish characters	2017
Kavitha et al.[14]	Image Processing	Tamil Palm Leaf characters	2016
Sachin S Bhat et al.[25]	Image Processing	Kanada Epigraphical script	2016
Venkata Krishna Kumar et al.[30]	Image Processing and NN	Tamil Epigraphical script	2014
Aladhahalli et al.[2]	Image Processing	Indian Document scripts	2017
Kavitha et al.[13]	Image Processing	Historical Documents	2016
Kakde et al.[19]	Image Processing	Devnagri characters	2016
Priyadarshni et al.[20]	Image Processing and NN	Image based characters	2016
Rajan et al.[22]	NN	Tamil Document characters	2009
Vellingiriraj et al.[31]	Image Processing	Tamil Brahmi and Vattezhuthu characters	2016
Mittal et al.[4]	Image Processing and NN	Video frame text	2017
Bolan et al.[7]	NN	Scene words	2017
Giuseppe et al.[8]	NN	Latin characters	2017
Jia yu et al.[11]	NN	English News programs	2017
Matko Saric et al.[16]	Image Processing and NN	English Scene Text	2017
Shu Tian et al.[27]	Image Processing and NN	English web video text	2016
Yang Zheng et al.[33]	Image Processing and NN	English Natural Scene image	2017

Section II describes the related literature survey for Ancient Stone Inscription Character recognition, Handwritten Character Recognition, Ancient Epigraphical and Palm Leaf character recognition, Degraded Historical Document text recognition and Video frames and Natural Scenes Text recognition. Section III describes convolutional neural network and Section IV contains Conclusion and Future work.

## II. RELATED WORKS

### A. Study on Ancient Stone Inscription Character Recognition

First, G. Bhuvanewari et al. [6] proposed this system for ancient character recognition (ACR) system for stone inscription using Positional metric which is used to solve the problem occurred in stone inscription. For better recognition rate Image based and Zone based normalized Positional Distance Metric (INPDM/ ZNPDM) feature were used. Nearest Neighbor were used for classification and recognition of characters. This work carried out 350 characters from 35 samples of 10 characters. It involves preprocessing which includes Noise removal, binarization, normalization and image thinning. Among INPDM, ZNPDM and structural + Regional feature vectors, INPDM and ZNPDM produced better accuracy rate of 84.8%.

G. Janani et al. [9] introduced to recognize Tamil inscription characters using Image processing techniques. The architecture of this system includes image acquisition, noise removal, binary conversion, morphological operation, connected component, feature extraction, segmentation and matching the characters. First the image has to be captured then the noise was removed using median filter and Gaussian noise was removed using Wiener filter. Binarization was done by using Otsu method. Morphological operation was performed for dilation and erosion to enlarge and compress the image. The unwanted region or space was removed in the connected component stage. Feature was extracted using bilinear interpolation technique then the images were segmented and to matching the characters.

N Jayanthi and S Indu [10] stated to recover an ancient inscription using Bag of Visual Words (BoVW) technique. This methodology includes inscription images, feature extraction, clustering, indexing and retrieved inscription images. First, the feature was extracted using Speed Up Robust Features(SURF), then k- mean clustering was used to cluster the closest vocabulary. Indexing was used to sort the list of files and finally, the indexed files were represented in the form of codebook. The dataset includes 300 inscription images from three categories; 72 images from English, 56 images from Tamil and 212 images from other languages.

K.G.N.D. Karunaratne et al. [12] proposed to identify the ancient Sinhala inscription using Optical Character Recognition (OCR) technique. The architecture of this methodology includes preprocessing, character recognition and postprocessing. Preprocessing includes binarization, boundary detection, segmentation and thinning. Character recognition involves feature extraction and classification. This proposed methodology includes three modules; template matching, ANN based OCR and CNN based OCR. Input

images should be in the format of bitmap or jpeg. Feature extraction used for template matching was SURF and for ANN based OCR was zoning techniques. Comparing three modules CNN based OCR has the better accuracy than ANN based OCR and template matching.

Rajakumar S and Subbiah Bharathi V [21] aimed to identify 7th century Tamil inscription characters using SIFT features and bag-of-key point representation. This proposed system contained digitization, preprocessing, feature extraction, classification and recognition. All images were resized equally and extracted, the extracted images were used to train feed forward backpropagation neural network which produce a better recognition rate. After preprocessing this method involved in detecting the interest point in SIFT and then constructing visual codebook using k-mean clustering and then bag-of-key points were constructed. Finally SVM classifier was applied. Preprocessing involves size-normalization, binarization and smoothing. This system produced 84% of accuracy using SIFT feature.

P. Rajan et al. [23] planned to identify ancient Tamil characters and to identify modern Tamil characters using contour-let transform which was a 3D technique. It involves preprocessing, in which first the images was brightened and binarized and the next process was smoothing. To smoothen the image, Fuzzy median and Gaussian filter were used and then finally thresholding were applied to extract the foreground from the background. The dataset were collected from various century stone inscriptions. Segmentation has two parts; they are top down and bottom up segmentation. If the pixels lay on the same object then the pixels belongs together in top down segmentation whereas if the pixels were locally coherent then it belongs together in bottom up segmentation. The planned effort has composed better accuracy for ancient character recognition.

L. Suriya Kala and P. Thangaraj [29] proposed to predict different ancient Tamil characters belong to different time period using Hybrid classifier. The proposed method involves input image, preprocessing, segmentation, feature extraction, character recognition and classification. All input images must be in the jpg format. Preprocessing of an image has three phases; noise reduction smoothing and filtering. Gaussian, laplacian and low filter smoothing used to smoothing the images. Bilinear filtering, bilateral filtering and median filtering were used to reduce the noise. Markov random fields, Active contour method with Markov random fields were used for segmenting the image. Chain code histogram of character contour, Chain code histogram of character contour and improved optical character recognition (IOCR) were used for feature extraction. Fuzzy Neural Network with k- mean, SVM and C mean were used for classification. Finally the segmented characters are compared with database and then exported to a text file to recognize the characters and it yields more accurate with low computational complexity. Algorithms, classification methods, Performance measure and filters used for ancient inscription character recognition were tabulated in Table II.

### B. Study on Handwritten Character Recognition

Antony P J and Savitha C K [3] determined to recognize handwritten south Dravidian Tulu script. Preprocessing, feature extraction, learning, classification and recognition, and mapping are the five modules to recognize handwritten Tulu characters.

TABLE II. ANCIENT INSCRIPTION CHARACTER RECOGNITION

<i>Paper</i>	<i>Algorithm/ Feature</i>	<i>Performance measure / Classification</i>	<i>Filter</i>
G. Bhuvanewari et al.[6]	INPDM, ZNPDM	Positional Distance Metric	Median
G. Janani et al.[9]	Mapping	Connected Components	Median Weiner
N. Jayanthi et al.[10]	K-mean	BoVW	Median
Karunarathe et al.[12]	SURF zoning	-	Mean
Rajakumar et al.[21]	BPN	Euclidian Distance	Bicubic Interpolation
Rajan et al.[23]	SVM	-	Fuzzy median
Suriya Kala et al.[29]	Chain code histogram	K-mean, C-mean, SVM	Bilinear Bilateral Median

First the documents were scanned in the form of image file format then the RGB image was converted to grayscale image and the noise were removed using filtering and finally segmentation was applied for lines and characters. Otsu thresholding were used for binarization and the feature extraction includes numbers of end points, joints, branches and global features. For classification Bayes decision rule, ANN, SVM can be used. The final result was written in notepad for editable use.

Saikat Roy et al. [26] presented to identify handwritten Bangla characters using deep neural network. To recognize compound characters, Deep Convolutional Neural Network (DCNN) was compared with supervised Layer wise Deep Convolutional Neural Network (SL-DCNN). The only difference was that cost function for supervised training was used instead of cost function for unsupervised training for error in classification. RMSProp algorithm was used with Rectified Linear (ReLU) activation function for input and hidden layers, and Softmax function was used for output layer. The dataset used was CMATERdb which included grayscale Bangla character images. From these 34,439 images were used for training and 8520 images for testing set. It generated 90.33% in accuracy for recognizing the character.

Xuefeng Xiao et al. [32] implemented to identify handwritten character of Chinese language using Convolutional Neural Network (CNN). To overcome existing problem of speed and storage capacity, Global Supervised Low- Rank Expansion (GSLRE) and an Adaptive Drop-Weight (ADW) methods were implemented which increased the pruning threshold dynamically. CNN has nine layers which contained 3755 neurons for final classification and size



of the filter was 3x3. Rectified Linear (ReLU) activation function was used and to normalize non-linear inputs, Batch Normalization was used. To detect the redundancy, Connection Redundancy Analysis (CRA) was used which helped in target pruning ratio. Dataset was taken from CASIA-HWDB1.0 and CASIA-HWDB1.1 which contained 2,678,424 samples. Finally the proposed methodology produced the accuracy of 97.30%.

Sk Md Obaidullah et al. [28] planned to detect the text from handwritten document images using numeral script identification framework. The dataset contain 4000 images from four languages namely Bangla, Devangari, Urdu and Roman. Input images involves preprocessing, feature extraction, classification and the final outcome of this work was 55 dimensional feature set. Preprocessing includes segmentation and binarization. Feature Extraction includes frequency domain and spatial domain feature, where frequency domain feature used daubechies wavelet transform technique and spatial domain used statistical measures such as mean and standard deviation. Classifier involves NBTtree, PART, LIB linear, SMO, Random Forest, MLP and simple logistic. Among these, MLP achieved better in identifying the scripts.

Rina D. Zarro and Mardin A. Anwer [24] implemented for recognizing online Kurdish character using Hidden Markov Model (HMM). This proposed method used harmony search algorithm to recognize the characters. Dataset has been collected from 10 users and it contained 24,960 characters. The characters are preprocessed using cubic Bezier curve interpolation and the characters were segmented. Feature extraction includes geometric feature extraction and HS recognition. Geometric feature extraction includes loop feature detection, reverse movement feature and dimensional shape feature. Recognition accuracy measure were classified into four category; HMM, HS, HMM- HS without penalty and HMM- HS with penalty. This scheme achieved 93.52 % recognition rate.

Prathima Guruprasad and Jharna Majumdar [18] had been successfully implemented to identify Handwritten Nandinagari characters using optimal clustering technique. Here Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Feature (SURF) were used for transform technique. K-mean, Partition Around Medoids (PAM) and Hierarchical Agglomerative clustering were used for comparison among clustering techniques. Dataset used was handwritten Nandinagari characters which contained 1040 characters. First, images were scanned then preprocessed to convert images to grayscale. Then feature transform techniques were applied to extract the features and then exact match point, maximum match point and dissimilarity ratio were computed. Finally, clustering technique was applied to grouping and performance of the cluster was measured. It has been proved that the hierarchical clustering was more suitable for SURF and SIFT feature than k- mean and PAM techniques.

Pranav P Nair et al. [17] aimed to predict Malayalam handwritten characters using Convolutional Neural Network (CNN). Character recognition process includes preprocessing, segmentation and recognition. Handcrafted feature was used

to extract the features. First the images were scanned and then scanned images were augmented and then preprocessed. Dataset contains first six characters from Malayalam language which was collected from 112 different users. Network contained convolutional layers, ReLU layers and fully connected layers, final layer contain softmax activation function. The proposed system provided better accuracy rate for Malayalam character recognition.

Apoorva Chaudhary and Roshan lal choker [5] proposed to identify Hindi handwritten numeric characters using Histogram of Oriented Gradient (HOG), Principal Component Analysis (PCA), K- Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithm. Dataset contained Hindi numeric characters written by 100 users. Feature extraction includes HOG and PCA. Classification includes KNN and SVM. Algorithms were compared with one another; HOG with KNN, HOG with SVM, HOG with ANN, PCA with SVM, PCA with KNN PCA with ANN. In this proposed work it has been proved that HOG was the best algorithm for feature extraction and KNN was the best algorithm for classification.

P E Ajimire [1] designed to verify handwritten Devanagari vowel characters using neural network. First the images were scanned and preprocessed and stored in database. After preprocessing, images are allowed to extract the feature using Histogram Oriented Gradient (HOG). Algorithms used were Support Vector Machine (SVM), Multilayer Perceptron (MLP), Principal Component Analysis (PCA) and Generalized Feed Forward (GFF) network. From these four algorithms, it has been proved that Support vector machine generated better accuracy of 100% for training, 85.74% for testing and 87.60% for cross validation data.

Luiz G. Hafemann et al. [15] planned to verify the handwritten signature using convolutional neural network. Input images were preprocessed to resize the original images. Then trained in CNN and the performance were measured by False Rejection Rate (FRR) and False Acceptance Rate (FAR). Dataset was taken from; GPDS, MCYT, CEDAR and Brazilian PUC-PR. GPDS were the publically available dataset which contain 24 genuine samples and 30 skilled forgeries samples. Proposed method achieved 1.72% of Equal Error Rate (ERR). Algorithms, classification methods, Performance measure and filters used for handwritten character recognition was tabulated in Table III.

### C. Study on Ancient Epigraphical and Palm leaf Character Recognition

Kavitha Subramani and Murugavalli [14] published to recognize Tamil palm leaf characters from historical documents. This system includes three processes: preprocessing, binarization and postprocessing. In preprocessing, input images were resized and converted to gray image and then noise were removed. In binarization, Otsu method was applied for encoding with shift algorithm. Finally in postprocessing, trimmed median filtering was used to remove noise such as white Gaussian noise and random impulse noise. Dataset were collected from GOML (Government Oriented Manuscript Library).

Sachin S Bhat and H.V. Balachandra Achar [25] designed to identify different time periods of Kannada epigraphical scripts. Preprocessing process includes image acquisition, noise removal, binarization, segmentation, classification and recognition. Global thresholding were used for binarization. First, the dataset were created and the images were resized to particular format. Secondly, match the test image with the characters in the datasets for that calculate the arithmetic mean and variance. Finally calculate absolute difference algorithm measure to match the characters and it produced overall accuracy of 80%. All the images were in the jpg format.

TABLE III. HANDWRITTEN CHARACTER RECOGNITION

<i>Paper</i>	<i>Algorithm/ Feature</i>	<i>Performance measure / Classification</i>	<i>Filter</i>
Ajimore [1]	HOG	SVM, MLP, GFF, PCA	Tanh Axon, Sigmoid
Antony et al.[3]	Zoning	Wavelet Transform	-
Apoorva et al.[5]	HOG, PCA	KNN, SVM	-
Hafemann et al.[15]	DCNN	FFR, FAR	Mean, Variance
Nair et al.[17]	CNN	-	Gaussian, ReLU, Softmax
Prathima et al.[18]	SIFT, SURF	PAM, Hierarchical Clustering	Mean
Zarro et al.[24]	Geometric feature	HMM	Bezier curve interpolation
Saikat et al.[26]	RMSProp	DCNN, SL-DCNN	ReLU, Softmax
Obaidullah et al.[28]	MLP	Mean, S.D	Frequency and Spatial Domain
Xuefeng et al.[32]	CNN	GSLRE	ReLU

S Venkata Krishna Kumar and Poornima T V [30] introduced a method to predict the Tamil epigraphical scripts belong to various period of time. This proposed methodology consists of various stages which include image acquisition, binarization, feature extraction, segmentation, classification and recognition. Transductive Support Vector Machine (TSVM) results higher accuracy compared to Support Vector Machine (SVM). Binarization involves global thresholding by using Otsu method. Median filter were used to remove noise from an image. Bilinear interpolation technique was used for feature extraction. TSVM was a semi supervised learning method which used for classification and it produced 94.66% of overall accuracy. Algorithms, classification methods, Performance measure and filters used for ancient epigraphical and palm leaf character recognition were tabulated in Table IV.

TABLE IV. EPIGRAPHICAL AND PALM LEAF CHARACTER RECOGNITION

<i>Paper</i>	<i>Algorithm/ Feature</i>	<i>Performance measure / Classification</i>	<i>Filter</i>
Kavitha et al.[14]	Shift algorithm	-	Trimmed Median filter
Sachin S Bhat et al.[25]	Absolute Difference	Mean Variance	-
Venkata Krishna Kumar et al.[30]	Bilinear Interpolation	SVM, TSVM	Median

#### D. Study on Degraded Historical Document Text Recognition

K. Rajan et al. [22] organized this application to classify Tamil language documents automatically using Vector Space Model (VSM) and Artificial Neural Network (ANN). The dataset were collected from CIIL corpus, Mysore and the dataset has 386107 images from five different categories. First unique words from each document were found and then calculate the frequency of occurrence, total frequency of occurrence and then sort all the words in ascending order then very high and very low frequency were removed and finally, invalid characters were removed. Backpropagation algorithm was used to train the network with hyperbolic tangent activation function. The structure of the network was 1000-25-5. The proposed methodology produced 93.33% for NN model and 90.33% for Vector Space Model.

Aladhahalli Shivegowda kavitha et al. [2] proposed to segment the characters from the degraded text using new methodology based on watershed model. Nearest neighbor criterion were used to extract the text lines and then the characters were segmented from lines. To segment characters from text line image, watershed algorithm was explored which identify non-linear space between two character components. Sobel and Laplacian were used to remove the background noise. The dataset used was own dataset which was collected from archeology survey of India, Mysore and Magazines with 500 text line images. Degraded text document used in this system was Indus document and other Indian scripts. Recall, precision and F-measure were used as a quantitative measure and the outcome was measured by True Positive (TP), False Positive (FP) and False Negative (FN) and it produced good result in segmenting the characters.

A.S. Kavitha et al. [13] developed to segment the text from historical document images. Optical Character recognition (OCR) was used to recognize the Indus script images. Proposed system used Sobel and Laplacian to enhance the low contrast image. To reduce the pixel width skeleton method were applied to enhance image and then the images were grouped using nearest neighbor clustering. Finally, text and non-text region were classified. The performance of this work was measured as recall and precision. Dataset were collected from archeology survey of India, Mysore and magazine which contained 500 images. Proposed system achieved better recognition for segmented text.

E.K. Vellingiraj et al. [31] determined to verify Brahmi and Vattezhuthu which was converted from ancient historical documents of inscription to Tamil digital text format. Tamil Unicode was used to convert the digital text format from

inscription and palm manuscripts. Brahmi and Vattezhuthu scripts were taken as input. The architecture of this system includes Image capture, preprocessing, feature extraction, character recognition and text conversion. Images were captured using high resolution camera and stored in the format of JPEG, then the image preprocessing were applied to the captured image which involves cropping, segmentation, resizing, thickening and binarization. To segment characters Grapheme extraction technique was used. Segmentation involves line, word and character segmentation. The dataset contained 5000 characters which include 11 vowels, 18 consonants and 227 consonantal vowels. The conversion accuracy rate for Brahmi was 91.57% and for Vattezhuthu was 89.75%, thus Brahmi characters generated more accuracy than Vattezhuthu.

Priyadarshni and J. S. Sohal [20] designed to improve the neural network for character recognition using Scilab. Kohonen Self Organizing Maps (SOM) was an unsupervised method used to train the network. The methodology followed by this system was image acquisition, image preprocessing, feature extraction, SOM and k-mean clustering process. Images should be in specified format of JPEG or BMT. Preprocessing and feature extraction included binarization, inverting, logical conversion, median filtering and dilation. Dataset of colored images were collected from web, further the samples are partitioned into training, validation and testing data. Some of the measures used in this methodology was sum-of-squares and Euclidean distance. First the images were preprocessed, and then the feature was extracted using SOM process. Secondly, SOM was fine-tuned by applying k-mean clustering. Finally, the characters were recognized and the text was displayed in Scilab console window must be same as stored characters.

Prashant M. Kakde and Dr. S. M. Gulhane [19] proposed to predict the devnagri characters which include Hindi, Marathi, Nepali and Sanskrit using the algorithm Support Vector Machine (SVM) and Particle Swarm Optimization (PSO). The input characters were fed into android mobile and recognize the characters in MATLAB software which display the output in computer screen. PHP language was used as a connection between Android mobile and MATLAB. Devnagri script contained 14 vowels and 33 consonants. Chain code histogram and moment based feature were used as a feature extraction. Kernel functions used by SVM were linear kernel, Radial basis function kernel and polynomial kernel. By comparing PSO and SVM, PSO produced much better accuracy of 90% than SVM. Algorithms, classification methods, Performance measure and filters used for degraded historical document text recognition were tabulated in Table V. Degraded Historical Document Text Recognition

Priyadarshni et al.[20]	Edge Detection	Euclidian Distance	Median
Rajan et al.[22]	TFIDF	-	Median
Vellingiriraj et al.[31]	Image Zoning	Density	-

#### E. Video Frames and Natural scenes Text Recognition

Bolan Su and Shijian Lu [7] aimed to predict words from scenes without segmenting the characters using Recurrent Neural Network (RNN). First the word images were converted to sequential feature using HOG feature. Secondly, two-multilayer RNN was trained with Long Short Term Memory (LSTM) to store it in memory. Finally, recognize the characters using Connectionist Temporal Classification (CTC). Comparing HOG and SIFT feature, it was proved HOG performed much better than SIFT. Sigmoid function was used as activation function to train the network. Four datasets were used; ICDAR, Robust Reading Competition datasets, Google Street View Text (SVT), Sign Recognition Dataset(SRD) and IIT5K dataset. The proposed work produced better recognition for words.

Yang Zheng et al. [33] designed to identify the text from natural scene images using cascaded method. Character and non-character components are two classifiers used to identify the text. Recursive local search algorithm was proposed to identify the wrongly classified characters. Convolutional Neural Network with entropy was treated together for cascaded method. Datasets used were ICDAT 2011, ICDAR 2013 and ICDAR 2015. First, original image was extracted using connected components. Second, extracted connected components were allowed for connected component filtration and then performed removal of repeating components. Finally, text line was constructed and verified. In this work English language characters was used for text detection. HOG and LBP were used as feature extraction. Precision, recall and F-measure were used to measure better performance.

Giuseppe Airo Farulla et al. [8] designed to segment touching characters for Latin language using fuzzy logic which combines fuzzy sets, membership function and fuzzy rules. Input images were preprocessed and segmented, then finally trained in ANN. Dataset was taken from GitHub repository named CCC dataset which contained 57,293 samples combined both handwritten and printed characters. Dataset were categorized into two; Dataset A which contain handwritten cursive characters and Dataset B which contain printed characters. Algorithm used was heuristic and PSO algorithm. It has been checked that 96.1% of touching characters were correctly segmented.

Shu Tian et al. [27] planned to extract the text from video using multiple frames. For this, proposed a unified framework called Tracking based Text Detection and Recognition (T2DAR) of Bayesian-based framework. It includes three major elements; text tracking, tracking based text detection and tracking based text recognition. For tracking based text recognition, an agglomerative hierarchical clustering algorithm and temporal over-segmentation technique were used. Dataset used was USTB-VidTEXT which is available

Paper	Algorithm/ Feature	Performance measure / Classification	Filter
Aladhahalli et al.[2]	WatershedAlgorithm	Recall, Precision, F-measure	Sobel, Laplacian
Kavitha et al.[13]	-	Recall, Precision	Sobel, Laplacian
Kakde et al.[19]	Kernel	PSO, SVM	-

publically. This work improved the performance rate of detection and recognition of text from web video.

Anshul Mittal et al. [4] designed for multi-oriented text recognition in video frames using sub pixel mapping function. This function was proposed to enhance the images, then from enhanced image feature was extracted using Histogram of Oriented Moment (HOM) and then Support Vector Machine (SVM) was used to identify text and non-text region. Finally, to recognize the text, Recurrent Neural Network (RNN) was used. Dataset used was ICDAR 2013 and IITR. Dataset ICDAR 2013 contained 24 videos which are captured at different location and the dataset IITR contained 500 images which are captured from railway station board images which includes five scripts; Roman, Devanagiri, Odia, Urdu and Telugu. Performance measures used were recall, precision and F-measure.

Jia yu et al. [11] aimed to figure out the topic information using Deep Neural Network (DNN) for story segmentation. Bag-Of-Words (BOW) and Bottleneck Feature (BF) were used to enhance the performance. Evaluation of this system was done by TextTiling and Normalized cuts (Ncuts). Words were segmented using Term Frequency- Inverse Document Frequency (tf-idf), BOW and Latent Dirichlet Analysis (LDA). Dataset was taken from Topic Detection and Tracking (TDT2) Corpus which carries 2,280 English broadcast news programs. Results were compared with; DNN, DNN with MTR (Multi Time Resolution) model and DNN with BNF (Bottleneck Feature) model.

Matko Saric [16] proposed to segment the text from natural scene images. For this, novel method was introduced for character candidate based extremal region(ER). First, character candidate were generated then the generated characters was classified using SVM classifier. After classifying, character candidates were grouped to eliminate the non-characters and finally characters were restored. Dataset used was ICDAR 2013 which contained 462 images. MSER algorithm was used to identify the character region which occurs more often. This work achieved low computational complexity and better performance. Algorithms, classification methods, Performance measure and filters used for Video frames and Natural Scenes text recognition was tabulated in Table VI.

TABLE V. VIDEO FRAMES AND NATURAL SCENES TEXT RECOGNITION

Paper	Algorithm/ Feature	Performance measure / Classification	Filter
Mittal et al.[4]	RNN, SVM	Recall, Precision, F-measure	HOM
Bolan et al.[7]	RNN	LSTM	Sigmoid
Giuseppe et al.[8]	PSO	Fuzzy set, Membership, and Rules	Gaussian, Salt and Pepper
Jia yu et al.[11]	BOW, tf-idf	DNN	-
Matko Saric et al.[16]	MSER Algorithm	Recall, Precision, F-measure	Extremal Region

Shu Tian et al.[27]	Agglomerative hierarchical clustering algorithm	Recall, Precision	Euclidean Distance
Yang Zheng et al.[33]	Recursive Local Search	Recall, Precision, F-measure	-

### III. CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN) is a feed forward neural network in machine learning which was applying to evaluate the images. It is also known as ConvNet. It contains number of convolutional layers and pooling layers for feature extraction and flattens and fully connected layer for classification in Fig. 2. ReLu activation function can be used to train fast in feature extraction task. For classification task, if it is binary value use sigmoid activation function and loss function for binary as cross entropy or if it is categorical use softmax activation function and loss function for categorical as cross entropy. If it is in same domain, feature extraction can be used or if it is in other domain, fine tuning can be used. Fine tuning dataset is different from training set and it takes lots of time to train the network and it loads all the images to train or freeze some layer to train the network whereas feature extraction dataset are almost same as training set and first, it extracts the features and then train the network. Dropout can be implemented to remove the dependency which removes the data randomly and freeze can be implemented to freeze the particular layer and to run the remaining layer. Mostly, CNN was implemented using Python and the order to define model:

(Number of channels, Image width, Image height)= Theano

(Image width, Image height, Number of channels)= Tensorflow

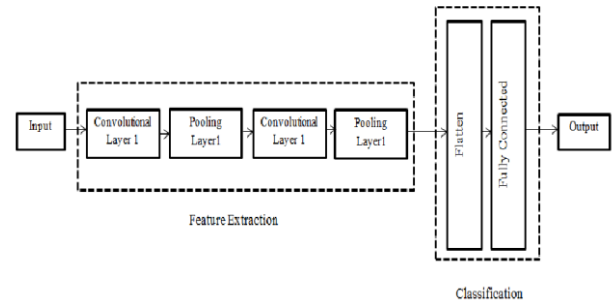


Fig. 1. Flowchart of Convolutional Neural Network.

### IV. CONCLUSION AND FUTURE WORK

Character Recognition plays a challengeable role in recognize many script languages such as Tamil, English, Malayalam, Hindi, Devanagari, Bangla, Chinese, Latin etc. Especially research work for Tamil script was more challengeable because of the complexity in Tamil language which contains circles, vertical, diagonal and horizontal lines. The more complexity is that ancient Tamil script. Ancient Tamil script will differ from modern Tamil script which is being used now-a-days. An ancient Tamil script will vary according to the centuries, so the recognition of any century's Tamil character is being difficult.

In future, the research work is going to concentrate on Ancient Tamil Inscription character recognition from different century's temple stone inscription in South India especially in Tamil Nadu.

## V. REFERENCES

- [1] G. P E Ajimire, "Handwritten Devanagari Vowel recognition using artificial neural network", International Journal of Advanced Research in Computer science, Vol. 8, No. 7, July 2017, pp. 1059 --1062.
- [2] Aladhahalli Shivegowda kavitha, Palaiahnakote Shivakumara, Govindaraj Hemantha Kumar and Tong Lu, "A new watershed model based system for character segmentation in degraded text line", International Journal of Electronics and Communications(AEU) 71, 2017, pp. 45 -- 52.
- [3] Dr. Antony P J and Savitha C K, "A framework for recognition of handwritten south Dravidian Tulu script", Conference on Advances in Signal Processing(CASP), June 2011, IEEE, pp. 7 -- 12.
- [4] Anshul Mittal, Partha Pratim Roy, Priyanka Singh and Balasubramanian Raman, "Rotation and script independent text detection from video frame using sub pixel mapping", J. Vis. Commun. Iamge R. 46, 2017, pp. 187 -- 198.
- [5] Apoorva Chaudhary and Mr. Roshan lal choker, "Handwritten Hindi Numeric character Recognition and comparison of algorithms", ICCCDSE( International Conference on Cloud Computing, Data Science and Engineering) , IEEE, 2017, pp. 13 -- 16.
- [6] Mrs. G. Bhuvanewari and Dr. V. Subbiah Bharathi, "An Efficient Positional algorithm for recognition of Ancient Stone Inscription Characters", International Conference on Advanced Computing(ICoAC), IEEE, 2015, pp. 1 -- 5.
- [7] Bolan Su and Shijian Lu, "Accurate Recognition of words in scenes without character segmentation using Recurrent Neural Network", Pattern Recognition 63, 2017, pp. 397 -- 405.
- [8] Giuseppe Airo Farulla, Nadir Murru and Rosaria Rossini, "A fuzzy approach to segment touching characters", Expert Systems with Applications , 2017, pp. 1 -- 13.
- [9] G. Janani, V. Vishalini and Dr. P. Mohan Kumar, "Recognition and Analysis of Tamil Inscriptions and Mapping using Image Processing Techniques", International Conference on Science Technology Engineering and Management(ICONSTEM), IEEE, 2016, pp. 181 -- 184.
- [10] N Jayanthi and S Indu, "Inscription Image Retrieval using Bag of Visual words", ICMAEM, 2017, pp. 1 -- 7.
- [11] Jia yu, Lei Xie, Xiong Xiao and Eng Siong Chng, "Learning distributed sentence representation for story segmentation", Signal Processing, 2017, pp. 403 -- 411.
- [12] K.G.N.D. Karunarathne, K.V. Liyanage, D.A.S. Ruwanmini, G.K.A. Dias and S. T. Nandasara, "Recognizing ancient Sinhala inscription characters using Neural network technologies", International Journal of Scientific Engineering and Applied Science(IJSEAS), Vol. 3, Issue:1, January 2017, pp. 37 -- 48.
- [13] A.S. Kavitha, P. Shivakumara, G.H. Kumar and Tong Lu, "Text Segmentation in degraded historical document images", Egyptian Informatics Journal, 2016, pp. 189 -- 197.
- [14] Kavitha Subramani and Dr. S. Murugavalli, "A Novel Binarization method for degraded Tamil palm Leaf image", International Conference on Advanced Computing(ICoAC), IEEE, 2016, pp. 176 -- 181.
- [15] Luiz G. Hafemann, Robert Sabourin and Luiz S. Oliveira, "Learning features for offline handwritten signature verification using deep convolutional neural network", Pattern Recognition 70, 2017, pp. 163 -- 176.
- [16] Matko Saric, "Scene text segmentation using low variation extremal regions and sorting based character grouping", Neurocomputing 266, 2017, pp. 56 -- 65.
- [17] Pranav P Nair, Ajay James and S Saravanan, "Malayalam handwritten character recognition using convolutional neural network", ICICCT (International Conference on Inventive Communication and Computational Technologies), IEEE, 2017, pp. 278 -- 281.
- [18] Prathima Guruprasad and Prof. Dr. Jharna Majumdar, "Optimal clustering technique for handwritten Nandinagari character recognition", International Journal of Computer Applications Technolgu and Research, Vol. 6, Issue:5, 2017, pp. 213 -- 223.
- [19] Prashant M. Kakde and Dr. S. M. Gulhane, "A comparative analysis of particle swarm optimization and support vector machines for devnagri character recognition: an android application", Procedia Computer Science 79, 2016, pp. 337 -- 343.
- [20] Priyadarshni and J. S. Sohal, "Improvement of Artificial Neural Network based character recognition system using Scilab", Optik 127, 2016, pp. 10510 -- 10518.
- [21] Rajakumar S and Dr. Subbiah Bharathi V, "7th Century ancient Tamil character recognition from Temple wall inscription", Indian Journal of Computer Science and Engineering(IJCSE), Vol.3, No.5, October, 2012, pp. 673 -- 677.
- [22] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel and B. Palaniappan, "Automatic classification of Tamil documents using vector space model and artificial neural network", Expert Systems with Applications 36, 2009, pp. 10914 -- 10918.
- [23] P. Rajan and S. Sridhar, "Identification of Ancient Tamil Letters and Its characters: Automatic date fixation based on Contour-Let technique", Association for Computing Machinery (ACM), 2017, pp. 40 -- 43.
- [24] Rina D. Zarro and Mardin A. Anwer, "Recognition- based online Kurdish character recognition using Hidden Markov Model and Harmony Search", Engineering Science and Technology, an International Journal, 2017, pp. 783 -- 794.
- [25] Sachin S Bhat and H.V. Balachandra Achar, "Character recognition and period prediction of ancient Kannada epigraphical scripts", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering (IJIREICE), vol.3, Issue 1, February, 2016, pp. 114 -- 118.
- [26] Saikat Roy, Nibaran Das, Mahantapas Kundu and Mita Nasipuri, "Handwritten Isolated Bangla compound character recognition: A new benchmark using a novel deep learning approach", Pattern Recognition Letters 90, 2017, pp. 15 -- 21.
- [27] Shu Tian, Xu-Cheng Yin and Hong-Wei Hao, "A unified framework for tracking based text detection and recognition from web videos", IEEE, 2016, pp. 1 -- 14.
- [28] Sk Md Obaidullah, Chayan Halder, Nibaran Das and Kaushik Roy, "Numeral script identification from handwritten document images", Procedia Computer Science 54, 2015, pp. 585 -- 594.
- [29] L. Suriya Kala and Dr. P. Thangaraj, "Advance algorithm based ancient Tamil character recognition by using MATLAB", International Journal of Scientific Research and Education, Vol. 4, Issue: 12, December, 2016, pp. 6096 -- 6098.
- [30] S Venkata Krishna Kumar and Poornima T V, "An efficient period prediction system for Tamil Epigraphical scripts using Transductive Support vector machine", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol.3, Issue: 9, September, 2014, pp. 7999 -- 8002.
- [31] E.K. Vellingiriraj, Dr. M. Balamurugan and Dr. P. Balasubramanie, "Information extraction and text mining of ancient Vattezhuthu characters in historical documents using Image Zoning", International Conference on Asian Language Processing (IALP), IEEE, 2016, pp. 37 - 40.
- [32] XuefebIng Xiao, Lianwen Jin, Yafeng Yang, Weixin Yang, Jun Sun and Tianhai Chang, "Building fast and compact convolutional neural networks for offline handwritten Chinese character recognition", Pattern Recognition 72, 2017, pp. 72 -- 81.
- [33] Yang Zheng, Qing Li, Jie Liu, Heping Liu, Gen Li and Shuwu Zhang, "A cascaded method for text detection in natural scene images", Yang Zheng, Qing Li, Jie Liu, Heping Liu, Gen Li, Shuwu Zhang Neurocomputing 238, 2017, pp. 307 -- 315.

# LangTool: Identification of Indian Languages for Short Text

Sreebha Bhaskaran  
Department of Computer  
Science & Engineering  
Amrita School of  
Engineering, Amrita  
Vishwa Vidyapeetham  
Bengaluru, India  
b\_sreebha@blr.amrita.edu

Geetika Paul  
Department of Computer  
Science & Engineering  
Amrita School of  
Engineering, Amrita  
Vishwa Vidyapeetham  
Bengaluru, India  
geetika2604.95@gmail.com

Deepa Gupta  
Department of Mathematics  
Amrita School of  
Engineering, Amrita  
Vishwa Vidyapeetham  
Bengaluru, India  
g\_deepa@blr.amrita.edu

Amudha J  
Department of Computer  
Science & Engineering  
Amrita School of  
Engineering, Amrita  
Vishwa Vidyapeetham  
Bengaluru, India  
j\_amudha@blr.amrita.edu

**Abstract**— Language Identification is used to categorize the language of a given document. This can categorize the contents and can have a better search results for a multilingual document. In this work, we classify each line of text to a particular language and focused on short phrases of length 2 to 4 words for 15 Indian languages. It detects that a given document is in multilingual and identifies the appropriate Indian language. The approach used is, the combination of n-gram technique and a list of short distinctive words. The results show the effectiveness of our approach over the synthetic data.

**Keywords**—n-gram; language identification; trigrams; accuracy

## I. INTRODUCTION

In an era of Internet, technologies and social media, there are many ways of communication. While messaging or commenting or posting on Twitter [1] or in Facebook or in Whatsapp, people feel comfortable using their own language. In a country like India which is rich in languages have 22 major languages, 13 different scripts and 720 dialects as per the survey reports [2]. Identification of language will have a major role when they use these languages.

A common tendency for people is to mix up two or three languages while commenting for the posts. When we have to translate such multilingual texts, first we have to realize that the sentences are in more than one language, then correctly identify the language used and finally translate it to a required language. This is another situation where the text identification comes in. Also, text is highly researched area in computer vision applications to model smart self-learning system, so identifying text from images [3],[4] also play a vital role. That is the OCR technology [5] uses language information and dictionaries during the process of optical character recognition. Other application of language identification, during translation [6],[7] of text from one language to another language, identifying the source language is a crucial one.

One of the major bottlenecks of language identification system is Discriminating between Similar Languages (DSL). The task of text categorization becomes more difficult as the number of languages increases and in Indian scenario, there are many languages with similar texture. The lengths of the text in a document also have a major role in text categorization. As the

length of the text decreases, similarity between languages increases and the size of training data is reduced.

In this paper, we choose for 15 Indian Languages and are grouped based on the similarity of text for each language used is as follows:

- Group 1 – {Gujarati (*Guj*), Kannada (*Ka*), Malayalam (*Mal*), Gurmukhi (*Gur*), Tamil (*Ta*), Telugu (*Te*), Urdu (*Ur*), English (*En*)}
- Group 2 – {Assamese (*As*), Bengali (*Be*), Oriya (*Or*)}
- Group 3 – {Hindi (*Hi*), Marathi (*Mar*), Nepali (*Ne*), Sanskrit (*Sa*)}

Group 1 classification is based on no ambiguous or no similar text in these set of languages. Group2 and Group3 have similar text pattern or same word with different meanings in each of these languages. Assuming that minimum of 2 - 6 word phrases is required for accurate result.

The paper describes a hybridized approach to classify phrases in a document under any of the mentioned languages. Section II describes the detailed study of language identification with different Indian languages which helped to form this problem. In section III, the proposed approach which is the combination of n-gram based and short words is described. The data description and evaluation measures are given in section IV. Results and analysis was discussed in section V. Conclusion and the future works are justified in section VI, which is followed by references.

## II. RELATED WORK

The language identification is the task of categorizing given text as a particular language based on certain features. These works can be broadly classified as linguistic models and statistical models. Linguistic model makes use of the linguistic rules. The developer of the system must be aware of all the rules associated with the languages under consideration. On other hand, the statistical model is system dependent. The system is trained to automatically classify the languages based on features that are unique to a language like short word method, bag of words method, n-gram, Support Vector Machine (SVM) etc.

One of the popular works in the text categorization has been done by Canvar and Trenkle [8] known as Textcat tool. The method uses n-gram profile to calculate a simple rank order statistical called as out-of place measure. An n-gram is a sequence of  $n$  consecutive characters extracted from a given string. If we ignore the white spaces and all the special character then, the word "Hello World!" can be represented in terms of bi-gram and tri-gram as follows:

bi-gram: He, el, ll, lo, Wo, or, rl, ld

tri-gram: Hel, ell, llo, Wor, orl, rld

To calculate the out-of-place measure, they used finding the difference between the ranks of various matched n-grams and adding them up to find distance. The text belongs to the language with least distance. The method given by Carvar and Trenkle [8] has the advantages as: it is independent of the language and the alphabet, it doesn't require knowledge about the language and the effect of spelling or syntactical error in the document is limited.

The modification of the work [8] has been given by Keselj[9], in which they worked on word and character level n-grams. The algorithm used calculates the dissimilarity between two profiles and returns a positive value. This means n-gram which is present in less number of languages has a higher weight and could be decisive in classifying the text.

Another simple, effective and widely used method for text categorization is Nearest Neighbors (KNN) [10], [11]. The basic idea is to convert the test sample (text) into weighted feature vectors. It tries to find the KNN from the complete set of training samples using cosine similarity measure. The biggest drawback with this system is the large amount of computation that eventually hampers the speed and performance of categorization. There are multiple approaches to KNN method such as lazy learning, eager learning.

Certain linguistically motivated models have also been proposed. The work by Grefenstette [12] and Johnson [13] is based on generation of language model using the short words such as prepositions, pronouns, determiners and conjunctions to determine the language of the document. A list of unique words is generated by tokenizing the sample document, selecting the words with 5 or less characters and retaining the ones with frequency greater than 3. The probability of the test data belonging to a particular language is calculated as the product of the probabilities of occurrence of tokens in the list of that language. This is faster than the n-gram based technique, since the training list is very small. But the quality of result is highly dependent on the occurrence of words from the list in the test data.

DueireLins and Goncalves [14] considered the syntactic structure of language (adverbs, articles, pronouns and so on) rather than the word structure and sequences. They have analyzed which set of words could be used to differentiate between the languages. Such a list of unique word model is highly dependent on the set of languages begin considered. The efficiency of this model may suit for one set of languages but may not work for similar set of languages.

Prager [15] proposed a hybridized approach for language identification by applying the vector space model based on the similarity computation with respect to the distance between the trained and test language model.

The recent trend of language identification is with Transliteration. [19], [20], [21] refers transliteration on very few languages and unambiguous text.

The n-gram model is slow but language independent whereas short word method is less time consuming as it uses less complex computation but is language dependent. This leads to hybridized approach for Indian language identification. Most of the work done for Indian languages [16], [17], [18] have considered either few languages or languages with dissimilar text. The proposed work is based on combination of n-gram model and short word method applied on 15 different Indian languages having similar text patterns, classified as Group 2 and Group 3.

### III. PROPOSED APPROACH

We started our model by collecting the corpus for all these 15 languages from different sources such as newsgroups, Wikipedia, dictionaries, etc. The approach used is a hybridized language detection, which combines the advantages of n-gram method and short word method as discussed in the related work section.

The detailed description about the proposed work is shown in Fig.1. This model has two phases: Training and Testing. Both phases requires data, for training phase, as major data collection is done from newsgroups and Wikipedia, it needs to be cleaned (along with the required language data it will have data with other language) and for testing phase, the given data may have punctuation marks, etc. So, both phases requires *Cleaned data*, for which the collected data needs to pass through Cleaning and Filtering stage where the removal of unwanted text like numbers, punctuation marks etc. While creating training set, we filter out other languages from our document, for instance, if it's Hindi document we remove words of other languages.

The data after cleaning and filtering stage, it then generates trigrams and its corresponding frequencies for both training and testing documents. In training phase, it also identifies unique short words having frequencies greater than the threshold frequency. Based on the distance calculated between the profiles if the unique language could be identified then display that language. Otherwise, it compares with the short words collected for each language; and displays the detected language.

#### A. N-gram Technique

This technique uses two phases called *training phase* and *testing phase*. Training phase consists of the following steps:

- Collect documents from different sources and concatenate them.

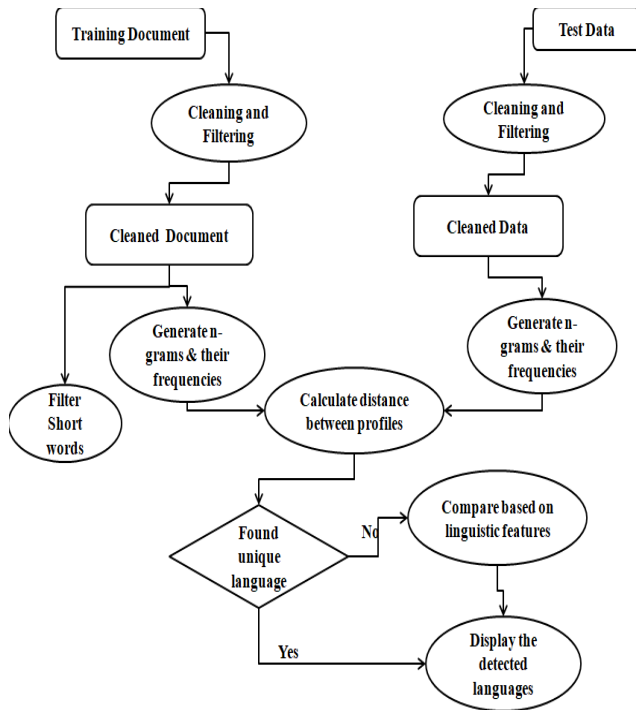


Fig. 1. Flow of proposed work

- The document is filtered to get rid of digits, special characters, extra white spaces, and punctuation.
- Select the n value. We have used  $n=3$  i.e. trigrams.
- Scan through the document to generate the n-grams.
- Store the n-grams in hash map. Hash maps are key-value pair, in this case keys are the trigrams and values are their frequencies. Keep incrementing the value with each occurrence of the particular n-gram.
- Sort them in decreasing order of their frequencies.

Repeat the above steps for all the languages and store the n-grams for all the languages in separate files.

The second phase, testing phase, uses the files produced during the training phase to determine the language of the given document. It consists of the following steps:

- Generate trigrams and compute their frequencies for the given input text following the steps described above.
- Compute the dissimilarity between the test document profile and the category's profile.
- Find the cumulative distance category wise adding the differences calculated in the previous step.
- The category with the smallest distance measure has highest probability of being the correct category.

If there are multiple languages with similar distance measure, then we apply the short words algorithm for the list of languages identified by the n-gram method.

### B. List of Short Words

When dealing with set of similar languages such as Group 2 and Group 3, there are high possibilities to have trigrams that are common in two or more languages in this set. In such cases,

there might be some ambiguity as to which language does the n-gram belong to. Even the cumulative distance might be very close. This could affect the accuracy of the result. To overcome this problem we need language specific features. Intuitively unique words such as determiners, conjunctions, and prepositions are good clues for guessing language.

Every Indian language follows a particular pattern which differentiates it from the other languages. A list of such differentiating words or features are collected and stored for each language in separate files. The test document is looked upon for pattern match in the set of similar languages identified.

We derive the attributes of language using the same corpus. This involves extracting the unique features for all the languages. The filtered corpus used for n-gram extraction is tokenized to get a list of words. The frequency of all the words is calculated. The words that have length less than or equal to four and frequency greater than threshold are retained.

Initially the given document is tokenized and the tokens are searched in the list of the languages identified by the n-gram method. The sentence belongs to a particular category if the token in the test document is found in the corresponding list. In case of no match, the language specific patterns are used to categorize the text. Our approach has been tested with 30 – 50 short phrases for Group 2 and Group 3 set of languages. TABLE I and TABLE II shows sample short words used to categorize text for Group 3 and Group 2 language sets.

TABLE I. SAMPLE SHORT WORDS FOR GROUP 3 LANGUAGES

Tokens	Hi	Mar	Ne	Sa
है	✓	✗	✗	✗
भी	✓	✗	✗	✗
और	✓	✗	✗	✗
छ	✗	✗	✓	✗
थियो	✗	✗	✓	✗
पनि	✗	✗	✓	✗
आहे	✗	✓	✗	✗
झाले	✗	✓	✗	✗

TABLE II. SAMPLE SHORT WORDS FOR GROUP 2 LANGUAGES

Tokens	As	Be	Or
□□□□	✗	✗	✓
□□□□	✗	✓	✗
□□□□	✗	✗	✓
□□□□□	✓	✗	✗
□□□□□	✗	✓	✗
□□□	✓	✗	✗

Each language has its own pattern, for example most of the Maphrases end with आहे. A list of such differentiating words or features are collected and stored for Group 2 and Group 3 set of languages. The test document is looked up for any pattern match in the set of languages identified.



#### IV. DATA DESCRIPTION AND EVALUATION MEASURES

The proposed model was trained for 15 Indian languages (Group 1, Group 2 and Group 3), by collecting data randomly from different source of newsgroups, Wikipedia, dictionaries etc. So, the size of the training dataset for each languages varied from 25 to 63495 (kilobytes) based on the availability. The training dataset in terms of number of words, number of unique words and number of unique trigrams and testing dataset for each language is briefed in TABLE III. The size of the training dataset used is not uniform for all languages because of the unavailability of data for those set of languages.

Testing has been done including and excluding similar languages. Accuracy measure (1) in percentage has been used as an evaluation parameter for analyzing the results.

$$Accuracy = \frac{\# \text{ correctly identified phrases (words)}}{\# \text{ phrases (words)}}(1)$$

For small phrase of 2 - 6 words, it is possible to have those words to be unique or not. In case of unique words, our tool can detect the language in which the phrase belongs too. But n-gram on the other hand is language independent. It is not restricted by a set of words. For example, consider the trigram “act” which could be present in “subtract”, “enact”, “acting”, “actor”, “fact” and so on. Although most of these words are not interrelated, but contain a common n-gram. All these words can be identified to belong to English language using just single trigram.

As we used hybrid model it gave better results when executed for Group 2 and Group 3 Indian languages, where different languages have similar script used, but the meaning of the word in each language is different. For example, □□□□□□ - Education in Hindi and Punishment in Marathi, □□□□□□ - Practice in Hindi, study in Marathi, □□□□□□ - attempt in Hindi and make fun of in Marathi etc.

TABLE III. TRAINING AND TESTING DATASET

Language	Training Dataset			Testing Dataset
	# Words	# Unique Words	#Unique trigrams	# of words
As	1508	804	1465	127
Be	286619	16847	11535	29040
En	9,30,730	37823	11048	58824
Guj	58555	16020	11725	4200
Gur	22596	5585	6330	6256
Hi	7,07,247	21239	15874	38720
Ka	6685	3242	4651	188
Mal	2,98,267	32976	16299	23200
Mar	11971	4799	5967	8280
Ne	28,60,354	220734	43171	25670
Or	1439	814	1719	162
Sa	20,475	8918	10788	4048
Ta	3,39,463	31527	8427	27664
Te	4,16,306	35,202	15695	31416
Ur	5,83,476	14689	10150	43524

#### V. EXPERIMENTAL RESULTS AND ANALYSIS

Main objective of our proposed work is to identify the language for short text with the phrase length to be 2 to 6 words. TABLE IV, show the results for Group 1 set of languages with phrase length of 2 words, where the accuracy is almost 100 percent but for Mal and Ur there was error because of unavailability of few words in training dataset.

In TABLE V, the result for Group 2 and Group 3 set of languages with phrase length of 2 words each is shown. The error rate for Be in Group 2 is high, as the text pattern is similar to As. The results for Group 2 is better with our proposed approach, as same words with different meanings are less when compared to Group 3. Group 3 is a set of language with same text pattern and same words with different meaning for example, पैसा, खाना, अन्न, दण्ड, शिक्षा, चन्द्र, गीत etc., so the error rate with 2 words phrase length is high as most of the words are getting mapped with other language in that set.

TABLE VI show the results of Group 3 set of languages with phrase length as 4 and 6 words, this shows a drastic improvement in accuracy as the word length increases. For Group 2 as the results were better with 2 word phrase length, it is prominent that it will show substantial improvement as the phrase length is increased. The accuracy can also be increased when all the unique words for each language are identified.

TABLE IV. RESULT FOR GROUP 1 LANGUAGES

Language	Accuracy %
En	100.00
Guj	100.00
Gur	100.00
Ka	100.00
Mal	98.00
Ta	100.00
Te	100.00
Ur	97.85

TABLE V. RESULT FOR GROUP 2 & GROUP 3 LANGUAGES

	Language	Accuracy % (Phrase length = 2 words)
Group 2	As	96.88
	Be	56.67
	Or	82.72
Group 3	Hi	58.75
	Mar	66.67
	Ne	23.66
	Sa	28.18

While analyzing the results, Group 1 languages were giving 100 percent accuracy keeping two as phrase length. Group 2 set of languages were also giving better accuracy keeping phrase length as two.

TABLE VI. RESULT FOR GROUP 3 LANGUAGES

Language	Accuracy % (phrase length = 4 words)	Accuracy % (phrase length = 6 words)
Hi	67.50	81.37
Mar	69.57	83.87
Ne	42.71	54.71
Sa	52.27	63.33

For Group 3 set of languages which has very similar text and words with different meanings as shown in Fig. 2, shows drastic improvement in accuracy for shorter phrase lengths.

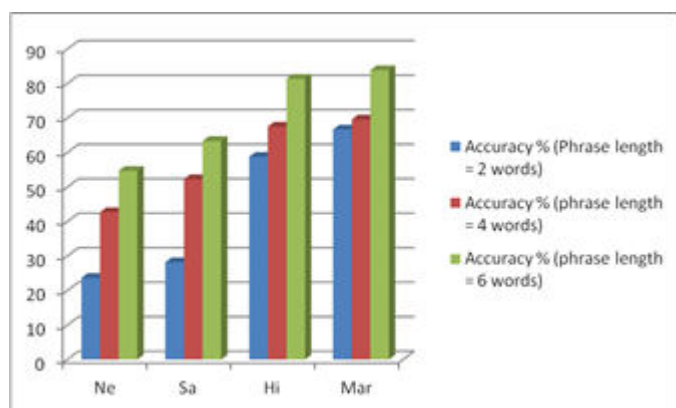


Fig. 2. Accuracy for Group 3 set of languages with different phrase length

## VI. CONCLUSION AND FUTURE WORKS

Though N-gram technique is the one which is most widely used to identify the languages, it has got certain drawback as the accuracy of the result depends on the quality of the training dataset and the execution time is higher when compared with other techniques like Bag of Words. So, in our model, it overcomes these drawbacks of N-gram by proper filtering of the collected data, before generating trigrams. As it is random collection of data from various different sources on many different topics, and by smartly selecting the threshold frequency to get rid of useless or erroneous trigrams. There is no way to reduce the execution time of n-gram technique. To overcome this issue, another famous technique list of short words is used in combination with the traditional n-gram technique.

Words are the privileged features in all the languages, it conveys more information than a sequence of characters (n-grams), and they should be treated as privileged units of the language detection tool. This is generally a small sized file. Thus matching through it is very fast as compared to scanning through n-grams.

In our work, we have developed a system to identify 15 Indian languages. The accuracy shows the system works fine with similar scripts when the minimum phrase length given is 2. This accuracy can be improved by increasing the number of words given for testing dataset but our goal is to identify the

language for smaller phrase length and in the test results shows better accuracy. As a future work, the words which were unavailable in training dataset could be updated whenever a new word occurs.

## REFERENCES

- [1] Venugopalan, Manju, Deepa Gupta, "Exploring sentiment analysis on twitter data," Contemporary Computing (IC3), 2015 Eighth International Conference on. IEEE, 2015
- [2] mhrd.gov.in/sites/upload\_files/mhrd/files/upload\_document/languagebr.pdf
- [3] Poonam, Salunkhe, Sreebha, Bhaskaran, Amudha, J., and Deepa, Gupta, "Recognition of Multilingual Text from Signage Boards," Sixth International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017 [This paper got Accepted and yet to be published]
- [4] Amudha, J., Naresh, Kumar, "Gradual Transaction Detection using Visual Attention System," Advances in Intelligent Informatics, 2014 pp. 111–122.
- [5] Deepa, Gupta, M. L. Leema, "Improving OCR by Effective Pre-Processing and Segmentation for Devanagari Script: A Quantified Study," Journal of Theoretical and Applied Information Technology. Publisher: ARPN. vol 52 No.2, 2013, pp 142–153.
- [6] K. Jaya, Deepa, Gupta, "Exploration of Corpus Augmentation Approach for English-Hindi Bidirectional Statistical Machine Translation System," International Journal of Electrical and Computer Engineering (IJECE). vol. 6, 2016, pp. 1059–1071. No. 3.
- [7] Deepa, Gupta, Aswathi, T., Rahul, Kumar, Yadav, "Investigating Bidirectional Divergence in Lexical-Semantic Class for English-Hindi-Dravidian Translations," International Journal of Applied Engineering Research. vol. 10, No. 24, 2015, pp. 8851–8884.
- [8] W. B. Cavnar, J. M. Trenkle, "N-gram –based text categorization," Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, Nevada, USA, 1994, pp. 161–175.
- [9] V. Keselj, F. Peng, N. Cercone, C. Thomas, "N-gram based author profiles for authorship attribution," Proceedings of the Pacific Association for Computational Linguistics, 2003, pp. 255–264.
- [10] P. Soucy, G. W. Mineau, "A simple KNN algorithm for text categorization," Proceedings 2001 IEEE International Conference on Data Mining. San Jose, CA, 2001, pp. 647–648.
- [11] Zheng, Wenbin, Yuntao Qian, Huijuan, Lu, "Text categorization based on regularization extreme learning machine," Neural Computing and Applications 22.3-4, 2013, pp. 447–456.
- [12] G. Grefenstette, "Comparing two language identification schemes," 3rd International conference On Statistical Analysis of Textual Data. , 1995
- [13] Hwong, N., Caswell, A., Johnson, D.W., Johnson, H, "Effects of cooperative and individualistic learning on prospective elementary teachers' music achievement and attitudes," Journal of Social Psychology. 133(1), 1993, pp. 58–64.
- [14] Rafael, Dueire, Lins, d. Paulo, Goncalves, "Automatic language identification of written texts," Proceedings of the 2004 ACM Symposium on Applied Computing, SAC '04., 2004, pp. 1128–1133, New York, NY., USA. ACM.
- [15] J.M. Prager.: Linguini, "Language identification for multilingual documents," Proceedings of the 32nd Hawaii International Conference on System Sciences. , 1999
- [16] B. Sinha, M. Garg, S. Chandra, "Identification and classification of relations for Indian languages using machine learning approaches for developing a domain specific ontology," International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), New Delhi, 2016, pp. 415–420.
- [17] R. Bhargava, Y. Sharma, S. Sharma, "Sentiment analysis for mixed script Indic sentences," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, 2016, pp. 524–529.

- [18] S. S. Prasad, J. Kumar, D. K. Prabhakar, S. Tripathi, "Sentiment mining: An approach for Bengali and Tamil tweets," 2016 Ninth International Conference on Contemporary Computing (IC3), Noida, 2016, pp. 1--4.
- [19] Dias Cardoso PM, Roy A, "Language Identification for Social Media: Short Messages and Transliteration," Proceedings of the 25th International Conference Companion on World Wide Web 2016 Apr 11 (pp. 611-614). International World Wide Web Conferences Steering Committee.
- [20] Banerjee S, Kuila A, Roy A, Naskar SK, Rosso P, Bandyopadhyay S, "A hybrid approach for transliterated word-level language identification: Crf with post-processing heuristics," Proceedings of the Forum for Information Retrieval Evaluation 2014 Dec 5 (pp. 54-59). ACM.
- [21] Gupta DK, Kumar S, Ekbal A, "Machine learning approach for language identification & transliteration," Proceedings of the Forum for Information Retrieval Evaluation 2014 Dec 5 (pp. 60-64). ACM.

# Dynamic Remote Data Auditing using Privacy Preserving Auditing Protocol in Cloud Environment

## Raja

Research scholar, Department of computer science and Engineering, Sathyabama University, Chennai Tamilnadu, India.  
rajajster@gmail.com

## Ramakrishnan

Professor and Head, Department of Computer Application School of information Technology, Madurai Kamarajar University, Tamil Nadu, India.  
ramkrishhod@gmail.com

## Hariharan

Professor, Department of Computer Science and Engineering, VeltechMultitech Engineering College, Chennai, Tamilnadu, India.  
sss.hariharan@gmail.com

## Ramprasath, Arunkumar

Assistant Professor, Department of computer science and Engineering, Madanapalli Institute of Technology, Andhra Pradesh, India  
mramprasath@gmail.com, saarun@gmail.com

**Abstract**—In recent year, cloud computing provide consistent, customized and quality service to the cloud user for securing the data in cloud storage. Currently, numerous business organization generate enormous amount of insightful data such as, Individual information, financial data and health records. Consequently, the digital data produced by the organization has increased correspondingly they avoid to store the data locally and they planned to outsource the data storage in cloud environment. On the other hand, the significant worry to the data owner is to provide the security and integrity for their outsourced data. Our proposed system take this issues as a challenging task and provide security to the out sourced data in cloud environment by using Remote Data auditing (RDA) Technique. In earlier days the majority of auditing techniques only focused on static data and not supported for dynamic data. In this paper, we proposed a professional RDA technique using Data Privacy Preserving Protocol for cloud storage system. Our system also design system model which support the dynamic data operation in the cloud environment. The experimental result shows that our proposed auditing protocol is secure and extremely efficient as compare to existing auditing techniques

**Keywords**— Remote Data Auditing, cloud computing, Privacy preserving protocol, Data integrity

## I. INTRODUCTION

Now a days, cloud computing has capability to computing resources on demand and offer a simple pay-as-you-go services model for customers. It also emerged as a new computing standard, and has gained more attractiveness in business environment. On the other hand, the data produced by the company and managing this information in local storage [1] is difficult process. Currently, numerous big companies have move their business services form local computing infrastructure to Amazon elastic computing cloud (EC2)[2] or cloud storage, which is a most important public Infrastructure as a services (IaaS) in cloud computing environment. To reduce the burden on the local storage all the company have a choice to outsource their data to cloud storage to dismiss the maintenance [3] and communication cost. Storing the data in cloud environment [4] is a significant service which helps the data owner to store their data in cloud. Nevertheless, this type of data hosting services introduces new security challenges [5] to the user. Though, its offer

numerous benefit to the data owners such as outsourcing data to remote server, hand over the management of data to entrusted cloud service provider which can direct to [6] loss of data controls. Infrequently accessed data have been discarded by the cloud and declare that the data are correctly stored in the cloud space. In the 2011, some business organization reported that the data was corruption in server of major cloud service provider and many instance of cloud services, such as Amazon S3 break down, Gmail mass detection [8].

There have been more than 535 data breaches happened in 2011 was reported by (PRC) Privacy rights clearing house such as Sony picture and online entertainment theft of medical data record and customer information. When intruder revoke the control over cloud server, they have capability to produce reply attack over the encoded data stored in cloud server. Due to this attack the integrity of the user data stored in the remote server admired to internal or external attack. To provide data integrity in cloud storage traditional integrity methods needs local copy of the data stored in cloud storage. Even though it's not possible for mobile user to down load large amount of data from cloud storage [9] which makes complexity to the mobile user to access the data remotely in cloud storage. [20] presented cloud storage system bear the data privacy preserving in cloud environment.

In this context, to verify data integrity in cloud storage system requires more efficient techniques to validate the integrity of the data. Many researchers have proposed several techniques to address the problem using Remote data auditing techniques which can efficiently, securely validate the data by generating challenges [10]. In general RDA can be classified into three different categories such as integrity based, recovery based [11], deduplication- based [12] methods. In existing RDA techniques focus on computational and communication cost of data owner which was the huge load for data owner. The design principal of Remote data auditing is to support dynamic operation on different application for that purpose the data owner incur different type of data structure (binary tree) to support dynamic operation in cloud environment.

Conversely, these data structure is not support the dynamic operation for large scale of data efficiently because of frequent update on the cloud data which leads high

computational cost to on the auditor. To overcome these problem in the cloud environment, our system propose Dynamic remote data auditing which support the dynamic operation by using privacy preserving auditing protocol. The significant objective of the auditing protocol is to protect the data privacy against auditor. The main contribution of the paper as follows: Development of dynamic remote data auditing techniques for outsourced data in cloud environment using data privacy preserving protocol. b) Design and overview of privacy preserving auditing protocol with dynamic data operation such as update, modify, insert with minimum computational cost. c) Proposed protocol implementation in real environment and the results shows protocol ability to provide better data integrity, security and performance as compare to the existing techniques. The rest of the paper is organized as follows: section 2 discussed about related works in the area of RDA and section 3 present common system model for remote data auditing, section 4 discuss proposed system model using data privacy preserving protocol, section 5 discuss about dynamic data operation in cloud environment.

## II. RELATED WORKS

In recent year, outsourcing the data in the cloud storage is an important service [13] in cloud computing which allows the data owner to reduce the local burden for storing the data. Numerous user starts to accumulate the data remotely in cloud storage which make the data owner to worry about the data loss in cloud because of security issues. To overcome these issues researcher have presented several study related to RDA schemas [14] to check the integrity and correctness of the outsourced data in the cloud storage. Several existing methods were reviewed using data integrity and discussed the advantages and disadvantages of methods. The very first provable secure schema discussed [15] to authenticate the data integrity in cloud without download the data form cloud. This method uses the RSA- based homomorphic verifiable tag to produce single tag using group of tag.

This methods acquire higher computational and communication cost because of using RSA numbering concepts. Proof of- Retrievability (POR) [16] in a newer type of RDA techniques which is used to check the data integrity and prevent from data losses by using forwarded correction techniques, remotely. The computational cost for POR method is high on client side which leads to perform data recovery and encryption process. To improve the security and efficiency of the POR method it uses BSL [17] homomorphic authentication techniques. This process allows the auditor to combined tags into fixed size in order to minimize the computational cost. However in cloud environment it's unfortunate to conduct dynamic remote data auditing because none of them (cloud service provider or data owner) give guaranteed for balanced auditing result. [21] Discuss the provable data possession in multi cloud storage environment for data verification and support scalability data migration.

In several Remote auditing methods dynamic data update operation is an important issue in cloud computing. During this operation data owner having permission update their data present in the cloud storage without retrieving outsourced file.

To improve the scalability and efficiency of the dynamic operation a new [10] RDA method was proposed which uses the symmetric key operation to defeat the problem in static RDA techniques. However the owner has to do pre-computation process for verification of data before uploading in to cloud storage, also data owner can only perform append, delete and modify operation but owner doesn't having permission to do dynamic update operation on the data which leads to re-computation of all the outstanding data and its acquire high computation cost on the data owner. [18] Has discussed the remote data ownership checking which allows the integrity checking or verification on the remote data in crucial information environment. [19] Present the design of dynamic provable data ownership and framework which support to store the updated data.

In our proposed system, we introduced new method and algorithm for dynamic remote data auditing using privacy preserving protocols which allow the data owner to perform dynamic operation to ensure the data integrity in the cloud environment. The system also discussed proof of correctness using some characteristics.

## III. SYSTEM MODEL FOR REMOTE DATA AUDITING

The following fig 1 shows the general system model for remote data auditing which consider the following components such as data owner, Cloud Storage Provider (CSP), third party auditor (TPA).

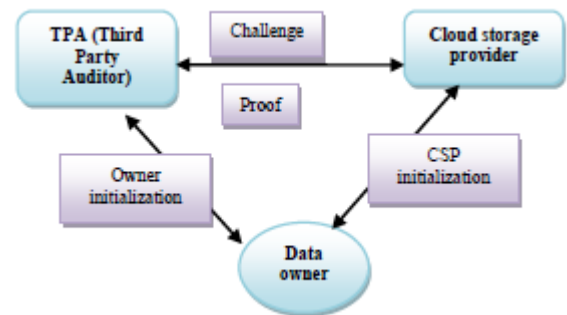


Fig 1. RDA system model

The enterprise or business person will be act as data owner used to upload the data into cloud storage and later he/she can able to modify or update the outsourced data. CSP is responsible for managing data in the cloud space hosted by the data owner. It also has a considerable amount storage space and computing resources for doing operation on the stored data. Third party auditor has sufficient skill set for performing the audit operation on the data also aids to reduce the computational complexity of data auditing process. Before discuss the proposed auditing protocol in details let us discuss the general notation going to used in these protocol which is listed in table 1

TABLE 1 Notations used in proposed system

Symbols	Meaning
$Stk$	Secret Tag key
$Ptk$	Public Tag key
$Shk$	Secret hash key
$D_c$	Data Component
$T_d$	Data tag sets
$N_c$	Number of blocks in each data component
$N_s$	Number of sector in each data block
$D_{c,info}$	Abstract information of $D_c$
$C_a$	Auditor challenge
$S_p$	Server proof

The proposed system model has the following function, such as Key generation, Tag generation, challenge, proof generation, Proof verification which helps to design a remote data integrity checking protocols.

**Key generation setup ( $\mu$ )  $\rightarrow$  ( $Stk, Ptk$ )** this function takes only security parameter as input  $\mu$  and produce output as pair of secret hash key and secret public key ( $Shk, Ptk$ ).

**Tag generation ( $D_c, Shk, Stk$ )  $\rightarrow T_d$**  The main goal of this function is to verify the data integrity. It takes Data component as  $D_c$ , secret tag key  $Stk$  and secret has key  $Shk$  as input and compute data component and make it publically known to everyone.

**Challenge ( $D_c$  info)  $\rightarrow C_a$**  its takes input as abstract information of data component and out puts the challenge message to data owner.

**Proof generation ( $D_c, T_d, C_a$ )  $\rightarrow S_p$**  its takes inputs as abstract information, data component, Auditor challenge and out put the server proof  $S_p$ .

**Verification ( $C_a, S_p, Shk, Ptk, D_c$  info)  $\rightarrow$  (Accept, Reject)** this verification takes as inputs the Auditor challenge, server proof, secret hash key, Public tag key, data component and abstract information of  $D_c$  and out the auditing results as accept or reject.

During the whole auditing process auditor should be truthful and interested about received data. The server could be not truthful and may begin the following attacks such as replace, replay and forge attacks.

**Replace attack:** To replace the already discarded data block and data tag ( $D_c, T_d$ ) the server chose the appropriate and unaffected data block and tag for challenge operation.

**Replay attack:** Without using the data owner information the server may generate the duplicate proof from previous information or other information for replay attack.

**Forge attack:** The data tag of the data block was forge by the server and mislead the auditor, when the owner secret tag key is reused for different version of the data.

A. Privacy preserving protocol for cloud environment

In this segment, we discuss the basic techniques applied in the proposed design of auditing protocol after that we

present the proposed algorithms and structure of auditing protocol for cloud storage system. In our system, data privacy is the major challenge in the design of the data storage auditing protocol. This reason behind is a) if the data is publically available means the auditor is easily attaining the data information by recovering the data blocks. b) If it is encrypted data means, the auditor can obtain the encrypted key through by using some special operation and can able to decrypt the data. In our proposed system the data privacy problem could be solved by generating the encrypted proof by challenging stamp by using By linearity property where the auditor can verify the correctness of the data by decrypting it.

In general, to conduct auditing service in cloud environment, auditor should have knowledge and capabilities. The computing viabilities for auditor are not as strong as cloud server. Since the performance of the system get reduced because of the huge auditing process done by the auditor. To overcome the issues, the proof of intermediate value verification will be computed by the cloud server and the auditor make use of this intermediate value to verify the proof. As a result, the computation load is reduced by delegating work to the server.

B. Algorithm

Let us consider a file ' $f$ ' having  $n$  data components as  $f = (f_{dc1}, \dots, f_{dcn})$  which each file has its own meaning and it can be updated dynamically by the data owner. There are two cases for performing encryption operation on the file. If the file is publically available means that owner need not to encrypt the data but for private data component, that data owner must do the encryption operation with its corresponding keys. Due to the security reason the data components ( $DC_i$ ) of each files are divided into  $N_{dc}$  data blocks as:

$$f = (DB1, DB2, \dots, DBn) \tag{1}$$

Security parameter has been used to reduce the data block size for provided more production to the out sourced data by the data owner. For instance, if the security level is set to 180bit then the data block size should be 30byte. This block size reduction will help to reduce storage overhead in real time process.

Data fragment techniques could be used to divide each data block into sectors and the size of the sector also have been reduced with some limitation using the security parameter. To reduce the number of data tags, its generated for each block which consist of 's' sectors. The size of the data block could be varied in real time storage system; different data block contains different size sectors. For instance, the frequently read data block  $DC_i$  which contain large number of sector  $S_i$  at the same time if the data block is regularly updated means the sector size is relatively small. In general, data component and constant number of sector for each data block can be consider for construction of auditing protocol.

Initially the data component  $D_c$  can be divided into  $n$  number of data blocks and each block split into 'S' sectors.

The sector for each data blocks will varies based on the data components, first its selects maximum number of sector 'SEmax' among all sector numbers  $S_i$ . Then we consider for each  $D_c$  with  $S_i < S_{Emax}$  which tell that the data block has  $< S_{Emax}$  sector by setting  $D_{cij} = 0$  for  $S_i < j \leq S_{Emax}$ . Since the size of the each sector is constant and equal to the security parameter 'p' also the data component can be computed as follows :

$$n = \text{sizeof}(D_c) / s \cdot \log p \quad (2)$$

The encrypted data component is denoted as  $DC = \{dcij\}$   $i \in [1, n], j \in [1, s]$ . Let us consider the multiplicative group  $MG1$ ,  $MG2 \dots MGt$  with same parameter  $p$  and  $E : MG1 \times MG2 \square MGt$  be the bilinear map. The generator of  $MG1$  and  $MG2 \in g_1$  and  $g_2$  respectively and the secure hash function  $H : \{0,1\}^* \rightarrow MG1$  that maps the DC information to a point in  $MG1$ . The proposed auditing protocol contains the Key generation tag generation, challenge and proof algorithm which help to build the frame work for data privacy preserving audit protocol.

**Key generation** randomly choose two number  $Stk, Shk \in \mathbb{Z}_p$  as secret key and secret hash key and produces output as public tag key, secret key and secret has key  $pkt = g^2 Stk \in MG2$ .

**Tag generation** algorithm first chooses  $s$  as random values as  $V_1, V_2 \dots V_s \in \mathbb{Z}_p$  and computes  $u_i g^1 X_j$  for all  $j \in [1, s]$ . for each data block  $DC_i (i \in [1, n])$ , the data tag  $DCT_i$  is computed as:

$$DCT_i = (h(Shk, W_i) \cdot \prod_{j=1}^s dc_{ij} V_j) \cdot skt \quad (3)$$

Where  $W_i$  indicate the concatenation operation which uses the data identifier FID and block number of data component to produces set of data tags as outputs.

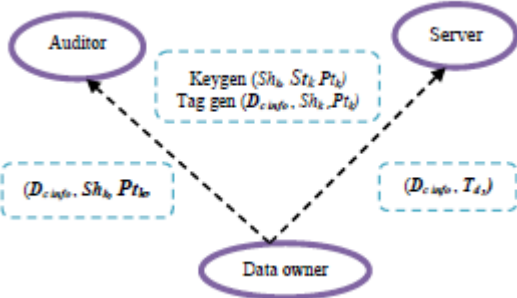


Fig 2. Owner initializations

**Challenge (Dc info)** :the input for this process is abstract information of data  $Dc info$  and select some data block to construct the challenge set  $q$  and generate random number for all data block. It output the challenge stamp  $C_s = Ptk$  by randomly choosing number from data block.

**Proof (Dc, Td, Ca)**  $\square P$  this process uses the challenge from the previous step and data component used as inputs. It contain tag proof  $tp$  and data proof  $dp$  and output as

$$P = (tp, dp).$$

**Verify (Ca, P, Shk, Ptk, Dc info)**  $\square 0/1$ . It initially computes identifier  $I_{challenge}$  value and of all challenge data block and calculate challenge has value  $H_{challenge}$  as follows :

$$H_{challenge} = \prod (Shk, w_i)^{i \in q} \quad (4)$$

it uses the following equation to verify the data proof from the server.  $DP.e(H_{challenge}, Ptk) = e(TP, g^2 r)$  (5)

The output of the above equation 5 holds 0 otherwise 1.

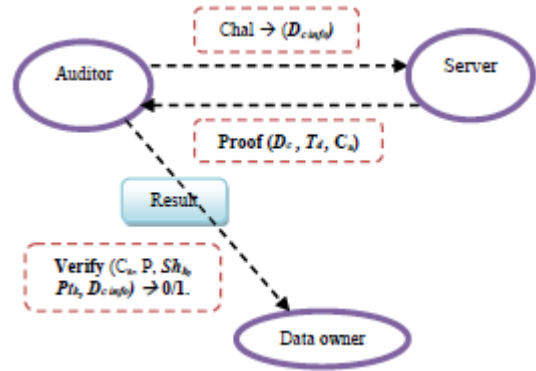


Fig 3. Audit conformation

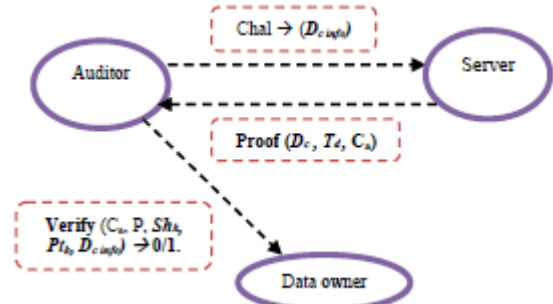


Fig 4. Sample auditing

### C. Audit protocol construction

The above diagram contains three phases for audit protocol construction such as initialization of the data owner or owner construction, Audit conformation and Trial auditing. In the begin or initial phase, data owner generates key and tags for data. Next the data has to be store in the server after wards the owner asks the auditor to contact the audit conformation to make sure that their data correctly stored on the server. After conformation received from the auditor the owner has rights to delete the local copy of the data. To verify or check the data integrity the auditor can periodically contact the trail auditing on the data.

Data owner initialization component initially run the key generation algorithm to generate secret-public tag key

(**Stk, Ptk**) and secret has key **Shk** , then it compute the data tag by running tag generation algorithm. After generating the data tags , the data owner send the each data components **Dc** and its corresponding tags to the server in concert with set of parameter. Finally, the data owner send the abstract information of data component, secret hash key and public tag key (**Dc, Shk, Ptk**) to the auditor for initialization process.

The next phase of the audit construction is conformation auditing which involves only two way communication such as challenge and proof. The main goal of this phase auditor to check whether the owner data is correctly stored in the server or not. The working process of the conformation auditing as follows:

First, auditor run the challenge algorithm ‘**Ca**’ (**CHALL**) to produce challenge for all data blocks in the data components ‘**Dc**’ and auditor send response message to server. After receiving the challenge response from the auditor, the server run the prove algorithm to generate proof ‘**Sp**’ and send back to the auditor. After receiving the proof by the auditor from the server it can run the verification algorithm to check the correctness of the challenge message and extract the audit result. Next, audit result should be send to the data owner and he will check whether the result is correct of not. If its correct then the owner convinced that data is correctly stored on the server and he/she can delete the local data.

Sample auditing is important phase in audit protocol construction, which could be contacted periodically by modifying test set of data blocks. This audits process depends on the service conformity between the data owner and server and how much data owner having trust on the server. During the sample auditing process if any data corruption is happen that could be calculated as follows by using probability function. For instance, each sector in the data block is corrupted every so often with probability ‘**P**’ on the server. The probability of detection of ‘t’ challenged data blocks in sample auditing is calculated as

$$P(t,s) = 1 - (1-p)^{t-s} \tag{6}$$

The equation 6 is used to detect any corrupted data block in sample auditing process. Correctness of the proof for privacy preserving auditing protocol can uses the standard following principal: which state that, the server only pass the audit challenge – response protocol if all the data blocks and tags or correctly stored. The proof verification equation can be written as follows:

$$DP.(eHchallengeptk) \tag{7}$$

if any of the data block or tags or corrupted or modified the server cannot pass the audit.

#### IV. SECURE DYNAMIC AUDITING AND SOLUTION

Data owner can dynamically update their data in cloud environment. In our proposed system the auditing protocol designs to support the static and dynamic archive data. Even

though, the dynamic operation will formulate the auditing protocol insecure by conducting the *reply* and *forge* attack by the server. During the reply attack, the server fails to update the owner data and he will use the oldest version of the data to contact the auditing. In case of *forge* attack the owner update the data to the current version and server may receive the information about forge data tag from dynamic operation by using this data tag he can pass auditing.

The solution for reply attack could be provided by introducing I Table or *Index Table* which is used to record the abstract information about the data tag. The I Table contains four major components such as, *index* component used to denote the current block number of the data clock **DBi** in data components. **BNi** it denotes the original block number of data blocks and **VS** indicate the version number **TS** denotes the time stamp used to generate the data tag. During the owner initialization Index table is created and managed by auditor. After completion dynamic operation, the data owner sends the update message to auditor for updating the I Table. Subsequent to the conformation meeting the auditor can send the result to data owner to ensure that owner data and abstract information on the auditor are both up to date. The above information completes the dynamic operation. In the case of forge attack, specially focus on modify the tag generation algorithm while generating the data tag **Td** for each data block **Bt** the data owner has to insert abstract information **Dc info** in to data tag. This operation helps for the server cannot get sufficient information for to forge data tag for dynamic operation.

##### A. Dynamic Operation

In general the dynamic audit construction protocol having for phase such as owner initialization, conformation auditing, sample auditing and dynamic auditing. The major difference in this auditing protocol is *tag generation* and *index table* creation during the first phase (OI). In the following phase we discuss the DAP (Dynamic auditing Phase) which consist *Data update*, *Index update* and *Update conformation*.

##### Data update

Data owner can perform three type of data update operation such as Modification, Insertion, and deletion. For every update operation there is a corresponding algorithm in dynamic auditing process which helps for future auditing phase to perform easy operations.

**Modification** (**DBi** ,**StkShk**) → (**MG modify**, **TSi**) the input to the algorithm is all new version of data block **DBi**, secret tag key **Stk**, and secret has key **Shk**. and generate the new version number **VNi**, new time stamp **TSi** and new data tag **DTi** for data block **DBi** which was generated using tag generation algorithm. This algorithm give the updated output as follows: **MG modify** = (**I**, **BNi**, **VS**, **TSi**). Final its send the updated message to auditor and new pair of data block and tag should send to server.

**Insert** (**DBi** ,**StkShk**) → (**MG modify**, **TSi**) it also take same parameter as input as same as modification algorithm.



Then insert new DBi data block before the ith position and generate original data block DBi, new version number VNi, and Time Stamp TSi. Next it uses tag generation algorithm to generate new tag DTi for new data block DBi and output the updated message as MG insert = (I, BNi, VSi, TSi). Now it can insert new pair of data block and tag as (DBi, DTi) on server and send the updated message to the auditor.

<i>D<sub>info</sub></i> Initial abstract information of data <i>D</i> .			
Index	BN <sub>i</sub>	VS <sub>i</sub>	TS <sub>i</sub>
I <sub>1</sub>	BN <sub>1</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>2</sub>	BN <sub>2</sub>	VS <sub>1</sub>	TS <sub>2</sub>
I <sub>3</sub>	BN <sub>3</sub>	VS <sub>1</sub>	TS <sub>3</sub>
I <sub>4</sub>	BN <sub>4</sub>	VS <sub>1</sub>	TS <sub>4</sub>
I <sub>n</sub>	BN <sub>n</sub>	VS <sub>1</sub>	TS <sub>5</sub>

After modification of <i>D<sub>2</sub></i> VS <sub>2</sub> , and TS <sub>2</sub> are updated			
Index	BN <sub>i</sub>	VS <sub>i</sub>	TS <sub>i</sub>
I <sub>1</sub>	BN <sub>1</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>2</sub>	BN <sub>2</sub>	VS <sub>2</sub>	TS <sub>2</sub>
I <sub>3</sub>	BN <sub>3</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>4</sub>	BN <sub>4</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>n</sub>	BN <sub>n</sub>	VS <sub>1</sub>	TS <sub>1</sub>

After modification of <i>D<sub>2</sub></i> VS <sub>2</sub> , and TS <sub>2</sub> are updated			
Index	BN <sub>i</sub>	VS <sub>i</sub>	TS <sub>i</sub>
I <sub>1</sub>	BN <sub>1</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>2</sub>	BN <sub>2</sub>	VS <sub>2</sub>	TS <sub>2</sub>
I <sub>3</sub>	BN <sub>3</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>4</sub>	BN <sub>4</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>n</sub>	BN <sub>n</sub>	VS <sub>1</sub>	TS <sub>1</sub>

After deletion before <i>D<sub>2</sub></i>			
Index	BN <sub>i</sub>	VS <sub>i</sub>	TS <sub>i</sub>
I <sub>1</sub>	BN <sub>1</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>2</sub>	BN <sub>3</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>3</sub>	BN <sub>4</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>4</sub>	BN <sub>5</sub>	VS <sub>1</sub>	TS <sub>1</sub>
I <sub>n-1</sub>	BN <sub>6</sub>	VS <sub>1</sub>	TS <sub>1</sub>

TABLES 2 index table updating after dynamic operation

**Delete (DBi)**  $\square \square \square$  MG delete This algorithm takes input as data block DBi and output the updated message as :MG delete  $\square$  (I, BNi, VSi, TSi) then it delete the pair of data clock and tag (DBi, DTi) form the server and send the updated message to auditor.

## V. SECURITY ANALYSIS FOR PRIVACY PRESERVING PROTOCOL

Security analysis is one of the important in cloud environment operation. In our proposed system we are prove that our auditing protocol can provide guarantee for data privacy under the security model. During the protocol design data privacy will be the important need in the cloud storage system. This could state as following theorem: In our

proposed auditing protocol, during the auditing process neither server or nor the auditor can obtain the information about data and secret has key.

**Verification:** In our process the owner will encrypt the data and the server cannot decrypt the with knowing the secret key of the server. The key information are kept it as secret and cannot infer that based on receiving information through the audit process is running. As a result, the data and keys are kept secret against the server in auditing protocol.

## VI. CONCLISION AND FUTURE ENHANCEMENT

In this paper we discussed about the dynamic remote data auditing using privacy preserving protocol which is used to perform the dynamic operation such as update, modify, insert and delete also provide the security over the data. We also tested our protocol with real data in the cloud environment and result gives data integrity and protect it from auditor. Next the security system model presented.

## REFERENCES

- [1] M. Ali, S.U. Khan, A.V. Vasilakos, Security in cloud computing: opportunities and challenges, Inf. Sci. 305 (2015) 357–383.
- [2] Amazon.com, Amazon elastic compute cloud (Amazon EC2). [Online]. Available: <http://aws.amazon.com/ec2/>.
- [3] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, Above the clouds: a view of cloud computing, Commun. ACM 53, pp. 50–58, 2010.
- [4] P. Mell and T. Grance, “The NIST Definition of Cloud Computing,” technical report, Nat’l Inst. of Standards and Technology, 2009.
- [5] T. Velte, A. Velte, and R. Elsenpeter, Cloud Computing: A Practical Approach, first ed., ch. 7. McGraw-Hill, 2010
- [6] W. Cong, R. Kui, L. Wenjing, L. Jin, Toward publicly auditable secure cloud data storage services, IEEE Netw. 24 (2010) 19–24.
- [7] S. Shamshirband, N.B. Anuar, M.L.M. Kiah, A. Patel, An appraisal and design of a multi-agent system based cooperative wireless intrusion detection computational intelligence technique, Eng. Appl. Artif. Intell. 26 (2013) 2105–2127.
- [8] T. Armerding, The 15 Worst Data Security Breaches of the 21st Century, in: COS Security and Risk, csonline, 2012.
- [9] M. Ali, R. Dhamotharan, E. Khan, S.U. Khan, A.V. Vasilakos, K. Li, A.Y. Zomaya, SeDaSC: secure data sharing in clouds, IEEE Syst. J. PP (2015) 1–10.
- [10] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, D. Song, Remote data checking using provable data possession, ACM Trans. Inf.Syst. Secur. 14 (2011) 1–34.
- [11] J.S. Plank, T1: erasure codes for storage applications, in: Proceedings of the Fourth USENIX Conference on File and Storage Technologies, San Francisco, 2005, pp. 1–74.
- [12] M. Sookhak, H. Talebian, E. Ahmed, A. Gani, M.K. Khan, A review on remote data auditing in single cloud server: taxonomy and open issues, J. Netw. Comput. Appl. 43 (2014) 121–141
- [13] P. Mell and T. Grance, “The NIST Definition of Cloud Computing,” technical report, Nat’l Inst. of Standards and Technology, 2009.
- [14] Q.A. Wang, C. Wang, K. Ren, W.J. Lou, J. Li, Enabling public audit ability and data dynamics for storage security in cloud computing, IEEE Trans. Parallel Distr. 22 (2011) 847–859.

- [15] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, D. Song, Provable data possession at untrusted stores, in: Proceedings of the fourteenth ACM Conference on Computer and Communications Security, ACM, Alexandria, Virginia, USA, 2007, pp. 598–609.
- [16] A. Juels, J. Burton, S. Kaliski, PORs: proofs of retrievability for large files, in: Proceedings of the Fourteenth ACM Conference on Computer and Communications Security, ACM, Alexandria, Virginia, USA, 2007, pp. 584–597
- [17] D. Boneh, C. Gentry, B. Lynn, H. Shacham, Aggregate and verifiably encrypted signatures from bilinear maps, Advances in Cryptology (EUROCRYPT), Springer, Berlin Heidelberg, 2003, pp. 416–43
- [18] F. Sebe, J. Domingo-Ferrer, A. Martinez-Balleste, Y. Deswarte, and J.-J. Quisquater, “Efficient Remote Data Possession Checking in Critical Information Infrastructures,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 8, pp. 1034-1038, Aug. 2008.
- [19] C. Erway, A. Kupcu, C. Papamanthou, and R. Tamassia, “Dynamic Provable Data Possession,” Proc. 16th ACM Conf. Computer and Comm. Security (CCS ’09), pp. 213-222, 2009.
- [20] C. Wang, S.S.-M. Chow, Q. Wang, K. Ren, and W. Lou, “Privacy-Preserving Public Auditing for Secure Cloud Storage,” Cryptology ePrint Archive, Report 2009/579, <http://eprint.iacr.org/>, 2009.
- [21] Y. Zhu, H. Wang, Z. Hu, G.-J. Ahn, H. Hu, and S.S. Yau, “Cooperative Provable Data Possession,” Cryptology ePrint Archive, Report 2010/234, <http://eprint.iacr.org/>, 2010.

# A Novel Approach to Optimize Hybrid Item-based Collaborative Filtering Recommendation Model using R

Abhaya Kumar Sahoo  
School of Computer Engineering  
Kalinga Institute of Industrial Technology  
Deemed to be University, Bhubaneswar, India  
abhayakumarsahoo2012@gmail.com

Chittaranjan Pradhan  
School of Computer Engineering  
Kalinga Institute of Industrial Technology  
Deemed to be University, Bhubaneswar, India  
chitaprakash@gmail.com

**Abstract**—Recommender system is the most information filtering based system which deals with the problem of information overload by filtering vital information from large dynamically collected information according to the user's choices, interest or item's behavior. Collaborative based filtering recommender system is one of best filtering approaches which is very effective in wide range applications. Item based collaborative filtering approach solves scalability and data sparsity issue with better accuracy. In this paper, we have developed optimized hybrid item based collaborative filtering recommendation model using binary rating matrix and Jaccard similarity. In this model, binary rating matrix is first examined and relationships among various items are identified by optimizing nearest neighbors as parameter. Then we use these relationships that help with recommendations for the user.

**Keywords**—Binary rating matrix; Item based Collaborative filtering; Jaccard Similarity; Optimized Parameter; Recommender Systems

## I. INTRODUCTION

Now-a-days, everything is available through the internet. When people are going to buy any kind of product through the internet, they first search for any reviews or comments about that product. At that time people may be confused whether that product is preferable or not based on comments. So Recommendation system provides a platform to recommend such a product which is valuable and acceptable for people. Such system is based on item characteristic, user profile filled on website and information related to products. This filtering based system collects a large amount of information dynamically from user's interest, ratings, choices or item's behavior, filters this information and provides vital information [1, 6].

Recommender system has the ability to predict whether a particular user would prefer an item or not based on the user's profile. This system can be implemented based on user's profile or item's profile. This paper explains about the item based collaborative filtering based recommendation system which provides valuable information to users based on the item's profile. Now-a-days different no. of blog forums are available in the different webs where people can give their opinions, reviews, blogs, comments about the items. After

getting ratings about any product by users, recommendation system makes decisions about users who don't give any ratings [6,8]. Number of e-business websites is taking the help of recommendation system to increase their revenue in the competitive market. Millions of users buy their products from online e-commerce websites. After buying products, they give their opinions or any comments about that product in the respective web forum. So, Generating revenue is the main goal of all entrepreneurs. Using this recommendation system process, we can increase our sales productivity in the market [4].

The rest of the paper is organized as follows: Section 2 describes about the overview of recommendation system. Section 3 presents the collaborative based filtering recommendation system. Section 4 presents proposed item-based collaborative filtering recommendation model using the R language. Section 5 shows the experimental result and section 6 contains the conclusions and future work.

## II. OVERVIEW OF RECOMMENDATION SYSTEM

### A. Preliminaries and Basic concepts of Recommendation System

In recommendation system, the two main entities play main role, i.e. users and items. Users give their preferences about certain items and these preferences must be found out of the collected data. The collected data are represented as a utility matrix which provides the value of each user-item pair that represents the degree of preferences of that user for specific items. In this way, these are mainly two broadcast categories of recommender engine algorithms: user-based and item-based recommenders. In user-based recommender system, users give their choices and ratings on items. We can recommend that item to the user, which is not rated by that user with the help of user-based recommender engine, considering similarity among the users. In item-based recommender system, we use similarity between items (not users) to make predictions from users. Data collection for recommender system is the first job for prediction [1, 6,8].

Phases of Recommendation System are:

**Information Collection Phase:** This phase collects vital information about users and prepares user profile based on user's attribute, behaviors or resources accessed by users. Without constructing well defined user profile, recommendation engine cannot work properly. Recommender system is based on inputs which are collected in different ways, such as explicit feedback, implicit feedback and hybrid feedback. Explicit feedback takes input given by users according to their interest on an item whereas implicit feedback takes user preferences indirectly through observing user behavior. Hybrid feedback can be collected as both explicit and implicit feedback [1].

**Learning Phase:** This phase takes feedback gathered in information collection phase as input and processes this feedback by using learning algorithm and exploits user's features as output [1].

**Prediction/Recommendation Phase:** Preferable items are recommended for users in this phase. By analyzing feedback collected in information collection phase, prediction can be made which is happening through model or memory based or observed activities of users by the system[1].

### B. Types of Recommendation System based on Filtering Techniques

An efficient recommendation technique is very necessary to provide useful recommendation to its individual users. This explains about three types of recommendation techniques which are mainly used for providing recommendation to users about the item. The following figure shows the hierarchy of recommender system based on different filtering techniques [1].

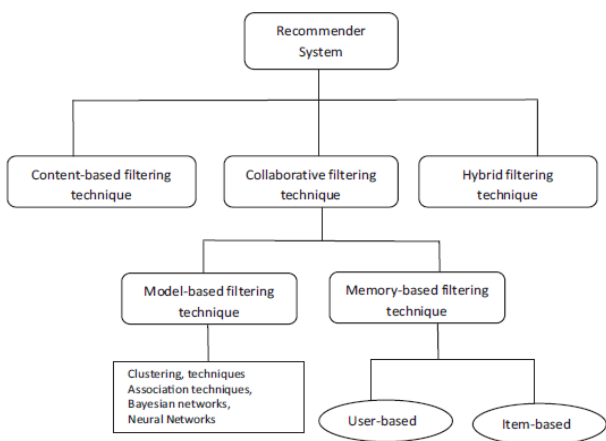


Fig. 1. Hierarchy of Recommender System based on Filtering

### Content based Filtering Recommendation System:

Content-based filtering technique focuses on the analysis of features and attributes of items to generate predictions. Content based filtering is usually used in case of document recommendation. In this technique, recommendation is done based on user profiles, which deal with different attributes of items along with user's previous buying history. Users give their preferences in the terms of ratings which are positive or negative or neutral in nature. In this technique, positive rated items are recommended to the user [1].

### Collaborative based Filtering Recommendation System:

Instead of considering features and attributes of items to determine their similarity, this approach uses user based ratings to find similarity between items. After collecting all user ratings, system compares these ratings with other users with the help of utility matrix and items are recommended to the user. We use different distance measure to approach like Jaccard's distance, cosine distance and Pearson's coefficient, etc. to find a user's similarity. This filtering method is usually used in e-commerce website to recommend items based on users' ratings [2].

**Hybrid Filtering Recommendation System:** This technique comprises above two methods to increase the accuracy and performance of recommendation system. The hybrid filtering technique can be achieved by using any of the following ways: building a unified recommendation system that combines both above two approaches, applying some collaborative filtering in content-based approach and utilizing some content-based filtering in the collaborative approach. This technique uses different hybrid methods such as cascade hybrid, weighted hybrid, mixed hybrid and switching hybrid according to their operations[1].

### III. COLLABORATIVE BASED FILTERING RECOMMENDATION SYSTEM

Collaborative filtering predicts unknown outcomes by creating user-item matrix of choices or preferences for items by users. Similarities between users' profile are measured by matching user-item matrix with users' preferences and interests. The neighborhood is made among groups of users. The user who has not rated to specific items before, that user gets recommendations to those items by considering positive ratings given by users in his neighborhood. The CF in recommendation system can be used either in prediction or recommendation. Prediction is a rating value  $R_{i,j}$  of item  $j$  for user  $i$ . This collaborative filtering technique is mainly categorized in two directions: memory based and model based collaborative filtering. The following figure explains about the whole process of collaborative filtering technique [5, 10].

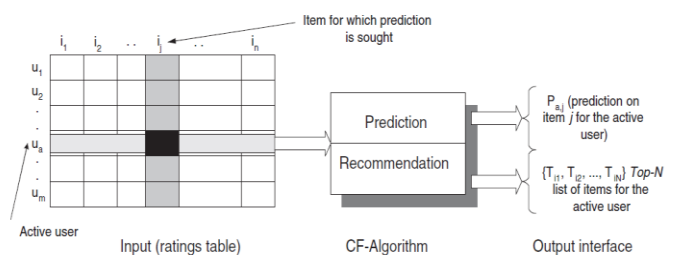


Fig. 2. Collaborative Filtering Technique

#### A. Memory based Collaborative Filtering

Item and user are two key factors in this filtering technique. So this technique comprises into two ways, such as item-based collaborative filtering and user-based collaborative filtering. Prediction is calculated by measuring similarity among the items [5]. This technique builds a model on item similarities by considering all items rated by an active user from user-item matrix, by which we can measure the similarity among target item and all retrieved items. Then we select  $k$  most similar

items and prediction is calculated by considering a weighted average of the active user rating on similar items k. Different mathematical methods are used to compute similarity among item and user. These are: correlation based similarity measure, cosine based similarity measures and Pearson's correlation coefficient. Pearson's coefficient can be defined as:

$$s(a, u) = \frac{\sum_{i=1}^n (r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i=1}^n (r_{a,i} - \bar{r}_a)^2} \sqrt{\sum_{i=1}^n (r_{u,i} - \bar{r}_u)^2}} \quad (1)$$

In the above equation,  $s(a, u)$  represents similarity between two users  $a$  and  $u$ .  $r_{a,i}$  and  $r_{u,i}$  denote rating on an item by user  $a$  and  $u$  respectively, whereas  $\bar{r}_a$  and  $\bar{r}_u$  are mean rating given by user  $a$  and  $u$  respectively, while  $n$  is the total no. of items in user-item matrix.

Cosine similarity can be defined as a vector space model which is based on linear algebra. This method measures similarity between two n-dimensional vectors based on angle between them. It is mainly used in information retrieval and text mining. The similarity between two items  $u$  and  $v$  can be denoted by:

$$s(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{|\vec{u}| * |\vec{v}|} = \frac{\sum_i r_{u,i} r_{v,i}}{\sqrt{\sum_i r_{u,i}^2} \sqrt{\sum_i r_{v,i}^2}} \quad (2)$$

Jaccard Similarity of sets A and B is  $|A \cap B| / |A \cup B|$ ; that is the ratio of the size of the intersection of A and B to the size of their union. It can be denoted by:

$$JS(A, B) = |A \cap B| / |A \cup B| \quad (3)$$

In user-based collaborative filtering technique, similarity between users is measured by comparing ratings on the same item. By calculating a weighted average of ratings of item by users, it predicts rating for an item by active user [6]. In this way, above three methods are used to measure similarity between two items.

### B. Model based Collaborative Filtering

This approach is based on previous ratings to learn a model which uses machine learning or data mining techniques. Different approaches like association rules, clustering, decision tree, artificial neural network, regression and Bayesian classifiers, etc. are used to classify user and item based on model [1].

**Association Rule Mining:** Association rule mining algorithms generate association rules which decide the relationship among items in a transaction. Association rule  $A \rightarrow B$  means item set A predicts item set B. This rule can be fitted to a recommendation model to predict about user and item [3].

**Clustering:** Clustering is a method which is used to partition a set of data into no. of clusters. A good clustering method means high intra-cluster similarity and low inter-cluster similarity. In recommendation system, users can participate in different clusters partially and degree of participation can be calculated by taking average across the clusters of participation [9].

**Decision Tree:** Decision tree makes a graph like tree structure which is constructed by considering the training data set in which class labels are known. This tree can be used to classify test data. The decision tree is one type of classifier which handles and classifies previous unseen examples.

**Artificial Neural Network:** ANN is a network of many connected neurons arranged in different layers. The weight and bias factors are associated with each and every neuron of each layer. Each neuron has a transfer function through which it measures input, process this input and gives output. This ANN is a classification technique to classify test data.

**Regression:** Regression is an analysis method where two or more variables are related to each other. One variable is dependent, whereas one or more are independent variables. This regression technique comprises prediction, curve fitting and hypothesis testing, which create relationships among variables.

**Bayesian Classifier:** Bayesian classifier is used to solve classification problem based on conditional probability and Bayes theorem. Bayesian classifier is used to predict the class through considering the probability of the class with respect to particular attribute by applying Bayes' theorem. This classifier is usually useful when users' preferences changes with respect to the time required building the model.

## IV. PROPOSED OPTIMIZED HYBRID IBCF BASED RECOMMENDATION MODEL USING R LANGUAGE

### A. R Language

R is a statistical computing language which provides facilities like data manipulation, data handling and data analysis, etc. It provides different statistical techniques such as linear and nonlinear modeling, time-series analysis, data mining approaches, machine learning concepts which are very required to develop a recommendation system. This language requires different packages which are open source to download and use for developing recommendation system [7].

### B. Data Exploration

We explore the dataset on which we build recommendation system. Here we use MovieLense dataset that tells about movie ratings. MovieLense is a realRatingMatrix object where each row corresponds to a user, each column to a movie and each value to a rating. After exploring the data, we find the nature of the data along with its dimension. We find vector\_ratings from exploring data by which we collect ratings of movies along with their occurrences. Then we remove movie and its occurrences from vector\_ratings whose rating is 0. In this phase, we explore the average ratings and find out heatmap of the rating matrix.

### C. Data Preparation

To build an efficient recommendation model, after completing data exploration phase, we first prepare data which should be a relevant and normalized form. After applying data normalization technique, the ratings of movies can be any number within a specified range. We convert this normalized matrix into Binary matrix where matrix contains 1 if the user rated the movie, otherwise 0. This Binary matrix can also

contain 1 if the rating is above or equal to a definite threshold value, otherwise 0. After creating Binary matrix, we build item-based collaborative filtering recommendation model.

#### D. Building Model using IBCF

Recommendation model is built using item-based collaborative filtering technique. This model is developed using the following steps:

- Training set and test set are defined. Training set includes users from which the model learns whereas test set includes users to whom we recommend movies.
- Recommendation model is built by taking inputs (training set as data, technique name as method, rating matrix and normalized method as parameters).
- Recommendation model is applied on the test set and items is recommended on the basis of Jaccard similarity.

#### E. Evaluation of Model

After building recommendation model, the model is evaluated by measuring the accuracy through different techniques like root mean square error (RMSE), mean squared error (MSE) and mean absolute error (MAE). Different performance parameters are used to measure accuracy of recommendation model. According to confusion matrix, we can find precision and recall. The following formulas for finding precision, recall and performance are given:

$$Precision = \frac{Total\ number\ of\ false\ positive}{Total\ number\ of\ positives} \quad (4)$$

$$Recall = \frac{Total\ number\ of\ true\ positive}{Total\ number\ of\ purchases} \quad (5)$$

$$Performance\ index = (precision * weight_{precision}) + recall * (1 - weight_{precision}) \quad (6)$$

#### F. Optimization of Model Parameters

In the evaluation of model, we can get better accuracy by optimizing model parameters, i.e. number of nearest neighbors and weight description. We can measure the performance depending on the number of neighbors by calculating a weighted average between precision and recall.

### V. EXPERIMENTAL RESULT

We worked on MovieLense dataset on which we built item based collaborative filtering recommendation system. The following figures explain about the result of each phase. Finally, we measured performance of recommender system through different performance parameters.

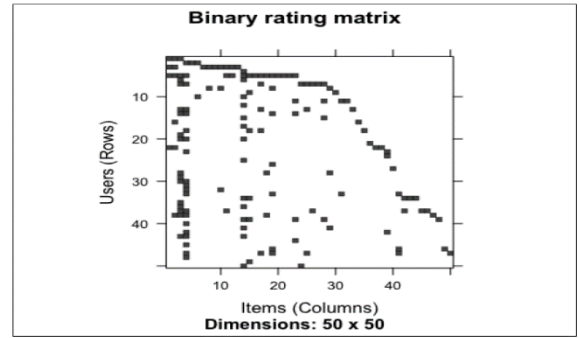


Fig. 3. Binary Rating Matrix

After normalization of relevant data, we converted rating matrix into binary form. The above fig. 3 shows a binary rating matrix which explains the ratings of users on different items in binary form.

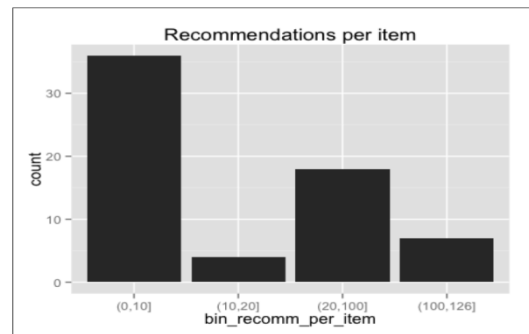


Fig. 4. Recommendations per Item

After building the recommender model, we explored the output in which we counted how many times it has been recommended. The above fig. 4. Shows most of the items have been recommended more than 100 recommendations. The distribution has a long tail.

TABLE I. PERFORMANCE PARAMETERS

nn	precision	recall	performance
4	0.1663	0.5935	0.3799
6	0.1769	0.621	0.399
8	0.1769	0.5973	0.3871
10	0.175	0.5943	0.3846
12	0.174	0.5909	0.3825
14	0.1808	0.6046	0.3927

After building the model, we should evaluate the accuracy of this model by measuring parameters like precision, recall and performance. We took the number of nearest neighbors, i.e. 30, number of fold parameters, i.e. 10 and rating matrix as input data. We used Jaccard's coefficient as a distance measure. In this experiment, we optimized number of neighbors parameter to achieve hybrid item-based collaborative filtering model. The above Table I shows result of precision, recall and performance value by taking different number of

neighbors. Here precision is the percentage of recommended items that have been purchased whereas recall is the percentage items that have been recommended and weight\_precision value is 0.5.

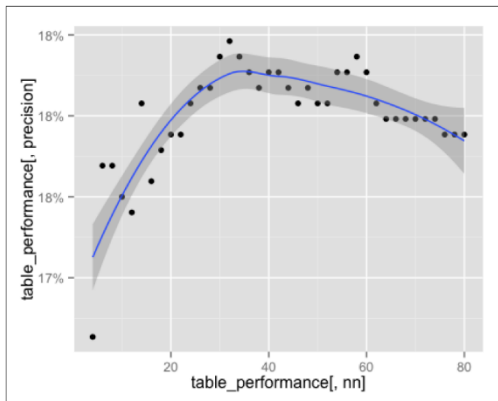


Fig. 5. Performance Graph between nn and precision

The fig. 5 shows a smoothed line that grows until the global maximum, which is around  $nn=35$ , slowly decreases. This index expresses the percentage of recommendations that have been successful, so it is useful when there are high costs associated with advertising.

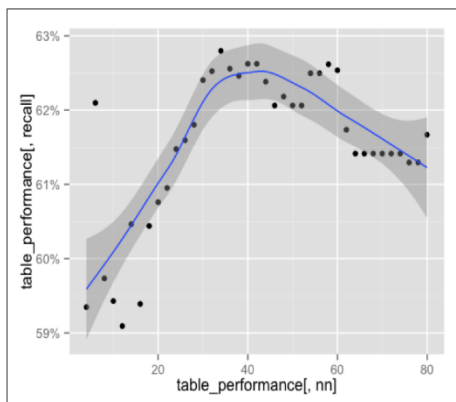


Fig. 6. Performance Graph between nn and recall

The fig. 6 shows that maximum recall is around  $nn=40$ . This index expresses the percentage of purchases that we recommended, so it is useful if we want to be sure to predict most of the purchases.

## VI. CONCLUSION

Item-based collaborative filtering recommendation system is better to use when users are far greater than the number of items. The performance of this system can be affected by data sparsity, cold start problem, shilling attack and privacy. So there is a great chance of the future research area. Here we built an optimized hybrid item-based collaborative filtering recommendation model by gaining high performance through optimized number of neighbors. In the future, we can build high performance based recommendation model by optimizing our IBCF algorithm on the basis of optimized IBCF parameters and improving the item description.

## REFERENCES

- [1] F.O. Isinkaye , Y.O. Folajimi, and B.A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," Egyptian Informatics Journal, ISSN: 1110-8665 , pp.261–273, 2015.
- [2] F. Ortega , A. Hernando , J. Bobadilla , and J. H. Kang, "Recommending items to group of users using Matrix Factorization based Collaborative Filtering," Information Sciences ,vol. 345, pp.313–324, 2016.
- [3] J. H. Jooa, S.W. Bangb, and G. D. Parka, "Implementation of a Recommendation System using Association Rules and Collaborative Filtering," Information Technology and Quantitative Management ,procedia computer science, vol.91, pp.944 – 952, 2016.
- [4] L. T. Ponnam , and S. D. Punyasamudram, "Movie Recommender System Using Item Based Collaborative Filtering Technique," International Conference on Emerging Trends in Engineering, Technology and Science , Thanjavur, Vol.1, pp.56-60, 2016.
- [5] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," ACM Digital Library International world wide web conferences, ISBN:1-58113-348-0, pp. 285-295, 2001.
- [6] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutierrez, "Recommender systems survey," Knowledge-Based Systems, vol.46, pp.109-132, 2013.
- [7] W. N. Venables, D. M. Smith, "An Introduction to R", 2017.
- [8] R. Burke, A. Felfernig, and M. H. Goker "Recommender Systems: An Overview," Association for the Advancement of Artificial Intelligence, ISSN 0738-4602, pp.13-18, 2011.
- [9] X. Ma, H. Lu , Z. Gan , and J. Zeng "An explicit trust and distrust clustering based collaborative filtering recommendation approach," Electronic Commerce Research and Applications, vol. 25, pp. 29–39, 2017.
- [10] P.A. Riyaz, and S. M. Varghese, "A Scalable Product Recommendations using Collaborative Filtering in Hadoop for Bigdata," International Conference on Emerging Trends in Engineering, Science and Technology, vol. 24, pp. 1393 – 1399, 2016.

# Urban Slum Extraction using GLCM based Statistical approach from Very High Resolution Satellite data

R.Prabhu, S.Umasree, R.Avudaiammal

Department of ECE,

St. Joseph's College of Engineering,  
Chennai, India

prabhutcece@gmail.com, umasree2908@gmail.com

R.A.Alagu Raja

Department of ECE,

Thiagarajar College of Engineering,  
Madurai, India

alaguraja@tce.edu

**Abstract**—This paper proposes a new technique to detect urban slums from urban buildings using a very high resolution (VHR) satellite data. Unlike buildings, vegetation and other features, urban slums lack in their unique spectral signatures. Thus, an accurate detection of slums using remote sensing data poses a real challenge to researchers and decision-makers. In this work, Gray level co-occurrence matrix (GLCM) based statistical feature extraction technique is proposed for detecting the urban slums. Development of this algorithm is motivated by modelling the urban slums using the GLCM features (Entropy, Energy, Angular Second Moment, Homogeneity, Contrast, and Correlation) for the whole image. Thus, to increase the accuracy, GLCM features can be applied to each block of the image of size 5\*5, 7\*7 and 9\*9 respectively. The very high-resolution data of Madurai city, South India acquired by World View 2 sensor (1.84m) proved the ability of proposed approach to identify urban slums from urban buildings and can generate higher classification accuracy (68.8%) than the object based image analysis approach.

**Keywords**—GLCM; informal settlement; urban building; very high resolution

## I. INTRODUCTION

Slum households are the group of individuals living under the same roof in an urban area who lacks tenure security, access to safe water, access to acceptable sanitation, durability of housing, and are overcrowded as defined by UN-HABITAT [1].

The intent of the United Nations Millennium Development is to have achieved a significant improvement in the lives of at least 100 million slum dwellers by 2020. In 2001, 924 million people, or 31.6% of the world's urban population, lived in slums and accounted for 43% of the urban population in developing countries. In 2007, the number of slum dwellers worldwide crossed one billion marks, with 1 of every 3 city-dwellers worldwide [2, 3, 4] lacking basic services such as clean drinking water and adequate living space. The explosive growth of the slum population creates the living conditions that contribute to poor health, high mortality, excessive crime and economic degradation with inadequate response mechanisms during times of natural disaster. A small area of at least 300 population or about 60-70 households of poorly built congested tenements, an unhygienic environment usually with inadequate

infrastructure and lack of proper sanitation and drinking water facilities are generally termed as slums [5]. India aims to be slum-free city by the scheme Rajiv Awas Yojana (RAY) with the budget of 12.7 billion Indian Rupees (INR) (USD 278 million).

One of the fundamental difficulties that authorities face when planning a response to the formation and growth of informal settlements is the lack of spatial and temporal data. Such data allows us to identify and quantify services and infrastructure, which are required to improve our understanding of settlement morphology, population distribution and emerging settlement patterns. Several reasons exist for the scarcity of data on informal settlements. Thus, there are no suitable methods to identify the spatial behavior of informal settlements.

Several studies have been addressed to identify slums from the remotely sensed data. One approach in identifying the slums, used image features like Histogram of Oriented Gradients (HoG), Line Support Regions (LSR), TEXTONS, and Object-Oriented Approach (OBA) at multiple scales along with a Decision tree to classify the informal from formal area. However, these approaches are classifying slums effectively, this uses pixel based approach for the computation [6,7,8]. Nursidiket *al.* has explored the use of Gabor filter and GINI index that automatically detect the slum area when the density of housing is very high. It gives a better result but has a low kappa value since non-slum areas may be categorized as slums due to the texture characteristics after filtering [9].

In recent years, the Gray-level co-occurrence matrix (GLCM) has become one of the popular approaches to analyze urban areas [10,11]. To improve the living condition of slum dwellers (especially within developing countries) standardized methods for slum extraction is necessary. In this paper, we present and discuss the utility of the Haralick's features based on the Gray level co-occurrence matrix (GLCM) to detect and extract the urban slums from urban buildings. In section 2 and 3, the study area and the proposed methodology are discussed. In section 4, experimental results are given and discussed before to conclude the paper with final remarks in section 5.



## II. STUDY AREA

In this research, we have chosen Madurai city as our study area. Madurai is the second largest city in the Indian state of Tamil Nadu and is the 25<sup>th</sup> populated city in India [10]. It is known as the Athens of the East, which is one of the ancient historic cities in the world. The municipal corporation of Madurai has an area of 52 km<sup>2</sup>, within an urban area now covering 178 km with an overall population of around 30 Lakhs people. Tamil Nadu Slum Clearance Board (TNSCB) and the city corporation has identified 196 slum areas in Madurai for clearance and relocation. Initially, 331 slums within the Madurai Corporation area were identified for relocation of the residents. But later, the officials scrapped 135 of them from the list as they were already developed. It is estimated that 40,000 families are living in the 196 slums. There are nearly 200 slums in the Madurai city along the Vaigai banks and railway tracks that are mostly concentrated in Karumbalai and Alwarpuram.

The image has been taken by the World View-2 satellite sensor in the year of June 2012, with the Resolution of 0.46 meter for panchromatic and 1.85 meters for multispectral data. The World View-2 sensor provides a high resolution Panchromatic band and four Multispectral bands (red, green, blue, and near-infrared bands) for enhanced spectral analysis mapping and observing applications, land-use planning, disaster relief, defense and intelligence, visualization and simulation. In this work, the Worldview-2 MSS (Multispectral data) of Madurai city, Karumbalai Slum dataset has been considered with the size 418\*393 as shown in Fig. 1.

Geographic location of our study area

Longitude 78 ° 8' 22.39" E - 78 ° 8' 52.83" E  
Latitude 9 ° 55' 56.71" N - 9 ° 55' 34.66" N



Fig. 1. World View-2 subset Image of Madurai

## III. PROPOSED METHODOLOGY

The steps involved in our research are categorized into three stages: Feature extraction, Classification and Accuracy Assessment. Feature extraction stage involves the computation of GLCM features (Entropy, Energy, Angular Second Moment, Homogeneity, Contrast, and Correlation) for each block of the

image with size 5\*5, 7\*7 and 9\*9 respectively. After computing the features, classification is performed using Decision tree classification algorithm. Finally, the classified result is verified with the ground truth data to assess the accuracy. The flow chart of the proposed methodology is depicted in Fig. 2.

Haralick's *et al.*[12,13] introduced the Gray level co-occurrence matrix (GLCM) technique, which is widely used in image analysis applications. GLCM is the tabulation of how often different combinations of gray levels co-occur in an image where each image is composed of pixels each with an intensity value.

GLCM is a two-dimensional array,  $P$ , in which both rows and columns represent a set of all possible brightness values. Equation (1) specifies a probability matrix for a displacement vector  $d=(dx,dy)$  and counting pairs of pixels having specific gray levels  $i$  and  $j$  such that

$$P_d(i, j) = n_{ij} \quad (1)$$

where  $P_d(i, j)$  is the probability of the  $(i,j)$ th element of the GLCM.  $n_{ij}$  is the number of occurrences of pixel values  $(i,j)$  at distance  $d$  in the image. The normalized Gray Level co-occurrence matrix (GLCM) is given on (2).

$$N(i, j) = \frac{P_d(i, j)}{\sum_{i, j=0}^{N_g-1} P_d(i, j)} \quad (2)$$

According to GLCM, texture statistics of the remote sensing data can be extracted using 14 textural features defined from the normalized probability. In this paper, 6 important features namely, Entropy, Energy, Angular Secondary Moment, Homogeneity, Contrast, and Correlation are selected for implementation. Once the GLCM is computed, rest of the texture measures can be produced. Various features used for GLCM texture measures are listed in Table I.

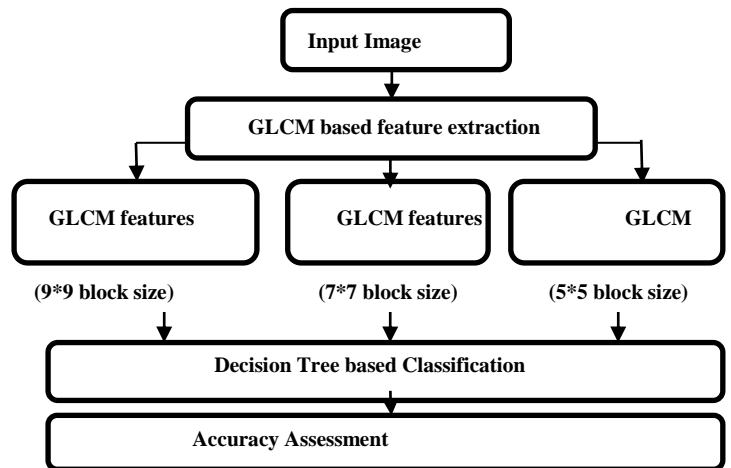


Fig. 2. Proposed Methodology

TABLE I. GLCM TEXTURE FEATURES

Features	Description	Formula
Entropy	Entropy is a Statistical measure of randomness that can be used to characterize the texture of the input image	$\sum_{i,j=0}^{N_g-1} P_{i,j} (-\ln P_{i,j})$
Energy and Angular Secondary Moment (ASM)	Provides the sum of squared elements in the GLCM. High values of Energy occur when the window is very orderly.	$Energy = \sqrt{ASM}$ $ASM = \sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_d^2(i, j)$
Homogeneity	It measures image homogeneity as it assumes larger values for smaller gray tone differences in pair elements. It measures the closeness of distribution of elements in GLCM to the GLCM diagonals.	$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_d(i, j) \frac{1}{1 + (i - j)^2}$
Contrast	Returns a measure of the intensity contrast between a pixel and its neighbor over the whole. Measures the local variations in the gray-level co-occurrence matrix. The weight continues to increase exponentially as (i-j) increases.	$\sum_{i=0}^{N_g-1} \sum_{j=0}^{N_g-1} P_d(i, j) * (i - j)^2$
Correlation	Returns a measure of how correlated a pixel is to its neighbor over the whole image Range= [-1 1] Correlation is 1/-1 for a perfectly positively or negatively correlated image.	$\sum_{i,j=0}^{N_g-1} P_{i,j} \left[ \frac{(i - \mu_x)(j - \mu_y)}{\sigma_x \sigma_y} \right]$

IV. RESULTS AND DISCUSSION

A. After GLCM Based Feature Extraction

GLCM features have been calculated for sample datasets of urban building and slum areas having the size of 9\*9, 7\*7 and 5\*5 respectively as shown in the Tables II, III and IV.

By comparing all the tables, (II-IV), it is found that, as the size of the blocks in an image decreases progressively, GLCM features namely Contrast, Energy, Angular Secondary Moment (ASM), and Dissimilarity decreases whereas Entropy and Homogeneity increases for slum areas and vice-versa for urban buildings.

On comparing each feature value of urban slums with the urban buildings, it has been shown that slums are having higher Entropy for blocks of size 9\*9 and 5\*5. But for the sample datasets of size 7\*7, entropy of the slum area is lower than the formal area. In order to overcome this, mean value is calculated for all the GLCM features of different blocks as given in Table V. From Table V, it is proven that slums are having Higher Entropy, Lower ASM, Lower Energy, Lower Homogeneity, Lower Correlation and Higher Dissimilarity than urban buildings for all the blocks of size 9\*9, 7\*7 and 5\*5.

TABLE II. TEXTURE MEASURES FOR SIZE 9\*9

Texture Measures	Contrast	Entropy	ASM	Energy	Homogeneity	Correlation	Dissimilarity
Slum Area 1	1751.50	3.3550	0.035	0.1886	0.0793	0.5962	31.1333
Slum Area 2	534.602	2.8467	0.095	0.3091	0.2889	0.7945	16.1333
Slum Area 3	429.433	3.4012	0.033	0.1816	0.0469	0.5849	17.2333
Slum Area 4	773.333	3.1330	0.055	0.2379	0.1854	0.5819	18.6667
Slum Area 5	1120.10	3.3550	0.035	0.1886	0.0172	0.8163	25.1333
Formal Area 1	974.083	4.8513	0.007	0.8888	0.0680	0.8597	23.3667
Formal Area 2	1456.03	4.7065	0.010	0.1039	0.1273	0.8645	24.1000
Formal Area 3	1691.30	4.1934	0.045	0.2130	0.2358	0.7654	27.1333
Formal Area 4	663.802	4.2341	0.078	0.2801	0.5172	0.7234	12.9667
Formal Area 5	1255.50	4.5345	0.063	0.2519	0.5443	0.5945	16.8000

TABLE III. TEXTURE MEASURES FOR SIZE 7\*7

Texture Measures	Contrast	Entropy	ASM	Energy	Homogeneity	Correlation	Dissimilarity
Slum Area 1	977.90	4.3150	0.018	0.1363	0.1109	0.6062	30.9335
Slum Area 2	857.50	4.1393	0.023	0.1536	0.1379	0.6344	23.6339
Slum Area 3	547.40	4.4998	0.011	0.1053	0.0279	0.6547	19.6337
Slum Area 4	38.333	4.1913	0.024	0.1577	0.1554	0.7517	21.9664
Slum Area 5	160.10	4.4844	0.011	0.1067	0.0472	0.7363	26.2333
Formal Area 1	173.32	4.0255	0.016	0.1276	0.0623	0.8549	26.5665
Formal Area 2	84.765	3.9098	0.011	0.1053	0.1256	0.9480	18.0005
Formal Area 3	491.80	3.3012	0.081	0.2856	0.2938	0.7370	25.9332
Formal Area 4	20.700	2.8550	0.153	0.3917	0.4030	0.8822	12.3665
Formal Area 5	105.60	2.3950	0.293	0.5416	0.5432	0.8556	5.7306

TABLE IV. TEXTURE MEASURES FOR SIZE 5\*5

Texture Measures	Contrast	Entropy	ASM	Energy	Homogeneity	Correlation	Dissimilarity
Slum Area 1	1029.40	4.6456	0.015	0.1236	0.1204	0.6507	27.9337
Slum Area 2	948.620	4.5696	0.017	0.1315	0.1155	0.5943	25.6335
Slum Area 3	766.666	4.8797	0.007	0.0842	0.0482	0.6246	21.6333
Slum Area 4	981.343	4.4413	0.027	0.1667	0.1657	0.6918	22.9666
Slum Area 5	1403.50	4.8742	0.007	0.0877	0.0363	0.6562	29.2333
Formal Area 1	1589.78	3.4389	0.033	0.1816	0.1059	0.8416	30.5667
Formal Area 2	525.845	3.2594	0.040	0.2000	0.1703	0.9581	15.0009
Formal Area 3	1799.80	3.1616	0.055	0.2347	0.2125	0.6465	29.9331
Formal Area 4	237.598	1.4952	0.453	0.6732	0.6712	0.6962	6.9666
Formal Area 5	102.131	2.6957	0.133	0.3652	0.3531	0.8615	7.2033

TABLE V. TEXTURE MEASURES FOR THE DATASET

Texture Measures		Contrast	Entropy	Energy	ASM	Homogeneity	Correlation	Dissimilarity
(9*9) Image	Slum Areas	921.789	4.462	0.2211	0.0422	0.0832	0.7001	21.8661
	Formal Areas	1208.09	3.238	0.3475	0.0374	0.2971	0.7562	20.7812
(7*7) Image	Slum Areas	1066.26	4.331	0.1319	0.0207	0.1095	0.6666	22.6423
	Formal Areas	794.980	3.261	0.2899	0.1141	0.2891	0.8556	17.6324
(5*5) Image	Slum Areas	1145.41	4.635	0.1187	0.0179	0.1161	0.6897	23.3959
	Formal Areas	850.626	2.752	0.3283	0.1434	0.3026	0.7961	17.9034

### B. Classification Of Image

The Worldview 2 dataset of Madurai city divided into blocks of size 9\*9, 7\*7 and 5\*5 respectively is depicted in Figs. 3-5. After dividing the image into blocks, GLCM features (Contrast, Energy, Entropy, Correlation, ASM, Dissimilarity, Homogeneity) are calculated on each 9\*9 block and all these GLCM features are compared with the average GLCM features (Contrast, Energy, Entropy, Correlation, ASM, Dissimilarity, Homogeneity) of size 9\*9 block for both urban building and slum classes.

If more than 4 features falling onto the average value of particular class, then that block can be categorized as the corresponding class. Thus, by comparing the values, each block can be classified as formal or informal area. Similarly, this concept can be extended to all blocks in an image. Hence, the

dataset can be classified as formal and informal area by using the above algorithm. The result shows that of the formal and informal areas. In order to have a better accuracy, the image is divided into 7\*7 and 5\*5 blocks and the same procedure is repeated. The results are shown in Figs. 6-8.

### C. Accuracy Assessment

The accuracy of the classified outputs is obtained by using the ground truth data. By comparing the samples of ground truth data (100 Pixels) with our classified output, the accuracy assessment is performed that is shown in Table VI.

Upon calculating the accuracy, it has been found that GLCM features of 5\*5 blocks produces a better producer accuracy for urban slums (68.8%), whereas the GLCM features of 7\*7 and 9\*9 blocks produces the producer accuracy of 54.45% and 60% respectively for urban slums.



Fig. 3. Blocks of size 9\*9

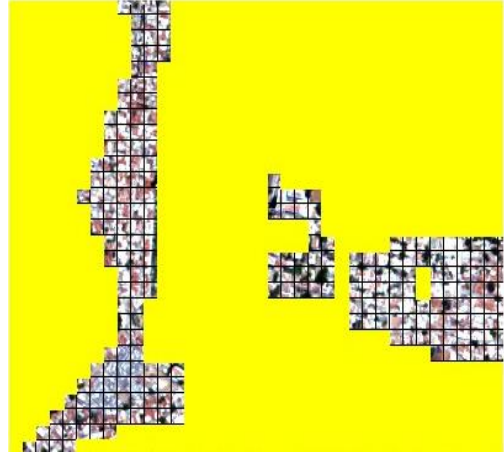


Fig. 6. Slum extraction with block of size 9\*9



Fig. 4. Blocks of size 7\*7

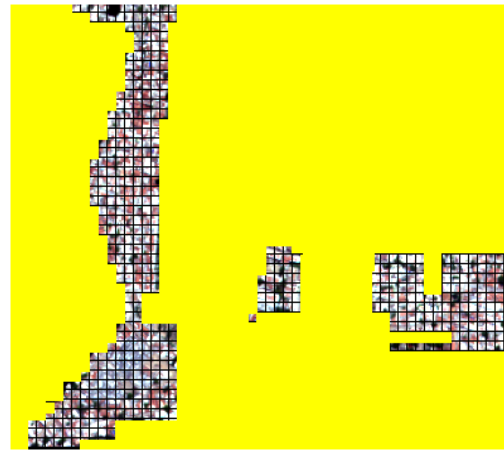


Fig. 7. Slum extraction with block of size 7\*7



Fig. 5. Blocks of size 5\*5



Fig. 8. Slum extraction with block of size 5\*5

TABLE VI. ACCURACY ASSESSMENT FOR OUR DATASET WITH BLOCKS OF DIFFERENT SIZE

Classified output		Urban Slums (pixels)	Others (Pixels)	User's Accuracy (%)	Producer's Accuracy (%)	Overall Accuracy (%)
Ground truth data (100 pixels)						
Block Size(9*9)	Urban Slums	48	52	48	<b>54.45</b>	54
	Others	40	60	60	46.45	
Block Size(7*7)	Urban Slums	54	46	54	<b>60.00</b>	59
	Others	36	64	64	41.00	
Block Size(5*5)	Urban Slums	62	38	62	<b>68.80</b>	67
	Others	28	72	72	65.50	

V. CONCLUSION

This project dealt the approach for identifying the slums (informal settlements) using statistical based feature extraction. The GLCM based statistical feature extraction approach produces the result with Lower ASM, Lower Energy, Higher Entropy, Lower Homogeneity, Lower Correlation and Higher Dissimilarity for the urban slum than urban buildings. The very high-resolution satellite data of Madurai city, South India acquired by World View 2 Sensor (1.85 m) proved the ability to identify slums from buildings. Finally, the accuracy has been calculated for all images of blocks of size 9\*9, 7\*7 and 5\*5 that produces 54.45%, 60 % and 68.8% of producer's accuracy respectively. Hence, it has been found that GLCM features of 5\*5 blocks produces a better result than other blocks (7\*7, 9\*9).The future work is to enhance the performance of the slum extraction algorithms.

REFERENCES

[1] UN-HABITAT(2003). Slums of the world: The face of urban poverty in the new millennium? Working paper. Nairobi, Kenya.  
 [2] UN-HABITAT(2003). Slums of the world: The face of urban poverty in the new millennium? Working paper. Nairobi, Kenya.  
 [3] UN-HABITAT(2009). Urban indicators guidelines e Monitoring the habitat agenda and the millennium development goals Slums target. Nairobi, Kenya: UNHABITAT.  
 [4] UN-Habitat, The Challenge of Slums: Global Report on Human Settlements 2003. London and Sterling, VA: Earthscan Publications Ltd., 2003  
 [5] Baltsavias, E.P.,&Mason, S. (1997)"Image-based reconstruction of informal settlements". International workshop 5e9. May, Ascona, Switzerland (pp. 87-96)  
 [6] Niebergall, S., Loew, A., & Mauser, W. (2008). "Integrative assessment of informal settlements using VHR remote sensing data the Delhi case study". IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1(3), 193-205.  
 [7] Jordan Graesser, Anil Cheriyaat, RangaRaju Vatsavai,Varun Chandola, Jordan Long, and Eddie Bright, "Imagebased characterization of Formal and Informal Neighborhoods in an Urban Landscape", IEEE Journal of selected topics in Applied Earth Observations and Remote Sensing, 2012.

[8] Sulochana Shekhar, "Detecting Slums from Quick Bird Data in Pune Using an Object-Oriented Approach", Remote Sensing and Spatial Information Sciences, September 2012.  
 [9] Nursidik Heru Praptono, Pahala Sirait, M. Ivan Fanany, and Aniat Murni Arymurthy, "An Automatic Detection Method for High Density Slums based on Regularity Pattern of Housing using Gabor Filter and GINI Index", ICACIS, 2013.  
 [10] Ryan Engstrom, Avery Sandborn, Qin Yu, Jason Burgdorfer and Douglas Stow & John Weeks, "Mapping Slums Using Spatial Features in Accra, Ghana",2015.  
 [11] Monika Kuffer, Karin Pfeffer, Richard Sliuzas, and Isa Baud, "Extraction of Slum Areas from VHR Imagery Using GLCM Variance", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2016.  
 [12] P. Mohanaiah, P. Sathyanarayana, L. GuruKumar, "Image Texture Feature Extraction Using GLCM Approach", International Journal of Scientific and Research Publications, Volume 3, Issue 5, May 2013  
 [13] Nourhan Zayedand and Heba A. Elnemr," Statistical Analysis of Haralick texture features to discriminate lung abnormalities", International Journal of Biomedical Imaging,2015

# Optimized Feature Selection in Dimensionality Reduction for Big data with Random Forest Algorithm

Dr.B.Vinayagasundaram

Department of Information Technology  
Madras Institute of Technology,  
Anna University  
Chennai, Tamil Nadu  
bvsundaram@mitindia.edu

Swathy R

Computer Center  
Madras Institute of Technology,  
Anna University  
Chennai, Tamil Nadu  
swathy.balamurugan@mitindia.edu

C.Roshini, D.Sarshini, G.Senthilkumar

Department of Information Technology  
Madras Institute of Technology,  
Anna University  
Chennai, Tamil Nadu  
roshinivcp@gmail.com

**Abstract**—Emergence of Big Data with high voluminous and data complexity has resulted in processing difficulty to acquire beneficial information. There exist various challenges associated with Big Data like data transfer, querying the data, privacy of the data, variety, etc., One such challenge selected is high dimensionality of the data which is impossible to visualize. In this paper we have concentrated in reducing the dimension of data based on most important data features using Random Forest algorithm. Random Forest is an ensemble learning method based on decision trees. This paper presents a Random Forest based optimized feature selection for dimensionality reduction in big data. The algorithm considers most important features in reducing the data dimensions and also removes the redundant data. The present work uses Optimal Trade Off algorithm for optimal selection of features which increases accuracy. Accuracy of the model is further increased by weighted voting method by reducing regression error. The proposed work has helped in reducing the system complexity by optimizing the feature selection and reduces data complexity in terms of its dimension.

**Keywords**—Dimensionality Reduction, Big data, Random Forest Algorithm, Feature Selection, Optimal Trade off

## I. INTRODUCTION

Big Data is a large collection of massively generated data that is difficult to be processed by using traditional Relational Database Management Systems (RDBMS). Big Data has redundant and unimportant information present, which increases the complexity of handling and processing. It has been found that 90 percent of world's data has been generated only in the past two years. Hence we can say that big data is emerging at high rate and there is a need to reduce the complications implicated in big data.

For a day, roughly 2.5 quintillion bytes of data are being generated. The size of the data is in Petabytes, Terabytes, etc. Among this, there exists the useful data which need to be processed according to the requirements. The application value of this data is becoming more important over time [1].

The Big Data age comes not only with benefits but also with certain challenges. With increased demand in business for processing the data in real-time, data mining must be brought into consideration. The major problem accompanying

big data is, mining valuable information from the massive data in an optimal manner. Additional challenges such as redundancy, privacy, high dimensionality and complexity must also be taken into account when mining valuable information from Big Data. One such problem taken for consideration is dimensionality reduction of Big Data.

The technique chosen is Random forest, an ensemble learning method based on decision tree model. It is used for classification and regression. This random forest algorithm is a suitable for mining big data[1] by generating decision trees and the results are generated from these decision trees. This technique has been used in reducing dimensions of Big Data.

The paper is organized as follows: Section I presents about introduction about Big data, complexity in processing and need for dimensionality reduction. Section II presents the literature review about various existing methods of dimensionality reduction. Section III gives the system framework of the proposed work. Section IV discusses the experimental design and setup. Section V presents the results and discussion. Section VI presents the conclusions of the proposed work and future works.

## II. LITERATURE REVIEW

Some of the articles by illustrious scholars on dimensionality reduction and feature selection are studied and discussed below.

Min et.al [3] designed the system for reducing the dimensionality of high dimensional images. Dimensionality reduction is achieved by registering two high dimensional images into a single image. Principal Component Analysis (PCA) [4] algorithm is proposed to reduce the dimension of the pair of images prior to registration into a single image. Unlike other algorithms, PCA algorithm based dimensionality reduction works directly on the noisy data set. The dimensionally reduced pair of image consists of more information. The dual image is combined to a single band image. Pair wise registration is performed to obtain two single banded images. PCA algorithm obtains the smallest registration error. But the PCA algorithm does not aim at producing high registration accuracy.

Several classification algorithms were developed to classify data sets. All classification systems performs dimensionality reduction in order to reduce the complexity of the algorithm. One such classification algorithm is Support Vector Machine (SVM). The drawback of SVM [6] which classifies data based on just one direction is overcome by RSVM which classifies based on multiple orthogonal directions. This model is proposed by Tao et.al, [5] helps in dimensionality reduction performed based on multilevel maximum margin features which increases the accuracy of the classification algorithm.

Yen et.al [7] proposed a multiple kernel learning based dimensionality reduction (MKL-DR) approach to increase the effectiveness of dimensionality reduction. But working with low dimensional data helps to analyze the intrinsic structure of the data. The traditional algorithm suffers from the limitation that the data are in different forms e.g., bag of figures [8] which reduces the efficiency of the algorithm. Kernel machine is trained with multiple kernel functions. Recent researches [9] [10] in kernel functions has shown that the accuracy of classification of data increases when SVM is worked along with kernel functions. MKL-DR with graph embedding [11] is used for dimensionality reduction which is one of the important methods among all DR methods. MKL-DR with graph embedding can be applied for all three learning techniques i.e., supervised learning, unsupervised learning and semi supervised learning. Pyramid matching kernel [12] is a kernelization technique which helps the DR methods to work efficiently if the data are not in vector form. The main aspect of MKL-DR is generalization.

Dimensionality reduction algorithms are more prone to noise and artifacts. Hamid et.al [13] proposes an out of sample extension framework for a global manifold algorithm. Manifold learning methods can be used to embed the high dimensional data into low dimensional data hence, reduce the dimensions of high-dimension ambient space.

Hao et.al [14] proposes incremental high order singular value decomposition (IHOSVD) for decomposing high dimensional data. Traditional dimensional reduction algorithms are time consuming for high dimensional data. IHOSVD method is incremental high dimensional reduction algorithm for extracting core information from data set.

Zhang et.al [17] proposes a weighted sparse graph based dimensionality reduction (WSGDR) for hyper spectral images. WSGDR encourages sparse coding to be local. WSGDR mainly focuses on the training pixels. The representation accuracy and robustness can be improved by integrating the locality and scarcity properties in WSGDR.

One of the methods to classify streaming data is proposed by Martin et.al [18] which is streaming random forest algorithm. Multiclass classification problem is solved using streaming random forest algorithm. Streaming random forest is an incremental streaming classification algorithm that achieves high classification accuracy. Streaming decision tree is constructed as the first step of the algorithm. Initially, a tree consists of a single frontier node, which develops into a tree as data arrives. Entropy refers to randomness which is calculated for detecting the concept drift as the data arrives. The nodes of

the decision tree are split based on the best split value. The streaming decision trees are merged to form a streaming random forest which produces efficient and accurate algorithm for classification of stream data. The streaming random forest algorithm repeatedly extracts subset of the record and each one is built into a tree of the ensemble.

Oda et.al [15] proposes a method for monitoring satellite data which is of high dimensionality. To reduce the dimensions, an integrated model of mixture of probabilistic principal components analyzers (MPPCA) [16] and categorical distribution (MPPCAD) is proposed. This model is a special case of Gaussian Mixture Model (GMM). This model is the joint probability density distribution of the continuous and categorical measurement variables. MPPCA method is used for modeling real or continuous data while categorical distribution is for categorical data.

From the above literary survey, the dataset is dimensionally reduced based on the gain ratio computing and importance value. It is also evident that in Parallel Random Forest method the features are selected in a random manner. In our work we propose an Optimal Trade Off algorithm for feature selection. As a result of optimal Trade Off it is found that the accuracy of the Random Forest method in Dimensionality reduction has increased compared to random feature selection. For further improvement in accuracy, weighted Voting method is also employed.

### III. SYSTEM FRAMEWORK

The Input is provided as training and testing dataset in the ratio of 75:25 [19]. The dataset is first partitioned into number of subset in order to reduce data complexity. Partitioning is done based on vertical data partitioning algorithm which compares and maps the input feature variable with that of target feature variable. Next, the partitioned data is dimensionally reduced based on Random Forest algorithm with Optimal Trade Off. This process involves Entropy calculation for the feature variable in the subset and the target feature variable. Then the Gain ratio and Importance value of the feature variable is evaluated and the feature variables are sorted based on importance value. Finally, the Optimal Trade Off algorithm is applied which involves selecting lesser number of features than that of the existing random selection method. As a result of Optimal Trade Off the accuracy has improved when compared with existing random selection method. In order to further improve the accuracy weighted voting method is employed.

#### A. Vertical Data Partitioning

In dimensionality reduction algorithm, the process of computing the importance value using the gain ratio takes up the majority of training process. The gain ratio computing requires only the current feature variable and the target feature variable instead of the entire set of feature variables (dataset). Vertical data partitioning is employed as a pre-processing step on the dataset. The training and testing subsets are divided into feature subsets.

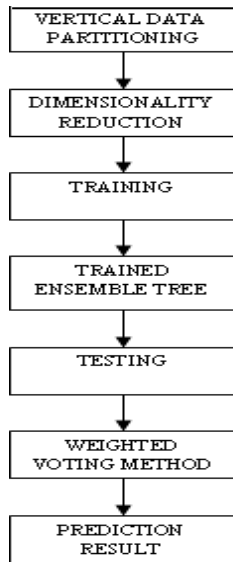


Fig. 1. System Framework

The training dataset has a size of  $N$  with  $M$  features variables, say  $y_0 - y_{M-1}$ . A feature subset has two attributes, a feature variable and target feature variable. The feature subset  $FS_j$  is represented by

$$FS_j = \begin{bmatrix} < 0, & y_{0j}, & y_{0(M-1)} > \\ < 1, & y_{1j}, & y_{1(M-1)} > \\ \dots & & \\ < (N-1), & y_{(N-1)j}, & y_{(N-1)(M-1)} > \end{bmatrix} \quad (1)$$

where  $j$  is the index of the records in the training dataset for each feature variable. The vertical data partitioning process [1] is represented in Fig.2.

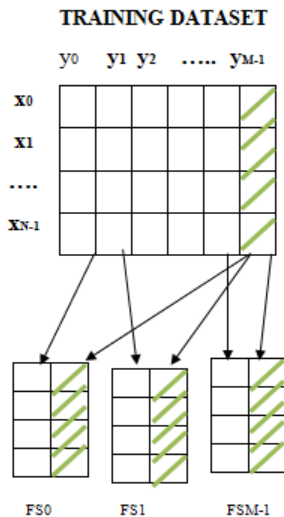


Fig. 2. Vertical Data Partitioning

The vertical data partitioning method is presented in Algorithm1

**Algorithm 1:** Vertical data partitioning

*Input:*

$S$ : Training dataset of size  $N$  and  $M$  feature variables say

$y_0 - y_{M-2}$ : input feature variable

$y_{M-1}$ : target feature variable

*Output:*

$FS_j$ : Feature subset

1. Identify feature variable  $y_j$  where  $j=0,1,\dots, M-1$ .
2. Map each input feature variable with target feature variable.
3. Generation of feature subset  $FS_j$

### B. Dimensionality Reduction

The dimensionality reduction method reduces the number of dimensions according to its importance. The training dataset of size  $S$  is assumed to have  $M$  feature variables. In the training process the gain ratio of the features variables are calculated and sorted in descending order. From this, the top  $k$  feature variables are selected. Here,  $k$  is much smaller than  $M$  ( $k \ll M$ ).

Then, the entropy of the target feature variable and each input feature variable is evaluated. The entropy of the target feature variable is defined as

$$Entropy(S_i) = \sum -p_a \log p_a \quad (2)$$

where  $x$  is the number of distinct values in the feature variable and  $p_a$  is the probability of the value in each of the feature variable.

The entropy of the input feature variable  $y_{ij}$  is defined as

$$Entropy(y_{ij}) = \sum Entropy(V_{ij}) \quad (3)$$

Where  $V_{ij}$  is a set of all values of  $y_{ij}$ .

The information gain (Gain) of each input feature variable is defined as

$$Gain(y_{ij}) = Entropy(S_i) - Entropy(y_{ij}) \quad (4)$$

The importance value (IV) of input feature variable is defined as

$$IV(y_{ij}) = G(y_{ij}) / \sum G(y_{ij}) \quad (5)$$

The importance values are sorted in descending order and the top  $k$  features are selected which will form a part of the new training dataset. The dimensionality reduction method is given in Algorithm 2.

### Algorithm 2: Dimensionality reduction algorithm

*Input:*

$S_i$ : The  $i^{\text{th}}$  training subset

*Output:*

$FS_j$ : Important feature variables

1. Calculate the entropy of the target feature variable,  $E(S_i)$
2. For every input feature variable  $y_{ij}$  in dataset  $s_i$ :



1. Calculate the entropy of the input feature variables,  $E(y_{ij})$ ;
2. Calculate the gain for the input feature variable,  $\text{Gain}(y_{ij})$
3. Calculate the self-split information  $I(y_{ij})$
4. Calculate gain ratio  $\text{GR}(y_{ij}) \leftarrow \text{Gain}(y_{ij})/I(y_{ij})$ ;
3. End for
4. Calculate the importance value of feature variable,  $\text{IV}(y_{ij})$ ;
5. Sort the  $\text{IV}(y_{ij})$  in descending order
6. Implement optimal trade off algorithm.
7. Return  $\text{FS}_j$

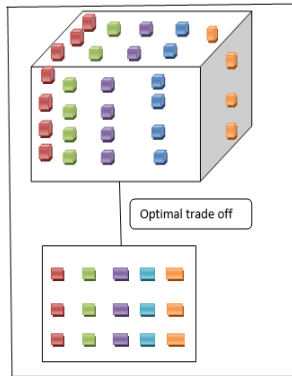


Fig. 3. Dimensionality Reduction

Fig. 3 provides a visual representation of dimensionally reduces dataset based on Random Forest with Optimal Trade Off.

### C. Optimal Trade Off

The optimal trade off algorithm involves selecting the most important features there by increasing the accuracy instead of selecting all the features.

The dimensionality reduction method and optimal trade off algorithm is represented in Fig. 4.

The optimal trade off algorithm improves the accuracy of the model there by reducing the dimensions as well. This algorithm involves adding a set of feature variables and training to form an ensemble forest. This trained random forest is tested for accuracy. The process continues until there is an increase in the accuracy or until the number of features is lesser than the original dimensions.

The optimal trade off method is presented in Algorithm3.

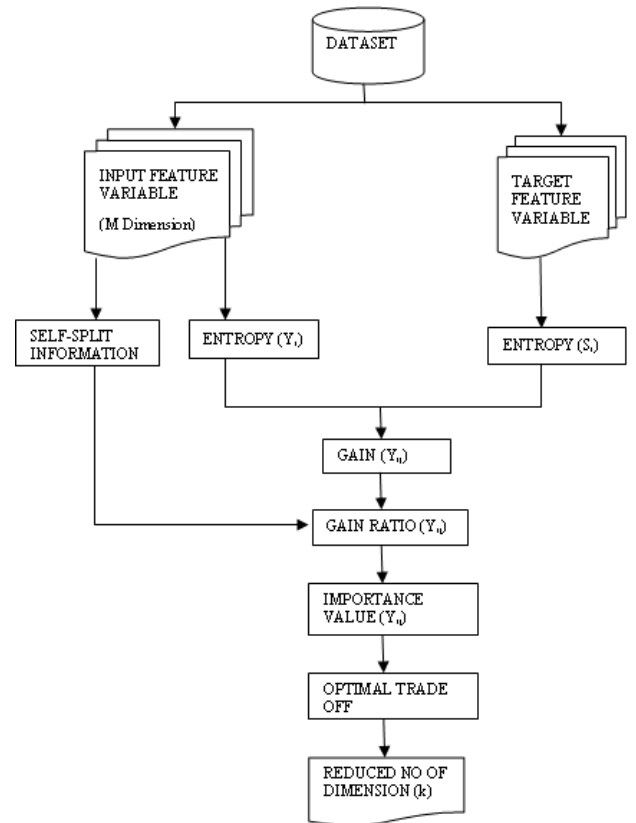


Fig. 4. Dimensionality Reduction and Optimal Trade Off

### Algorithm 3: Optimal trade off

$Ac_0$ : Accuracy of existing model  
 $NF_0$ : Number of features (reduced dimension) in existing model  
 $NF_n$ : Number of features in the proposed work  
 $Ac_n$ : Accuracy of proposed work

1. Sort the features in descending order of Importance value
2. Initialize  $NF_0$  to 0
3. Initialize  $x$  to 1
4. while  $x$ 
  - 4.1. if  $NF_n \leq NF_0 \& Ac_n \geq Ac_0$ 
    - 4.1.1.  $x=0$
  - 4.2. else if  $Ac_n > Ac_0$ 
    - 4.2.1. Increment  $NF_n$
    - 4.2.2.  $x=0$
  - 4.3. else
    - 4.3.1. Increment  $NF_n$
  - 4.4. End if
5. Store  $NF_n$  features in  $\text{FS}_j$
6. Return  $\text{FS}_j$

#### D. Weighted Voting

In the testing process, the results are based on direct voting method. However this will result in classification or regression error which will reduce the accuracy. This can be overcome by weighted voting method. The weighted voting method is devised to improve the classification or regression accuracy. The voting weight of a tree is regarded as the accuracy for classification or regression.

The classification accuracy (CA) is defined as the ratio of number of votes to the correct class to number of votes to both correct class and error class.

The weighted voting method is represented in Algorithm 4.

**Algorithm 4:** Weighted voting method

RF<sub>trained</sub>: Trained Random Forest after Dimensionality Reduction

TD: testing data set

O(TD): prediction result for the testing dataset TD

1. For each data d in TD
  - 1.1. For each tree  $T_i$  in RF<sub>trained</sub>
    - 1.1.1. Predict the result  $\alpha_i(d)$
  - 1.2. end For
2. end For
3. Calculate the classification or regression accuracy
4. For each data d in TD
5. if classification then
  - 5.1. final result,  $O_c(d) \leftarrow \text{Majority}[CA_i * \alpha_i(d)]$ ;
  - 5.2.  $O(TD) \leftarrow O_c(d)$ ;
6. else if regression then
  - 6.1. final result,  $O_r(d) \leftarrow \text{Average}[CA_i * \alpha_i(d)]$ ;
  - 6.2.  $O(TD) \leftarrow O_r(d)$ ;
7. end if
8. end For
9. return O(d)

#### IV. EXPERIMENTAL DESIGN AND SETUP

The work is done using Python 3.6. Anaconda framework supports Python 3.6 which provides an environment to work with Big Data. The dataset used is GISETTE. It is a handwritten digit recognition problem for identifying confusing digits like '4' and '9'. The digits are normalised and centred to disambiguate 4 and 9. The training and testing dataset are in the ratio of 75:25 [19]. Number of real features are 2500, Number of probes are 2500 and Total of 5000 attributes are considered for processing.

#### V. RESULTS AND DISCUSSION

The modules involved in the proposed system are vertical data-partitioning, dimensionality reduction, optimal trade off

and weighted voting method. The output is shown in terms of accuracy obtained after testing the dataset with random forest algorithm with optimal trade off and random forest with random selection method. The proposed work reduces the system complexity and data complexity there by increasing the accuracy and reducing the dimensions of big data. Thus Optimal Trade Off method selects the optimal feature which improves accuracy. The Pythagorean forest is used to visualize the trees that form the random forest for the dataset before and after reducing the dimensions of the big data.

Fig.5 below shows the accuracy in testing using random forest with optimal trade off of the model before and after dimensionality reduction and weighted voting. It is also evident that as the number of features increases; the accuracy of the model has increased.

```

number of features:
5000
ACCURACY :
0.954

percentage of features:
25.0
number of features:
1250
ACCURACY AFTER DIMENSIONALITY REDUCTION
0.943
ACCURACY AFTER WEIGHTED VOTING
0.94

percentage of features:
30.98
number of features:
1549
ACCURACY AFTER DIMENSIONALITY REDUCTION
0.954
ACCURACY AFTER WEIGHTED VOTING
0.96
    
```

Fig. 5. Accuracy of Model with Gisette dataset

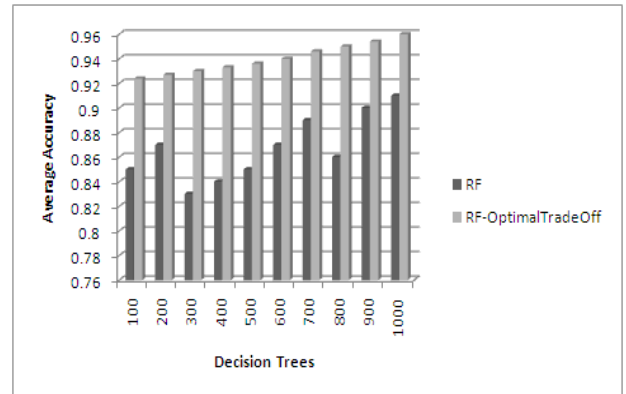


Fig. 6. Average Accuracy of RF vs RF-Optimal Trade Off

Fig.6 shows the average accuracy of the Random Forest model and Random Forest with Optimal Trade Off model with increasing number of decision trees. The results show that the average accuracy in RF with Optimal Trade Off has increased when compared with RF with random selection method.

Fig.7 and Fig.8 shows the pictorial representation of the Random Forest model with Optimal Trade Off using Pythagorean forest before and after dimensionality reduction.

It is evident that increase in the number of features, increases the size and complexity of the random forest.

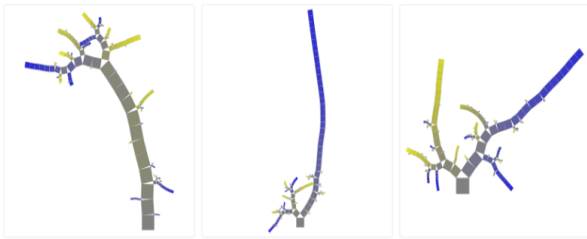


Fig. 7. Pictorial Representation. (Random Forest before Dimensionality Reduction)

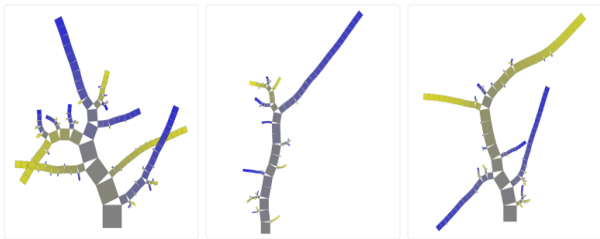


Fig. 8. Pictorial Representation. (Random Forest after Dimensionality Reduction)

## VI. CONCLUSION AND FUTURE WORK

The proposed work for Dimensionality reduction, Random Forest with Optimal Trade Off has been successfully implemented and has shown improved accuracy when compared with Random Forest with Random Selection. Thus, the system complexity is reduced by optimal feature selection process and dimension of the data is also reduced thus reducing data complexity. The accuracy is also further improvised by weighted voting technique. This will help the analyst to work on big data by overcoming the challenges imposed by it.

As a future work, proposed method will be trained and tested with various datasets. Optimization techniques like, task parallel optimization can so be implemented and parallelizing training process of the random forest may be done which will reduce the data transmission cost and the performance of the algorithm can be further improved.

## REFERENCES

[1] J. Chen, Z.Tang, K.Bilal, S.Yu, "A Parallel Random Forest Algorithm for Big Data in a Spark Cloud Computing Environment", IEEE Trans. Parallel And Distributed Systems, Vol. 28, No. 4, April 2017.  
 [2] X. Cai, J.Wei, G. Wen, Z.Yu, "Local and Global Preserving Semisupervised Dimensionality Reduction Based on Random Subspace

for Cancer Classification", IEEE Journal of Biomedical and Health Informatics vol: 18, March 2014 .  
 [3] M. Xu, H. Chen, and P. K. Varshney, "Dimensionality reduction for registration of high-dimensional data sets," IEEE Trans. Image Process., vol. 22, no. 8, pp. 3041–3049, Aug. 2013.  
 [4] C. E. Mandujano and S. Mitra, "Cross-power spectrum phase for automated registration of multi/hyperspectral data-cubes for efficient information retrieval," in Proc. IEEE Southwest Symp. Image Anal. Interpretat., Apr. 2002, pp. 111–115.  
 [5] Q. Tao, D. Chu, and J. Wang, "Recursive support vector machines for dimensionality reduction," IEEE Trans. Neural Netw., vol. 19, no. 1, pp. 189–193, Jan. 2008.  
 [6] N.Cristianini and J.Schawe-Taylor, An Introduction to Support Vector Machines. Cambridge, U.K.: Cambridge Univ. Press, 2000.  
 [7] Y. Lin, T. Liu, and C. Fuh, "Multiple kernel learning for dimensionality reduction," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 6, pp. 1147–1160, Jun. 2011.  
 [8] A. Berg, T. Berg, J. Malik, "Shape Matching and Object Recognition Using Low Distortion Correspondences", Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 26-33, 2005.  
 [9] F. Bach, G. Lanckriet, M. Jordan, "Multiple Kernel Learning Conic Duality and the SMO Algorithm", Proc. Int'l Conf. Machine Learning, 2004.  
 [10] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, M. Jordan, "Learning the Kernel Matrix with Semidefinite Programming", J. Machine Learning Research, vol. 5, pp. 27-72, 2004.  
 [11] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, S. Lin, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 40-51, Jan. 2007.  
 [12] K. Grauman, T. Darrell, "The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features", Proc. IEEE Int'l Conf. Computer Vision, pp. 1458-1465, 2005.  
 [13] Hamid Dadkhahi, F. Duarte, M. Marlin, "Out-of-Sample Extension for Dimensionality Reduction of Noisy Time Series" IEEE Trans. Image Processing vol.26, Issue: 11, Nov. 2017.  
 [14] L. Kuang, F. Hao, L. T. Yang, M. Lin, C. Luo, and G. Min, "A tensor-based approach for big data representation and dimensionality reduction," IEEE Trans. Emerg. Topics Computer., vol. 2, no. 3, pp. 280–291, Apr. 2014.  
 [15] T.Y airi, N. Takeishi, T. Oda, Y. Nakajima, N. Nishimura, N. Takata, "A data driven health monitoring method for Housekeeping data based on probabilistic clustering and dimensionality reduction" IEEE Trans. Aero. and Electronics systems, vol.53, no.3, June 2016.  
 [16] M. Tipping, C. Bishop, "Mixtures of probabilistic principal component analyzers", Neural Comput., vol. 11, pp. 443-482, 1999.  
 [17] W. W. He, H. Zhang, L. Zhang, W. Philips, W. Liao, "Weighted Sparse Graph Based Dimensionality Reduction for Hyperspectral images", IEEE Geoscience and remote sensing letters, vol. 13, no.5, May 2016.  
 [18] Hanady Abdulsalam, David B. Skillicorn, Patrick Martin, "Classification Using Streaming Random Forests", IEEE transactions on knowledge and data engineering, vol. 23, no. 1, January 2011.  
 [19] Jithendra Kumar Rout, Anmol Dalmia, Kim-Kwang Raymond Choo, Sambit Bakshi, Sanjay Kumar Jena, "Revisiting Semi-Supervised Learning for Online Deceptive Review Detection", IEEE Access, vol.5, January 2017.

# Analysis and Estimation of Brain Tissue Atrophy Using Magnetic Resonance Images

N. Ahana Priyanka

Department of Electronics Engineering  
MIT Campus, Anna University  
Chennai, Tamil Nadu  
India  
ahanachellian@gmail.com

G. Kavitha

Department of Electronics Engineering  
MIT Campus, Anna University  
Chennai, Tamil Nadu  
India  
kavithag\_mit@annauniv.edu

**Abstract**—Cognitive impairment results in decline of intellectual and social skills. Structural and texture changes in MR brain images determine the abnormalities in cognitive impaired (CI) subjects under different stages of progression. The abnormality in brain tissue exhibits variations in atrophy pattern among individuals. Hence, there is a need to study these atrophy variations to analyse the abnormalities in CI subjects. Segmentation and analysis of the atrophy in these structures can lead to accurate ways of CI detection. In this work, the MR brain images are skull stripped and analyzed using Robust Brain Extraction (ROBEX) Tool and Statistical parameter mapping (SPM 12). The images are obtained from Information extraction from Images (IXI) database and Minimal Interval Resonance Imaging in Alzheimer's disease (MIRIAD) database. The ROBEX tool achieve better skull stripped brain tissues compared to Brain Extraction Tool (BET) and Brain Surface Extractor (BSE). Further, an attempt has been made to examine volume variation of gray matter (GM), white matter (WM) and Cerebrospinal Fluid (CSF) in the skull stripped T1 weighted Magnetic Resonance (MR) brain image for normal and CI subjects, using SPM toolbox (CAT12). The WM and GM texture variation are observed with the help of Local Binary Pattern (LBP). The extracted features from normal and CI subject are analysed by Naive bayes and random forest classifiers. The experiment result shows that CI subjects have brain volume atrophy when compared to normal. In addition, statistical analysis determines the specific regions of atrophy in the brain. Naviebayes classifies normal and CI subject more accurately than random forest. Further, Naviebayes shows better variation in WM for CI subject than normal using LBP. Thus, this frame work could aid to study the interrelationship between structure and texture changes in brain tissue effectively.

**Keywords**—ROBEX (Robust Brain Extraction), CSF (cerebrospinal fluid), Local Binary Pattern (LBP) Segmentation, Atrophy, Cognitive impaired (CI)

## I. INTRODUCTION

Neurodegeneration is an irreversible brain disorder. It is characterized by a severe memory atrophy and deterioration of cognitive functions. It occurs due to the accumulation of protein beta-amyloid outside neurons and protein tau inside neurons in the brain. It has been estimated that more than 36 million people are affected by this disorder worldwide and is expected to double every 20 years. Neurodegenerative disorders such as Dementia, Alzheimer's disease (AD), vascular disease,

dementia with Lewy bodies (DLB), frontotemporal lobar degeneration (FTLD) and mild cognitive impairment (MCI) [1] affects the cognitive functions of brain. Such disorders lead to the reduction and damage of brain regions such as temporal lobes, parietal lobes, hippocampus, brain stem and amygdale. The diagnostic strategy for cognitive decline can be developed with the combination of clinical assessment and analysis of pathological changes [2]. The clinical syndrome is characterized by progressive global impairments in cognitive skills [3]. Many patients show varying levels of behavior disturbance at later stages of illness.

The diagnoses of cognitively impaired subjects is carried out by different criteria such as behavioral, psychiatric symptoms, cognitive test results, imaging data, laboratory findings and the subjects ability to manage daily activities. Current medication treatment seems to slow down the progression. The presence of cognitive impairment in functional, behavioral and quality of life is reported based on the clinical rating scales. There exist various assessment scales [4] to validate subjects cognitive skills such as Memory Impairment Screen and Mini-Mental State Examination. It assesses cognitive orientation, registration, recall, concentration, memory, attention, calculation, language and visual construction. These scales are used to diagnose the rate of cognitive decline and observe the change corresponding to the subjects.

A definitive diagnosis [5] requires histopathological confirmation from the brain images. Computed tomography (CT) is used to identify the treatable symptoms of cognitive decline. Disease-specific alterations in brain function can be assessed by functional MRI (fMRI). Structural gray matter (GM) and white matter (WM) properties are increasingly investigated by Diffusion Tensor Imaging (DTI). The functioning of various tissues and organ is provided by nuclear imaging technique such as positron emission tomography (PET) [6]. The blood flow inside the tissue and organ can be analysed by single photon emission computer tomography (SPECT). These methods are used to analyse neurobiological activities of brain and identify the early bio-markers for diagnosis of cognitive decline. Magnetic Resonance Images (MRI) examines information about variations in tissue pattern and can be utilized for predicting the abnormality [7]. It produces no radiation and excellent soft tissue contrast. It is

more suitable for classification of brain tissues (GM, WM and CSF) related to brain disorder.

Cognitive decline also causes volume differences in brain. The biological markers [8] are complicated and expensive to analyse. Tissue segmentation helps to understand the structure variation in brain [9]. Hippocampus and parahippocampal gyrus are major bio-markers [10] to find atrophy. Gray matter atrophy majorly occurs in the hippocampus, medial temporal lobes, left amygdale, left thalamus, left medial temporal pole and cornuammonis region of the left hippocampus [11]. Deep white matter and gray matter losses are present in the right laterobasal amygdala and fascia dentate region of the right hippocampus. Cortical atrophy was found in the anterior temporal, posterolateral temporal, and dorsolateral prefrontal regions of the left hemisphere. There exists correlation between rate of atrophy and change in neuropsychological test scores. In order to understand the atrophy and control in the disease progression an efficient framework is required.

In this work, the normal and CI subjects are skull stripped using ROBEX tool. The skull stripped results are compared with BSE and BET tools. The results are validated with similarity and performance measures. Then volume variation of brain tissues such as gray matter (GM), and white matter (WM) and cerebrospinal fluid (CSF) are analysed using SPM. These diagnostic variations are also studied by statistical analysis. Volume atrophy shows significant structural difference in brain regions and identifies the region of most declines. Further, tissue pattern changes are analysed using LBP features. These features are analyzed using Naviebayes and random forest classifier to study the differentiation between normal and CI subjects. Section 2 describes the methods and material used in this work. The significant results of this work are discussed in Section 3.

## II. METHODS AND MATERIALS

There are different types of tissues in brain and brain disorder causes change in them. Medical images show anatomical regions of interest and are generally characterized by numerical image features. There is a need to develop efficient image processing tools to assist clinical experts for better diagnosis. These include techniques for automated image segmentation, feature based classification or content based image retrieval. The decision of most appropriate feature for particular disorder is challenging. The proposed method visualizes the degree of atrophy and identifies the affected region. This analysis of atrophy variation results in improved understanding of neurodegenerative disorders and their interaction. Figure 1 shows the proposed flow diagram for the detection of atrophy in CI subjects.

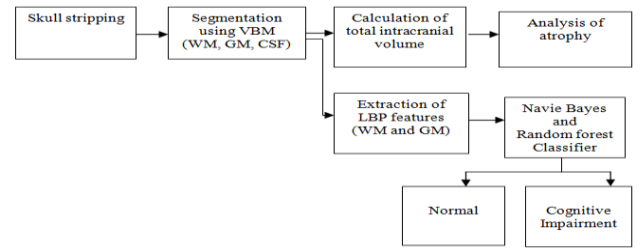


Fig. 1. Overall block diagram of the proposed method

The proposed method defines the changes in brain atrophy and its correlation with different brain tissues. The raw images are subjected to skull stripping. After delineation of skull, segmentation is carried out to determine the volume of different brain tissues. From the obtained volume of brain tissues, total intracranial volume is calculated. Further, statistical analysis is carried out to show the amount of atrophy occurrence in brain tissues. The regions of atrophy in the brain regions are highlighted. In order to analyse the tissue pattern changes LBP features are extracted from the segmented region. The LBP features are used to distinguish the CI subjects from the normal. Naviebayes and random forest classifier have been used for classification in this work.

### A. Database

The MR images are considered from the Information extraction from Images (IXI) and Minimal Interval Resonance Imaging in Alzheimer's disease (MIRIAD) database for this work. IXI dataset provides image metadata and demographic details of the images. The Minimal Interval Resonance Imaging in Alzheimer's disease (MIRIAD) dataset [12] is a series of longitudinal volumetric T1 MRI scans of 46 mild-moderate Alzheimer's subjects and 23 controls. It consists of 708 scans that include sequences at intervals of 2, 6, 14, 26, 38 and 52 weeks, 18 and 24 months from baseline. In addition it provides information on gender, age and Mini Mental State Examination (MMSE) scores.

### B. Skull Stripping

The space between the skull and brain tissue is generally term as subarachnoid space. This space is small which protect the brain from shock and damage. The preprocessing is difficult due to the closeness in pixel intensity. This requires preprocessing in the form of skull stripping. Skull stripping delineates the non brain tissue in MR brain images. Here ROBEX is used to eliminate the non brain region. ROBEX [13] is a learning-based brain extraction system. It does not require parameter tuning [14]. In this work, this tool is used to perform automatic skull stripping process and it is robust to intensity variations. Skull stripping is required to obtain non brain matters from brain tissues by delineation of skull. It preserves the prominent bio-markers such as cerebellum and brainstem that are highly influential in cognitive impairment. The default atlas and the properties of this tool support robust and accurate delineation of skull. The internal blocks used in a ROBEX tool to segment the whole brain are shown in Figure 2.

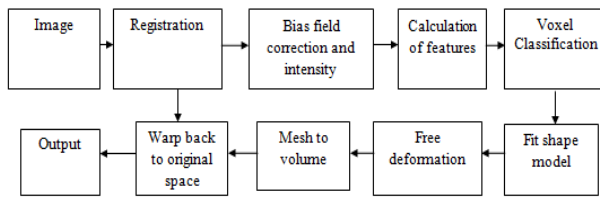


Fig. 2. Segmentation of Whole brain using ROBEX

Skull stripping is performed on the T1 MR image and the resulting mask is propagated to other channels. Initially, the input images are registered based on the template. The signal intensity is standardized and bias correction is performed. This tool is based on hybrid approach which combines generative and discriminative model. Landmarks of brain surface are identified using triangular mesh. Random forest classifier is used to discriminate the voxels that are located on the brain boundary using gaussian mixture and gradient features. The classifier is trained to find the contour of brain. Graph cut method refine the contour to extract the final brain tissue. Point distribution model and free deformation are used to balance the variation in brain shape. This prevents the extraction of required brain tissue from an image. Inverse affine transformation help to get the binary volume of brain which is warped back to the space of the original scan. The tool produces an output mask with improved intensity standardization, higher boundary probability volume and better final segmentation.

### C. Segmentation

Segmentation of different brain regions such as GM, WM and CSF is carried out by Statistical Parameter Mapping (SPM12) with CAT12 toolbox [15]. It performs tissue segmentation, bias correction, and spatial normalization all in the same model [16]. The assessment of brain anatomy includes volumetric and morphometric measurements. Xjview is used to view the resulted data. T1 MR brain images are normalized and segmented into GM, WM and CSF. This tool provides Total Intracranial Volume (TIV) that can be used to analyse the volume of GM, WM, and CSF.

Morphometry is performed on the basis of local deformations or voxelwise GM attenuation. Volume based morphometry (VBM) is an adaptation [17] of the SPM technique. It permits the investigators to quantitatively examine the whole-brain structural changes in a variety of conditions. VBM allows a voxel wise comparison of regional GM and WM concentrations in spatially normalized images.

Regional GM changes can be assessed by cortical thickness measurement. The prior probability that any voxel contains gray or white matter can be determined using a probabilistic atlas of tissue types. This prior probability is then combined with the data from the image to determine the tissue class. Using this approach, two voxels with identical intensities can

be identified. Intensities are modeled by a Gaussian Mixture Model[18].

These segmented GM and WM regions were then used to obtain a more accurate result using Diffeomorphic anatomical registration through exponential lie algebra (DARTEL) [19]. This model computes based on a group template and warps an individual's tissue. Hence this segmentation combines spatial normalization, bias field correction and tissue segmentation. The segmented brain tissues of normal and CI subjects are further analyzed and discussed in following section.

### III. RESULT AND DISCUSSION

Totally MRI T1 images of 20 subjects are considered and evaluated in all three views namely sagittal, coronal and axial. In this work, input images are in NIFTI file format. ROBEX is used to skull strip the images obtained from MIRIAD and IXI database. The extraction is carried out in all three views for 10 normal and 10 CI subjects from both the database. Voxels are detected based on brain boundary by means of random forest. The advantage in proposed brain tissue extraction method is that there is no need of specific tuning. Parameter tuning is frequently required in other methods such as BET [20] and BSE [21].

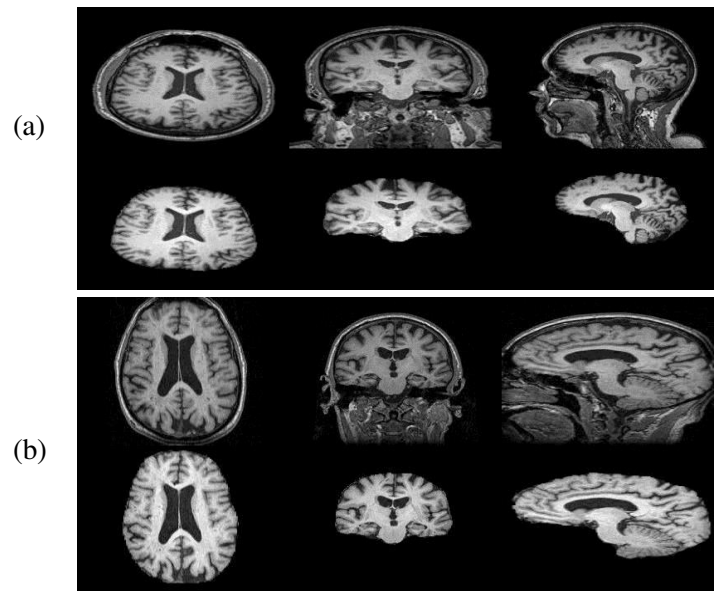


Fig. 3. Typical brain extraction results: (a) Normal and (b) Cognitive impairment for three different views (axial, coronal and sagittal)

The skull stripped results and the corresponding input images in all three views are shown in Figure 3. In Figure 3 (a) and (b) represents the skull stripping results of normal and cognitively impaired subjects. The presented results show the efficacy of the proposed skull stripping methodology using ROBEX.

The results obtained from automated skull stripping tool ROBEX is validated using different similarity and performance

measures such as Jaccard index, Dice coefficient, accuracy and sensitivity. The raw brain MR images in axial, sagittal and coronal view are skull stripped and the corresponding similarity and performance measures are shown in Table I. Detection success of delineation is calculated by comparing the skull stripped result with the manually obtained ground truth. The ground truth provided by experts is used to validate the obtained result. ROBEX produces a high value for all similarity and performance measures. It is able to provide a better delineation of brain tissue from skull region with the aid of a hybrid approach. The other methods such as BET and BSE provide low similarity measures due to similar tissue contrast, under segmentation and over smoothing.

TABLE I. NORMALIZED SIMILARITY AND PERFORMANCE MEASURE FOR DIFFERENT VIEWS USING ROBEX, BET AND BSE

Axial View				
Methods	Normalized values			
	Jaccard	Dice	Accuracy	Sensitivity
ROBEX	0.832	0.919	0.936	0.879
BET	0.510	0.792	0.801	0.493
BSE	0.458	0.618	0.757	0.443
Coronal View				
Methods	Normalized values			
	Jaccard	Dice	Accuracy	Sensitivity
ROBEX	0.835	0.935	0.964	0.930
BET	0.619	0.773	0.921	0.655
BSE	0.458	0.561	0.879	0.583
Sagittal View				
Methods	Normalized values			
	Jaccard	Dice	Accuracy	Sensitivity
ROBEX	0.849	0.921	0.942	0.923
BET	0.604	0.756	0.884	0.646
BSE	0.625	0.776	0.889	0.662

The Jaccard index, Dice coefficient, accuracy and sensitivity are significantly higher for ROBEX based skull stripped image. Sagittal view results in higher (0.849) Jaccard index than axial and coronal view. High Dice index, accuracy, and sensitivity are observed as 0.935, 0.964, 0.930 and 0.962 respectively for coronal view. This is because non brain regions such as spinal cord and mid neck region are not considered by ROBEX for segmentation.

#### A. Analysis of Total Volume Variation

The brain tissues are segmented from the skull stripped images. Prior to segmentation smoothing is carried out to remove the noise and strengthen the local features in the brain. Segmentation is carried out by DARTEL algorithm. The algorithm work based on registration, deformation and wrapping of image. Initially image is registered to generate both forward and backward deformation. The next step is to group the brain tissue based on tissue probability templates. The template is then regenerated by applying the inverse of deformations to the images and averaging. Finally the brain mask warps an individual's tissue. The volume of GM, WM and CSF are computed from the segmented results.

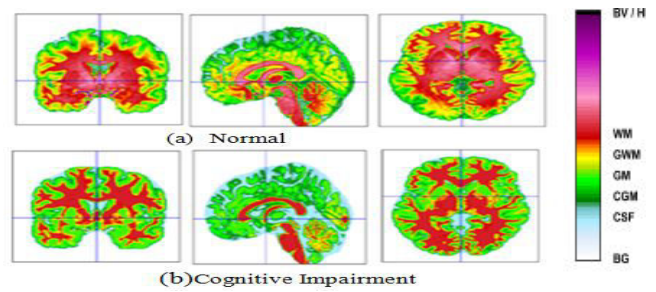


Fig. 4. Segmented white matter, gray matter and CSF for (a) Normal and (b) Cognitive Impairment images

The segmented tissue is mapped with template tissue patterns based on the tissue probability map. It is prescribed by International consortium for brain mapping (ICBM). Figure 4 shows a segmented normal and cognitive impairment brain tissue. Each colour in the colour bar corresponds to a specific brain tissue. The segmented regions are indicated by different intensity scale. Red colour represents the presence of white matter (GWM), yellowish red colour indicates gray white matter (GM), gray matter is viewed by light green colour, green colour points to the cortical gray matter (CGM) and cerebrospinal fluid is represented by blue colour. There is a change in the distribution of white matter and gray matter reduction in the distribution of normal and CI subjects in all the views. The result shows that atrophy of cognitive impairment has produced relative volume variations compared to normal subjects.

TABLE II. EVALUATION OF WHITE MATTER (WM), GRAYMATTER (GM) AND CSF FOR NORMAL SUBJECTS

S.No	Absolute Volume (cm)			Relative Volume (%)			TIV(M3)
	Tissues						
	CSF	GM	WM	CSF	GM	WM	
1	254	599	501	17.5	44.9	37.6	1334
2	364	560	483	26.2	39.6	34.2	1412
3	273	704	606	16.5	44.9	38.6	1569
4	325	679	533	19.7	45.0	35.3	1509
5	233	599	501	17.5	44.9	37.6	1334
6	264	546	461	21.3	42.6	36.1	1280
7	370	693	635	19.9	41.8	38.3	1658
8	258	704	567	16.6	46.2	37.2	1524
9	297	682	568	22.5	42.2	35.2	1614
10	264	767	587	16.3	47.4	36.3	1618

The evaluated WM, GM and CSF measures are shown for normal and cognitive impairment in Table II and Table III respectively. The absolute volume, relative volume and total intracranial volume (TIV) are calculated for normal and cognitive impairment. The absolute volume corresponds to contrast transfer function that is vital to obtain high resolution structure in an image, relative volume defines total occurrence of brain tissue and TIV represents combined volume of GM, WM and CSF.

TABLE III. EVALUATION OF WHITE MATTER (WM), GRAY MATTER (GM) AND CSF FOR NORMAL SUBJECTS

S.No	Absolute Volume (cm)			Relative Volume (%)			TIV (m3)
	Tissues						
	CSF	GM	WM	CSF	GM	WM	
1	431	644	556	26.4	39.4	34.0	1635
2	423	646	564	25.9	39.4	34.4	1637
3	362	482	406	28.9	38.5	32.4	1253
4	363	479	407	29.0	38.3	32.5	1252
5	362	480	406	29.0	38.4	32.4	1251
6	363	472	401	29.3	38.1	32.3	1239
7	366	466	401	29.6	37.7	32.5	1237
8	448	453	417	33.9	34.3	31.6	1321
9	448	458	415	33.8	34.6	31.4	1324
10	459	448	414	34.7	33.8	31.2	1328

The average absolute volume for CSF, GM, and WM are observed as 297cm, 657cm and 500 cm respectively for normal subjects. The cognitive impairment shows high absolute CSF volume (403 cm) and a lower absolute GM (503 cm) and WM (439 cm) volume. The tissue loss in cognitive impairment contributes to lower GM and WM volume. In the case of normal subject the relative volume was found to be 19.8%, 43.6% and 36.5%. However there is an increase in relative CSF volume (30.5%) and a reduction in relative GM (37.3%) and WM (32.4%) volume in cognitive impairment. Further the average TIV for normal and cognitive impairment is calculated as 1504 cm<sup>3</sup> and 1347 cm<sup>3</sup> respectively. The high occurrence of white matter, gray matter or CSF has resulted in significantly lower TIV in cognitive impairment.

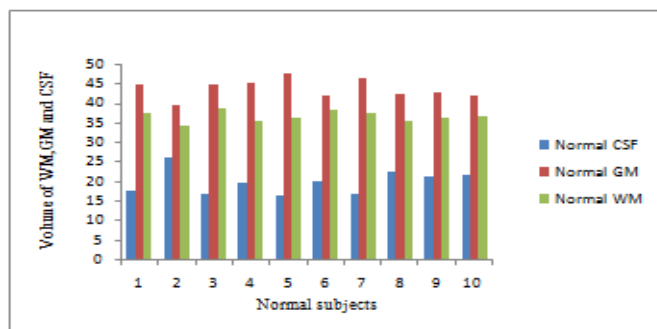


Fig. 5. Distribution of GM, WM and CSF in normal subjects

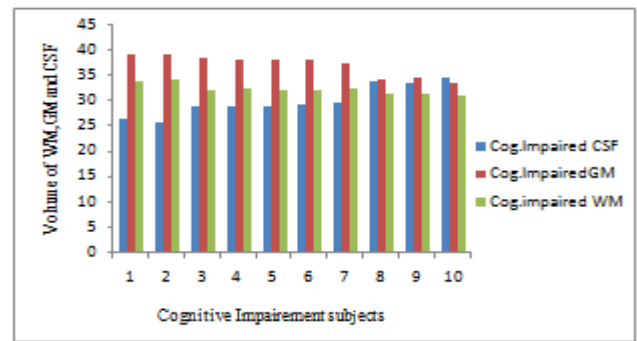


Fig. 6. Distribution of GM, WM and CSF in cognitive impaired subjects

Figure 5 and 6 illustrates the volume distribution of brain tissues for normal and CI subjects respectively. From the figure it is evident that in majority of cases there is a reduction in gray matter volume for cognitive impairment compared to normal. Further there is an increase in CSF volume for cognitive impairment compared to normal. There exists a similar level of mapping in brain tissue for normal and cognitive impairment (gray matter) in second subject. However the same level of volume does not mean the same distribution of tissue. The volume variation is due to high occurrence of any one of the brain tissue. In order to evaluate a significant difference in brain tissue, statistical analyses need to be observed. This is suitable for initial episode treatment and to provide the treatment during disease progress.

The average time taken to segment the WM and GM for normal was observed to be 12.88 mins. Further, cognitive impaired subject was found to be 14.90 mins for segmentation. The determination of atrophy alone is insufficient to diagnosis the severity of cognitive disorder. In order to identify the significant volume variation in the brain, statistical analysis is considered. The statistical analysis provides accurate degree of regional atrophy. The analyses are performed on both the data sets using significant two tail t-test. TIV is considered as a covariate input and false wise error is given at a rate of 0.05 for t-test. The false wise error defines the criteria for the uncorrected search volume in cluster level. The t-test produces significant variation with a statistical threshold of 0.05 with the p value  $p < 0.05$ . A lower p value indicates highly significant variation in the dataset. Thus the resulted p value indicates that there exists a significant difference between normal and cognitive impairment that assures guaranteed brain tissue atrophy in affected subjects.

Various clusters are obtained from t-test that determines the amount of localized atrophy. Figure 8 (a), (b) and (c) shows the alteration in brain functional connectivity based on two-sample t-tests for various p values. In the result of t-test each cluster is visualized by different intensities. The intensity scale ranges from positive to negative values. The range determines the activity of brain. In this work, activity is nothing but an occurrence of atrophy in the brain region. The positive value indicates the average activity of cluster. The negative value indicates the specific activity of cluster. The positive range is indicated by yellow color and negative range represent by red



color. Each cluster determines the volume variation. The clusters are formed based on Gaussian intensity function.

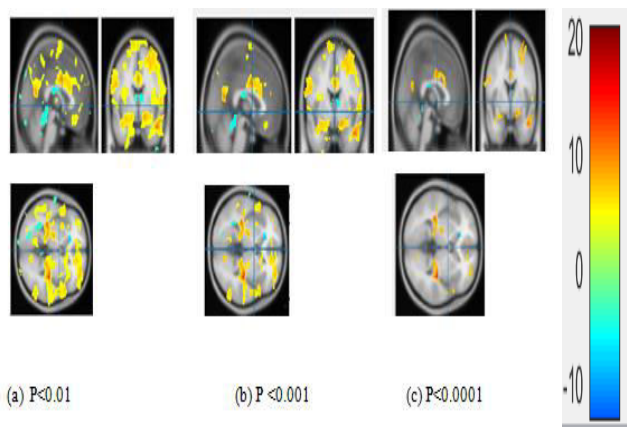


Fig. 7. Functional network connectivity alterations analysis based on two-sample t-tests for different p value. (a)  $p < 0.01$ , (b)  $p < 0.001$  and (c)  $p < 0.0001$

Totally 64 clusters are obtained for the considered dataset. These clusters determine the activation of voxel. The region of activation of particular interest is obtained corresponding to different p values. The p value of 0.1 shows formation of large number of clusters with lower intensity. It indicates regions of less reduction in brain tissue. The images obtained with  $p < 0.001$  shows average atrophy of brain tissue indicated by lower number of clusters. Finally  $p < 0.0001$  represent higher intensity regions of maximum brain atrophy occurred. The quantitative analysis defines the atrophy in specific brain region. Maximum atrophy has occurred in left Cerebrum, White Matter, Sub-Gyral, Temporal Lobe, Occipital Lobe, Middle Temporal Gyru, Superior Occipital Gyrus, Gray Matter, brodmann area, Parietal Lobe and Precuneus. The heavy subject atrophy variations in cerebrum and gyrus region are related to cognitive failure. The result states the atrophy occurs in volume irrespective of normal or cognitive impairment. The atrophy can be viewed heavily in cognitive areas. The system analyses the difference in structure and indicates the area in brain where major atrophy is observed.

### B. Analysis Of Atrophy in WM and GM

Texture analysis of GM and WM was carried out by Local Binary Pattern (LBP) based features. LBP observes the distinctive micro texture pattern change which is suitable for classification [22]. In LBP, the binary label of each pixel is obtained by comparing with the center pixel [23]. LBP operator assigns value to a pixel by comparing it with surrounding 8 pixels. LBP computes 256 bin histogram for each image to describe the texture. In uniform pattern, the length of the feature vector for a single cell reduces from 256 to 59 features. Non uniform patterns are assigned to a single bin in uniform pattern LBP. The goal of this feature is to identify prognostic information about WM and GM.

Random forest (RF) refers to an ensemble of trained decision trees used for classification problem. Each tree in the forest is independently trained with random samples, and it is combined together to construct a group of trees. These trees provide classification between the groups. Naïve bayes classifier [24] assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature. Comparison of normal and CI subjects is carried out by Naive bayes and random forest classifier. The normalized values of precision, recall and F-measure of WM and GM for normal and CI subject are shown in Table IV and Table V.

TABLE IV. NORMALIZED PERFORMANCE MEASURES OF NAIVE BAYES

Normalized Value				
Tissue	Subjects	Precision	Recall	F-measure
White Matter	Normal	1	0.94	0.947
	Cognitive Impaired	0.909	1	0.952
Gray Matter	Normal	0.833	0.5	0.625
	Cognitive Impaired	0.643	0.9	0.75

TABLE V. NORMALIZED PERFORMANCE MEASURES OF RANDOM FOREST

Normalized Value				
Tissue	Subjects	Precision	Recall	F-measure
White Matter	Normal	0.818	0.8	0.842
	Cognitive Impaired	0.889	0.9	0.852
Gray Matter	Normal	0.98	0.5	0.667
	Cognitive Impaired	0.667	1	0.8

The LBP features for an image are given as input to the classifier independently to study the performance measures. The classification of normal and cognitive impaired exhibit a reliable variation using naive bayes and random forest in terms of WM and GM. Results show that Naive bayes classifier performs well with LBP features. Naive bayes provide 95% accuracy in white matter and 75% accuracy in gray matter. Similarly random forest provides 85% accuracy in white matter and 70% accuracy in gray matter. The accuracy of random forest is relatively low when compared with Naive bayes. Relatively the atrophy differences are well diagnosed by Naive bayes using LBP features. WM precision, recall and F-measure for CI provide a better result than GM in Naive bayes. Various study states that changes in WM [25] contribute more to progression of neurodegenerative disorder. Thus, this framework could be used to find the atrophy in WM to diagnose the psychiatric diseases.

## IV. CONCLUSION

In this work, an attempt has been made to analyze brain tissue changes in normal and CI subjects from MR brain images. Initially, the MR brain images are skull stripped. The similarity and performance measures obtained using the skull stripping tool ROBEX is better compared to BET and BSE. The contribution of this work shows the structure and texture variation in MR brain tissue for normal and CI subjects. The structural changes have been observed in brain for normal and the CI subjects using SPM. The proposed method is suitable for finding the volume variation in brain tissue and to

determine the region of atrophy. This atrophy in brain causes the functional network connectivity alterations. It is observed that there exists reduction in volume of brain in cognitive subjects compared to normal. The observation shows that there is a decrease in volume of GM, WM and increase in volume of CSF for cognitive impairment. Computation time to segment the brain tissue is high for cognitive impaired subject than normal. The regions of atrophy are observed using clusters. Similarly, in NaviebayesWM texture variation for cognitive impairment is well diagnosed using LBP features. Hence Naviebayes performs better than random forest. These results show a lower total volume of brain for cognitive impairment than normal subjects. WM alteration is observed for cognitive impairment. Thus this framework can be used to better understand the tissue changes in normal and CI subjects effectively.

#### REFERENCES

- [1] ChuanchuanZheng, Yong Xia, Yongsheng Pan and Jinhu Chen, "Automated identification of dementia using medical imaging: a survey from a pattern classification perspective", *Brain Informatics* Vol.3, pp.17–2, 2016.
- [2] JonathanE.Peelle, Rhodri Cusack and Richard N.A. Henson, " Adjusting for global effects in voxel-based morphometry: Gray matter decline in normal aging," *NeuroImage*, Vol.60, pp.1503–1516, 2012.
- [3] ElisabetLondos, "A Diagnostic Model for Dementia in Clinical PracticeCase Methodology Assisting Dementia Diagnosis", *Diagnostics*, vol. 5, pp.113-118, 2015.
- [4] Bart Sheehan,"Assessment scales in dementia",*Therapeutic Advances in Neurological Disorders*, Vol.5(6), pp.349–358, 2012.
- [5] Wei Huang and Guang Chen,"A Novel Functional MRI-based Immersive Tool for Dementia Disease Severity Prediction via Spectral Clustering and Incremental Learning", *International Conference on Orange Technologies*, pp.169-172, 2015.
- [6] S.Mueller,D. Keeser,M.F. Reiser,S. Teipel and T. Meindl, " Functional and Structural MR Imaging inNeuropsychiatric Disorders, Part 1: ImagingTechniques and Their Application in MildCognitive Impairment and Alzheimer Disease", *American Journal of Neuroradiology*, Vol. 33(10), pp.1845–50, 2012.
- [7] Hardeepkaur and Jyoti Rani,"MRI brain image enhancement using Histogram equalization Techniques", *International conference on wireless communication, signal processing and Networking*,pp.770-773, 2016.
- [8] Jon Snaedal,GisliHolmar, Johannesson, Thorkell Eli, Gudmundsson, Nicolas, PeturBlin, AsdisLiljaEmilsdottir, Bjorn Einarsson and KristinnJohnsen," Diagnostic Accuracy of Statistical Pattern Recognition of Electroencephalogram Registration in Evaluation of Cognitive Impairment and Dementia", *Dement Geriatric Cognitive Disorder*, Vol.34, pp.51–60, 2012.
- [9] Tanabe JL, Amend D, Schuff N, DiSciafani V, Ezekiel F, Norman D, Fein G and Weiner MW, "Tissue segmentation of the brain in Alzheimer disease", *AJNR Am J Neuroradiology*,vol 18(1), pp. 115-23, 1997.
- [10] Michael P. Harms, Lei Wang, John G. Csernansky and Deanna M. Barch, "Structure–function relationship of working memory activity with hippocampal and prefrontal cortex volumes", *Brain Structure and Function*, Vol.218, pp.173–186, 2013.
- [11] C. Echa'varri , P. Aalten , H. B. M. Uylings ,H. I. L. Jacobs , P. J. Visser ,E. H. B. M. Gronenschild ,F. R. J. Verhey and S. Burgmans, " Atrophy in the parahippocampalgyrus as an early biomarker of Alzheimer's disease", *Brain Structure and Function*, Vol.215, pp.265–271, 2011.
- [12] Ian B. Malone, David Cash, Gerard R. Ridgway,David G. MacManus,SebastienOurselin, Nick C. Fox and Jonathan M. Schott , "MIRIAD—Public release of a multiple time point Alzheimer's MR imaging dataset", *NeuroImage*, Vol.70, pp.33-36, 2013.
- [13] Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M. Thompson, and ZhuowenTu," Robust Brain Extraction Across Datasets andComparison With Publicly Available Methods", *IEEE Transactions On Medical Imaging*, Vol. 30( 9), pp.1617 - 1634 , 2011.
- [14] Shaswati Roy and Pradipta Maji, "A simple skull stripping algorithm for brain MRI".*Advances in Pattern Recognition*, INSPEC Accession Number-14949082,2015.
- [15] Gunther Helms, "Segmentation of human brain using structural MRI", *Magnetic Resonance Material Physics* , Vol.29, pp.111–124, 2016.
- [16] John Ashbumer T and KarlJ.Friston, "Unified segmentation", *Neuroimage*, Vol.26, pp.839-851, 2005.
- [17] Tinu Varghese, R. SheelaKumari, P.S Mathuranath and Albert Singh, "Performance Comparison of Voxel based Morphometry and K means algorithm on longitudinal MR images", *The Institute of Science and Technology*, pp.1-4, 2012.
- [18] On Tsang, Ali Gholipour, Nasser Kehtarnavaz, Kaundinya Gopinath, Richard Briggs, and Issa Panahi, "Comparison of Tissue Segmentation Algorithms in Neuroimage Analysis Software Tools", *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*,pp.3924 – 3928, 2008.
- [19] Colloby SJ, Firbank MJ, Vasudev A, Parry SW, Thomas AJ and O'Brien JT ,"Cortical thickness and VBM-DARTEL in late-life depression", *Journal of Affective Disorders*, Vol133, pp.158-64, 2011.
- [20] D. W. Shattuck and R. M. Leahy, "BrainSuite: An Automated Cortical Surface Identification Tool," *Medical Image Analysis*, vol. 6, pp.129–142, 2002.
- [21] S. M. Smith, "Fast Robust Automated Brain Extraction," *Human Brain Mapping*, vol.17, pp.143–155, 2002.
- [22] Min-Joo Kang, Jung-Kyung Lee, Je-Won Kang, "Combining random forest with multi-block local binary pattern feature selection for multiclass head pose estimation", *PLoS ONE* 12(7): e0180792,2017.
- [23] Taha H. Rassem, AbdulRahman A. Alsewari and Nasrin M. Makbol, "Texture Image Classification using Wavelet Completed Local Binary Pattern Descriptor", *The Seventh International Conference on Innovative Computing Technology*, pp.15-20, 2017.
- [24] W. R. Shankle, S. Mani, M.B. Dick, M.J. Pazzani, "Simple Models for Estimating Dementia Severity Using Machine Learning", *NCBI, Studies in Health Technology and Informatics*, vol1, pp.472-476,1998.
- [25] Christopher M. Filley,"White matter dementia", *Therapeutic Advances in Neurological Disorders*, vol5(5) , pp.267–277, 2012.

# A Multifamily Android Malware Detection Using Deep Autoencoder Based Feature Extraction

Teenu S. John, Tony Thomas, Md. Meraj Uddin

*Indian Institute of Information Technology and Management-Kerala*

Thiruvananthapuram, Kerala, India teenu.john@iitmk.ac.in, tony.thomas@iitmk.ac.in, meraju@iitmk.ac.in

**Abstract**—With the advent of mobile devices and the open nature of the operating system, Android has gained popularity over other mobile platforms in recent years. Studies have shown that Dalvik opcode based static malware detection is highly efficient and accurate in Android mobile platforms. This is because of its limited resource consumption and its ability to detect obfuscated malware variants. The  $n$ -grams constructed with large values provide more accuracy. But when the value of  $n$  in an opcode ngram increases, the dimension of the data also increases and the data becomes more sparse which is called the curse of dimensionality problem. The existing feature extraction methods are not at all sufficient to model the underlying complexities of the data and does not take into account the highly informative opcode  $n$ -grams. So the feature extraction should reduce the dimensionality and preserve the most significant opcode  $n$ -grams for classification. The proposed method employs deep learning using deep autoencoders to extract the highly significant opcode  $n$ -grams for malware detection. Deep autoencoders using deep learning can learn the complexities and nonlinearities of the dataset, thus reducing the computational overhead caused by the curse of dimensionality problem of  $n$ -grams. The results show that feature extraction using deep autoencoders outperform the conventional feature extraction techniques like information gain and TF-IDF. The proposed technique can be used to detect multifamily Android malware variants with 98% accuracy.

**Keywords**—Android, malware detection, deep autoencoder, RBM.

## I. INTRODUCTION

Now a days Android operating system is gaining popularity because of its ease of use and ease of application development. Because of its open nature, malware authors are also finding their way to enter the market, exploiting various vulnerabilities in the platform. The applications available in official Android application store are verified by a security scanner called Bouncer [1] before making it available to the users. Recent studies show that malware authors can adopt complex techniques to bypass its detection capability [2]. The conventional signature-based system used by antivirus scanners only detects malwares whose signature are predefined. The drawback of the signature-based detection system is that it should predefine all the signatures of the known malwares beforehand. If an unknown variant comes, then the system fails to detect it.

Security researchers have developed a variety of techniques to detect malwares that evade antivirus scanners. The existing techniques can be classified into static and dynamic analysis methods. Static method inspects the code without executing the application while dynamic analysis monitors the application in

runtime. The disadvantage of dynamic detection is high resource consumption and processing overhead. In addition to that, malware may exhibit different behaviours in the sandbox and real environments and can result in incorrect behaviour logs. Moreover, malware can also delay its malicious behaviour execution to evade dynamic detection. The advantage of the static method is that they are fast and easy to implement and consumes fewer resources. Most of the static method detection techniques use permissions for detection [3] [4]. The drawback of permission-based malware detection is that it only examines the manifest file for malware detection. The permission-based detection system can be easily evaded by obfuscation techniques. Obfuscation can be done by repacking application components, disassembling and reassembling, changing package names and encryption. The developers use a variety of tools to prevent malware authors from reverse engineering their application. The popular tools are Dexguard, Proguard, ApkProtectetc [5] [6]. Dexguard encrypts the strings while Proguard renames the classes, variables and methods using random strings. Unfortunately, these tools are widely used by malware authors than the application developers for obfuscation.

Since static analysis using permissions can be easily evaded by obfuscation technique, opcodes that contain useful information about the program behaviour can be used effectively to detect malware [7]. Malware authors usually modify their behaviour by changing the existing malicious code or by merging the different parts of the malicious code to produce a new malware. This means that the same family of Android malware shares similar opcode patterns. Since opcodes are closely related to Android application codes, a feature vector using opcodes prove to be efficient in detecting various malware families. There are several works that show the effectiveness of opcodes for malware detection [8] [9].

Since the same Android malware family shares similar opcode sequences, opcode  $n$ -grams provide a useful measure to detect mobile malware. The  $n$ -grams constructed with large values provide more accuracy. But when the value of  $n$  increases, the dimension of the data also increases and the data becomes more sparse. This is called the curse of dimensionality problem [10]. So the objective should be to extract highly relevant opcode  $n$ -grams for malware classification, without making  $n$  large. We refer to the highly relevant opcode  $n$ grams as the feature of the Android application.

Inorder to mitigate the drawbacks of the feature extraction employed in the previous works, the method proposed here adopts deep autoencoders using RBM for feature extraction that learns the most informative opcode  $n$ -gram from unlabelled data. Feature extraction using deep autoencoders

provide excellent results in speech recognition and image processing. The advantage of deep autoencoders using deep learning is that it can learn the implicit representation of data more precisely. Deep learning can perform unsupervised learning using many hidden layers that learn the complex structure of data. The deep autoencoders using RBM (Restricted Boltzmann Machine) can thus extract highly informative opcode  $n$ -grams for malware detection with a reduced feature set.

Each family of malware shares some common behaviour like the type of attack, malicious payload installation mode and the activation strategy adopted by them. The samples taken are botnets and trojans that perform malicious activity. Most of the malware samples taken are obfuscated samples that evade detection. The proposed method is evaluated using various feature extraction techniques employed in the literature to show the effectiveness of the method.

This work aims to detect Android malware using opcode  $n$ -grams with reduced feature sets. The proposed method is able to classify malicious applications without expert intervention and hand engineered features. The existing opcode based malware detection in Android requires high dimensional feature vectors for training the classifier. When the feature vectors are excessive, a classifier may encounter extra overhead of processing irrelevant high dimensional opcode  $n$ -grams. The method employed here reduces the computational overhead by extracting only informative opcode sequences thereby improving the accuracy of the classifier.

## II. RELATED WORKS

There are several static analysis techniques for detecting Android malwares. Most of them use permissions [11] [12] [13] for detecting malicious applications. The permission based malware detection system is easily prone to obfuscation techniques and proves to be less efficient in detecting various obfuscated malwares. Arp et al. [14] used static detection based on API calls, network addresses, permissions and other features for malware detection using SVM. The detection of malware becomes complex when the number of features to be analysed becomes large. Some of the technique uses either static taint analysis [15] or similarity of files [16] for detecting anomalous application.

In contrast with the complex features for malware detection, Dalvik opcodes can be used to detect malware from a benign application. There are several works that prove the effectiveness of opcodes for malware detection [17] [18]. Most of the opcode based malware detection technique takes either opcode histograms [19] or frequencies of opcodes [7]. In opcode based malware detection, dimensionality reduction is very important since there are thousands of opcode  $n$ -grams. The existing opcode based detection technique uses information gain [20] or relative difference [21], PCA [22], TF-IDF [23] etc. to extract most relevant opcodes. These techniques are not at all sufficient to extract the relevant opcodes because of high feature permutations of opcode  $n$ -grams. The information gain ignores the feature dependencies and selecting the threshold value for determining the most informative  $n$ -grams are not properly chosen in most of the cases. The disadvantage of TFIDF is that it assumes that the

counts of different  $n$ -grams provide independent evidence of similarity. Hence a more accurate efficient feature extraction technique that extracts the most informative opcode  $n$ -grams for malware detection is required.

As the number of  $n$ -gram opcode increases, the feature extraction should extract only the most significant opcodes and eliminate the less significant ones. Yerima et al. [17] use neural networks using CNN for dimensionality reduction. However, their method requires opcodes to be projected to an embedding space. Moreover training a CNN for classification task is much hard.

## III. PROPOSED METHOD

The Android application developers can misuse the coding idioms by either copying vulnerable code or reusing the sample code obtained from the benign applications. Android games and applications can also be repackaged with modified code to evade copyright protection. Despite the obfuscation adopted by the malware authors, the raw Dalvik opcodes of the application can be used to detect vulnerabilities, since they are closely related to the application codes. So a feature vector using Dalvik opcodes can be used effectively to detect vulnerabilities and obfuscations in an application software.

The method proposed in this paper detects multi-family Android malware using opcode  $n$ -grams. The malware families are grouped based on the type of attack, type of malicious payload installation, and the mode of activation used. The  $n$ -gram opcodes of malicious and benign applications are constructed with  $n = 3$ .

For feature selection, deep autoencoders using RBM is proposed in this paper that learns the most informative features from unlabeled data. The problem with  $n$ -grams is that it suffers from the curse of dimensionality problem. As the number of feature vectors increases, the computational overhead associated with the processing of  $n$ -gram opcodes also increases. The advantage of deep autoencoders is that it takes the non-linearities and complexities of the data and produces a low dimensional compressed representation of the input. The compressed features contain the most informative features of the data that can be given to the classifiers for malware detection. The effectiveness of feature extraction using deep autoencoders is compared with conventional feature extraction techniques like information gain, TF-IDF and the results are evaluated.

### A. Pre-processing

An Android apk file consists of manifest, resource and Dalvik executable files. The opcodes are obtained by disassembling .dex files using blacksmali that generate a set of smali files. Each smali files has classes and methods used by the apk file. The methods contain opcodes and multiple operands. From each smali files, the operands are eliminated and the unique opcodes are then extracted. The opcodes are then constructed as  $n$ -grams for classification.

### B. Deep Learning

Unlike other traditional neural networks, deep learning uses many hidden layers that capture the implicit representation of the data. This is because it can model the complexities of the

data accurately unlike other shallow learning architectures. Deep learning using Deep Belief Nets (DBN) exhibit promising results in image processing and speech recognition [24]. DBN are composed of many layers of RBM and it is trained using unlabelled data. Since DBN can learn from unlabelled data, it can be used for classification tasks when the number of labelled data is less. They are generative models that provide a joint probability distribution over observed data, unlike discriminative models that learn the data with conditional probabilities for classification. Other than classification tasks, DBN can also be used as deep autoencoders that learn the most informative features from unlabeled data.

### C. RBM (Restricted Boltzmann Machines)

It is a stochastic energy based model that associates an energy function to each variable so that learning can be done by modifying it [25]. RBM consists of two layers: a hidden and a visible layer. By adding more hidden layers we can increase the efficiency of RBM. The units in different layers are fully connected between each layer, with no connection between the neurons in the same layer.

Let  $X$  be the opcode  $n$ -grams given as an input to the visible layer. Each hidden layers computes the best features that describe the characteristics of malicious and benign opcode  $n$ -grams. Each input vector is given as binary features for learning. There is an energy associated with hidden and visible layers [26].

The energy between visible and hidden layers is given by,

$$E(v, h; \theta) = -v^T W h - b^T v - a^T h \quad (1)$$

$$= \sum_{j=1}^D \sum_{k=1}^F W_{j,k} v_j h_k - \sum_{j=1}^D b_j v_j - \sum_{k=1}^F a_k h_k \quad (2)$$

where  $b, a$  are the bias vectors of visible and hidden layers,  $W$  is the weight between the visible and hidden layers,  $v, h$  are the input of the visible and hidden layers and  $\theta = \{W, a, b\}$  is the model parameter. The hidden bias vectors helps to produce the activations or outputs in the forward pass and visible bias vectors helps to produce activations in the back propagation pass.

The high energy signifies low probability of occurrence of a particular  $n$ -gram. Therefore if bias vector is set as negative, then the corresponding hidden as well as visible vectors should be set as 0 and if the bias vector is set as positive, the visible and hidden vectors are set as 1.

The joint probability distribution describes the occurrence of input vector given the hidden vectors and the occurrence of hidden vectors given the input vectors. In this case, it is the occurrence of a particular  $n$ -gram given its hidden layer configurations and vice versa. It is given by

$$P(v, h, \theta) = \frac{1}{Z(\theta)} \exp(-E(v, h; \theta)), \text{ where} \quad (3)$$

$$Z(\theta) = \sum_v \sum_h \exp(-E(v, h; \theta)). \quad (4)$$

$Z(\theta)$  is called the partition function. Computing the joint probability distribution is intractable because of the normalizing constant  $Z(\theta)$  which requires the exponential sum of the energies.

The probability of a visible vector  $v$  is defined as the probability of observing a data  $x = v$ , given the weights  $W$ . It is obtained by marginalizing the hidden vectors and is given by

$$P(v; \theta) = \frac{1}{Z(\theta)} \sum_h \exp(-E(v, h; \theta)) \quad (5)$$

In RBM the conditional probability of  $h$  given  $v$  and  $v$  given  $h$  is easy to compute. The conditional probability of  $h$  given  $v$  is the product of the conditional distribution of each individual unit of  $h$  given the value of  $v$ . It is derived from Equation 3 and is given by,

$$p(h | v; \theta) = \prod_j P(h_k | v). \quad (6)$$

Similarly the conditional distribution of  $v$  given the value of  $h$  is given by

$$p(v | h; \theta) = \prod_i P(v_j | h). \quad (7)$$

The conditional distribution of  $h_k$  given  $v$  is given by,

$$p(h_k = 1 | v) = g\left(\sum_i W_{j,k} v_j + a_k\right). \quad (8)$$

Where  $g(x) = \frac{1}{(1 + \exp(-x))}$  is the logistic function.

$$\text{Similarly } p(v_j = 1 | h) = g\left(\sum_k W_{j,k} h_k + b_j\right) \quad (9)$$

The training phase consists of three phases:- forward propagation, back propagation and weight updations. In the forward propagation phase, the probabilities of hidden units are taken and then a positive gradient is computed. In back propagation phase, the visible units are reconstructed from the hidden units and the negative gradient is computed and the weight matrix is updated by calculating the difference between positive and negative gradients.

The derivative is used to find the optimal configuration to the model. It helps to decide whether the model needs to increase or decrease the weights so as to maximize the correct learning procedure of opcode  $n$ -grams. Let  $E_{P_{data}[.]}$  be the data dependent expectation that depends on the input value of  $n$ -gram,  $E_{P_{model}[.]}$  be the models expectation and  $V_n$  be the values of the visible vectors. The derivative of the log likelihood with respect to the weights is obtained from Equation 5 and is given by

$$\frac{1}{N} \sum_{n=1}^N \frac{\partial \log P(V_n; \theta)}{\partial W_{j,k}} = E_{P_{data}}[v_j, h_k] - E_{P_{model}}[v_j, h_k] \quad (10)$$

Equation 10 describes the change of visible vectors with respect to the weights.

$$\frac{1}{N} \sum_{n=1}^N \frac{\partial \log P(V_n; \theta)}{\partial a_k} = E_{P_{data}}[h_k] - E_{P_{model}}[h_k] \quad (11)$$

is the change of visible vector with respect to the hidden bias and

$$\frac{1}{N} \sum_{n=1}^N \frac{\partial \log P(V_n; \theta)}{\partial b_j} = E_{P_{data}}[v_j] - E_{P_{model}}[v_j] \quad (12)$$

is the change of visible vectors with respect to the visible bias vector. The computation of  $E_{P_{model}[.]}$  takes exponential time because this computation is intractable as explained before in the same section.

Hence learning is done by contrastive divergence [27],

$$\Delta W = \alpha (E_{P_{data}}[vh^T] - E_{P_T}[vh^T]), \quad (13)$$

where  $\Delta W$  is the weight updation,  $\alpha$  is the learning rate and  $P_T$  represents a distribution defined by running Gibb's chain.

#### D. Deep autoencoders for Feature Selection

The pretraining of RBM is a greedy layer wise procedure that trains one layer at a time using unsupervised data. The deep autoencoder are built by stacking trained RBMs. The deep autoencoder consists of encoding units and decoding units. The encoding units produce the data in the encoded form with reduced features and the decoding units decodes the data back to the original input. This means that the central layer of RBM contains the best feature that describes the malicious and benign opcode  $n$ -grams.

It extracts the higher level dependencies between input vectors so as to capture the non-linearities of the data. The lower layers encode the input while the upper layer decodes it. Each layer refines the previously learned features. After pre-training, deep autoencoders use error back propagation method for fine tuning. The output of the central layer of the deep autoencoder is the new set of opcode  $n$ -gram features that can be given as an input to the classifier. The dimension of top hidden layers is less than input dimension so as to obtain the reduced feature set. Deep autoencoder fine tuning is done with a learning rate of  $10^{-5}$  for 6 iterations. The output of the hidden layers is set as 150,100,50 units. Fig 1 and 2 shows the deep autoencoder training steps and the architecture of deep autoencoders. Figure 2 shows the architecture of deep autoencoder. The dimension of the hidden layer is set as 200,300 and 50. The raw opcode  $n$ -grams are given as inputs to the visible layer. The dimension of the central hidden layer is less than the outer hidden layers so as to obtain a reduced feature set. The encoding layer produces a low dimension of the input vector while the decoding layer gradually reconstructs the input. The pretraining phase sets all the weights so as to obtain a more accurate feature set. The output of the central layer is then given as an input to the classifier.

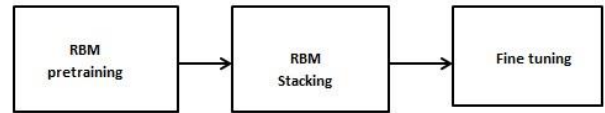


Fig. 1. AE Training Steps

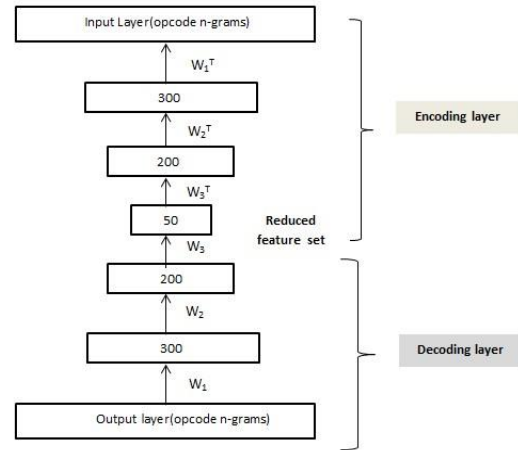


Fig. 2. Architecture of AE

#### IV. ANDROID MALWARE FAMILIES

Each family of Android malware shares some common behaviour like the type of attack, malicious payload installation mode and the activation strategy. Most of them are botnets and trojans that perform malicious activity. The description of different malware families is given below.

Table 1 shows the malware families, activation mode and type of obfuscation used to evade detection.

**DroidKungfu:** DroidKungfu repackages with benign applications and it adds a new service and a new receiver. The receiver will be notified that the system has finished booting in order to launch the attacks. It also uses string encryption, and native payload exploits and kills antivirus scanners. It uses two root exploits called Rage Against Cage and Exploit that are encrypted.

TABLE I ANDROID MALWARE FAMILIES

Android Malware Family	Activation mode	Obfuscation	Type
Droiddream	MAIN	String renaming	Botnet
Opfake	Repackaging	Polymorphic algorithm	Trojan
Gingermaster	Boot,MAIN, PCKG	Stringrenaming, encryption	Trojan
Airpush	Scheduling	String renaming	Adware
Geinimi	Boot,sms	Bytecode obfuscation	Trojan
Bankbot	Restart after killing	Renaming,string encryption	Banking Trojan
Dowgin	Net,pkg, sys	Encrypt network flow	Adware
Aples	Boot, MAIN	Lock the device	Ransomware
Droidkungfu	Boot, battery,sys	String and native payload encryption	Trojan

**Opfake:** This family uses server side polymorphism to evade detection by antivirus scanners. Each time a different version of the file is downloaded. Polymorphism is done by file reordering, variable data changes and insertion of dummy files. Each polymorphic variants share the same opcode despite the polymorphic behaviour.

**GingerMaster:** This malware contains a root exploit called GingerBreak. This root exploit is packed into the application as a png file. After obtaining root privileges, the malware can make the device as a botnet. The malware can send device information to a remote server without the user’s knowledge.

**Airpush:** This family of Android malware is an adware that uses aggressive advertisement and uses renaming to evade detection. The malware belonging to this family sends the device information such as IMEI numbers, phone model, kernel version, network operator information to a remote server. The activation is done by the host application.

**Geinimi:** This family of malware has botnet like capabilities and can perform bytecode obfuscation. Geinimi can send SMS and personal information to a remote server and can download files without the user’s knowledge.

**Bankbot:** This is a malware that is a banking Trojan aimed to steal credit card information. It appears as an Adobe application. Bankbot asks to provide confidential information such as credit card details and personal information to verify Google Accounts. The collected information is transmitted to a remote server. The malware monitors its detection and restarts its activity, even after it is removed.

**Dowgin:** It is an adware that is bundled with Android applications. It can evade dynamic analysis by encrypting the network flow. It also downloads and requests installation of new applications.

**Aples:** This is a family of ransomware that demands payment for unlocking the phone. The application pretends like an antivirus scanner and locks the device. If the system is rebooted, a message by FBI is seen demanding payment and the device will be locked if the victim fails to pay the ransom.

**DroidDream:** It is a botnet that takes root privileges and downloads malicious programs without the user’s knowledge.

## V. EXPERIMENTS AND RESULTS

The dataset is built with 1500 benign and 1500 malicious Android applications. The trusted applications are collected from Google Play [28] and the malicious applications are taken from the AMD [29] dataset. A python script is used to extract unique opcodes from the dataset that eliminates operands and unwanted character strings from the smali files.

The opcode  $n$ -grams are constructed with  $n = 3$ . More than 10000 distinct opcode  $n$ -grams were obtained after the preprocessing step and the opcode  $n$ -grams with highest information gain with 300,200 and 150 features were taken for evaluating the model.

A Random Forest classifier is used for detecting malware families with the information gain feature selection method. The experiments are also evaluated with SVM and Decision

Tree but gave less detection rate. The number of hidden units in the output is set as 50,100 and 150 for deep autoencoder.

The number of hidden units are chosen based on the highest true positives obtained after the experiments. An unlabeled data set is prepared for training an RBM.

The RBM’s are initialized with random weights and trained for 70 iterations. The learning rate  $\alpha$  is set to be 0.001. The encoded data samples are taken from the deep autoencoder and are given to the Random forest for classification. The training of RBM is performed on an AMD Radeon™HD 7970 GPU.

Table II shows the top five opcode 3-grams extracted by autoencoder.

TABLE II TOP FIVE OPCODE 3-GRAMS EXTRACTED BY AUTOENCODER FOR DOWGIN AND BANKBOT FAMILIES

Dowgin	BankBot
add-int/lit aget-object array-length	return-void add-int/lit aget-char
if-eq if-eqz if-ge	nop rem-int/lit return
iget object iput object move-result-object	add int/lit aget-char aget-object
return-object return-void sget-object	array-length check-cast const/1
move-exception move-object move-result	goto if-ge if-gt

Table III and IV shows the detection rate using information gain and deep autoencoder feature selection.

For information gain feature selection, the opcode  $n$ -grams of malicious applications and benign applications are taken and the entropy is calculated.

The entropy of a variable  $X$  is given by

$$H(X) = -\sum_i P(x_i) \log_2(P(x_i)) \quad (14)$$

where  $P(x_i)$  is the probabilities for all values of  $X$ . The entropy of  $X$  given the values of  $Y$  is defined by

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j)) \quad (15)$$

$y_j$  where  $y_j$  is the probabilities for all values of  $Y$ . Information gain of  $X$  given  $Y$  is defined by

$$I(X|Y) = H(X) - H(X|Y) \quad (16)$$

Let  $Y$  and  $Z$  be any two opcode  $n$ -grams of an application. Then  $Y$  is considered as a more informative opcode  $n$ -gram that indicates the behaviour of the application than  $Z$  if

$$I(X|Y) > I(X|Z) \quad (17)$$

The highly informative opcode  $n$ -grams are chosen as feature vectors for evaluating the classifier with dimensions 200,150 and 300. The results are shown in Table III.

TABLE III MALWARE DETECTION WITH INFORMATION GAIN FEATURE SELECTION

Android Malware Family	Features	TPR	FPR	Precision	Accuracy	Recall	F-measure
Droid dream	200	0.979	0.012	0.987	96.37	.983	.985
Opfake	150	0.959	0.068	0.934	94.65	.983	.958
Ginger Master	300	0.963	0.015	0.984	96.08	.996	.990
Airpush	150	0.954	0.014	0.986	94.44	.989	.987
Geinimi	200	0.953	0.024	0.976	93.87	.982	.979
Bankbot	200	0.980	0.016	0.984	96.98	.989	.987
Dowgin	100	0.956	0.013	0.986	94.43	.987	.987
Aples	150	0.977	0.014	0.985	96.10	.983	.984
Droid-kungfu	200	0.954	0.079	0.923	94.17	.981	.952
Avg accuracy %					95.0%		

Bankbot Families

The autoencoders are implemented with 50, 150 and 100 units for the central layer. The architecture and the training of autoencoder is described in Section III.

The classification accuracy is determined by the number of True positives(TPR), False positives(FPR), Precision and Fmeasure.

*True Positive Rate(TPR)*: True Positive Rate is the number of malicious applications classified as malicious.

$$TPR = \frac{TP}{TP + FN} \quad (18)$$

*False Positive Rate(FPR)*: False Positive Rate is the number of benign application classified as malicious.

Android Malware Family	Features	TPR	FPR	Precision	Accuracy	Recall	F-measure
Droid dream	100	0.994	0.011	0.990	97.73	.983	.986
Opfake	100	0.999	0.014	0.986	98.25	.983	.985
Ginger Master	150	0.990	0.013	0.987	98.79	.998	.992
Airpush	100	0.990	0.013	0.988	97.95	.989	.988
Geinimi	100	0.999	0.015	0.986	98.22	.983	.984
Bankbot	150	0.990	0.011	0.989	97.95	.989	.989
Dowgin	50	0.988	0.001	0.999	97.53	.987	.993
Aples	100	0.998	0.013	0.987	98.12	.983	.985
Droid-kungfu	150	0.984	0.046	0.956	98.47	1.00	.977
Avg accuracy %					98.10%		

TABLE IV MALWARE DETECTION WITH DEEP AUTOENCODER FEATURE SELECTION

$$FPR = \frac{FP}{FP + TN} \quad (19)$$

*False Negative Rate(FNR)*: False Negative Rate is the number of malicious applications classified as trusted.

*Precision*: Precision is the number of True Positives over the number of True Positives and number of False positives.

$$Precision = \frac{TP}{TP + FP} \quad (20)$$

*Recall*: Recall is the number of true positives over the number of true positives and the number of false negatives.

$$Recall = \frac{TP}{TP + FN} \quad (21)$$

*F - measure* : It is the harmonic mean of precision and recall.

$$F - measure = 2 \cdot \frac{Precision}{Precision + Recall} \quad (22)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (23)$$

The results in Table III and Table IV show that compared with information gain, the proposed deep autoencoder based feature selection gives better results with low dimensional feature vectors. It is seen that by using deep autoencoder feature selection the average accuracy of malware detection is about 3% higher than that of the information gain feature selection. This is obtained by the encoding layers of the autoencoder that outputs a low dimensional feature vectors of the corresponding opcode  $n$ -grams. The objective of the decoding layer is to reconstruct the raw opcode  $n$ -grams from the reduced feature vectors. After fine-tuning, the central layer of the autoencoder outputs the reduced, highly informative opcode  $n$ -grams that can be given as an input to the classifier. The efficiency of the proposed method is also compared with TF-IDF feature selection method on two malware families Opfake and Geinimi and the results can be seen in Table IV and V. For each opcode  $n$ -grams the normalized term frequency(TF) and TF inverse document frequency (TF-IDF) is computed. The normalized TF is calculated by dividing the frequency of  $n$ -gram in an application by the frequency of the most frequently occurred  $n$ -gram in the application. TF-IDF combines the term frequency (frequency of opcode  $n$ -gram in an application) and the frequency of opcode  $n$ -gram in the particular Android family (document frequency). The top 250 opcodes that have high TF-IDF is taken for evaluating the performance. It is seen that even with more number of features, the classification accuracy is less. This is because of the poor feature extraction method that selects the  $n$ -gram opcodes for classification.

This is because deep learning can model the implicit data representations more precisely. The advantage of deep autoencoder is that it can work on unlabelled data. As the



number of opcode  $n$ -gram increases deep autoencoder helps to extract highly informative  $n$ -grams thereby reducing the dimensionality and improving the accuracy of the classifier.

TABLE V FEATURE EXTRACTION USING TF-IDF ON MALWARE FAMILIES

Android Malware Family	No of Features	TP	FP	Precision	Accuracy	Recall	F-Measure
<b>Opfake</b>	250	0.963	0.015	.984	.974	.963	.973
<b>Geinimi</b>	200	0.953	0.023	.975	.965	.953	.964

## VI. CONCLUSION AND FUTURE WORK

This paper proposed a multifamily Android malware detection using opcode  $n$ -grams with deep autoencoder feature extraction method. Since opcodes are related to the application code, it can be used to effectively detect malware that uses obfuscation to evade detection. As the number of  $n$ -gram opcode increases, it suffers from curse of dimensionality problem. So it is necessary to reduce the dimension by effectively extracting opcodes that contain relevant information. The method proposed here utilizes deep learning to effectively learn the complexity and nonlinearity of the data. Deep autoencoder using deep learning is used for feature extraction that reduces the dimensionality of  $n$ -gram opcodes thus improving the classification accuracy. The results are compared with the traditional feature extraction techniques like information gain, and TF-IDF to prove the effectiveness of the method. The results show that deep autoencoder outperform traditional feature extraction techniques thereby improving the accuracy of classifiers.

The future work aims to discover robust features for malware detection by denoising autoencoders that helps to reconstruct the input after adding some random noise. Data poisoning attacks are popular in malware detection classifiers using deep learning where the adversary manipulates the data instances to misclassify the samples. Future work also aims to adopt a defensive strategy against an attacker that uses various data augmentation techniques to evade detection.

## REFERENCES

[1] "A look at google bouncer:trendlabs security intelligence," <http://blog.trendmicro.com/trendlabs-security-intelligence/a-look-at-google-bouncer/>, accessed:2017-10-21.

[2] "How rtf malware evades static signature-based detection," <http://www.fireeye.com/blog/threat-research/2016/05/how-rtf-malware-evad.html>, accessed:2017-10-21.

[3] N. Peiravian and X. Zhu, "Machine learning for android malware detection using permission and api calls," in *Tools with Artificial Intelligence (ICTAI), 2013 IEEE 25th International Conference on*. IEEE, 2013, pp. 300–305.

[4] U. Pehlivan, N. Baltaci, C. Acarturk, and N. Baykal, "The analysis of feature selection methods and classification algorithms in permission based android malware detection," in *Computational Intelligence in Cyber Security (CICS), 2014 IEEE Symposium on*. IEEE, 2014, pp. 1–8.

[5] "The most advanced mobile app security software for android," <https://www.guardsquare.com/en/dexguard>, accessed:2017-5-19.

[6] "Obfuscation in android malware, and how to fight back," <https://www.virusbulletin.com/virusbulletin/2014/07/obfuscation-android-malware-and-how-fight-back>, accessed:2017-3-6.

[7] A. Shabtai, R. Moskovitch, C. Feher, S. Dolev, and Y. Elovici, "Detecting unknown malicious code by applying classification techniques on opcode patterns," *Security Informatics*, vol. 1, no. 1, p. 1, 2012.

[8] H. Divandari, B. Pechaz, and M. V. Jahan, "Malware detection using markov blanket based on opcode sequences," in *Technology, Communication and Knowledge (ICTCK), 2015 International Congress on*. IEEE, 2015, pp. 564–569.

[9] J. G. de la Puerta, B. Sanz, I. Santos, and P. G. Bringas, "Using dalvik opcodes for malware detection on android," in *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 2015, pp. 416–426.

[10] "Curse of dimensionality," [https://en.wikipedia.org/wiki/Curse\\_of\\_dimensionality](https://en.wikipedia.org/wiki/Curse_of_dimensionality), accessed:2017-10-21.

[11] P. P. Chan and W.-K. Song, "Static detection of android malware by using permissions and api calls," in *Machine Learning and Cybernetics (ICMLC), 2014 International Conference on*, vol. 1. IEEE, 2014, pp. 82–87.

[12] K. A. Talha, D. I. Alper, and C. Aydin, "Apk auditor: Permission-based android malware detection system," *Digital Investigation*, vol. 13, pp. 1–14, 2015.

[13] X. Liu and J. Liu, "A two-layered permission-based android malware detection scheme," in *Mobile cloud computing, services, and engineering (mobilecloud), 2014 2nd IEEE International Conference on*. IEEE, 2014, pp. 142–148.

[14] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. Siemens, "Drebin: Effective and explainable detection of android malware in your pocket." in *NDSS*, 2014.

[15] W. Huang, Y. Dong, A. Milanova, and J. Dolby, "Scalable and precise taint analysis for android," in *Proceedings of the 2015 International Symposium on Software Testing and Analysis*. ACM, 2015, pp. 106–117.

[16] J.-w. Jang, J. Yun, A. Mohaisen, J. Woo, and H. K. Kim, "Detecting and classifying method based on similarity matching of android malware behavior with profile," *SpringerPlus*, vol. 5, no. 1, p. 273, 2016.

[17] N. McLaughlin, J. Martinez del Rincon, B. Kang, S. Yerima, P. Miller, S. Sezer, Y. Safaei, E. Trickett, Z. Zhao, A. Doupe' et al., "Deep android malware detection," in *Proceedings of the Seventh ACM Conference on Data and Application Security and Privacy*. ACM, 2017, pp. 301–308.

[18] I. Santos, F. Brezo, X. Ugarte-Pedrero, and P. G. Bringas, "Opcode sequences as representation of executables for data-mining-based unknown malware detection," *Information Sciences*, vol. 231, pp. 64–82, 2013.

[19] B. B. Rad and M. Masrom, "Metamorphic virus variants classification using opcode frequency histogram," *arXiv preprint arXiv:1104.3228*, 2011.

[20] Q. Jerome, K. Allix, R. State, and T. Engel, "Using opcode-sequences to detect malicious android applications," in *Communications (ICC), 2014 IEEE International Conference on*. IEEE, 2014, pp. 914–919.

[21] G. Canfora, A. De Lorenzo, E. Medvet, F. Mercaldo, and C. A. Visaggio, "Effectiveness of opcode ngrams for detection of multi family android malware," in *Availability, Reliability and Security (ARES), 2015 10th International Conference on*. IEEE, 2015, pp. 333–340.

[22] P. Okane, S. Sezer, and K. McLaughlin, "Detecting obfuscated malware using reduced opcode set and optimised runtime trace," *Security Informatics*, vol. 5, no. 1, p. 2, 2016.

[23] A. Shabtai, R. Moskovitch, Y. Elovici, and C. Glezer, "Detection of malicious code by applying machine learning classifiers on static features: A state-of-the-art survey," *information security technical report*, vol. 14, no. 1, pp. 16–29, 2009.

[24] C. Huang, W. Gong, W. Fu, and D. Feng, "A research of speech emotion recognition based on deep belief network and svm," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[25] "Restricted boltzmann machines (rbm)," <http://deeplearning.net/tutorial/rbm.html#id1>, accessed:2017-3-9.

[26] R. Salakhutdinov, "Learning deep generative models," *Annual Review of Statistics and Its Application*, vol. 2, pp. 361–385, 2015.

- [27] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.
- [28] "<https://play.google.com/store?hl=en>, accessed:2017-10-21.
- [29] <http://amd.arguslab.org/>, accessed:2017-3-9.

# Anomaly Based Intrusion Detection System using Classifiers

NarasimaMallikarjunan K

Assistant Professor

Department of Computer Science and Engineering  
Thiagarajar College of Engineering  
Madurai, Tamilnadu  
arjunkambaraj@tce.edu

Aarthi M

PG Scholar

Department of Computer Science and Engineering  
Thiagarajar College of Engineering  
Madurai, Tamilnadu  
aarthisree002@gmail.com

**Abstract**— Nowadays analyzing unsuspecting network traffic has become an important research topic to protect organizations from intruders. It is a big challenging to accurately found out threats due to the high volume of network traffic. In the existing system, to detect whether network traffic is normal or abnormal we need lots of information about the network. When lot of information is involved in the identification process the relationship between different attributes and the important attributes consider for classification plays an important role in the accuracy. Information gain selection process is used to provide a rank for features. Based on the rank, the most contributed features in the network is found and used to improve the detection rate based on the features selection. In this work, we have analyzed performance of two classifiers namely, Bayesian and Lazy classifiers. In Bayesian classifier has two algorithms namely, BayesNet, and Naïve Bayes. In lazy classifier has two algorithms namely, IBK and Kstar. Analyzing performances of Bayesian and lazy classifiers by applying various performance factors. From the performance analysis, lazy classifier is more efficient than Bayesian classifier.

**Keywords**— Information gain selection, Kstar, Bayesian, BayesNet, Classification, IBK, Naïve Bayes, Lazy.

## I. INTRODUCTION

Attacker activities are still now rapidly increase using an intrusion detection system detects the intruder on the network. It can be classified into Signature Based Intrusion Detection System, Anomaly Based Intrusion Detection, Wireless Intrusion Detection System (WIDS), Host Based Intrusion Detection System (HIDS), Network Intrusion Detection System (NIDS), and Network Behavior analysis (NBA).NIDS monitoring all devices incoming and outgoing traffic in the network. It is used to monitor all the passing traffic on the subnet and passing traffic is compare to the library of known attacks. Based on known attacks identified behavior of network and inform to the administrator for protecting network. HIDS work on the networks but separately run on host or device. It analyzing both incoming and outgoingpackets from the device or in suspicious traffic is detected send alert message to the administrator. It is collecting a snapshot of an existing system file and check for the pervious system file snapshots. If system files are altered or deleted send notification message to the administrator. NBAdetermines network traffic to identify attacks that generate unusual traffic flows, such as DDos attack and some formof malware, virus etc. WIPS is used to analysis a wireless

network for suspicious traffic with help of wireless network protocol. A Signature based Intrusion Detection Systems references a database of previous attack signatures and known system vulnerabilities. Intrusion Detection Systems is stored evidence of an intrusion or attack. Each intrusion leaves signatures behind for examples failed attempt to run an application, nature of data packets, login failures, accessing for file and folder etc. These signatures are called footprints and it can be used to identified and prevent the same attacks in the future. Based on these detection system identify intrusion attempts. The disadvantages of this detection Systems are signature database must be frequently updated and maintained. This Detection Systems may be failed to identify a unique attack. An Anomaly Based Intrusion Detection Systems [7] references a baseline or learned pattern of normal system activity to identify active intrusion attempts. Nowadays, many computing infrastructures find it very difficult to prevent unauthorized access and attacks. We have to transfer packets from source to destination as data sharing is one of the key functionality of networking .while transmitting the data in lot of attacks are available to modified the original data and transfer altered data to destination. In some cases attacker send irrelevant request to system which may overwhelm the system and damage the system behavior. To find the behavior of system is complex and gathered lot of information about the network traffic. Information based on network traffic can be classified into three types 1) some features contribute to predict the behavior of network 2) some features do not give any impact 3) some feature gives confusion for predicting the behavior of thenetwork. So collect the contributed feature of network traffic to easily find out the behavior of the network.

## II. WEKA

WEKA is a machine learning tool, thereby forming a basis of data mining. It was developed in 1997 by professionals of University of Waikato, New Zealand. Such accumulation of data mining algorithms was issued under the General Public License. WEKA [1] executes calculations for information preprocessing, data preprocessing, association rules, visualization, regression, clustering. For such calculations, it comprises of 49 data preprocessing, 3 association rules and nearby 76 order classifications. The entire calculation is performed with the utilization of a Graphical User Interface (GUI) known by explorer that helps in investigating situations from information contained in the dataset. It consists of modules like Knowledge Flow, which is a java based interface

and is useful for running machine learning tests, experimenter which is used for making analysis and organizing diverse systems for testing. In addition, various dataset formats like .csv, .arff, .data etc are supported by WEKA in order to extract the relevant data from the crude data.

### III. LITERATURE REVIEW

SY Ji, BK Jeong, et al., [1] it is comprised of three steps as (1) understanding hidden patterns from network traffic data by making reliable rules to identify network abnormal behavior, (2) creating a predictive model to determine attack categories, and (3) integrating visualization analytic tools to conduct an interactive visual analysis and examine the identified intrusions. In this paper, analysis performance of three algorithms [10] Naive Bayes, Support Vector Machine (SVM) and Neural Network (NN). The result is Support Vector Machine (SVM) accurately detect the behavior of the network compare to remaining Classifier.

Kaushik H. Raviya et al., [2] it is analysis performances on three classification techniques which are Decision tree, K-nearest neighbor and Bayesian network respectively. The aim of this work is to identify the best technique [9] from the above three techniques. There is creating tree model for a direct relationship between execution time and the volume of data records and also there is creating the model for an indirect relationship between execution time and attribute size of the data sets.

Vaithyanathan V. et al., [3] This work has been carried out to make a performance evaluation of Multilayer Perceptron, J48, Naive Bayes Updatable, and BayesNet classification algorithm. J48 algorithm is based on decision tree. Classification and Naive Bayes algorithm is based on probability is used to classify each item in a set of data into one of predefined set of groups or classes. The paper creating comparative evaluation of classifiers BayesNet, J48, Multilayer Perceptron, and Naive Bayes Updatable in the Weather datasets, context of Labour and Soyabean. The results in this paper demonstrate that the efficiency of Naive Bayes and J48 is good.

Anand Kishor Pandey et al., [4] WEKA is an open framework programming tool consisting of various inbuilt classification algorithms like J48, Random Forest, Decision Tree, Random Tree, Naive Bayes, Simple Naive Bayes, Naive Bayes, Decision Stump, etc. However, the comparison of these algorithms has been made with the help of approaches that include correctly classified, incorrectly classified, Accuracy and many others parameters.

Nahla Ben Amor et al., [5] We concentrate three levels of attack granularities depending on whether dealing with entire attacks, or comprising them in four main categories or just focusing on normal and abnormal behaviors. In the entire experimentations, we compare the performance of decision tree with one of well-known machine learning techniques which is Naive Bayes network. Moreover, we compare the good performance of Bayesnet [8] with respect to existing best results performed on KDD'99.

### IV. METHODOLOGY

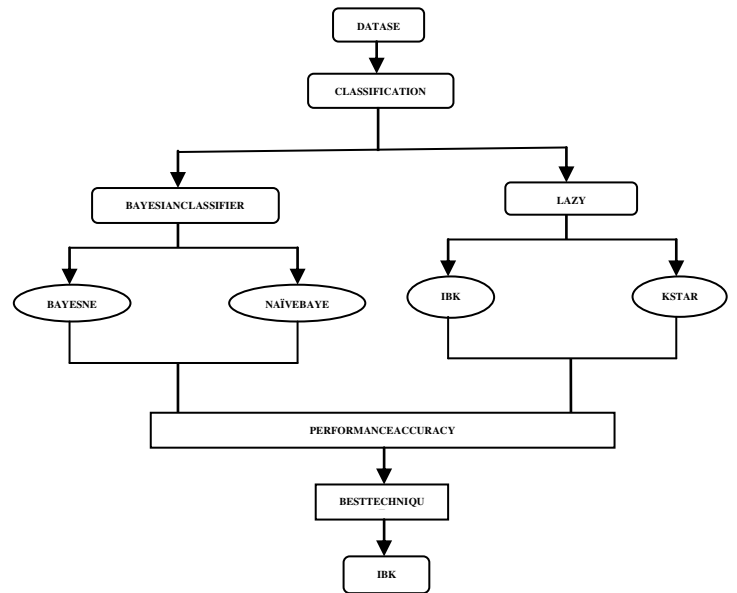


Fig. 1. Overview of Working Process

In Fig. 1 describe overview of working process, give input to dataset then select classifier either Bayesian or lazy. Its each comes under two classifier. Calculate accuracy for four classifier and compare to give best classification technique.

#### A. Dataset

There are important issues in the dataset that showed in statistical analysis which highly affects the performance of the systems, and outcomes in a very poor estimation of anomaly detection approaches. NSL-KDD [6] is describe, which consists of selected records of the complete KDD data set. It contains 33,800 record and 41 attributes. The advantages of NSL-KDD dataset are

- i. It has better reduction because no duplicate record in the test set.
- ii. Classifier will not produce any biased result because no redundant records in the train set.
- iii. The percentage of records in the original KDD dataset is inversely proportional to the number of selected record from each difficult level group.

#### B. Features Selection

Feature selection is also called data dimension reduction in predictive analytics, it refers to the process of identifying the few most important variables or parameters which help in predicting the results. Feature selection [11] should be one of the main concerns for a Data Scientist. Accuracy and generalization power can be leveraged by a correct feature selection based on entropy and information gain. Increasing interpretability of the model. The main advantage for feature selection is applied for dataset to reduce the training time and evaluating time.

TABLE I. ATTRIBUTES FOR EACH DATA SET AND DATA FORMATS

Type	Attributes
Nominal Attributes	is_guest_in, su_attempted, protocol_type, service, flag, land, logged_in, root_shell, is_host_login.
Continuous Attributes	srv_count, error_rate, dst_error_rate, error_rate, srv_error_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_error_rate, dst_host_srv_error_rate, dst_host_error_rate, dst_host_srv_error_rate, dst_bytes, dst_host_count
Discrete Attributes	srv_count, count, num_access_file, num_shells, num_file_creations, num_root, num_compromised, num_failed_logins, num_outbound_cmds, dst_host_srv_count, hot, urgent, wrong_fragment, dst_host_count

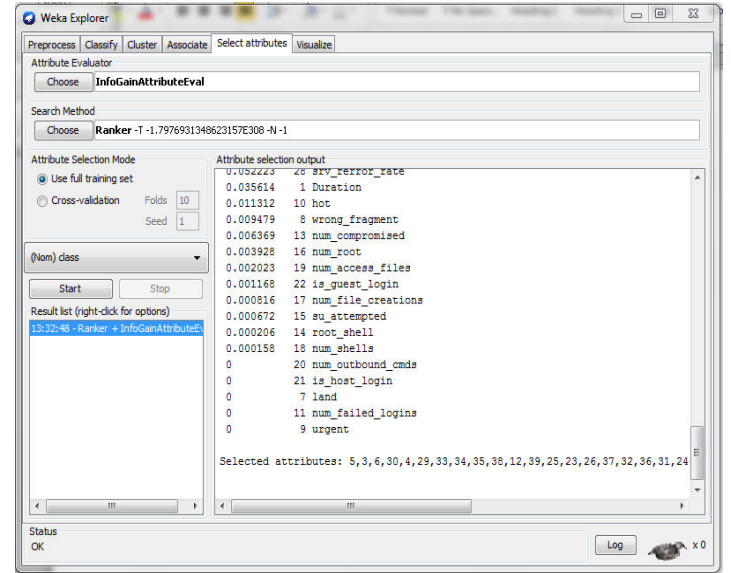


Fig. 2. Information gain Process

The important for features selection:

- i. It enables the machine learning algorithm to train faster.
- ii. It reduces the complexity of a model and makes it easier to interpret.
- iii. It improves the accuracy of a model if the right subset is chosen and reduces over fitting.

Weka supporting many feature selection techniques that help to outcomes of features selection.

### C. Information gain attribute evaluation

In feature selection [12] is classified into two portions Search Method and Attribute Evaluator. The search method is the technique by which to try or navigate different combinations of attributes in the dataset in order to arrive on a short list of chosen features. The attribute evaluator is the technique by which each attribute in your dataset also called a column or feature is examines in the context of the output variable (e.g. the class). Some Attribute Evaluator techniques [18] required the use of specific Search Methods. For example shown Fig.2, the CorrelationAttributeEvaluation technique used in the next section can only be used with a Ranker Search Method that evaluates each attribute and lists the results in a rank order.

$$H(\text{Category}) = -\sum(P_i \cdot \log_2(P_i)) \quad (1)$$

$$\text{InfoGain}(\text{Category}, \text{Attribute}) = H(\text{Category}) - H(\text{Category} | \text{Attribute}) \quad (2)$$

Where,  $P_i$  the probability of the Category  $i$  in the dataset.  $\log_2$  the base 2 logarithm Weka natural logarithm of base  $e$  is used, but generally we take  $\log_2$ . Entropy basically measures the **degree of "impurity"**. The closest to 0 it is, the less impurity there is in your dataset.

## V. CLASSIFICATION

In this work, we have analyzed performance of two classifiers namely, Bayesian and Lazy classifiers. In Bayesian classifier has two algorithms namely, BayesNet, and Naïve Bayes. In lazy classifier has two algorithms namely, IBK and Kstar.

### A. Bayesian Classifier

Bayesian networks are used to representing powerful probabilistic, and their use for classification has received valuable attention. Bayesian algorithms [13] predict the class based on the probability of belonging to that class. A Bayesian network is a graphical model probability relationship among a set of variables features.

### B. BayesNet

BayesNet refers to Bayesian networks made in nominal attributes called numeric ones are prediscritized and no missing values called any such values are replaced globally. Bayes Nets or Bayesian networks [13] are graphical representation for probabilistic relationships among a set of random variables. Given a finite set  $\text{Var} = \{\text{Var}_1, \dots, \text{Var}_n\}$  of discrete random variables where each variable  $\text{Var}_i$  may take values from a finite set represented by  $\text{Val}(\text{Var}_i)$ .

$$P(\text{Var}_1, \dots, \text{Var}_n) = \prod_i (P(\text{Var}_i | \text{Pa}(\text{Var}_i))) \quad (3)$$

### C. Naïve Bayes

The probabilistic Naïve Bayes classifier is implementing for classification. Naïve Bayes Simple using the normal distribution to model numeric attributes. The Naive Bayes algorithm is fully based on conditional probabilities. Naive Bayes using Bayes' Theorem [14] it is a formula that calculates a probability by counting the frequency of values and combinations of values in the historical data. Bayes Theorem evaluates the probability of an event occurring given the probability of another event that has already occurred.

## VI. EXPERIMENTAL RESULTS

$$P(\text{hyp}|d) = (P(d|\text{hypo}) * P(\text{hyp})) / P(d) \quad (4)$$

Where,

**P(hyp|d)** is called the posterior probability, it give the probability of hypothesis h given the data d.

**P(d|hyp)** is give the probability of data d given that the hypothesis hyp was true.

**P(hyp)** is called the prior probability of h, it give the probability of hypothesis hyp being true.

**P(d)** is give the probability of the data.

### D. Lazy Classifier

Lazy Classifier stored all the training instances and does no exact work until classification time. The main advantage for this method [16] is gained in employing that the target function will be evaluating locally such as in the k-nearest neighbor algorithm. Because the objective function is approximated locally for each query to the system, lazy learning systems at a time solve multiple problems and deal successfully with changes in the problem.

### E. IBK

IBK is also called k-nearest-neighbor classifier that refers to the same distance metric. The number of nearest neighbors can be specified explicitly in the object editor or evaluating automatically using leave-one-out cross-validation focus to an upper limit given by the specified value. A kind of different search algorithms [17] can be used to increase speed up the task of searching the nearest neighbors.

### Algorithm

K- Nearest neighbour algorithm  
 Training  
 Build the set of training examples  $D$ .  
 Classification  
 Given a query instance  $x_q$  to be classified,  
 Let  $x_1, x_k$  denote the  $k$  instances from  $D$  that are nearest to  $x_q$   
 Return  
 $F(x_q) = \arg \max \sum \delta(v, f(x_i))$

### F. KStar

The  $K^*$  algorithm can be describe as a method of cluster analysis which mainly aims at the partition of 'n' observation into 'k' clusters in which each observation belongs to the cluster with the nearest mean. We can describe  $K^*$  algorithm [15] is an instance based learner which is used for entropy as a distance measure.

$$K^*(x_i, y) = -\ln P^*(x_i, y) \quad (5)$$

Where  $P^*$  means probability of all transformational paths from instance  $y$  to  $x$ . It can be useful to understand the probability that  $y$  will arrive at  $x$  via a random walk in feature space.

### A. Correctly and Incorrectly Instance Classification

From table, it can be concluded that accuracy of classifications algorithms like BayesNet, NaiveBayes, IBK and Kstar for the correctly classified instances is more as compared to the incorrectly classified instances. Fig. 3 shows the bar graph of the correctly and incorrectly classified instances of the taken algorithms.

TABLE II. CORRECTLY AND INCORRECTLY INSTANCE CLASSIFICATION AMONG VARIOUS ALGORITHM

Algorithms	Correctly Instance Classification	Incorrectly Instance Classification
BayesNet	97.40	2.8
Naive Bayes	95.58	4.43
IBK	99.34	0.65
Kstar	99	1

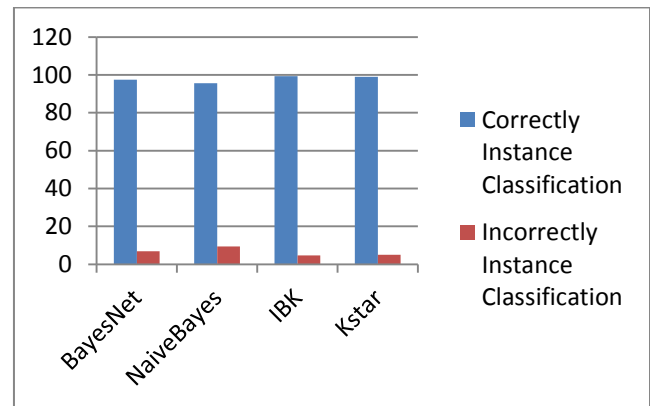


Fig. 3. Bar Graph of Correctly and Incorrectly Instance Classification of Algorithms

### B. Errors

#### a) Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

Mean Absolute Error Computes accuracy of attributes and average magnitude of errors. Actually it is calculate the average of absolute values between predicted observation and absolute observation. Root mean squared value also computes the average of magnitude of errors. From the experiment implemented with the used dataset as shown in table and Fig. 4.

TABLE III. ROOT MEAN SQUARED ERROR (RMSE) AND MEAN ABSOLUTE ERROR (MAE)

Algorithms	Root Mean Squared Error	Mean Absolute Error
BayesNet	0.1543	0.0282
NaiveBayes	0.1937	0.0494
IBK	0.0807	0.0066
Kstar	0.1343	0.0348

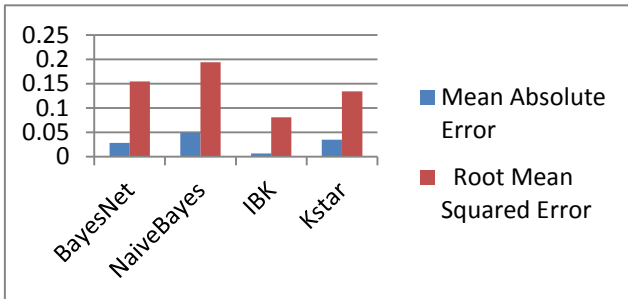


Fig. 4. Bar Graph of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)

b) Root Relative Squared Error (RRSE) and Relative Absolute Error (RAE)

These are the errors show performance of every experiment. Absolute error shows physical error and relative error shows information about how much efficient a particular attribute relatively measured. Table 4 and Fig.5 shown the dataset taken in this experiment to compared four algorithm.

TABLE IV. ROOT RELATIVE SQUARED ERROR (RRSE) AND RELATIVE ABSOLUTE ERROR (RAE)

Algorithms	Root Relative Squared Error	Relative Absolute Error
BayesNet	30.9376	5.6561
NaiveBayes	39.4328	9.9298
IBK	16.1701	1.322
Kstar	26.8804	6.9664

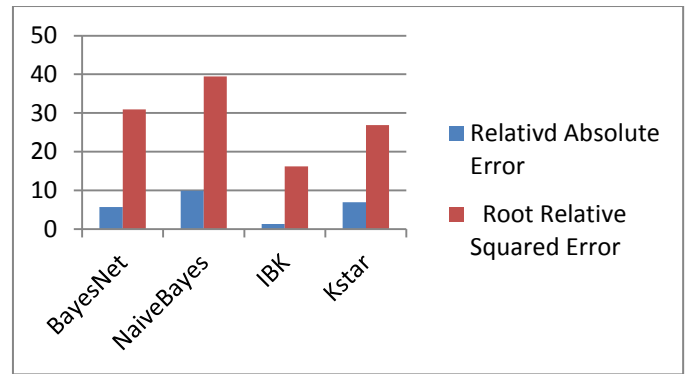


Fig. 5. Comparison graph of RAE and RRSE

C. Accuracy Measurement

The accuracy of algorithms is shown in table 5 and Fig. 6. It is measured with the help of parameters such as, precision, recall, TP rate, FP rate and F-measure.

TABLE V. VALUES OF TP-RATE, FP-RATE, PRECISION, RECALL, F-MEASURE FOR ALGORITHMS

Algorithms	Precision	Recall	TP Rate	FP Rate	F-Measure
BayesNet	0.974	0.974	0.974	0.028	0.997
NaiveBayes	0.956	0.956	0.956	0.046	0.95
IBK	0.993	0.993	0.993	0.007	0.995
Kstar	0.11	0.98	0.99	0.02	0.99

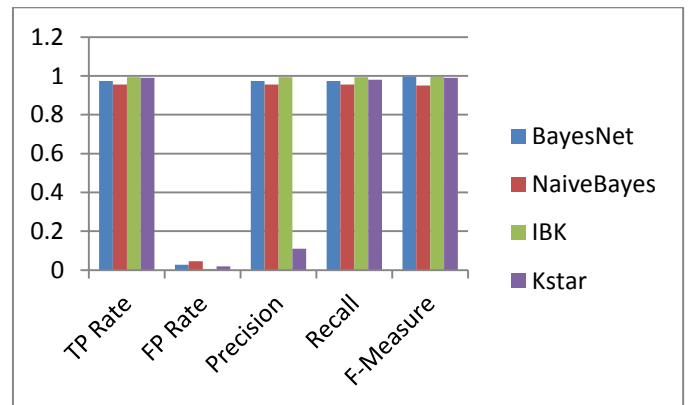


Fig. 6. Graphical Representation of Precision, Recall TP-Rate, FP-Rate, F-Measure values

TABLE VI. ACCURACY OF MEASUREMENT OF ALGORITHMS

Algorithms	Accuracy
BayesNet	97.2
NaiveBayes	95.4
IBK	99.3
Kstar	98.9

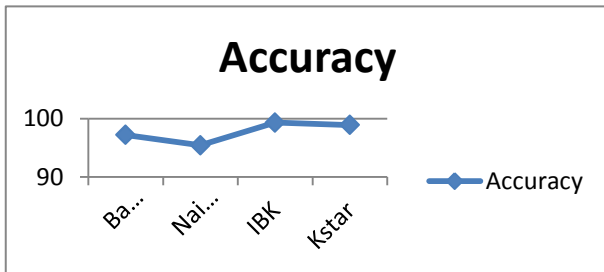


Fig. 7. Comparison of Algorithm Accuracy

### VII. CONCLUSION

In this work, we have analyzed performance of two classifiers namely, Bayesian and Lazy classifiers. In Bayesian classifier has two algorithms namely, BayesNet, and Naïve Bayes. In lazy classifier has two algorithms namely, IBK and Kstar. Analyzing performances of Bayesian and lazy classifiers by applying various performance factors. From the performance analysis, lazy classifier is more efficient than Bayesian classifier. The further extension of this work is to integrate the host based intrusion detection system and network based intrusion detection system for better detection. Developing a new IDS schemes for detecting novel attacks rather than individual instantiations.

### REFERENCES

[1] SY Ji, BK Jeong, S Choi and DH Jeong, "A multi-level intrusion detection method for abnormal network behaviors" ELSEVIER: Journal of Network and Computer Applications, vol.62, pp.9-17, 2016.

[2] Huang L, Milne D, Frank E, Witten IH, "Learning a concept-based document similarity measure", Journal of the Association for Information Science and Technology, pp.1593-608, 2012.

[3] Vaithyanathan V , Rajeswari K , Kapil Tajane and Rahul Pitale, "Comparison of different classification techniques using different datasets", International Journal of Advances in Engineering & Technology, May 2013.

[4] Sharma TC, Jain M, "WEKA approach for comparative study of classification algorithm", International Journal of Advanced Research in Computer and Communication Engineering. April 2013.

[5] Amor NB, Benferhat S, Elouedi Z, "Naive bayes vs decision trees in intrusion detection systems", ACM symposium on Applied computing, vol.14, pp. 420-424, March 2013.

[6] <http://nsl.cs.unb.ca/NSL-KDD/>

[7] Aljawarneh S, Aldwairi M, Yassein MB, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model", Journal of Computational Science, March 2017.

[8] Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, Vázquez E, "Anomaly-based network intrusion detection: Techniques", systems and challenges. computers & security, pp.18-28, March 2009.

[9] Alaei P, Noorbehbahani F, "Incremental anomaly-based intrusion detection system using limited labeled data", IEEE: International Conference, pp. 178-184, April 2017.

[10] Van NT, Thinh TN, Sach LT, " An anomaly-based network intrusion detection system using Deep learning. In System Science and Engineering", IEEE International Conference pp. 210-214, July 2017.

[11] Pan, Zhiwen, Salim Hariri, and Youssif Al-Nashif. "Anomaly based intrusion detection for building automation and control networks." In Computer Systems and Applications pp. 72-77, 2014.

[12] Sequeira, Karlton, and Mohammed Zaki, "ADMIT: anomaly-based data mining for intrusions." International conference on Knowledge discovery and data mining, pp. 386-395. ACM, 2002.

[13] Vasudeo SH, Patil P, Kumar RV, "IMMIX-intrusion detection and prevention system", IEEE: International Conference, pp. 96-101, 2013.

[14] Bartlett M, Bate I, Cussens J, "Learning Bayesian networks for improved instruction cache analysis" IEEE: International Conference, pp. 417-423, 2011.

[15] Garcia EK, Feldman S, Gupta MR, Srivastava S, "Completely lazy learning" IEEE Transactions on Knowledge and Data Engineering, pp:1274-85, 2010.

[16] Al-Jarrah O, Arafat A, "Network Intrusion Detection System using attack behavior classification" IEEE: International Conference on Information and Communication Systems (ICICS), pp. 1-6, 2014.

[17] Salunkhe UR, Mali SN, "Security Enrichment in Intrusion Detection System Using Classifier Ensemble", Journal of Electrical and Computer Engineering, 2017.

[18] Alipour H, Al-Nashif YB, Hariri S, "IEEE 802.11 anomaly-based behavior analysis", IEEE: International Conference on Computing, Networking and Communications, pp. 369-373, Jan 2013.



# Reliability Aware Self-Test Schedule for Aging Systems

Harini Sriraman

School of Computing Science and Engineering,  
VIT Chennai campus, Vandalur-Kelambakkam Road,  
Chennai, India  
harini.s@vit.ac.in

Pattabiraman Venkatasubbu

School of Computing Science and Engineering,  
VIT Chennai campus, Vandalur-Kelambakkam Road,  
Chennai, India  
pattabiraman.v@vit.ac.in

**Abstract**—With increasing transistor density, processor circuits are more prone to hardware failures. In the recent processor cores, early aging has become a major design constraint. To ease this situation, many on the field hardware age monitoring tools have been proposed. One bottleneck for these age monitoring tools is the frequency of self-tests. In this paper, we propose a smart scheduling algorithm for these self-tests such that unnecessary test schedules will be avoided. The proposed method also predicts faults in-time, before it affects the system state. In addition to saving redundant self-tests, power dissipation due to these tests are also saved by using the proposed algorithm. Since the proposed algorithm takes feedback from the system with respect to the current work load, test scheduling becomes a reliability aware efficient one.

**Keywords**—Self-Test, Hardware Faults, Test Scheduling, Reliability Aware, Aging of Circuits

## I. INTRODUCTION

The ever increasing computing power is met by increasing transistor density per chip. This is especially true for servers used in the computing power hungry domains like cloud and big data. As this trend of increased transistor density continues for technology nodes beyond 22nm, the circuits become more hardware failure prone. These can be contributed due to the factors like, on-chip power scaling, thin oxide layers and temperature loops that may result in thermal cycling. Major type of hardware fault mechanisms that causes these aging related problems are, Electron Migration (EM), Stress Migration (SM), Time Dependent Di-electric Break Down (TDDB), Thermal Cycling (TC), Negative Bias Temperature Instability (NBTI). To handle these faults on the field many hardware fault detection and repairing algorithms and tools have been proposed. Major bottleneck in these, on the field, fault handling techniques is the scheduling of the fault detection methods. The most popular on the field fault detection is based on test signatures. These techniques use a fixed frequency to perform self-test. The main disadvantage of fixed frequency self-tests is the unnecessary need for testing when there are no errors. Sometimes these tests may also miss faults and the faults may affect the system state due to the fixed test scheduling intervals.

## II. RELATED WORK

In the recent times, with technology nodes beyond 22nm, self-testing on the field for permanent faults is gaining momentum. This is primarily due to the fact that reliability has become a major design constraint for the new processor designs. To perform self-test, the important requirements are listed in Table I.

TABLE I. REQUIREMENTS FOR SELF-TESTING TECHNIQUES



As one of the major requirement for self-test is optimal test scheduling, in recent times many smart test scheduling techniques have been proposed. The discussions about these techniques are explained in this section. Different categories of self-test are listed in Table 2.

TABLE II. TYPES OF FAULT DETECTION APPROACHES AND ATTRIBUTES

Criteria	Attribute of testing method	Terminology
Testing mode	<ul style="list-style-type: none"> <li>• Concurrent with normal system operation</li> <li>• As a separate activity</li> </ul>	On-Line Testing  Off-Line Testing
Source of stimuli	<ul style="list-style-type: none"> <li>• Within System itself</li> </ul>	Self-Testing
Level of Test	<ul style="list-style-type: none"> <li>• Gate level pins</li> <li>• System blocks</li> </ul>	Gate level test  Component-level test
Storage of stimuli	<ul style="list-style-type: none"> <li>• Retrieve from storage</li> <li>• Generated during testing</li> </ul>	Stored Pattern Testing  Comparison testing

The recent trends have shown that many smart test scheduling are on the rise. It includes, power-aware test scheduling [1], [2], [3] performance aware test scheduling, workload aware test scheduling. Many of testing is done on processor's core components as explained in [5], [6]. In recent times for memory and un-core components are also included for testing using the above methods [4], [5]. With recent trend to minimize power dissipation and maintain temperature, many power aware, temperature [7],[8] aware test scheduling have been proposed. In the recent times use of software for performing self-test [9], [10] and [11] are on the rise. The test scheduling in this paper is reliability aware which takes care of power, temperature and voltage. In recent papers, energy efficiency [12] and variability tuning [13] have been used for self-testing. As can be seen, much work has been done on self-tuning logics whereas the self-test scheduling still has much scope for improvement. In this paper, we present an energy aware, reliability aware self-test scheduling algorithm.

### III. SYSTEM MODEL

Let us consider the time 't' is divided in to small intervals  $\Delta t$  where  $\Delta t$  is the execution time of the workloads. The splitting of time in to smaller quantum is explained in Fig 1. Consider the system model as shown in Fig 2. The circuit diagram of Self-Test hardware control flow is given in Fig 3.

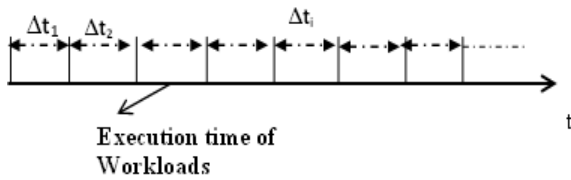


Fig. 1. Execution time 't' divided in to  $\Delta$

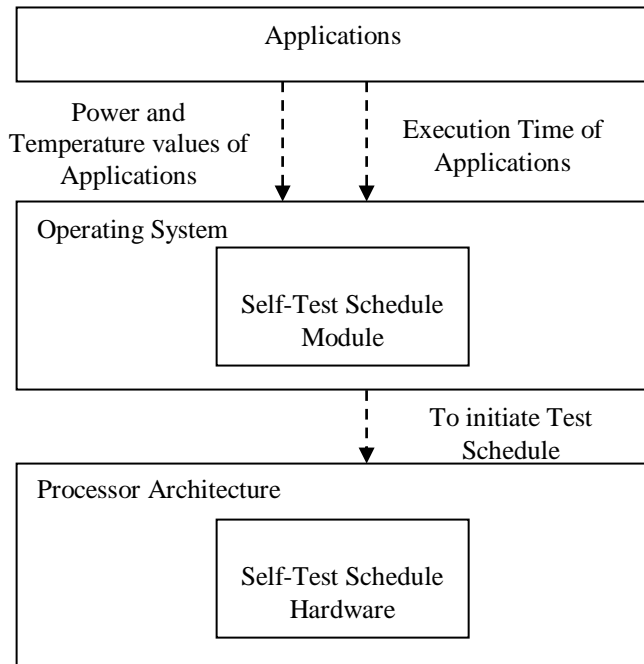


Fig. 2. System Model of Proposed Self Test Scheduling

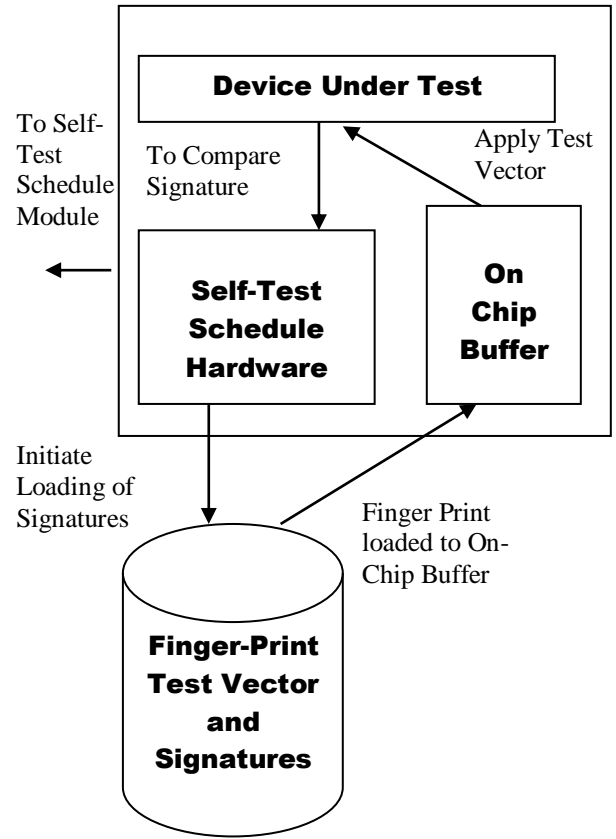


Fig. 3. Self-Test Hardware Control Flow

In the above system, the MTTF instantaneous will be calculated for each workload by fetching the power and temperature values. With instantaneous MTTF, cumulative failure rate is calculated. Let us call it Cumulative EETF. Based on these values the test scheduling is scheduled based on the proposed algorithm given in subsection A.

#### A. Proposed Algorithm

**DECLARE GLOBAL**

Test\_Schedule Frequency//Variable to tell the frequency of test schedule  
 i //Variable to point to workload  
 WTTFi //Weighted Expected Time To Fail  
 L // To find the aging Ratio  
 T //Gives the cumulative time that has elapsed  
**END GLOBAL**

**MAIN ROUTINE**

**INITIALIZE** Test\_Schedule Frequency // A predetermined value

**FOR** all workloads **DO**

**ASSIGN** i=i+1

**GET** power(P) and temperature(Temp) values for workload 'i'

    IMTTF<sub>i</sub> = **CALL** INSTANTANEOUS\_MTTF (P,Temp)

    WTTFi = **CALL** WEIGHTED\_ETTF (Δti, Δtj, IMTTF<sub>i</sub>, IMTTF<sub>j</sub>)

$L = \sum [((\Delta t_i / WTTFi) * 10)^9]$   
      $T = \sum \Delta t_i$  for all values of 'i'

**IF** (L/T > 1) **THEN**

        WTTF = WTTF + f1(L/T)

**ELSE**

        WTTF = WTTF - f2(L/T)

**ENDIF**

END FOR  
END MAIN

```
FUNCTION INSTANTANEOUS_MTTF (P,Temp )
BEGIN
// Instantaneous MTTF calculations using Black's Equation
IMTTF=kf(P,a,Temp) // 'a' is the activity factor
RETURN IMTTF
END
END FUNCTION
```

```
FUNCTION WEIGHTED_ETTF (Δti, Δtj, IMTTFi, IMTTFj)
WMTTFi=Σ[(Δti)*IMTTFi, Δtj*IMTTFj] / (Δti + Δtj)
RETURN WMTTFi
END FUNCTION
```

### B. Implementation

The implementation of the above algorithm is done in the Linux operating system. A virtual machine was created with Linux operating system. A script for the above logic was executed for various parallel benchmark workloads. The time overhead due to the proposed script is measured. To simulate the real time environment the above steps are performed for 107 iterations. Power and temperature tools were used to extract the real time values for the workloads executed. The tool work flow is shown in Fig4. The test scheduling values calculated during the above iterations are shown in Figure 5. The execution time overhead for the above proposed algorithm is shown in Fig6. For the implementation, temperature and power tools used are lm sensor and powerstat respectively.

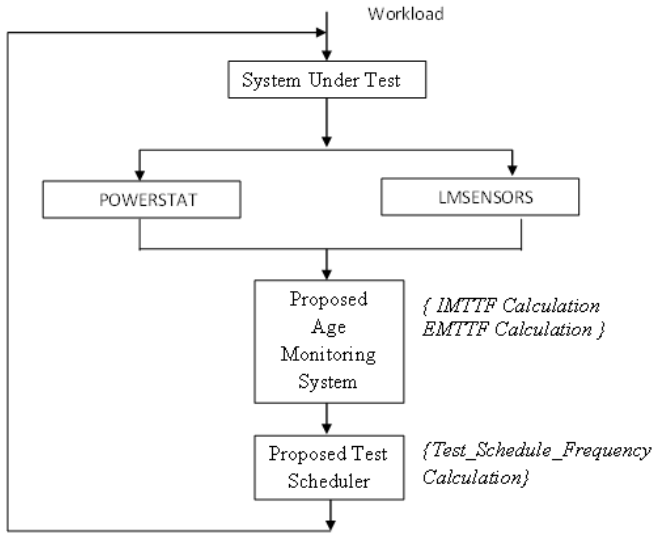


Fig. 4. Overall Implementation Flow

As given above the test scheduling results are in a closed loop to get feedback from the current workload and accordingly change the values the test frequency.

## IV. RESULTS AND ANALYSIS

The results of the above implementation can be measured in terms of test coverage, test time and area cost and test power. The details of the same are explained in this section.

### A. Test Scheduling Pattern

For our proposed algorithm, we generated random fault values throughout the lifetime of the processor. We considered the lifetime of the processor follows weibull distribution. Various test scheduling pattern for random fault insertion based on weibull-distribution fault rate is given in the following Fig 5.

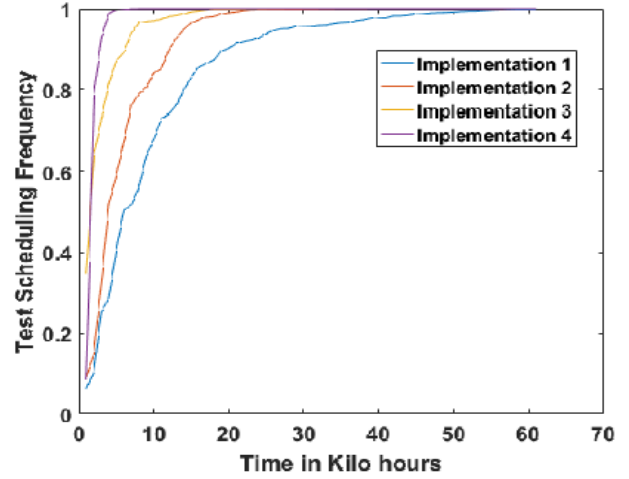


Fig. 5. Multiple Implementations of Test Scheduling Frequency

As we can see, in the initial hours, the test scheduling frequency is very less. As the aging factor increases, accordingly test scheduling frequency also increases. The benefits in terms of time and power are shown in Fig 6 and Fig 7 respectively.

### B. Time Overhead

TABLE III shows the time overhead components of the proposed algorithm and its corresponding values. The values in the TABLE III are calculated for workloads whose Δtis greater than equal to 100 μs.

TABLE III. TIME OVERHEAD

Component	Time Overhead
lmsensor	0.4 μs
Power stat	1.3 μs
IMTTF calculation	0.03 μs
WMTTF Calculation	0.07 μs
Aging Factor Calculation	0.03 μs
Total Time Overhead/100 μs workload	1.83 μs

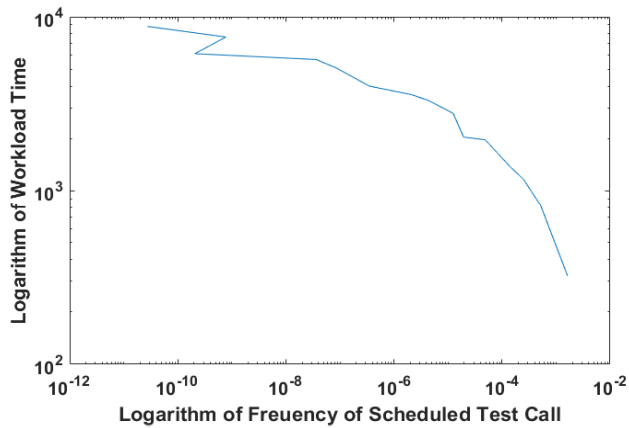


Fig. 6. Time Overhead due to Test Scheduling

From Fig 6, it is clear that as the time of workloads increases, the time overhead for test scheduling becomes significant. The power benefits of the proposed algorithm are shown in Fig 7.

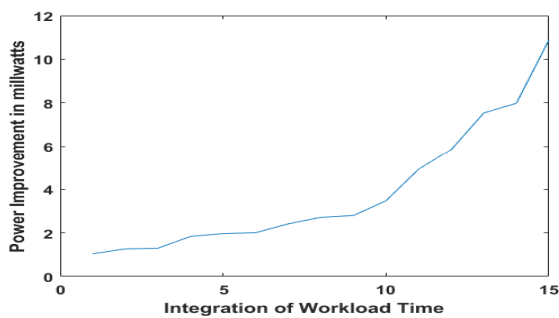


Fig. 7. Improvement in power dissipation for proposed test scheduling

As can be seen from Fig5, Fig 6 and Fig 7, the benefits of the proposed test scheduling logic diminishes its overheads. Especially for most part, the technique uses software based calculations, cost overhead due to the proposed test scheduling will be much lesser than other techniques that uses costly testing equipments.

## V. CONCLUSION & FUTURE WORK

As aging becomes a major reliability constraint and in recent times, reliability is a major design constraint, self-testing logics for hardware failures on the field are gaining much popularity. One of the main bottlenecks in such systems is the test scheduling. This paper proposes a reliability aware test scheduling using software that results in better performance, improved area overhead and decreased power dissipation. The benefits are clearly represented in the charts in section 4. This software based scheduling can also be extended for systems on

data center environments so that hardware refresh in data centers will become cost effective.

## REFERENCES

- [1] Haghbayan, M.H., Rahmani, A.M., Miele, A., Fattah, M., Plosila, J., Liljeberg, P. and Tenhunen, H., 2016. A power-aware approach for online test scheduling in many-core architectures. *IEEE Transactions on Computers*, 65(3), pp.730-743.
- [2] Sheshadri, V., Agrawal, V.D. and Agrawal, P., 2017. Power-Aware Optimization of SoC Test Schedules Using Voltage and Frequency Scaling. *Journal of Electronic Testing*, 33(2), pp.171-187.
- [3] Rittner, F., Ristic, M., Glein, R. and Heuberger, A., 2017, July. Automated test procedure to detect permanent faults inside SRAM-based FPGAs. In *Adaptive Hardware and Systems (AHS), 2017 NASA/ESA Conference on* (pp. 16-23). IEEE.
- [4] Theodorou, G., Kranitis, N., Paschalis, A. and Gizopoulos, D., 2014. Software-based self-test for small caches in microprocessors. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(12), pp.1991-2004.
- [5] Bernardi, P., Ciganda, L.M., Sanchez, E. and Reorda, M.S., 2014. MIHST: A hardware technique for embedded microprocessor functional on-line self-test. *IEEE Transactions on Computers*, 63(11), pp.2760-2771.
- [6] Lin, C.W. and Chen, C.H., 2016, October. A processor shield for software-based on-line self-test. In *Circuits and Systems (APCCAS), 2016 IEEE Asia Pacific Conference on* (pp. 149-152). IEEE.
- [7] Zhang, Y., Peng, Z., Jiang, J., Li, H. and Fujita, M., 2015, March. Temperature-aware software-based self-testing for delay faults. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition* (pp. 423-428). EDA Consortium.
- [8] Zhang, Y., Peng, Z., Jiang, J., Li, H. and Fujita, M., 2015, March. Temperature-aware software-based self-testing for delay faults. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition* (pp. 423-428). EDA Consortium.
- [9] Riefert, A., Ciganda, L., Sauer, M., Bernardi, P., Reorda, M.S. and Becker, B., 2014, March. An effective approach to automatic functional processor test generation for small-delay faults. In *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014* (pp. 1-6). IEEE.
- [10] Sabena, D., Reorda, M.S. and Sterpone, L., 2014. On the automatic generation of optimized software-based self-test programs for VLIW Processors. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 22(4), pp.813-823.
- [11] Scholzel, M., Koal, T. and Vierhaus, H.T., 2014, May. Systematic generation of diagnostic software-based self-test routines for processor components. In *Test Symposium (ETS), 2014 19th IEEE European* (pp. 1-6). IEEE.
- [12] Mintarno, E., Skaf, J., Zheng, R., Velamala, J.B., Cao, Y., Boyd, S., Dutton, R.W. and Mitra, S., 2011. Self-tuning for maximized lifetime energy-efficiency in the presence of circuit aging. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 30(5), pp.760-773.
- [13] Gupta, P., Agarwal, Y., Dolecek, L., Dutt, N., Gupta, R.K., Kumar, R., Mitra, S., Nicolau, A., Rosing, T.S., Srivastava, M.B. and Swanson, S., 2013. Underdesigned and opportunistic computing in presence of hardware variability. *IEEE Transactions on Computer-Aided Design of integrated circuits and systems*, 32(1), pp.8-23

# Optimized Channel Awareness Routing For Congestion Avoidance by Dynamic Queue Space Management in MANET

R.Sundar

Research Scholar

Dept. of Computer Science and  
Engineering

Sathyabama University, Chennai.  
apcesundar@gmail.com

Dr.A.Kathirvel

Professor

Dept. of Computer Science and  
Engineering

M.N.M Jain Engineering College,  
Chennai  
ayyakathir@gmail.com

S. Narayanan

Asst.Professor

Dept. of Computer Science and  
Engineering

Adhiparasakthi College of engg,  
Vellore.  
sakthi.sakthinarayanan@gmail.com

**Abstract—OBJECTIVES:** A mobile ad hoc network (MANET) is a self-organized, autonomous collection of mobile nodes forming a dynamic network and communication over wireless links. In a MANET, efficient load balancing done through a proper congestion avoidance mechanism. **METHODS:** Novel Dynamic queue spaces management congestion avoidance (DQSMCA) algorithm is proposed for better load sharing. A node may behave as a critical or non-critical node depending on its available bandwidth, node energy level and its signal strength. In our proposed system, only non-critical nodes are allowed to participate in route discovery and transmission process. **FINDING:** Depending on the packet incoming rate and service rate, a node is dynamically labeled as a finite or infinite queue space node. When service rate is higher than the arrival rate, the queue length is assumed as infinite length otherwise it is finite length. When all nodes in a path are infinite queue length then we can send more number of packets otherwise average data packets may be sent. **CONCLUSION/APPLICATION:** As a result of DQSMCA, it reduces the congestion and increases throughput. Moreover, while routing we need to consider the packet lifetime in order to reduce path loss. The network load is restricted to certain limits to maximum throughput.

**Keywords—**Load Sharing, dynamic queue space, critical node, Packet Lifetime.

## I. INTRODUCTION

In MANET the nodes are in random mobility and they are dynamically configured themselves to form a topological network without any centralized authority. Any number of nodes may be dynamically added or removed from the network without any additional requirements.

If a signal path is selected for packet transmission, all the packets are routed in the same path in which the energy level of the nodes present in the path may be drained. It may lead to node failure when the energy level is completely drained. The multipath transmission will eliminate the above drawback. The packets are routed among multiple paths instead of a single path, it reduces the packet transmission delay, increases the packet delivery ratio and node life time.

In a multipath transmission, at whatever point congestion in the system is more prominent than the limit of the system then the blockage happens. It will lead to packet loss during transmission period, each and every node in the MANET is provided with inter queue spaces where incoming packets are stored and processed one by one. When these available queue spaces are filled by incoming packets and there is no available empty space then further incoming packets are discarded.

In our proposed system using channel awareness routing we developed congestion avoidance and load balancing algorithm analyzing available bandwidth, energy level and signal strength node may behave as a critical or non-critical node [1]. The only non-critical node participating in the route discovery process and multipath can be established using incoming data rate and service rate depends on finite or infinite queue size. The proper load balancing based to avoid the complete congestion and also during the path loss is eliminated. This algorithm network performance is improved as throughput increases by restricting the load to a certain level in MANET.

The rest of the article is organized as follows: In this Section II we talk about the foundation and related work. The Section III insights about the Problem distinguishing pieces of proof of existing frameworks. In our proposed methodology of Dynamic Queue Spaces Management based on a Congestion Avoidance algorithm in Section IV. Examination and results are performing is given in area V. Conclusion and future upgrade exhibited in Section VI.

## II. BACKGROUND AND RELATED WORKS

In a MANET [2], versatile hubs convey over remote channels with no base station. In this system, because of portability hubs subjectively leave or join the system. It uses many routing algorithms. In the existing system, the reactive routing algorithm is used, where it communicates the control parcels through the Route Request and Route Reply message. At the point when the hub pronounces that the dynamic course is broken or congested, the occasionally clock is lapsed in the every way. The AOMDV (Ad hoc On Demand Multipath

Distance Vector) directing calculation repair the course of communicating the RREQ parcel to locate a backup way to go to goal. It takes more time and high congestion and overhead of this route repair to find an alternative route.

In this proposed [3] Multipath load balancing technique for congestion control (MLBCC) performs multipath load sharing based on selection of gateway node. These portal hubs are chosen by the hopeful hubs relies upon the estimation of connection cost and additionally way cost. These door hubs proficiently adjust the heap among the various paths. The Energy of the portal hubs will be depleted and furthermore the way misfortune will influence to decrease the throughput.

An In this user planned discuss the algorithm [4] an every node a transmission procedure performed by distinguishing the clog and maintained a strategic distance from to locate the option method for transmitting information. At the point when the medium investigation, every node to play out the procedure and send the information is more noteworthy than the assignment of the accessible time and energy, it takes more postponement. The accessible transfer speed shared by the quantity of clients on the every way or overwhelming burden information will be prepared in an every hub are not proposed in this method.

In author projected concepts [5], these techniques are discussed with each intermediate nodes are checked the channel capacity and estimation of queue length depends on the data are communicated to the transmitter and receiver. Using this capacity and length of the queue is to be calculating the estimate the data rate to be travelling in the each intermediate node, it reaches the final node. Here mobile nodes are highly dynamic, each time change the data rate and it can't be fixed resource utilization not considered.

In author proposed [6] concepts a load can increase and decrease depends upon the capacity of the network, data rate and power of the node. Reducing the traffic in distributed various path, the loads are shared by providing a better throughput and packet accept the delivery of rate are increased.

In user design algorithm [7] a path has connected by multiple intermediate nodes in that node check the constraints based the links are established without the noise of the signal, available bandwidth and power constraints of the node. Route discovery based on this status of the node connected by link and the path is established. The user not discussed load balancing depends on inter queue space are analysis in a node, it reduce the congestion.

In [8] proposed a Multipath route discovery based on signal strength, energy level and malicious nodes are detected to removing the packet loss and also improve the network performance. But user not discussed rate and available queue space based on to provide better load sharing in a multipath, it will distribute data packets at a transmission period of time.

In author discussed early congestion detection techniques [9] in each node calculates the queue status using early detects and congestion and indicate alarm message to be shared the entire neighbor node. Here dynamic movement mobile ad hoc

network on demand route discovery process the dynamic change the queue to be increased and decrease not considered.

In author explained the concepts [10] Load sharing based on this multiple paths is in the minimum round trip time taken for complete process of transaction to travel packet start and end of the node without delay and more throughputs. Each path has various round trip time can be provided dynamically moves the node occurred heavy traffic. Load sharing in the each path depends on the available resources shared by a number of users and reduces the traffic in the entire capacity of the network are not discussed.

### III. PROBLEM IDENTIFICATION

- How to route discovery selecting the non-critical only? Not sufficient numbers of non critical node is available, then how the route is discovery?
- When the transmission process performed non critical nodes are route discovery at movement of time any one of the node move to the critical, then what happen?
- Since every packet has to be transmitted sequentially from source to destination, there occurs more traffic due to the large number of control packets.
- How the queue size is dynamically changed due to the heavy traffic and low traffic?
- How the base level of asset in the portable specially appointed system effectively utilized as a part of the whole limit.

### IV. PROPOSED WORK

In this proposed system, there is no need of selection gateway node, because we only use infinite nodes for transmission that has sufficient queue space. The multipath that was established in Dynamic Queue Spaces Management, Congestion Avoidance (DQSMCA) will have better link and path quality it may not fail in near future. Only think are we wanted to distribute the load sharing among the multiple paths in round robin fashion. This system uses channel awareness routing we developed congestion avoidance depends on inter queue space which involves the load balancing and non critical nodes are selected before transmission of period, it will reduce the path loss and increase the packet life time in a each best path.

Only non-critical nodes are participating in the route discovery process and multiple path can be established using this node based on the incoming data rate and the service rate depends on finite or infinite queue size are decided to reduce the traffic with amount of data are processed. In mobile adhoc network minimum level of resources that are sufficiently used to improve throughput and the network performance.

#### A. A Classification Of Non Critical And Critical Node

Every Node in MANET is categorized as Non Critical or Non Critical Nodes based on following criteria. If a node has sufficient amount energy level and signal strength a Non Critical Node and it is suitable for further transmission. If a node is not having either sufficient amount of energy level or signal strength, then the node is labeled as critical nodes in algorithm1.

Input Algorithm 1: node [0], node [1]...node [n-1] LIST OF NODE IN MANET

---

Algorithm1 for Sufficient Node Detection:

---

AEL-Average Energy Level  
 ASS-Average Signal Strength  
 N- Available number of nodes in MANET.  
 Array Node [] <- {node [0], node [1],...,node[n-1]};  
 //LIST OF NODES IN MANET  
 Array Critical\_Node[M],Non\_Critical\_Node[N];  
 M<-Maximum number of critical node <= n  
 N<-Maximum number of Non critical node <= n  
 Initialize i=>0,j=>0,k=>0  
 While(i<n)  
     Begin  
     If((node[i]>=AEL)&&(node[i]>=ASS))  
     Non\_Critical\_Node[j]=node[i];  
     j++;  
     else  
     Critical\_Node[k]=node[i];  
     K++;  
     i++;  
     End

---

OUTPUT ALGORITHM1:

1.OUTPUT1:Non\_Critical\_Node[0],Non\_Critical\_Node[1],...,Non\_Critical\_Node[j-1] //LIST OF NON CRITICAL NODES

2.OUTPUT2:Critical\_Node[0],Critical\_Node[1]...Critical\_Node[K-1] //LIST OF CRITICAL NODES

#### B. B. INTER QUEUE SPACE ANALYSIS:

In this we are going to select only the non critical nodes and their nodes are further labeled as finite and infinite queue space nodes based on algorithm2.

If the packet arrival rate in a node is lesser than the node service rate then the node is labeled as infinite\_queue\_space otherwise labeled as finite\_queue\_space node. In infinite\_queue\_spacethe quantity of bundles in the interface queue has reduced over a period of time and these nodes are willing suitable for packet transmission. Where as infinite queue space nodes the number of packets in the interface queue may increase in near future.

InputAlgorithm2:

Non\_Critical\_Node  
 [0],Non\_Critical\_Node[1],...,Non\_Critical\_Node[j-1] //LIST OF NON CRITICAL NODES

Algorithm2 for Inter Queue Space Analysis

---

1.Non\_Critical\_Node[0],Non\_Critical\_Node[1],...,Non\_Critical\_Node[j-1];//List of all Non\_Critical\_Node.  
 2.ArrayInfinite\_queuespace\_node[l],  
 Finite\_queuespace\_node[M];  
 3.Initialize i=0,l=0,m=0  
 4.While (i<j)  
     Begin  
     If(Arrival rate (Non\_Critical\_Node[i])<Service rate (Non\_Critical\_Node[i]))  
     Infinite\_queuespace\_node[l]=Non\_Critical\_Node[i];  
     //Label Non\_Critical\_Node[i] as Infinite\_queuespace\_node.  
     l++;  
     else  
     Finite\_queuespace\_node[M]=Non\_Critical\_Node[i];  
     //Label Non\_Critical\_Node[i] as finite\_queuespace\_node  
     M++;  
     End  
     i++;

---

Output Algorithm2:

Output1: Non\_Critical\_Node[i]as Infinite\_queuespace\_node.

Output2:Non\_Critical\_Node[i] as finite\_queuespace\_node

#### C. DYNAMIC QUEUE SPACE

##### MANAGEMENTCONGESTION AOMDV:

In this proposed DQSMC-AOMDV, instead of broadcasting the RREQ (Route Request) to all neighbor nodes only few infinite queue space neighbor nodes are selected and RREQ is multicast to these nodes. The above process will produce multiple numbers of best paths from the source to the destination and it also reduces the routing overhead Algorithm3.

As a final step we are sharing the load by distributing the available data packets among the multiple best paths in a round robin manner.

Input Algorithm3: Non\_Critical\_Node[i] as Infinite\_queuespace\_node.

---

Algorithm 3 DYNAMIC QUEUE SPACE MANAGEMENT CONGESTION AOMDV

---

Start of Source Node(S)

Do

(Neighbor Node=Next\_hop\_Node(S))

If((neighbornode==infinite\_queue\_spacenode)&&(next\_hop\_node!=destination\_node))

Send AOMDV RREQ Packet to Neighbor node

Repeat step2 by assigning S=Neighbor Node

Else if (Next\_hop\_node==Destination\_node)

Respond to forwarding the packet back to the initialize node and discovery, multiple best path between initialize to the receiver node

Call DQSMC\_Load Balancing (Best\_path[n])

Else

Reject current Neighbor node move to next neighbor node

Until (all neighbor nodes are visited)

### OUTPUT ALGORITHM3:

Best\_path[0], Best\_path[1].....Best \_ path [n-1];//Let n number of the best path that are find the Algorithm 3

### D. DQSMC\_LOAD BALANCING (BEST\_PATH (N)):

The multipath that was established in DQSMCA Algorithm will have better link and path quality and may not fail in near future. Only think are we wanted to distribute the load among multipath in round robin fashion.

By using Algorithm 3 we will get the multipath best path between source to a destination that has infinite queue space and non critical nodes. Now we need to perform load sharing fast transmission and energy conservation.

Best\_path[0], Best\_path[1].....Best \_ path [n-1];//Let n number of the best path that are find the Algorithm 3

Data \_packets [0], Data\_packets [1]..... Data\_packets [m-1];  
//Let m number of packets to be transmitted between source to destination.

Initiate k=0, j ← 0

While (k<m)

Send Data\_packets [k] in Best \_path [j]

k← k+1; // send the next packet

j← j+1 mod n; // move to next best available multipath in round robin fashion.

## V. EXPERIMENTAL RESULTS

### A. Simulation model and parameters:

Ns-2[11] is simulate the dynamic queue spaces Management congestion avoidance algorithm AOMDV. In the simulation finds the non critical neighbor nodes are selected instead send the all nodes a route discovery by reducing the routing overhead and reduces the congestion. The neighbor node it finds the queue size based reducing the traffic. Depends upon the queue size are found, based on incoming and the service rate to reduce the traffic with available resources are shared amount of data to be processed. In

summarizing the Table1 the simulation and parameters are designed.

### B. B. Performance metrics:

In AOMDV PROTOCOL Route discovery in distributing a packet along the multiple path to reach the destination. Traffic analysis must be done before transfer the data to avoid the congestion. The intermediate node analysis of energy level, signal strength, available bandwidth in Fig1, Fig2 and Fig3.

TABLE I.

S.No	Factor	Significance
1	Number of nodes	100
2	Area Size	850 X 650
3	MAC protocol	802.11
4	Radio Range	250 m
5	Antenna	Omni directional antenna
6	Simulation Time	150 Sec
7	Traffic Source	CBR
8	Routing protocol	AOMDV,MLBCC,DQSMC
9	Packet Size	1200 bytes
10	Mobility Model	Random Way Point
11	Rate	100 KB, 200 KB, 300 KB
12	Maximum number of packets in queue	250
13	Speed (m/Sec)	2m/Sec

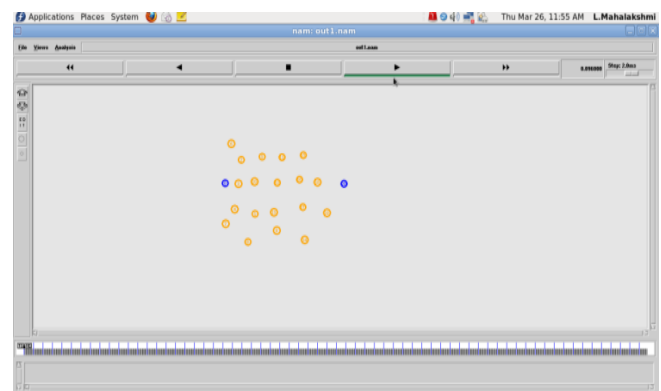


Fig 1. Node creation

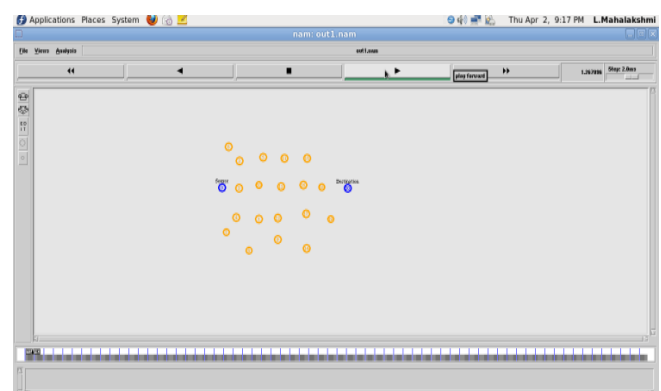


Fig 2. Node Movement



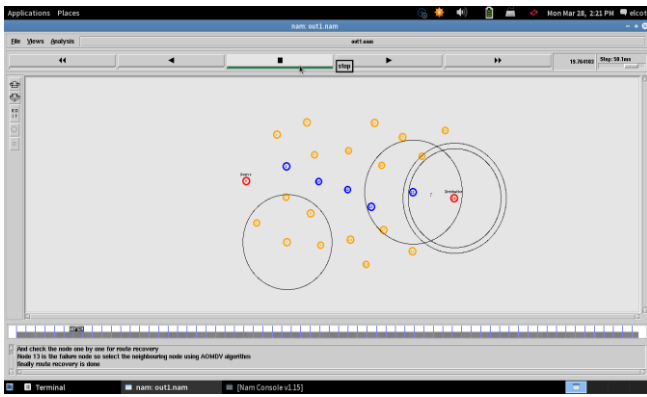


Fig3. Identifying Non Critical Node

The simulation parameters are formulated in Dynamic Queue Spaces Management, Congestion avoidance (DQSMCA) algorithm is delay, packet drop ratio and throughput.

$$Queue\ Delay[12] = \frac{number\ of\ packet}{available\ bandwidth} \quad (1)$$

In this equation (1) based non critical node are separated by the number of packets send by available bandwidth on the entire channel capacity. An entire capacity of channel shared by a number of users in inter queue space depends on the load balancing factor is reducing the delay. .

1. A number of user shared by the capacity of channel processed a large number of packets, it takes more delay.
2. Each packet shared by a number of users in its available bandwidth and also inter queue space depends on reducing the delay.

In AOMDV the number of packets is load shared by the all the multipath transmission, but here will not consider the availability of resource constraints based shard by number of packets at a transmission period of time. A AOMDV is 9% are increasing the delay when compare to MLBCC, and a DQSMCA is 1.33% are decreasing the delay when compare to MLBCC.

$$Packet\ Drop\ Ratio[13] = \frac{SHSDR}{RHRDR} \quad (2)$$

Here equation (2) Sender Host Sent Data Rate(SHSDR) is high compare the Receiver Host Receive Data Rate(RHRDR), it takes more delay and also fixed queue size becomes a packet are loss Fig 4 and 5.

Queue space dynamically changes depends on load of data and data rate. A sender arrival rate and receiver service rate based on finite and infinite queue are allocated to reduce the packet drop ratio and delay compared to MLBCC and AOMDV.

$$Packet\ Accept\ Ratio[14] = \frac{NP(LS)*DQS}{RABW} \quad (3)$$

Where components are formula 3, Packet Accepts Ratio (PAR), Number of Packets (NP), Load Sharing (LS), Depends

on Queue space (DQS), Restrict the Available Bandwidth (RABW)

In this equation (3) DQSMCA a network load to restrict to certain level depends on the available bandwidth to improve maximum throughput are achieved compare MLBCC and AOMDV Fig 6.

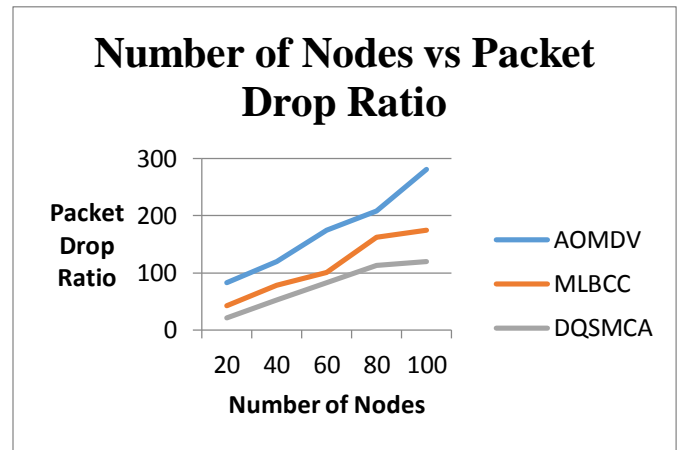


Fig4. Packet drop ratio vs. number of nodes in DQSMCA-AOMDV

The proper load balancing based inter queue space to avoid the complete congestion and also during the path loss is eliminated the availability neighbor nodes are selected the non\_critical node to improve packets accept ratio. A DQSMCA is 1.27% are increasing the packet accept ratio when compare to MLBCC, and an AOMDV is 3.12% are decreasing the packet accept ratio when compare to MLBCC.

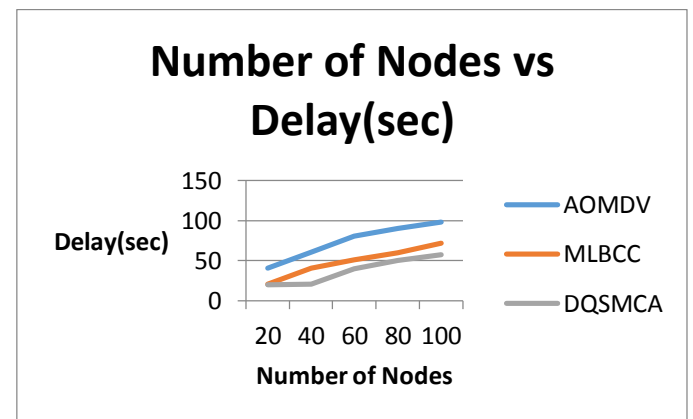


Fig 5. Delay (Sec) vs. Number of nodes in DQSMCA-AOMDV

## Packet Accept Ratio vs Number of Nodes

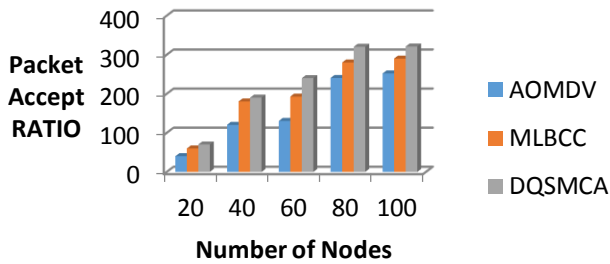


Fig 6. Packet Accept Ratio vs. number of nodes in DQSCALB-AODV

### VI. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, implements load balancing in an efficient manner. Our load balancing algorithm eliminates the path loss that occurs due to failure of a node. Since we are considering energy level and signal strength, the node availability is assured throughout the transmission period. Future there is no packet loss that occurs, when every the congestion happens; we examine the interface queue space and ensure the availability of queue space for storing incoming packet throughout the transmission period. In a future enhancement these concept average data packets may be sent in a finite queue space to better distribute data packets at the moment of transmission period.

#### REFERENCES

[1] Rajan, D., & Poellabauer, C. (2011). Cooperative energy management in distributed wireless real-time systems. *Wireless Networks*, 17(6), 1475-1491.

[2] Ranjan, P., & Velusamy, R. L. (2015, February). Optimized local route

repair and congestion control in Mobile Ad hoc Network. In *Computing and Communications Technologies (ICCT), 2015 International Conference on* (pp. 328-333). IEEE

- [3] Mallapur, S. V., Patil, S. R., & Agarkhed, J. V. (2015, November). Load balancing technique for congestion control multipath routing in mobile ad hoc networks. In *TENCON 2015-2015 IEEE Region 10 Conference* (pp. 1-6). IEEE.
- [4] Raza, I., Hussain, S. A., Qayyum, S., & Raza, M. H. (2010, October). Congestion aware nodes (CAN) based scheme for mobile ad hoc networks. In *Emerging Technologies (ICET), 2010 6th International Conference on* (pp. 348-352). IEEE.
- [5] Soundararajan, S., & Bhuvaneshwaran, R. S. (2012, May). Multipath load balancing & rate based congestion control for mobile ad hoc networks (MANET). In *Digital Information and Communication Technology and its Applications (DICTAP), 2012 Second International Conference on* (pp. 30-35). IEEE.
- [6] Ali, M., Stewart, B. G., Shahrabi, A., & Vallavaraj, A. (2012, March). Congestion adaptive multipath routing for load balancing in Mobile Ad hoc Networks. In *Innovations in Information Technology (IIT), 2012 International Conference on* (pp. 305-309). IEEE.
- [7] Tripti Sharma and Dr. Vivek Kumar (2016, May). congestion aware link cost routing for MANETS *International Journal of Computer Networks & Communications (IJCNC)* Vol.8, No.3, DOI: 10.5121
- [8] Jagadeesan, D., Narayanan, S., & Asha, G. (2015). Efficient load sharing using multipath channel awareness routing in mobile ad hoc networks. *Indian Journal of Science and Technology*, 8(15).
- [9] Kumaran, T. S., & Sankaranarayanan, V. (2011). Early congestion detection and adaptive routing in MANET. *Egyptian Informatics Journal*, 12(3), 165-175.
- [10] Naseem, M., & Kumar, C. (2015). Congestion-aware Fibonacci Sequencebased multipath load balancing routing protocol for MANETS. *Wireless Personal Communications*, 84(4), 2955-2974.
- [11] Network Simulator, <http://www.isi.edu/nsnam/ns>.
- [12] [https://en.wikipedia.org/wiki/Queueing\\_delay](https://en.wikipedia.org/wiki/Queueing_delay)
- [13] <http://blog.performancevision.com/eng/earl/links-between-latency-throughput-and-packet-loss>
- [14] <https://www.cisco.com/c/en/us/about/security-center/network-performance-metrics.html>

# A Review of Hierarchical Fuzzy Text Clustering

Seema Wazarkar, Bettahally N. Keshavamurthy  
 Department of Computer Science and Engineering  
 National Institute of Technology Goa  
 Ponda, India.  
 wazarkarseema@nitgoa.ac.in

Amrita Manjrekar  
 Department of Technology  
 Shivaji University  
 Kolhapur, India.

**Abstract**— Hierarchical fuzzy text clustering is a hybrid technique for clustering which is devised from a combination of number of techniques such as hierarchical clustering, fuzzy clustering, expectation-maximization approach, and similarity measure page rank algorithm. Initially, data is pre-processed, and then similarity measure page rank algorithm is applied to voluminous and high dimensional dataset. Inclusion of hierarchical clustering is advantageous due to hierarchical structure of the text data. Sometimes single phrase may be related to more than one topic hence fuzzy clustering is useful here as this algorithm has a property which allows placing one object into multiple clusters. The algorithm will be useful for different applications such as extraction of information from articles, news extraction, social network analysis, recommender systems or medical domain etc.

**Keywords**— Text Clustering; Hierarchical Clustering; Fuzzy Logic

## I. INTRODUCTION

In today's world, as increase in number of users for computer application huge amount of digital data is coming in front of us. Handling of these large amount of data is not an easy task hence data mining techniques have large scope in a research. Data mining is a process which is carried out to extract the latent patterns and interesting information from large dataset by examining it. Data mining is used to know something new about given data or hidden things in a data which is useful for handling it efficiently and easily. But, there are some issues related to implementation in data mining which are need of human interaction, model should be useful for databases used in future, all the objects present in a dataset should be fit in model, correct interpretation of results, etc. Data mining has two models predictive and descriptive [1]. Predictive model predicts values of data using experimental results obtained from algorithm which applied on different kinds of data. On the other side, descriptive model discovers relationship present in data as well as patterns in it. There are different tasks present for various activities in data mining such as classification, clustering, association mining, etc. These tasks belong to one of the above model such as classification, regression, prediction and time series analysis belong to predictive model and clustering association mining, summarization, sequence discovery belong to descriptive model.

Clustering is a task of making groups (it is also referred as clusters) on the basis of similarity and dissimilarity among

given objects in dataset. In data clustering two rules are important. First, maximize the inter cluster dissimilarity and second, maximize the intra cluster similarity. Clustering is an unsupervised approach which does not use the labeled data as classification. Clustering and classification are different kinds of grouping techniques. Hierarchical and partitional are the two main and most common kinds of clustering approaches.

Hierarchical clustering approaches creates numbers of nested clusters. Those are arranged in hierarchical manner using hierarchical tree called as dendrogram (Shown in Fig. 1). Hierarchical clustering approaches are divided into two subtypes i.e. agglomerative and divisive clustering. Agglomerative approach clusters objects with series of nested partitions i.e. from set of individual clusters to a single cluster. It is also referred as bottom up approach. Divisive approach starts clustering from single cluster to number of clusters. It is also called as top down approach. [2] Due to hierarchical structure of the text data in a document, this type of approaches are helpful in text data clustering.

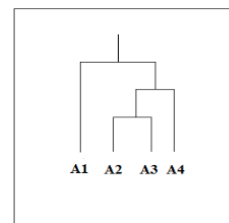


Fig. 1. Sample dendrogram

Partitional Clustering methods assign every object to mutually exclusive clusters or in other words disjoint clusters. It is also called as non-hierarchical clustering because it doesn't create a hierarchy of clusters. K-means clustering algorithm is most popular and usually used for clustering, which is one of the partitional clustering techniques. This approach works around centroid which is also known as mean, hence its name is K-means. Here, 'k' is number of clusters present in a dataset. It is necessary to know the number of clusters present in a dataset before initiating the implementation of partitional algorithms.

As most of the real world problems are fuzzy in nature, it is beneficial to use of fuzzy logic concepts. Fuzzy clustering technique results clusters with overlapping/soft boundaries. However, crisp or traditional clustering techniques have exact boundaries as shown in Fig. 2. Assume that we want to divide data from articles under different titles, but it is possible that some amount of data from an article is related to more than

one title. In this situation fuzzy logic concepts play an important role. It has provision to place data into multiple clusters with related titles. Hence, this approach is capable to handle the vagueness in given dataset.

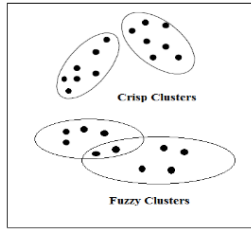


Fig. 2. Illustration of crisp and fuzzy clusters

In Section II, existing literature related to the hierarchical fuzzy clustering is reviewed. Then, discussion on reviewed literature and future directions are provided in Section III. Finally, conclusion is given as Section IV.

## II. HIERARCHICAL FUZZY TEXT CLUSTERING

In 2014 [3], combination of hierarchical and fuzzy clustering is applied first time to cluster the text data from articles. Text data is provided in the form of xml files. This xml files are generated from articles related to travelling which are present in corpus ONAC- Open National American Corpus. Page Rank algorithm is used to measure the similarity. Divisive hierarchical clustering is incorporated to split the data into number of groups and present it in hierarchical structure. An Expectation–Maximization (EM) is an iterative process, where the model depends on the unobserved latent variables. [4] An expectation (E) step uses a function which compute the probabilities of cluster membership. In Maximization (M) step, those probabilities are used to re-estimate the parameters. The body of literature discussed below is about its basic components hierarchical fuzzy text clustering i.e. basic algorithms involved in it.

Along with it, we will see few other aspects of text data analysis through the survey of papers related to text key-phrase extraction and clustering. Antecedently unknown information can be devised by text mining using methods from natural language processing as well as data mining. Considering frequency of terms or words is only useful for a document level clustering. Sometimes, number of presence of two terms in the document can be same, but it is possible that contribution of one term can be more to the meaning of its sentences than the other term. In this case, it is necessary to search term which is a conceptually more important. Shehata, Karray and Kamel introduced a new concept-based mining model which is used to analyze the terms at sentence level, document level and corpus levels. This model can effectively distinguish insignificant terms according to the meaning of sentence and significant terms that contribute more to the meaning of the sentence. [5]

Algorithm for new key-phrase extraction from a single document is introduced by Claude Pasquier. Based on the semantic similarity clustering of the sentences from document is done and then Latent Dirichlet Allocation is applied. Prior to the sentence clustering abbreviation expansion, sentence

detection, term identification, matrix creation and dimensionality reduction is carried out. Limitation for this algorithm is that it utilizes only the information accessible using a single document. [6]

Sentence clustering for multiple document summarization is carried out by Johanna Geiß with the help of latent semantic analysis. Hierarchical agglomerative clustering algorithm is applied to cluster the sentences along with Latent Semantic Analysis where parameters such as inter alia, type of vocabulary, optimal numbers of dimensions, size of the semantic space are investigated. This approach is finally compared with the simple word matching method of the traditional vector space model. Approach given by the author i.e. Latent Semantic Analysis produces better quality sentence clusters than simple word matching method of the traditional vector space model for Multi-Document Summarization. [7]

### A. Hierarchical Clustering

Naughton, Kushmerick and Carthy investigated the task of making groups of the text data from a news documents/articles which is related to the similar event with the help of clustering techniques such as agglomerative hierarchical clustering. For experimentation corpus of news documents which describe events happened in the Iraqi War is used. Their research emphases on combining detailed information of events collected from multiple sources to come up with concise explanation. Average, complete or single link based agglomerative hierarchical clustering are used for sentence level clustering. [8]

N. Rajalingam and K. Ranjini applied both the hierarchical algorithms (agglomerative and divisive) with three linkages on the database having information of the victims of Tsunami in Thailand. Initially, similarity measure for numeric data, binary data and string data is discussed because database used for experimentation has numeric, string, and binary type of data. Here, Euclidean Method for numeric data, simple matching Sokal & Michener distance measure for the binary data and Levenshtein Distance for string data is used to find the distance between objects present in a database. In single linkage, distance between the nearest members from the different clusters is calculated. In Complete linkage, distance among the farthest apart members is computed. Average linkage considers the distances among all pairs and takes average of all these distances. After analyzing the hierarchical algorithms which are mentioned above, author comes with the result that the divisive algorithm is faster than the agglomerative clustering algorithm and string data type requires more time than for the data having other data types. In case of binary field, the execution time required for the two combined binary fields is marginally larger or less equal to the time required for the single binary field. If the size of records get doubled, the running time get maximized by six times approximately. [9]

Guo, Shao and Hua experimented hierarchical text clustering method for four dimensions of cognitive situation i.e. spatiality, temporality, activity and protagonist. Author followed steps given below:

- Step 1: Sentence selection
- Step 2: Parsing of the sentence
- Step 3: Extracting cognitive situation dimensions
- Step 4: Constructing cognitive situation vectors
- Step 5: Constructing cognitive situation matrices
- Step 6: Compare cognitive situation matrices
- Step 7: Clustering tree construction

Two aspects, inner class and cross-class are used in the experiment. Better results are obtained using cross-class clustering as compared to inner-class clustering. [10]

Moshe Looks and et al have proposed a streaming hierarchical partitioning algorithm to extract meaningful data as well as useful associations, relationships, and groupings from voluminous data streams. With the help of cosine-theta measure similarity between a document vector and centroid is calculated. This algorithm is able to improve the ability to discover concepts. Hierarchical algorithm is compared with the most popular K-means algorithm where author found that in this case the hierarchical partitioning algorithm is superior to the K-means. This algorithm is also applied on the hardware i.e. Field Programmable Gate Arrays implemented for floating point calculations. It is useful to reduce the resources required while implementing the floating point arithmetic. [11] Optimizations implemented here are given as follows:

- Bitmap is used to pack 4K dimensional array of 8-bit byte.
- 32-bit registers are used to implement 32 dimensional vector sum.
- Instructions and 32-bit registers are used to compute multiple dot products.

As divisive hierarchical clustering approach is useful while developing hierarchical fuzzy clustering, pseudo code for it is provided as follows.

<p><b><i>Pseudo code for hierarchical clustering (divisive)</i></b></p> <ol style="list-style-type: none"> <li>1. Start with one cluster having all the given objects.</li> <li>2. Split the cluster based on the dissimilarities found among objects using subroutine algorithm (e.g. fuzzy clustering)</li> <li>3. Repeat step 2 until all objects get assign to their cluster.</li> </ol>
--

**B. Fuzzy Clustering**

Basis of study of terms, either words or phrase is very important in sentence level text clustering which is used in most of the common techniques of text mining. Some sentences may be associated to multiple themes from given document. Therefore, use of relational fuzzy clustering approach is advantageous. Relational fuzzy method allows objects to be a part of more than one cluster. This relational fuzzy clustering method experimented on relational data, i.e. data present in the form of square matrix consists pair-wise similarities among data objects. Page Rank method is applied as measure of general graph centrality. Basic concept of the Page Rank method is to find out important nodes from a graph by considering global information recursively which is calculated by using the complete graph. Node is a representative of a sentence in a graph and are weights of the edges represents a similarity among sentences. To contract a complete relational fuzzy clustering algorithm, data is used within an expectation-maximization model. This algorithm has

provision to identify overlapping clusters of conceptually (semantically) connected sentences. Thus, it is widely applied to accomplish a variety of text mining tasks. Andrew Skabar and Khaled Abdalgader proposed fuzzy relational eigenvector centrality-based clustering approach to deal with issues mentioned above. [4]

R. N. Davé and S. Sen devised a non-euclidean fuzzy relational data clustering algorithm which is applied for numerical examples. This algorithm is advantageous due to features like quick convergence, robust against outliers and ability to deal with all types of relational data, including non-Euclidean. Author also discussed a noise-clustering concept and new interpretation of the noise class. Relational techniques are extended for noise clustering. [12]

Deng, Hu, Chi and Wu proposed a fuzzy based text clustering technique where fuzzy C-means clustering and the edit distance algorithm is taken into consideration. This algorithm gives the more stable results and improves accuracy as compared to the traditional fuzzy C-means clustering approach. [13]

By using feed-forward neural network with supervised learning, semantic similarity is extracted and then fuzzy relational clustering method is used to partition objects in the dataset into clusters by P. Corsini, B. Lazzarini, and F. Marcelloni. Experimentation is done on two synthetic bi-dimensional datasets, the famous Iris dataset and synthetic dataset having 2-D images such as a tree, a house, an airplane and a car. Fuzzy relational clustering algorithm is adopted to make groups of objects which are more similar to each other and not so similar to objects in different clusters. By using proposed method, high number of correctly classified objects are gained using a less number of points of the dataset to train the neural network. [14]

Raghu Krishnapuram proposed fuzzy-medoids clustering algorithm and robust fuzzy-medoids clustering algorithm for web mining applications. Cosine measure is used to find similarity between two sessions. Document clustering, snippet clustering and mining of user profiles from access logs are carried out during experimentation. Comparison of both the algorithms is carried out which results that fuzzy-medoids clustering algorithm is more efficient. [15] [16]

Hierarchical fuzzy clustering is generated by using fuzzy clustering (as given below) as a subroutine algorithm in divisive hierarchical clustering algorithm.

<p><b><i>Pseudo code for fuzzy clustering</i></b></p> <ol style="list-style-type: none"> <li>1. Initially, select “k” fuzzy partitions from the given objects by using the membership matrix. Elements of membership matrix provide the score of membership of object for a particular cluster.</li> <li>2. With the help of membership matrix, compute the value of a fuzzy criterion function</li> <li>3. Reassign objects to clusters to reduce the value of fuzzy criterion function and re-compute the membership matrix.</li> <li>4. Repeat step 2 and 3 until convergence i.e. until the values of membership matrix do not change significantly.</li> </ol>
---

### III. DISCUSSION AND FUTURE DIRECTIONS

From data mining field, clustering is a very useful technique which helps in handling the huge unlabeled datasets. Many algorithms are present for clustering, but some of them such as hierarchical clustering and fuzzy clustering which are more suitable for text clustering according to the characteristics of it. Information related to similarity measure for computing distance between objects is also discussed. Evolution of hierarchical fuzzy clustering is represented through the year of algorithm discovery in Fig. 3.

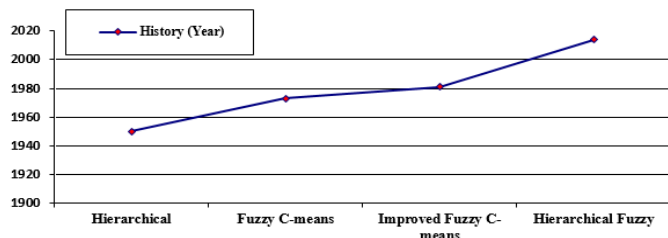


Fig. 3. Evolution of hierarchical fuzzy clustering

#### Challenges in handling text data:

- Unstructured in nature
- Ambiguous
- Multilingual
- High computational cost (in terms of time as well as memory)

TABLE I. TABLE STYLES

Clustering Approach	Characteristics
Hierarchical	Able to handle tree structure, Data can be extracted at different levels
Fuzzy	Deals with ambiguous/overlapped data, Able to assign single object to multiple clusters
Hierarchical Fuzzy	Characteristics of both the approaches (Hierarchical and Fuzzy) as mentioned above

In hierarchical fuzzy clustering, rough clustering can be used as an alternative to fuzzy clustering which will be helpful to reduce the computational cost. By using this information regarding hierarchical fuzzy text clustering, research in this field can be extended in different directions such as social text analysis, online recommendations, news extraction, research article extraction, etc.

### IV. CONCLUSION

Voluminous data is available everywhere due to presence of large number of Internet and computer application users. This requires advancement in clustering techniques. Hence hierarchical fuzzy text clustering is proposed which possess basic concepts in hierarchical clustering and fuzzy clustering.

Due to fuzziness and hierarchical structure of the text data hierarchical clustering and fuzzy clustering is advantageous. This algorithm is also useful for tasks in web mining such as social data (text data such as blogs, messages, etc.) analysis in future.

### REFERENCES

- [1] Margaret H. Dunham, "Data Mining Introductory and Advanced Topics," Pearson Education 2006.
- [2] Xu, Don Wunsch, "Clustering. IEEE Press Series on Computational Intelligence," John Wiley & Sons, INC. Publication 2008.
- [3] Seema V. Wazarkar and Amrita A. Manjrekar, "Text clustering using HFRECCA and rough K-means clustering algorithm," In International Conference on Advances in Computer Engineering & Applications 2014; vol. 15, no. 40.
- [4] Andrew Skabar and Khaled Abdalgader, "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm," IEEE Transactions on Knowledge and Data Engineering 2013; 25(1): 62-75.
- [5] Shady Shehata, Fakhri Karray and Mohamed S. Kamel, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering," IEEE Transactions On Knowledge And Data Engineering 2010; 22(10): 1360-1371.
- [6] Claude Pasquier, "Task 5: Single document keyphrase extraction using sentence clustering and Latent Dirichlet Allocation," In Proceedings of the 5<sup>th</sup> International Workshop on Semantic Evaluation, ACL 2010; pp. 154-157.
- [7] Johanna Geiß, "Latent semantic sentence clustering for multi-document summarization," Technical Report from University of Cambridge 2011.
- [8] Martina Naughton, Nicholas Kushmerick and Joe Carthy, "Clustering sentences for discovering events in news articles," In ECIR 2006; pp. 535-538.
- [9] N. Rajalingam and K. Ranjini, "Hierarchical Clustering Algorithm - A Comparative Study," International Journal of Computer Applications 2011; 19(3): 0975 – 8887.
- [10] Yi Guo, Zhiqing Shao and Nan Hua, "A Hierarchical Text Clustering Algorithm with Cognitive Situation Dimensions," In 2<sup>nd</sup> IEEE International Workshop on Knowledge Discovery and Data Mining 2009.
- [11] Moshe Looks, Andrew Levine, G. Adam Covington, Ronald P. Loui, John W. Lockwood, Young H. Cho, "Streaming Hierarchical Clustering for Concept Mining," In IEEE Aerospace Conference 2007; pp. 1-12.
- [12] Rajesh N. Davé and Sumit Sen, "Robust Fuzzy Clustering of Relational Data," IEEE Transactions on Fuzzy Systems 2002; 10(6): 713-727.
- [13] Jiabin Deng, Juanli Hu, Hehua Chi, Juebo Wu, "An Improved Fuzzy Clustering Method for Text Mining," In 2<sup>nd</sup> International Conference on Networks Security Wireless Communications and Trusted Computing 2010; pp. 65-69.
- [14] Paolo Corsini, Beatrice Lazzarini, and Francesco Marcelloni, "A Fuzzy Relational Clustering Algorithm Based on a Dissimilarity Measure Extracted From Data," IEEE Transactions on Systems, Man, and Cybernetics 2004; 34(1): 775-781.
- [15] Raghu Krishnapuram, "Low-Complexity Fuzzy Relational Clustering Algorithms for Web Mining," IEEE Transactions on Fuzzy Systems 2001; 9(4): 595-607.
- [16] M.-S. Yang, "A Survey of Fuzzy Clustering" Mathematical and Computer modelling 1993; 18(11): 1-16.

# Fuzzy-knowledge-inferred Edge Directed Image Enhancement Technique

Reshmalakshmi C.

Research Scholar, Kerala University  
Assistant Professor, Dept. of Electronics  
Marian Engineering College, Trivandrum  
reshmasujeer@gmail.com

Sasikumar M.

Head of Department, Dept. of Electronics  
Marian Engineering College, Trivandrum

**Abstract**—In this paper, we propose a novel image interpolation technique called fuzzy-knowledge inferred edge directed image interpolation termed as FEDI for image enhancement. We use similarity measures to capture the neighborhood characteristics of each pixel to estimate the edge and orientation of the image pixels. The missing pixels are then calculated by means of estimated pixel knowledge. Compared to new edge directed interpolation (NEDI) technique and other versatile approaches, proposed method requires less computational effort and provides relatively good image quality with higher PSNR and SSIM. Extensive experimental results on benchmark images demonstrated with subjective and objective tests reveal that the proposed FEDI technique achieves elevated performance in image upscaling and image enhancement.

**Keywords**—edge detection; fuzzy logic; image enhancement; image interpolation; orientation estimation

## I. INTRODUCTION

Imagerescaling technique or image interpolation is one of the active challenging areas of image processing. Interpolation technique is indispensable not only because of the widespread resolution enhancement necessity in the field of digital photography, medical imaging, consumer electronics, computer vision and remote sensing, but also the importance of identifying missing pixels in the image. The basic idea of interpolation technique is to obtain a high-resolution (HR) image from down-sampled low-resolution (LR) image with preserved image quality [1]. Researchers have come up with various interpolation methods such as non-adaptive approach [3] and [4], learning-based approach [17] - [21], and edge directed approach [5]- [12]. Non-adaptive approaches, such as bilinear, bicubic [3], or cubic spline method [4], generally use a polynomial based kernel for the entire image to predict the unknown pixel values. These methods are computationally simple and relatively less complex. Unfortunately, this technique tends to produce blurred edges and annoying visual artifacts. Whereas, learning-based approach [17] -[20] exploits relation between HR and LR image patches to interpolate the image details. A dataset of HR and LR image patches are generated from edge details and are then used to recover HR image of given LR image. Even though learning approach shows improved performance [21] in interpolation, the use of dataset makes it computationally expensive and less reliable.

On the other hand, edge-directed interpolation [5]-[12] use different methods to capture the edge information and then tune a function to interpolate unknown pixels. Estimation of edge orientation [5], [6] from detected edge direction helps to interpolate unknown pixel values. One of the explicit methods called NEDI [7] makes use of covariance of LR image to reconstruct HR image. But inaccurate estimation of edge direction and small edges in the image makes an inaccurate estimation. To improve the performance of NEDI, Tam et al. suggested modified NEDI (iNEDI) [8] by adaptively varying the image window size. An iterative curvature based interpolation (ICBI) method is discussed in [9] which use the second order derivative of the image to design and estimate the unknown pixel. The use of curved edge in this approach helps to produce a perceptually appealing image. The least square error norm is replaced by moving least square error norm in regularized local linear regression (RLLR) algorithm [10] proposed by Liu et al. The RLLR algorithm results in sharp edges which sometimes results over enhancement. Minimum mean squared error and directional estimation are combined to interpolate unknown pixel value in directional filtering and data fusion (DFDF) method [11]. However, DFDF fails to recover the small-scale edges in the image. By combining the principle of local adaptation and global consistency for better performance, a Bayesian inference technique is projected in [12]. The quality of all edge directed interpolation methods depend on the edge detection and are still a critical challenge in image processing.

Uncertainty occurs when an image is downsampled to construct LR image. Proper mapping of this uncertainty leads to an efficient upscaling technique. It has been also proved that fuzzy-knowledge based algorithm successfully models uncertainty in image processing [2]. Edge-directed interpolation techniques make use of image edges or other image features to find the missing pixels. Hence, strong and weak edge identification is a critical step. Explicit edge detection methods have been proposed in the fuzzy logic domain [13]-[15] to map uncertainty even under high noise level. An edge-adaptive algorithm called edge-based line average (ELA) designed using fuzzy inference approach [14] reinforce edges without much increase in computational complexity. Fuzzy direction oriented interpolation [15] designed on Sugeno model outperforms the de-interlacing system; nevertheless, the use of dataset increases

the design complexity. Furthermore, ELA uses the weighted averaging technique to interpolated unknown pixels in the image.

The motivation of this work is to make use of fuzzy inference based edge and orientation detection method in the area of edge directed image interpolation technique for better image quality. The main features of our method are that we make use of image pixel gradients in nine different directions and modified bilinear function to significantly improve the computational efficacy and interpolated image quality. Experimental results show that FEDI outperforms existing image interpolation methods in subjective as well as objective measurements.

The rest of the paper is organized as follows: Section II introduces the proposed FEDI approach. The implementation results along with evaluation measures are presented in Section III. Finally, Section IV draws a conclusion.

## II. FUZZY-KNOWLEDGE-INFERRED EDGE DIRECTED INTERPLATION

The input image obtained by downsampling the reference image [7] - [11] may be affected by uncertainty. This section presents a proposed technique called FEDI that improves PSNR and image quality compared to state-of-art interpolation methods including NEDI approach. The flow of proposed approach is shown in Fig. 1. Initially, edge detection is performed in the fuzzy-knowledge domain with a threshold value selected in such a way that weak edges are also identified to improve the image quality. Secondly, a vague orientation of edge pixel is calculated using eight gradients obtained from the neighborhood characteristics. A fuzzy inference evaluation framework is introduced to find out both edge and orientation of image pixel. Finally, the unknown pixels are calculated by employing different manipulation functions separately for both smooth and sharp region of the image.

### A. Edge Detection method

An edge is a result of a change in illumination intensity in an image. The edge is clearly seen when luminance difference is above a predefined level. False or inaccurate detection of an edge on a fixed threshold causes artifacts or even poor interpolation. Alternatively, fuzzy approach applies heuristic knowledge to overcome these drawbacks. Fuzzy rules are applied on image gradients calculated in four directions; horizontal, vertical and diagonal on the neighboring pixels. So, the first step is to find horizontal, vertical, and two diagonal gradients denoted as  $G_x$ ,  $G_y$ ,  $G_l$  and  $G_r$  for each pixel in the image using equation (1). Where  $I(i, j)$  is the intensity of pixel at coordinates  $(i, j)$ . All the edge pixels (both strong and weak edges) other than smooth regions in the image are identified and mapped to varying degree of membership function depending on the edge strength. The antecedents (four

gradients) and consequent (edge) of fuzzy edge detection approach are characterized by zero-mean Gaussian and triangular membership function respectively. These membership grades and linguistic rules detect edges of an

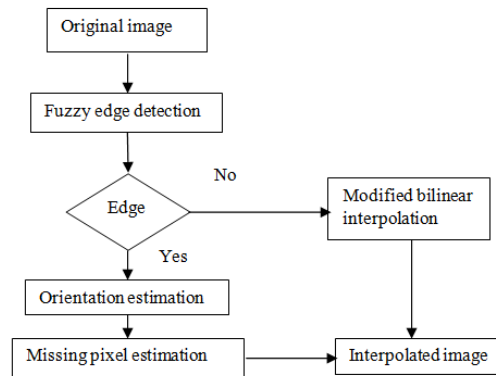


Fig. 1. Proposed workflow

image in a better way even in the presence of noise.

### B. Orientation Estimation method

Existing explicit methods propose precise orientation estimation of each edge pixel for image interpolation. Direct use of these orientations limits the interpolation process. Thus we propose to utilize the fuzzy inference based orientation estimation. In which, instead of finding the exact orientation of each edge pixel, vague orientation is calculated from gradients of a neighboring pixel. More clearly, the eight gradients denoted as  $G_x$ ,  $G_{2r}$ ,  $G_r$ ,  $G_{6r}$ ,  $G_y$ ,  $G_{6l}$ ,  $G_l$ ,  $G_{2l}$  which are located at eight different angles such as  $0^\circ$ ,  $22.5^\circ$ ,  $45^\circ$ ,  $67.5^\circ$ ,  $90^\circ$ ,  $112.5^\circ$ ,  $135^\circ$ ,  $157.5^\circ$  respectively are calculated using equations (1) and (2) for each edge pixels in the image. Among eight gradients and angles, the angle of the maximum gradient is selected as the orientation of the edge pixel. The arrangement of calculation of eight gradients at a different pixel location is shown in Fig. 2. The use of fuzzy orientation estimation reduces the staircase effect in edge directed image interpolation. Furthermore, proposed interpolation technique calculates the orientation of edge pixel which has a maximum deviation from the center pixel. Pixel-wise orientation estimation is done for the entire image edges to get a better understanding of its neighboring pixel characteristics.

$$\begin{aligned}
 G_x(i, j) &= I(i, j-1) - I(i, j) \\
 G_y(i, j) &= I(i-1, j) - I(i, j) \\
 G_l(i, j) &= I(i-1, j-1) - I(i, j) \\
 G_r(i, j) &= I(i+1, j+1) - I(i, j) \quad (1)
 \end{aligned}$$

### C. Proposed FEDI method



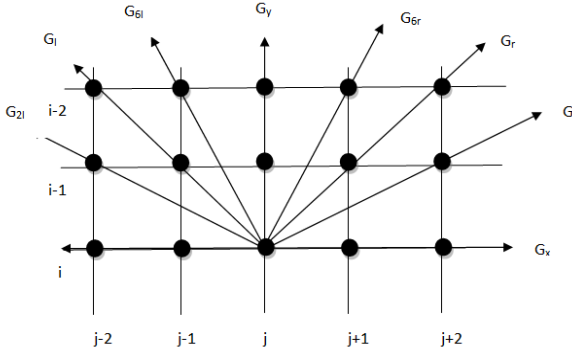


Fig. 2. Gradient calculation arrangement

As discussed in the above section, the strength of image edge obtained through image gradients and orientation helps to identify the path and direction of interpolation. The original LR image is decomposed into the sharp and smooth region before interpolation. FEDI method proposes to employ the idea of conventional and rhombic bilinear interpolation technique to find the unknown pixels. To estimate the unknown pixels in the smooth region, we start by finding the middle pixels ( $m$ ) by utilizing the four neighbors of known adjacent pixels in a conventional method. Then the vertical and horizontal pixels ( $v$  and  $h$  respectively) are calculated from newly generated ' $m$ ' pixels and known adjacent pixels. This arrangement is shown in Fig. 3. The calculation of pixels to be interpolated is illustrated in the following equations (3)-(5).

$$\begin{aligned}
 G_{2l}(i, j) &= I(i-1, j-2) - I(i, j) \\
 G_{2r}(i, j) &= I(i-1, j+2) - I(i, j) \\
 G_{6l}(i, j) &= I(i-2, j-1) - I(i, j) \\
 G_{6r}(i, j) &= I(i-2, j+1) - I(i, j) \quad (2) \\
 m(x, y) &= (1-\Delta v)(1-\Delta h)I(i-1, j-1) + (1-\Delta v)(\Delta h)I(i-1, j) \\
 &+ (\Delta v)(1-\Delta h)I(i, j-1) + (\Delta v)(\Delta h)I(i, j) \quad (3) \\
 v &= (1-\Delta v)(1-\Delta h)m(x, y) + (1-\Delta v)(\Delta h)I(i, j) \\
 &+ (\Delta v)(1-\Delta h)I(i-1, j) + (\Delta v)(\Delta h)m(x+1, y+1) \quad (4) \\
 h &= (1-\Delta v)(1-\Delta h)I(i, j-1) + (1-\Delta v)(\Delta h)m(x+1, y) \\
 &+ (\Delta v)(1-\Delta h)m(x, y) + (\Delta v)(\Delta h)I(i, j) \quad (5)
 \end{aligned}$$

Where  $\Delta v$  and  $\Delta h$  represent the vertical and horizontal distance between known and unknown pixels (the minimum distance of known neighboring pixel is set as 1). The pixel estimation utilizing conventional and rhombus-shaped pixels, protect the texture quality of the image while interpolation. Whereas, to estimate unknown pixel in the sharp region, FEDI employ a gradient-based pixel manipulation function. The interpolation along eight gradients measured between pixels located at eight angle helps to identify the characteristics of missing pixel in the interpolation process. More clearly, as illustrated in Fig. 3, pixels to be interpolated are located on the line of different angle drawn from center edge pixel. For example, if any of measured gradients is comparatively low, it means that the pixel to be interpolated is required to have the same property of pixels of the particular gradient. A gradient

matrix ' $G$ ' is formed from gradients to interpolate the unknown edge pixels. Use of these gradients detracts the knowledge of the orientation of edge pixel which shows the direction of interpolation. Let ' $p$ ' denotes the pixel to be interpolated and  $G_l$  represents the image matrix that takes part in image interpolation process, the following equations (6) and (7) helps to interpolate edge pixels by utilizing the idea of gradient scaling. The image interpolation size is selected to  $3 \times 5$  which makes it easier to estimate the unknown pixels.

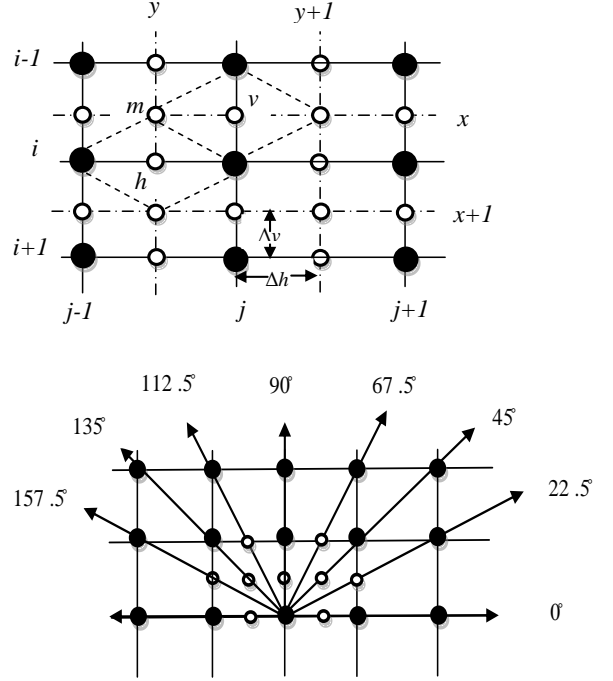


Fig. 3. Interpolating smooth (above) and sharp (below) region of image pixels: '●'-known pixel '○'- pixel to be interpolated calculation arrangement

$$p(x, y) = I(i, j) * G(i, j) + G_l(6)$$

$$G = \frac{1}{9} \begin{bmatrix} 0 & G_{6l} & 0 & G_{6r} & 0 \\ G_{2l} & G_l & G_y & G_r & G_{2r} \\ 0 & G_{x-1} & 0 & G_x & 0 \end{bmatrix} \quad (7)$$

The use of square and rhombic bilinear interpolation kernel and gradient matrix significantly improves the interpolated image quality and reduces the computational complexity compared to other interpolation techniques.

### III. EXPERIMENTAL RESULTS

In this section, experimental results are presented to evaluate the performance of proposed interpolation technique called FEDI using 10 commonly used test images *Baboon*, *Cameraman*, *Clock*, *Doll*, *Glaze*, *Lena*, *Peppers*, *Tiffany*, *Wheel*, and *Zebra*. Several of these test images are selected from online archive and others are classic test images. The RGB test images with 8 bit per channel as shown in Fig. 4 are

down-sampled by a factor of 2 to obtain LR image as in [7]-[11] and [16]. It is impossible to select one interpolation technique among numerous image interpolation techniques developed during recent decades. Hence, the proposed FEDI technique is compared with state-of-art image interpolation techniques such as Bicubic interpolation [3], new edge directed interpolation (NEDI) [7], the improved NEDI [8], directional filtering and data fusion [11] and iterative curvature based interpolation [9]. Among all these techniques, NEDI and DFDF sound better. Built-in MATLAB function is



Fig.4.10 test images. From left to right and top to bottom: *Baboon, Cameraman, Clock, Doll, Glaze, Lena, Peppers, Tiffany, Wheel, and Zebra*

used to perform bicubic interpolation and source codes of other interpolation techniques are provided by respective authors. Both objective and subjective evaluation framework and algorithm implementation for abovementioned interpolation techniques are performed on the MATLAB.

### A. Objective Evaluation

To evaluate the objective quality of interpolated images, we propose to calculate average quality metric peak signal-to-noise ratio (PSNR) and perceptual quality metric structural similarity (SSIM) [17]. The image to be interpolated (LR) is obtained by down-sampling the reference image (HR) by a factor of 2 without anti-aliasing. Image upscaling is performed by implementing the algorithm on all the three RGB channels and comparison is made between upscaled and the reference image.

All the test images are enlarged to  $2 \times$  factors and TABLE I summarizes the PSNR and SSIM of interpolated images. From the table, it is observed that PSNR of the proposed method is improved considerably. Proposed method exactly outperforms the local data fusion method DFDF by average PSNR gain of 6.6883dB. Compared to NEDI and iNEDI, FEDI achieves higher average PSNR of 1.1344dB and 1.1135dB respectively. Among all the interpolation techniques; Bicubic method, the globally accepted

interpolation technique, performs better than explicit edge directed approaches such as NEDI and improved NEDI. Nevertheless, proposed FEDI approach significantly improves PSNR of interpolated image by 0.0310dB compared to Bicubic. Furthermore, the average structural similarity measure SSIM is significantly higher for proposed FEDI method compared to all other higher values of PSNR and SSIM demonstrate that the proposed FEDI approach can effectively preserve both texture and details of the image during interpolation.

### B. Subjective Evaluation

Objective measures help to ensure the quality of proposed method among other interpolation techniques. But it is necessary to evaluate perceived quality of the interpolated image to have a better understanding. Fig. 5 illustrates the perceived quality of the enlarged image on different approaches. It is observed that Bicubic interpolation introduces blurred and jaggy edges that annoy visualization. On the other hand, the visualization quality of the enlarged image is improved in NEDI method. For few images, DFDF performs even better than NEDI. However, the reconstructed image quality is comparatively lower than proposed FEDI approach. The ICBI method ensures better-perceived image quality in the sharp and smooth region of the rescaled image. Proposed FEDI shows better visual improvements in the sharp regions compared to existing methods such as ICBI. So it is clear that perceived quality of FEDI interpolated image is superior to existing edge directed interpolation methods.

TABLE I. COMPARISON ON OBJECTIVE MEASURES OBTAINED FOR DIFFERENT METHODS.

Images	Measures	Bicubic [3]	NEDI [7]	iNEDI [8]	ICBI [9]	DFDF [11]	Proposed FEDI
Baboon	PSNR	18.3507	17.3301	17.5420	<b>19.4997</b>	16.9828	18.5352
	SSIM	0.711	0.6494	0.6532	0.7314	0.6637	<b>0.7379</b>
Cameraman	PSNR	22.3695	21.7337	21.8224	22.6866	17.1176	<b>22.7006</b>
	SSIM	0.7906	0.7594	0.7619	0.8077	0.6489	<b>0.8502</b>
Clock	PSNR	27.1640	26.3521	24.833	24.4691	18.0794	<b>29.2750</b>
	SSIM	0.9386	0.9226	0.9024	0.9369	0.8507	<b>0.9556</b>
Doll	PSNR	<b>27.0170</b>	26.7634	26.9888	22.5723	15.0019	25.9991
	SSIM	<b>0.9088</b>	0.9049	0.9018	0.9033	0.7339	0.9078
Glaze	PSNR	17.2249	16.8542	16.9337	<b>17.3525</b>	11.8032	16.4154
	SSIM	0.7548	0.7206	0.7441	0.7118	0.5305	<b>0.7633</b>
Lena	PSNR	22.1171	22.6937	22.861	23.4691	15.4687	<b>23.8254</b>
	SSIM	0.9091	0.9181	<b>0.9610</b>	0.9169	0.7913	0.9218
Peppers	PSNR	32.1377	29.5664	29.995	29.264	24.6545	<b>32.2185</b>
	SSIM	0.9850	0.9751	0.9801	0.9702	0.9652	<b>0.9851</b>
Tiffany	PSNR	<b>27.5456</b>	26.151	26.226	27.1786	19.873	27.537
	SSIM	<b>0.9624</b>	0.9496	0.9499	0.9561	0.9140	0.9622
Wheel	PSNR	<b>22.1703</b>	20.4676	21.0112	20.1547	14.7791	20.0100
	SSIM	<b>0.8146</b>	0.7364	0.7891	0.7851	0.5509	0.7011
Zebra	PSNR	<b>19.5863</b>	16.7373	16.6457	18.005	15.3501	19.4772
	SSIM	<b>0.8317</b>	0.7005	0.7017	0.8110	0.7208	0.8224
Average	PSNR	23.5683	22.4649	22.4858	22.4651	16.9110	<b>23.5993</b>
	SSIM	0.8606	0.8236	0.8361	0.8547	0.7389	<b>0.8607</b>



Fig. 5 Comparison of different edge directed image interpolations with 2x factor for some test images: From left to right: LR image, Bicubic, NEDI, iNEDI, DFDF, ICBI and proposed FEDI.

#### IV. CONCLUSION

This paper proposed an explicit edge directed image interpolation technique called FEDI by utilizing the idea of fuzzy inference. Both the edge detection and orientation estimation performed in fuzzy domain improved the quality of interpolated image compared to the existing techniques. Extensive experimental results on benchmark test images conveyed the significance of proposed method over the state-of-art interpolation techniques in terms of both subjective and objective quality. In the future, we preferred to add an optimization technique to improve the interpolated image quality. Furthermore, we want to speed up the technique for better enhancement.

#### REFERENCES

- [1] A. K. Katsaggelos, R. Molina, and J. Mateos, *Super Resolution of Images and Video*. San Rafael, CA: Morgan & Claypool, 2007.
- [2] E.E.Kerre and M.Nachtegael., *Fuzzy Techniques in Image Processing*, Physica- Verlag. A springer- Verlag Company, 2000.
- [3] R.G. Keys, 'Cubic convolution interpolation for digital image process', *IEEE Trans. Acoust., Speech Signal Process.*, vol.29, no. 6, pp. 1153-1160, Dec. 1981.
- [4] H.S. Hou and H. Andrews, 'Cubic splines for image interpolation and digital filtering', *IEEE Trans. on Acoust.,Speech Signal Process.*, vol.26, no. 6, pp. 508-517, Dec-1978.
- [5] Q. Wang and R. K. Ward, 'A new orientation- adaptive interpolation method', *IEEE Trans. Image Process.*, vol.16, no.4, pp.889-900, Apr.2007.
- [6] D. D. Muresan, 'Fast edge directed polynomial interpolation', *Proc. IEEE Int. Conf. Image Process.(ICIP)*, vol. 2. Genova, Italy, Sep. 2005, pp. II-990-II-993.
- [7] X. Li and M. T. Orchard, 'New edge-directed interpolation', *IEEE Trans. Image Process.* vol.10.no. 10, pp.1521-1527, Oct.2001.
- [8] W.S. Tam, C.W. Kok, and W. C. Siu, 'Modified edge - directed interpolation for images', *J. Electron. Image.* vol.19, no.1, pp.013011-1-013011-20, Jan./Mar. 2010.
- [9] A. Giachetti and N. Asuni, 'Real-time artifact-free image upscaling', *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 413- 423, Jan. 2014.
- [10] X. Liu, D. Zhao, R. Xiong, S. Ma, W. Gao and H. Sun, 'Image interpolation via regularized local linear regression', *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3455-3469, Dec-2011.
- [11] L. Zhang and X. Wu, 'An edge guided image interpolation algorithm via directional filtering and data fusion', *IEEE Trans. Image Process.*, vol. 15, pp. 2226-2238, Aug. 2006.
- [12] X.Liu, D.Zhao, J.Zhou, W.Gao, and H.Sun, 'Image interpolation via graph-based Bayesian label propagation', *IEEE Trans. Image Processing*, vol.23, no.3, pp.1084-1096, Mar.2014
- [13] C. Tao, W. Thompson, and J. Taur, 'A fuzzy if-then approach to edge detection' *Proc. IEEE 2nd Int. Conf. Fuzzy Systems*, vol.2, pp. 1356-1360, 1993.
- [14] P.Brox, I.Baturone, S.S-Solano and J.G-Rios, 'Edge- Adaptive Spatial Video De-interlacing Algorithm Based on Fuzzy Logic' *Proc. IEEE Int. Conf.*, pp. 375-385, 2014.
- [15] G. Jeon and J. Jeong, 'Designing Takagi- Sugeno Fuzzy Model-Based Motion Adaptive Deinterlacing System', *Proc. IEEE Int. Conf.*, pp. 1013-1023,2006.
- [16] Z.Wang, A.Bovik, H. Sheikh, and E. Simoncelli, 'Image quality assessment: from error visibility to structural similarity', *IEEE Trans. Image Process.* vol.13, no.4 pp.600-612, Apr.2004.
- [17] D. Glasner, S. Bagon, and M. Irani, 'Super-resolution from a single image', *Proc. IEEE Int.Conf. Comput. Vis. (ICCV)*, Kyoto, Japan, Sep/Oct. 2009, pp. 349-356.
- [18] G. Freedman and R. Fattal, 'Image and video upscaling from local self-examples', *ACM Trans. Graph.* vol.30, no. 2, 2011, Art.ID 12.
- [19] K.S.Ni and T.Q.Nguyen, 'Image super-resolution using support vector regression', *IEEE. Trans. Image Process.*, vol.16, no.6, pp.1596-1610, Jun.2007.
- [20] L.Wang, and S.Xiang 'Fast direct super-resolution by simple functions,' *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sydney, NSW, Australia, Dec.2013, pp.561-568.
- [21] C.Y.Yang, M.H.Yang, G.Meng, H.Wu and C.Pan, 'Edge directed single image super-resolution via adaptive gradient magnitude self-interpolation,' *IEEE Trans. Cir. Sys. For Video Tech.*, vol.23, no.8, Aug.2013, pp.1289-1299.

# IoT Based Charging Station for Electric Vehicles

Chellaswamy C, Archana V, Arunraj J, Bhagirathi S  
Professor, Department of ECE, Rajalakshmi Institute of technology  
Chennai, India

chella\_info@yahoo.co.in, Indiaarchanavenkateshwaran@gmail.com, aarunraj90@gmail.com,  
Indiabhagirathisivamani@gmail.com

**Abstract**—Electric vehicles are becoming popular nowadays as they could help us to reduce the power consumption of a device, pollution free, and saves energy. The main drawback of electric vehicle (EV) is its storage system and can't able to run the vehicle for longer distance in the same charge. In this paper an internet of things based charging station (IoT-CS) for electric vehicles is proposed. The supply is generated using the renewable sources such as wind and solar photovoltaic (PV) modules. The proposed system considers the travel success ratio, travel pattern, and the allocation of charging station for both the inside city travel and the urban travel. The IoT-CS automatically updates the status of the charging station in the cloud. The EVs can access the data from the cloud and decides which station is suitable to recharge for the remaining charge present in the vehicle. The simulation result has been done and it shows that the IoT-CS can easily accommodate in EVs thus increasing its usage and creating pollution free environment.

**Keywords**—Internet of things, charging station, electric vehicle, renewable energy

## I. INTRODUCTION

The transportation sector faces various challenges such as energy demand, an increase of oil price, and pollution in the environment. To reduce the pollution created by the transportation, it is necessary to increase the usage of EVs. To recharge the battery packs of EVs recharging stations are familiarly used. These stations can be powered with the help of renewable sources such as solar PV array and wind energy systems. An autonomous driverless vehicle usually accesses an automatic recharging station using D-lite path algorithm and vector field histogram (VFH). The D-lite algorithm is a heuristic search based on partial information is known environment VFH utilizes a statistical representation of robots environment through histogram. Once the vehicle goes into the vicinity automatically it will identify the recharging station. The vehicle is self-adjusted to an angle of 360 degrees to reach the station from the point of vicinity [1]. Plug in a hybrid vehicle cannot go the longer distance in a single charge and it leads to switching to fuel. Fuel creates more an environmental pollution and to overcome this problem satisfiability modulo theory is used for efficient navigation management [2, 3]. Plug-in Hybrid and Electric Vehicles (PHEV) is introduced in urban and suburban areas. Electric grid distribution system is used to provide required energy and power to PHEVs and study the impact. An optimal energy management strategy used for the recharging stations which are built in the urban and suburban area. Securing the data from unauthorized users by encryption and decryption under

the IOT domain is proposed by Morris et al. [4]. Cheaper modules that offer higher security have begun to replace the expensive ones. Through the information cannot be decrypted, it can be modified or deleted by the unauthorized users. There are enough mechanisms to improve the security of the data when external systems are willing to apply them [5, 6].

Interoperability is the ability to exchange usable data between two systems and to invoke their services using proper input parameters. Interoperable adaptation has been given as a key solution for dynamism in semantic interoperability in IoT. Smart office automation has been given as an example for the architecture of interoperable adaptation. Swarup et al. suggest an idea on building automation prototype and addresses the issues of scalability [7]. Crowd dynamics management in IoT helps in the congestion management of transportation. It also focuses on the applications such as transportation, shopping, and disaster evacuation. From the feedback factor, the output of the crowd dynamics management is effectively used [8]. Monitoring systems which use IoT can continuously gather data and is comprised of 4 layers: sensing, network, application, and service resource. This system integrates the data processing techniques for the production of steel cast. Heterogeneous data and multiple communication protocols are dealt in a large amount in data processing. The hardware and software capabilities can be integrated more flexible in this architecture [9]. Human beings waste a lot of energy resources in daily life. Some applications have energy control but it is not very efficient. An intelligent energy control system is used to control the peak-time, energy limit, and user control of an IoT based wireless smart socket. This approach saves up to 43.4% of energy for one day [10]. IoT in industries allows ubiquitous interaction and collaborative automation between machines. Efficient data interactions could be hindered by the technical gap between practical machine operation and IOT technique. Data-oriented system architecture towards flexible data interaction between machines, customized M2M protocol, flexible data structure, and information tracking is the major technical problems to investigate [11].

The proposed work comprises of two phases: the primary stage is demonstrating EV trip habitude, trailed by displaying the electrical energy present in the EV Batteries at the start of every outing. Thus, Service Range of Charging Station (SRCS), (which is the range that can be covered by the EV) and the trip success model (TSM) of each trip to evaluate the model. The evaluation technique can be accomplished by considering the volatility of PEV trip practices and the volatility of the electrical energy accessible in EVs batteries. The second stage is to pick the best CS set considering the

evaluated SRCS of the TSR levels from the initial stage. The designating optimizing issue will be made as the Maximum Coverage Problem (MCP) and the displayed model for giving-out CSs will be in standard form, so it ought to be compatible for various transportation systems (in-city ride and out-city ride), and evaluation reports of contrasting conditions will be displayed for various system formats. An Android based application has been created and the proposed system is verified.

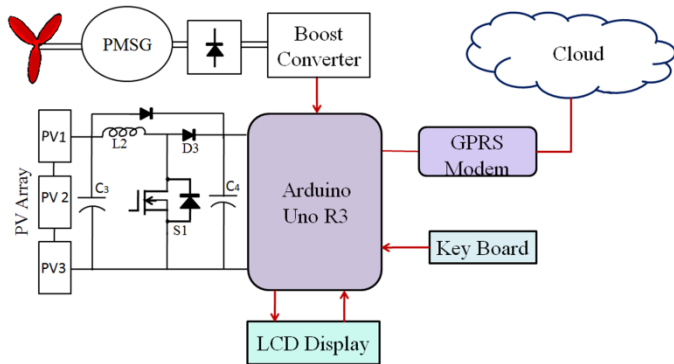


Fig. 1. Configuration of the proposed charging station

The principal responsibilities of this paper are encapsulated below:

- The proposition of an allocation model of the charging stations should satisfy the drivers' requirement: The displayed model for giving-out CSs will be in generic form, so it ought to be compatible for various transportation systems such as in-city and out-city rides.
- The evolution of the TSM Model demonstrated in the lights of MCP: The displayed model will be a beneficial system for allocating purposes, as well as to evaluate the existing CS from a driver's perspective.
- Demonstrating the assortment of outing mileages and the vitality of staying electric extents: Travel overview information is exploited to assuage the truancy of genuine PEV driving data that won't be accessible preceding evident PEV entrance levels. The inflation of information that indicates trip purposes and client disposition will upgrade the model, enabling it to be more plausible.

The remnants of this paper is structured as follows: Section 2 describes an overview of the proposed EV charging system with trip success model and travel pattern. Section 3 describes the simulation results. Finally, conclusions are given in section 4.

## II. PROPOSED SYSTEM

The proposed system consists of a controller for monitoring and controls the remaining charge present in the EV and identifies the charging stations present along the travel area. The main focus of the study is to optimally choose the location of charging stations and estimate the successful trip based on the EV travel. The CAN bus system is used to pass

the battery status information and the command instruction between the electric vehicle charging control and vehicle battery packs [12]. The connection between the vehicle sensors and main monitoring node is done by CAN bus which is a popular vehicle control network. The speed of electric vehicles and the driving range is random and dispersed.

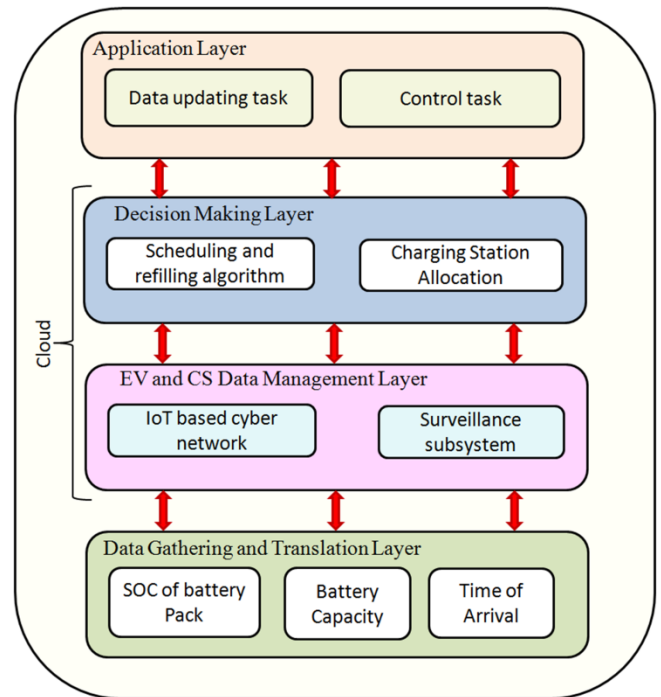


Fig. 2. Structure of IoT-based track monitoring system

The charging stations are the stations with a function similar to the petrol bank i.e. they provide the battery charging based on the capacity and they also provide the battery maintenance for electric vehicles. The block diagram of EV charging station is shown in Fig. 1. The equipment such as a charger for the battery, charging rack, battery box testing, and maintenance equipment that are typically configured. They are located usually near the charging stations. Usually, the EVs are charged in both the slow and fast charging mode. In order to bring more reliable environment perception, the perception layer collects all the information from onboard battery monitoring, vehicle sensors, and network equipment. The awareness of charging swap services to the electric vehicle user can be given by the information perception layer. The battery management system (BMS) will give the battery status parameters through the CAN bus interface during the process. The mobile wireless sensor network (WSN) in stations can be built by micropower technology and to receive context-aware telemetric services such as GPRS, CDMA, and 4G technologies can be used. The proposed architecture of IoT-CS consists of 4 layers: 1) the data gathering and translation layer 2) the EVs data management layer 3) the decision making layer for EV charging and 4) the application layer are shown in Fig. 2. The EV data management layer forms an IoT-based cyber network by connecting the controller which is present in the charging station. The information about all the charging stations is easily updated through the cyber network.

The main advantages are 1) update the information about the EVs such as SOC, arrival time, and the capacity in the cloud and the information about CS 2) The CS can understand the number of EVs arrival and the EVs can understand about the status of CS. The decision making layer decides the performance of the device and schedule for charging of EVs.

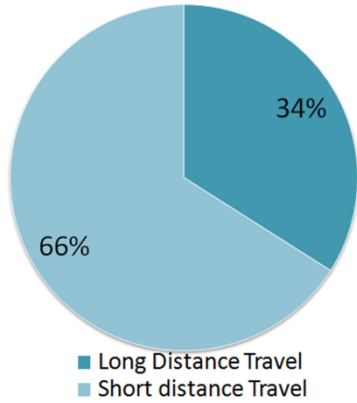


Fig. 3. The Probability of short and long travel [14]

#### A. Trip Success Model

This subdivision exhibits the trip success model (TSM) of the EVs in particular vicinity. This exhibited model evaluates the charging station arrangements in lights of two segments: the administration scope of charging stations and the trips completed effectively by the EVs. The transportation network is formed as a small cluster, and each of these modules ought to be secured by no less than one CS. The separation between CSs will be the main deliberation that impacts the level of EV trips finished adequately. Not exclusively does the separation between CSs impact the TSM of EV trips, there are likewise two different components: the EV day by day trip separations and the measure of vitality staying in the EV's battery toward the begin of each travel. Consequently, the success model has split in to exhibit the accountability of EV travel examples and EVs staying electrical vitality. Therefore, the TSM demonstrate, evaluates the satisfactory of SRCS.

#### B. Model of Travel Pattern

Here In order to create Virtual Travel distance (VTD), the survey data of travel is used to model the travel pattern. The survey data includes different trips of vehicles such as trucks, cars, etc. In this paper, the vehicle under public sector only considered in North America based on the national household survey data [13]. The data is initially collected in 2009 and a new version was conducted on 2016 [14]. The VTD can be easily obtained from the actual trip data and it can be classified into i) short trips and ii) long trips. The short trip indicated inside the city and long trip indicated the rural travel and it is classified into different time interval based on the starting time of the travel. The probability of short and long travel based on the national household survey is shown in Fig. 3. The maximum likelihood technique is used to estimate the closet data value of each classification. The probability distribution function (PDF) of the actual travel data ( $f_T$ ) with the travel distance  $D$  can be expressed as:

$$f_T(D) = \frac{1}{\sigma D \sqrt{2\pi}} \exp\left(-\frac{(\ln(D)-\mu)^2}{2\sigma^2}\right) \quad (1)$$

where  $\sigma$  and  $\mu$  are the standard deviation and mean of the PDF of actual travel respectively. The cumulative distribution function,  $F_T$  of the actual travel is expressed as:

$$F_T(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \frac{\exp\left(-\frac{(\ln(D)-\mu)^2}{2\sigma^2}\right)}{D} dD \quad (2)$$



Fig. 4. The Italian highway charging stations

#### C. Charging Station Allocation Model

In this section, charging station allocation is modeled using a maximum location coverage (MLC) problem. The charging stations of Italian highway is considered and shown in Fig. 4. The objective function for maximum coverage of charging station can be defined as:

$$OF = \text{Max} \sum_{i=1}^M DT_i B_i \quad (3)$$

where  $DT_i$  and  $B_i$  represents the demand of transportation depends on the  $i^{\text{th}}$  location and the decision variable ( $B_i = 1$  the demand of transportation covered the  $i^{\text{th}}$  location and  $B_i = 0$  it is not covered). The location constraints are

$$L_{i,j} = (|X_i^{cs} - X_j^{cs}| + |Y_i^{cs} - Y_j^{cs}|) \forall i \neq j \quad (4)$$

$$K_{i,j} = \begin{cases} 1 & \text{if } L_{ij} \leq SRCS \\ 0 & \text{if } L_{ij} > SRCS \end{cases} \quad (5)$$

$$CS_i * CS_j \leq (1 - L_{i,j}) \forall i \neq j \quad (6)$$

where  $L_{i,j}$  is the Manhattan distance between the transportation node in the network,  $K_{i,j}$  set of nodes near to the charging station,  $CS_i$  decision variable ( $CS_i = 1$  station is located at node  $i$ ,  $CS_i = 0$  station is not located in  $i$ ).

#### D. SOC Estimation

To study and understand the internal chemical process of the battery is important to estimate the SOC of the battery. The charging and discharging behavior of battery will vary depending on the temperature and loading scenario. A simplified model used to estimate the SOC based on [15, 16] and it has enough accuracy, simple, and low computational

cost. The flow diagram of the SOC estimation model is shown in Fig. 5.

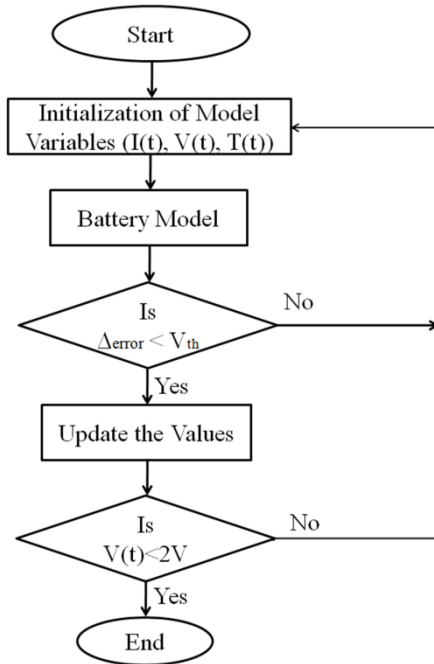


Fig. 5. Flowchart of SOC estimation

### III. SIMULATION RESULT AND IMPLEMENTATION

The CS consists of PV array, wind energy conversion system, a bidirectional DC/AC inverter to the grid, unidirectional DC-DC converters, and EV charging points. The total number of EVs charging per hour during a day is shown in Fig. 6 and the following constraints have to be considered: 1) Hourly loading is considered for each EVs 2) 2 kWh energy has been supplied to each EV for charging. The power requirement per hour can be calculated by multiplying the total number of the vehicle with the energy supplied (ie. 2 kWh × total number of the vehicle). The power requirement per day for charging EVs is shown in Fig. 7. From Fig. 7 one can easily understand that the energy requirement is less than that of the battery capacity of EVs.

In this paper both the renewable energy sources such as solar PV module and wind energy has been considered and the charging station charging profile is shown in Fig. 8. The wind energy is available throughout the day and the generation capacity depends on the speed of the wind. From Fig. 8 one can easily understand that the solar PV array [17] can generate the power from 6 am to 5 pm and the generated power is less than that of the wind-generated power. The probability distribution function of the travel distance of EV is shown in Fig. 9. Fig. 9 shows that the probability of occurrence of travel distance from 1-10 km covers 60.3%, 15-20 km covers 8%, 20-30 km around 12 %, 30-50 km around 13.5 %, 50-70 km around 3.8 %, and 70-100 km covers around 2.4 % for the proposed area.

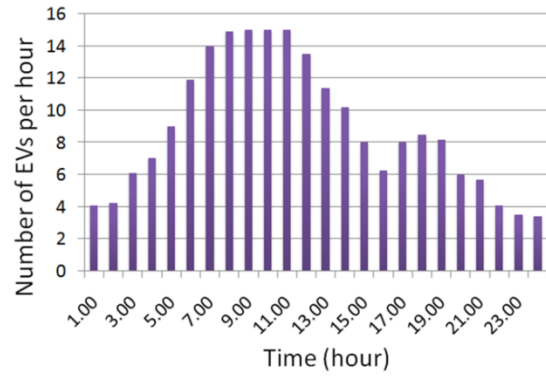


Fig. 6. Hourly charging of EVs during a day

The consequences of the TSM model, which is portrayed in the earlier section, are exhibited in this section. As appeared in Fig. 5, the virtual EV travels separations from the travel pattern model are contrasted with the electric vitality remaining assessed by the RDR model. If EV's RDR is sufficiently huge to cover the EV's VTD, it is considered as being finished effectively. If not, the EV's RDR is contrasted with the distance to the closest charging station, and the trip is considered as being finished effectively if the EV's RDR can cover the separation to the CS; generally, the trip is considered as a failed trip. MCS is used to acquire the TSR for various SRCS. The SRCS increases in predefined steps (i.e., 10 km), and the results of the MCS demonstrate the relation between the TSR and diverse SRCS.

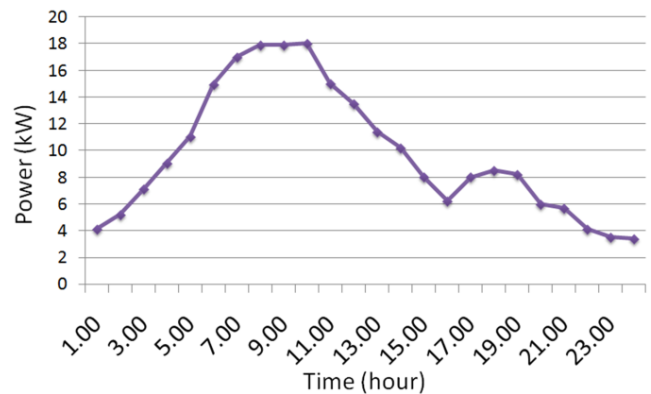


Fig. 7. Estimated power requirement for daily charging

The travel separations of PEVs that used in the trip success ratio (TSR) demonstrate are obtained from the contingent probabilities in the system utilizing MCS. Since there are two sorts of trip distance (city-highway), there are two pdfs are utilized: The virtual travel distance inside the city (VTDin-city) and virtual travel distance for out-city (VTDout-city). Table 1 demonstrates the parameters of the best-fit PDFs got from the travel design display for both the virtual distance of city and out-city. Table 1 additionally demonstrates the SOC, PDFs, and their parameters in both the city and out-city travel. The probability technique is utilized to acquire the best fitted pdfs as appeared in Table 1.



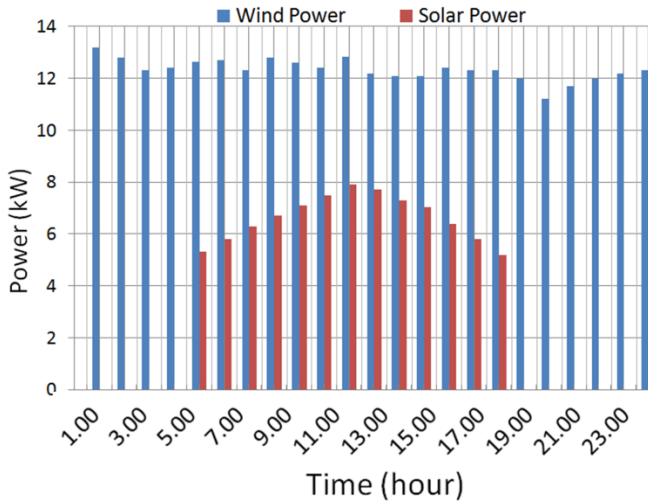


Fig. 8. Charging profile of the charging station

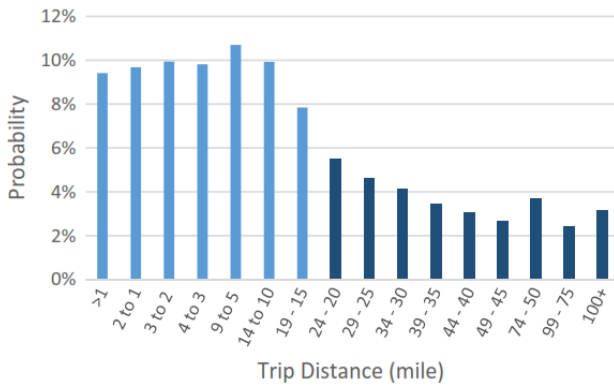


Fig. 9. probability distribution function of travel distance

The TSR and different SRCS relationships for the in-city and out-city cases are shown in Fig. 10 and Fig. 11 respectively. In the presence of a cloud network, all EVs can observe the results and at least 92% of in-city trips could be completed successfully since the in-city trip are very short and can be completed easily. However in the absence of cloud network at least 78% of all highway trips can be completed successfully by utilizing the previous data which is stored in the electric vehicles and remaining 22% cannot complete the charging.

TABLE I. THE PDF AND PARAMETER VALUES FOR DIFFERENT TSR INPUTS

Input	Pdf	Parameter values	
VTD <sub>in-city</sub>	Weibull distribution	$\alpha_1=1.7345$	$\beta_1=100.24$
VTD <sub>out-city</sub>	Weibull distribution	$\alpha_2=1.7315$	$\beta_2=97.267$
SOC <sub>in-city</sub>	Normal distribution	$\mu_1=0.5257$	$\sigma_1=0.1793$
SOC <sub>out-city</sub>	Normal distribution	$\mu_2=0.6347$	$\sigma_2=0.1684$

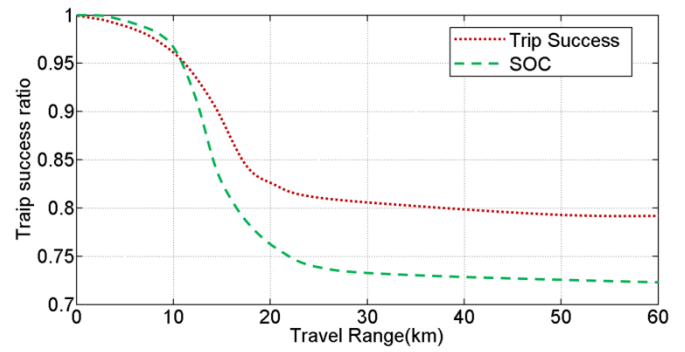


Fig. 10. Trip success and SOC for in-city travel

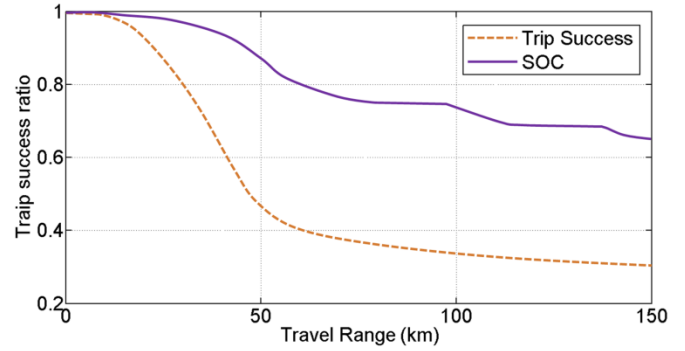


Fig. 11. Trip success and SOC for out-city travel

The National Household Survey Data were taken in the year 2009 [18], says that only 20% of daily trips are considered out-city trips and the remaining 80% of daily trips are considered in-city trips. From the above statements, we can also get the number of failed trips from the sample results. Therefore, even if the TSR level for out-city is lower than the TSR level in-city that does not mean the corresponding number of failed trips is lower. For example: Let us consider there are 700 EVs in the system and each one conducts three trips per day has been performed based on [18], there will be about 65,00 in-city trips and 250 out-city trips daily. Hence, if the out-city TSR increases from 95% to 98%, that will decrease the number of failed trips from 150 trips to only 60. However, increasing the TSR level in-city by 3% will increase the failed trips by 360, which is about four times that of the out-city ones. Thus, the TSR level in the In-city and out-city cases has different representations in terms of trip numbers (see Fig. 10).

The comparison between current electric vehicle charging method and the proposed system is carried out and the following conditions have been considered. The charging station (CS) can be divided into the three different categories such as:

Category 1: There is no communication between the charging station and the electric vehicle in this case. The function of these EV Charging stations is same as the function of the fuel charging stations that are currently in use. There is no information provided to the EV user regarding the location of the charging station.

TABLE II.COMPARISON OF CURRENT CHARGING TECHNIQUES AND THE PROPOSED METHOD

<i>Parameters</i>	<i>Category 1</i>	<i>Category 2</i>	<i>Category 3</i>
1.Communication between the EVs and the CSs	No Communication	Communication between the EVs and the CSs is present	Communication between the EVs and the CSs is present
2.Number of CSs that can be linked	No linkage since there is no communication	Only 1 CSs station can be linked to the EV at a time.	An EV can be linked to many number of charging stations at a time
3.Function	Similar to the function of petrol charging stations.	The information regarding the location of the CSs are provided to the EVs only if they are present inside the zone of that charging station.	More number of charging stations can be linked to each and every EV using IOT and cloud technology
4. Information received	No information can be received as there is no communication	Provides the location and distance of the charging station within the zone of which the EVs are present	Provides the location and distance of various charging stations and also the traffic on each of these charging stations.
5.Advantages	Less beneficial as the EV users do not get any prior information regarding the location of the charging stations.	It is more advantageous than category 1 type of CSs but less advantageous than those of category 3.	This is best suited when compared to the other two types of CSs.

Category 2: These type of charging stations provide the information to the user regarding their location. The only

exception is that the charging stations have a zone of some unit distance around them and the EVs can be linked to the Charging stations only if they are present inside that zonal area.

Category 3: In this category, the electric vehicles can be linked to the charging stations irrespective of the distance or any other factors. For this to happen, we make use of IOT based communication in which the EV users have the direct access to the clouds and can get the information of the electric vehicles location from the cloud and linkage is made possible to various charging stations.

When developing our EV charging system, Django 6 and Nginx 7 were used to create the web server. The Django is a free Python-based open source web framework. It is mainly for the rapid development of Web applications and development. Therefore, Django is used to implement the background and provide APIs that users can call. Nginx is one of the most popular open source web servers. It is also used as a load balancing web service. JSON (JavaScript Object Notation) 8 is a lightweight data exchange format. This is suitable for people to read and write. At the same time, the machine is easy to analyze and generate. Therefore, our web server provides an API with JSON data format. Our system runs on the server to provide background functionality developed by the Django Framework. The site runs on Nginx server. The Nginx server is also responsible for balancing user requests. The server receives the data from the terminal. The data is assigned to the cluster to be processed. The processing result is returned to the smartphone through the server. Our website uses MySQL as the database. It mainly contains the following relational tables.

(1) contact list: This form is used to store information from the administrator, administrator name, and password, including the administrator ID, and other administrators can manage users, systems, and online push information.

(2) user table: user table is for ordinary users. It contains information from the user, for example: as user ID, user name, user age, user's height, weight and user password. Users can manage their own personal information.

(3) task list: the table contains a specific charging methods. It contains the username, method of charging (slow, Fast), type of charging, and so on.

(4) Computer node table: The computer node is stored in this table. It contains basic information about each computer node, CPU types, such as hostname, IP address, GPU type, and memory management.

(5) log table: log table records system log. It contains usernames, charging time logs, mode of charging, task logs, and more.

In the design of our mobile application, the message push process is a very important feature of user-to-server interactions. In our system, we use the JPUSH 12 suite for push messaging. First, we need to register an account. Then download the SDK and add the library to our project and create an application. After creating the application, we get "AppKey" and "Master Secret" for clients such as smartphones, web browsers. We can also download a simple client program to test the message push service. In our system, message push is mainly used for the following aspects:

- SOC reminder: In order to complete the electric car program, our system prompts the user timely charging status.
- Health Reminder: When our system detects health risks such as storage system, engine, wheels, etc., we will remind the user.
- Charging Station reminder: this type of message from the recommendation system. Our system sends course information to similar users.
- Plugshare: This app helps you find charging stations anywhere, even internationally. It can even help you find some peculiar charging stations, including people's homes. There are many electric car enthusiasts charging their home charging station. They

often leave their cell phone numbers asking for a call or text message to let them know you're coming. Plugshare There is other non-residential sites that will surely be useful if you're away from home.

- Open the toll map: This app can also help you find the charging station. This gives detailed information about the charging station type pictures and other details so you can easily find it. This app also gives you a better direction charging station than other apps because it has a built-in navigation feature.

As more and more people buy electric cars, these applications will have to become more and more important to determine the location of electric car charging points to help people travel long distances and know they can charge cars. In 2007, Google launched Android mobile operating system. With open-source, simple and universal advantages, Android is the global leader in the smartphone market. So our application is implemented on the Android platform. This application was developed using Android Studio 11. The start page of our application is shown in Figure 12. This is the first page after the user logs in. It mainly includes five aspects, which are the five main features of our electric vehicle application.

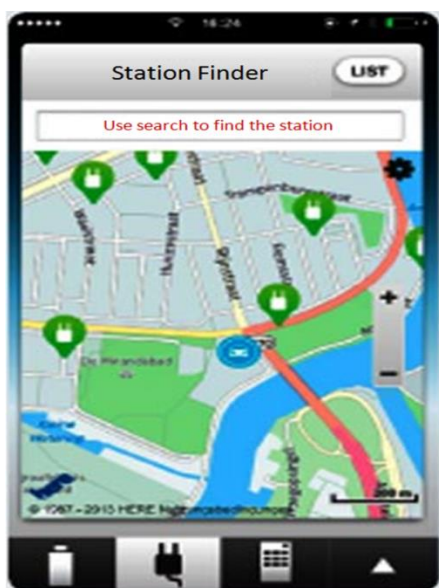


Fig. 12. Android application for charging station searching

#### IV. CONCLUSION

In this paper, a new scheme for EV charging station allocation has been proposed. The power required for charging the batteries of EVs has been done by the renewable sources. Charging station allocation scheme is considered different parameters such as battery capacity, SOC of battery, and trip classes. For batteries more than 54kWh there is no need of CSs for lesser distance and thus 97% of the trips can be finished successfully. In spite of making use of a single service range, this model helps in allocating CSs and update all the possibilities about the charging station in the cloud. This model is very helpful in measuring how the CS share the

data functions in a better way in meeting the demand of EV so as to make the suitable decision on the basis of the remaining sources. The output of this model is related to the driver need instead of electrical utilities. The IoT-CS is verified in two different scenarios such as in-city and out-city ride. The results show that the proposed IoT-CS provide better results and easy to share the information about the CS for providing good service of EVs. A mobile app has been developed for the proposed system and verified.

#### REFERENCES

- [1] Chaomin Luo, Yu-Ting Wu, Mohan Krishnan, Mark Paulik, Gene Eu Jan, Jiyong Gao, "An Effective Search & Navigation model to an Auto – Recharging station of Driverless vehicles," IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS), pp. 100-107, 2014.
- [2] Mohammad Ashiqur Rahman, Qi Duan, Ehab Al-Shaer, "Energy Efficient Navigation Management for Hybrid Electric vehicles on Highways," ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS), pp. 21-30, 2013.
- [3] Chellaswamy C, RameshR, Visveswar RauC.Y, "A supervisory control of a fuel free electric vehicle for green environment," IEEE International Conference on Emerging Trends in Electrical Engineering and Energy Management, pp. 387-393, 2012.
- [4] Morris Brenna, Federica Foiadelli, Dario Zaninelli, "Integration of Recharging Infrastructure for Electric vehicles in Urban Transportation system," pp. 1060-1064, 2012.
- [5] Juan D. Parra Rodriguez, Daniel Schreckling, Joachim Posegga, "Addressing Data Centric Security Requirements for IOT based system," 2016 International Workshop on Secure Internet of Things (SIoT), pp. 1-10, 2016.
- [6] Chellaswamy C, Ramesh R, "Future renewable energy option for recharging full electric vehicles," Renewable and Sustainable Energy Reviews, vol. 76, pp. 824–838, 2017.
- [7] Swarup Kumar Mohalik, Nanjangud C. Narendra, R Badrinath, Mahesh BabuJayaraman, Chakri Padala, "Dynamicsemanticinteroperability of control in IoT based systems: Need for adaptive middleware," IEEE 3rd World Forum on Internet of Things (WF-IoT), pp. 199-203, 2016.
- [8] Yuichi Kawamoto, Naoto Yamada, Hiroki Nishiyama, Nei Kato, Yoshitaka Shimizu, Yao Zheng, "A Feedback control based crowd Dynamics Management in IOT System," IEEE Internet of Things Journal, vol. 4(5), pp. 1466-1476, 2017.
- [9] Feng Zhang, Min Liu, Zhou Zhou, Weiming Shen, "An IOT Based online monitoring system for continuous steel casting," vol. 3(6), 1355-1363, 2016.
- [10] Kun Lin Tsai, Fang Yie Leu, Hsun You, "Residence Energy Control Slm based on wireless smart socket & IOT," IEEE Access, vol. 4, 2885-2894, 2016.
- [11] Zhipeng Wu, Zhaozong Meng, John Gray, "IoT based techniques for online M2M Interactive Itemized Data Registration and offline Information Traceability in a Digital Manufacturing system," IEEE Transactions on Industrial Informatics, vol. 13(5), pp. 2397-2405, 2017.
- [12] Chellaswamy C, Ramesh R, "Investigation of wind energy potential and electricity generation for charging the batteries of electric vehicles," ARPN Journal of Engineering and Applied Sciences, vol. 11(3), pp. 1966-1977, 2016.
- [13] Darabi Z, Ferdowsi M. "Aggregated impact of plug-in hybrid electric vehicles on electricity demand profile," IEEE Trans Sustain Energy, vol. 2(4), pp. 501-508, 2011.
- [14] Swain S. What's different in 2015–2016 compared to past NHTS?, Jan. 10, 2016. Available: <<http://nhts.ornl.gov/trb/2016/Workshop-Westat.pdf?>>[accessed:2016].
- [15] Junfu Li, Qingzhi Lai, Lixin Wang, Chao Lyu, Han Wang, "A method for SOC estimation based on simplified mechanistic model for LiFePO4 battery," Energy, vol. 114, pp. 1266-1276, 2016.

- [16] Chellaswamy Chellaiah, Balaji T.S, Muhuntharaj C. Design of a Fuel Free Electric Vehicle Using Fuzzy Logic for Pollution Control. International Conference on Modeling Optimization and Computing. Procedia Engineering, vol. 38, pp. 1547-1558, 2012.
- [17] Chellaswamy C, Ramesh R, "Parameter extraction of solar cell models based on adaptive differential evolution algorithm", Renewable Energy, vol. 97, pp. 823-837, 2016.
- [18] Administration FH. 2009 national household travel survey; 2010.

# Impact of Performance Analysis of Varied Subjects on Overall Result: An Empirical Discourse of Educational Data Mining

Mudasir Ashraf

Department of Computer Science  
University of Kashmir  
Jammu and Kashmir, India  
mudasir04@gmail.com

Dr. Majid Zaman

Department of Computer Science  
University of Kashmir  
Jammu and Kashmir, India

Dr. Muheet Ahmed

Directorate of IT&SS  
University of Kashmir  
Jammu and Kashmir, India

**Abstract**—Pedagogically to improve students' performance is primarily a challenging task as academic performance centers on various prime indicators and other miscellaneous academic factors. The association between these factors surface in an intricate way which is largely unexplored from the perspective of educational data mining procedures. Therefore, it becomes imperative to investigate and explore data generated from different educational sources using various data mining methods to unearth various hidden patterns. The main objective of this study was to analyse and identify factors from educational data which have considerable impact on improving the student's performance. Moreover, a novel attempt has been made to discover the influence of individual subjects and select demographic factors on overall performance through the application of various mining techniques vis-à-vis Linear Regression and Multiple Regression. The findings by and large have clearly confirmed that competence in English subject has paramount significance from the perspective of performance in overall result.

**Keywords:** Educational data mining, Pre-processing, Regression, Prediction, coefficient matrix.

## I. INTRODUCTION

EDM is an application of Data mining (DM) that endeavours to estimate the educational data issues by undertaking existing techniques of DM into consideration [1]. Educational data mining (EDM) contemplates to interdisciplinary study that deliberates on the development and improvement of diverse methods to ascertain the academic information produced from heterogeneous sources. The mined data can be suitable to upgrade teaching, learning experiences and accordingly aid in refining the institutional effectiveness. In yester years, it has witnessed impulsive growth in both the fields of software and databases associated with student's information which primarily signify their learning process [2], and this has proved to be a gold mine in the direction of academic research [3].

EDM is a subset of DM which encompasses the data that comes from educational background and centers on the development of various techniques and ascertaining various patterns that are exclusive in nature [4]. The identified patterns can be useful for academic stakeholders for decision making, to ameliorate student's performance and to devise healthier

teaching and learning strategies. EDM processes raw data coming from educational systems into effectual knowledge that can possibly have a considerable impact on academic strength [5]. EDM is also inexhaustible in model designs, methods and algorithms to investigate academic data [6].

To determine the student's performance is always considered a tricky task for the reason that his/her performance is based on number of parameters such as character, educational environment, demographics, emotional and other variables. The relationships among these fields are not apparently implicit as they are usually related in intricate non linear mode. Furthermore, a variety of data mining techniques in the realm of EDM have been consumed and applied to explore educational data and identify variables responsible for better academic achievements, but there are still deficiencies and it calls for the application of latest data mining tools.

The entire paper comprises of 7 sections. Section 1 provides general introduction of EDM and Section 2 presents the overview of the studies conducted in the related and associated field to underscore various research possibilities. While as Section 3 primarily focuses on the Conceptual Framework for the current study. Furthermore, Section 4 is based on objectives of the study and Section 5 provides detailed discussion of the research methodology adopted for carrying out this research endeavour. Section 6 encompasses of reports on experimental results. Finally, in Section 7 an attempt has been made to conclude the findings of the study and accentuate limitations and future directions.

## II. LITERATURE REVIEW

Romero et al (2013) applied Multiple Regression Model (MRM) and support vector machine (SVM) to forecast by and large the individual academic performance of students [7]. Also, Kotsiantis (2012) employed regression method to predict the student's marks in a distance learning system [8]. Noah, Barida and Egertib (2013) studied various parameters associated with the student using regression, k-means and neural networks to identify weak candidates for the purpose of performance enhancement [9]. Moreover, Bichkar (2014) described regression as statistical method of forecasting

students performance based on fields acquired from dataset [10].

Apart from the statistical measure viz Regression Model employed by a number of researchers, Jothi and venkatalakshmi (2014) made new strides and used clustering technique for analysing and predicting student's performance to improve the student's success rate [11]. Sheik and Gadage (2015) endeavoured to investigate the learning behaviour of various models with the help of various open source tools to get an insight of how these models train, evaluate and predict the performance of students [12].

Junco et al. (2011) applied ANOVA to calculate the impact on learning outcome and student engagement using twitter [13]. Stafford et al. (2014) examined the wiki activity indicators and the final grades of the students [14]. The results showed that there was significant correlation among the two variables and students who were engaged with wiki activities acquired better overall score. Giovanella et al. (2013) also investigated vigorous participation of students in various social media applications such as wiki, blog, Delicious and Twitter as a promising learning performance indicator [15].

Suyal and Mohod (2014) investigated students who necessitate special attention by studying the relationship among different attributes using association rule mining [16]. Baderwah and pal (2011) used mining technique namely decision tree on fields such as class test, attendance, assignment and semester marks for early detection of students who are at risk [17]. A number of efforts have been done in this direction and a research team comprising of Jeevalatha, Ananthi, and Saravana (2016) applied decision tree on undergraduate students dataset covering a set of factors such as Communication skills, higher secondary marks and undergraduate marks for performance assessment and placement selection [18]. In addition, Baker and Yacef (2009) conducted a survey on various techniques applied in the field of EDM. They came to the final conclusion that considerable work has been done in the direction of prediction rather than relationship mining [19].

From the review of previous studies, it is more obvious that the prime focus has been on predicting the students performance based on various attributes visa-a-vis family income, mid-term score, assignment, attendance and so on while using different data mining techniques such as decision tree, neural network, association rule mining, SVM and so on. However, there has been considerable gap in the realm of investigating the relationship between student's individual subjects and the overall results. Hitherto, there have been only modest attempts to employ Linear and Multiple Regression in EDM and there is still deficiency of literature/research endeavours in this context. Furthermore, novel techniques need to be applied in the direction of educational datasets to discover useful and imperative knowledge and facts from educational background. In addition, exploitation of deep learning methods in EDM would be a novel concept which can further assist in improving the performance of students.

### III. A Conceptual Framework For Our Study

To study and examine the impact of individual subjects on the overall result, a framework depicted in figure 1 has been proposed. In this framework, there are 4 individual subjects acting as input variables and one output variable in the form of overall result. Furthermore, rural/urban factor acts as a moderator for assessing the impact of relation on the overall result and can be seen in the conceptual model.

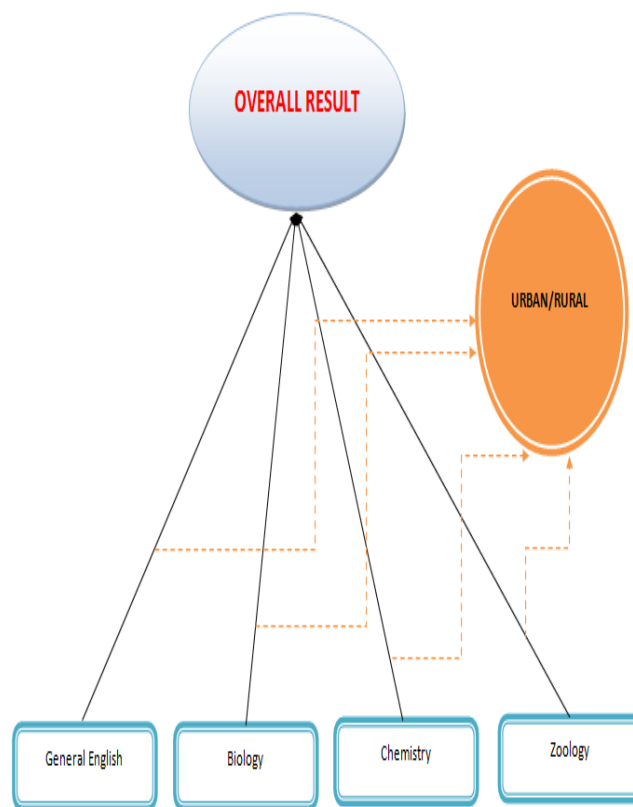


Fig 1 shows the conceptual framework

### IV. OBJECTIVES OF OUR STUDY

Analyzing big academic data can be extremely advantageous and provide insights about students teaching and learning behaviours based on which academic stakeholders can bring reforms in education and accordingly frame out apt strategies. In this direction, various data mining techniques have been applied to discover productive patterns from academic datasets. However, relational analysis among individual subjects and their impact on the overall result has not been investigated yet and rightly calls for the cerebral attention. In the current study, a novel attempt has been made to analyse the relationship between various academic subjects and consequently identify the factors responsible for student's performance in overall result and in respective subjects as well. Secondly, demographic effect on student's performance has also been examined to get a comprehension revision of

rural and urban education which would further assist stakeholders to improve the academic system.

### V. RESEARCH METHODOLOGY

In this research, techniques such as linear Regression and Multiple Regression, have been applied to discover interesting patterns from educational dataset. Primarily, an effort was made to pre-process the data while removing missing values, adding and deleting few columns and extracting some columns from existing dataset. Having done all the data purification, appropriate mining techniques have been applied for analysing and extracting meaningful information which would be useful in the direction of making concrete and decisive decisions for both students and the academic institutions.

#### A. Data Acquisition

The dataset used in this research was collected from University of Kashmir and it comprised of all the colleges from the Kashmir Division including colleges in North, South and Central Kashmir. The total number of colleges in the present study that were taken under investigation was 24 and important parameters like demographics and college code were derived from registration number during pre-processing phase. Furthermore, the dataset comprised of 9 attributes and 28991 records of Bachelor of Science (BSC) students at an initial stage. Among all the streams of BSC, we investigated only one stream comprising of subjects (English, Biology, Chemistry and Zoology). The total number of frequencies after extraction of records for the said stream turned out to be 1793. The dataset contained details of students including registration number, name, parentage, course code, capacity, subjects, total marks, total obtained and overall result. Table1 provides the snapshot of our dataset before pre-processing.

TABLE 1 shows the raw dataset 1 before pre-processing

REGISTRATION NO.	NAME	PARENTAGE	COURSE CODE	CAPACITY	SUBJECTS	TOTAL MARKS	TOTAL_OBT	RESULT
230-KN-2009	MUKHTAR	MANZOR			GE(27,33) HS(36,28) PS(35,35) UR(43,44)	600	281	PASS

42662-A-2009	SHOWKAT	MOHD YOUSUF DAR				BA	REGULAR	GE(36,39) GG(23,18;15,13) HS(27,39) PS(30,32)	600	272	PASS
2905-KR-2010	SHAHAR	MOHD HUSAIN				BA	REGULAR	GE(30,33) EC(29,19) PS(20,40) SO(27,29)	600	233	PASS
2342-KR-2009	MANSOOR	AHMAD HUSAIN				BA	REGULAR	GE(36,38) HS(32,28) PS(40,35) UR(34,30)	600	273	PASS
2963-KR-2010	MOHID	SHEIKH HUSIAN				BA	REGULAR	GE(40,35) HS(44,42) PS(50,50) UR(48,30)	600	339	PASS
2836-KR-2010	GHULAM	GHULAM ALI			GE(27,31) AR(27,25) ED(27,9) UR(39,48)	BA	REGULAR		600	233	REAR
2849-KR-	MOHAMMAD	MOHAMMAD BAQIR			GE(31,30) AR(28,30) ED(45,	BA	REGULAR		600	288	PASS

2010	Q A S I M				35) UR(47, 42)			
2851- K R- 2010	A Q E E B  J A V E E D D SAIN	MOH D HUS SAIN	BA	RE GU LA R	GE(34, 35) AR(22, 38) ED(48, 36) UR(48, 43)	600	304	PAS S
2886- K R- 2010	M O H D  A SAY ED	MOH D SAY ED	BA	RE GU LA R	GE(19, 35) EC(39, 27) ED(36, 36) MA(19, 5)	600	216	R MA
163- K S- 2010	S H A H Z A D A  A K H T E R AR	GH MOH D NAJ AR	BA	RE GU LA R	GE(30, 19) HS(32, 36) PS(37, 32) UR(40, 52)	600	283	PAS S ST
230- K N- 2009	M U K H T A R  A H M A D D W A N I	MAN ZOO R AHM AD WAN I	BA	RE GU LA R	GE(27, 33) HS(36, 28) PS(35, 35) UR(43, 44)	600	281	PAS S

of the data pre-processing, tasks such as data cleaning, data transformation, data extraction and so on were performed. Under data pre-processing, the following attributes were removed and subsequently various fields were extracted from the dataset.

- Name, parentage and registration were removed from the dataset.
- Null values and fields with less significance were also dropped.
- Fields such as English, Biology, chemistry and zoology were extracted from an attribute subjects using SQL 2008.
- Demography was derived from an attribute registration number.
- Additionally discretisation was performed on demographic filed as per the requirement of data processing.

After performing data pre-processing, a sum of 23 attributes were generated from the original data source against 9 attributes before pre-processing.

#### A. Description Of Variables

TABLE 2 displays the possible variables in our dataset with description

S.No	Fields	Description
1	Demography	Rural, Urban
2	GE_Paper A (General English Paper A)	0-75 (Possible Values)
3	GE_Paper B (General English Paper B)	0-75 (Possible Values)
4	GE_TOT (General English Total)	0-150 (Possible Values)
5	BO_Paper 1( Biology Paper 1)	0-50 (Possible Values)
6	BO_Paper 2(Biology Paper 2)	0-50 (Possible Values)
7	BO_INTERN( Biology Internals)	0-25 (Possible Values)
8	BO_PRACT (Biology Practical's)	0-25 (Possible Values)
9	BO_TOT (Biology Total)	0-150 (Possible Values)
10	CH_Paper 1(Chemistry Paper 1)	0-34 (Possible Values)
11	CH_Paper 2(Chemistry Paper 2)	0-33 (Possible Values)
12	CH_Paper 3(Chemistry Paper 3)	0-33 (Possible Values)
13	CH_Extern (Chemistry External)	0-25 (Possible Values)

#### B. Data Pre-Processing

This step is said to be the significant stage in the process of data mining. Generally factual data is incoherent, noisy and incomplete [20]. Therefore, data has to be selected and transformed into a more consistent and reliable state. As a part



14	CH_INTERN (Chemistry Internal)	0-25 (Possible Values)
15	CH_TOT (Chemistry Total)	0-150 (Possible Values)
16	ZO_Paper 1 (Zoology Paper 1)	0-50 (Possible Values)
17	ZO_Paper 2 (Zoology Paper 2)	0-50 (Possible Values)
18	ZO_INTERN(Zoology Internals)	0-25 (Possible Values)
19	ZO_PRACT (Zoology Practicals)	0-25 (Possible Values)
20	ZO_TOT (Zoology Total)	0-150 (Possible Values)
21	TOT_Marks (Total Marks)	600
22	TOT_OBT (Total Obtained)	0,<600 (Possible Values)
23	Overall Status	Pass, Fail

## VI. RESULTS AND DISCUSSION

In this study, a combination of various techniques such as linear regression and multiple regression on real dataset acquired from university of Kashmir have been employed. The results achieved after application of above techniques have been illustrated in the following subsections followed by suitable interpretations.

### A. Linear Regression

After running linear regression across various predictor variables including General English, Biology, Chemistry and Zoology on dependent variable TOT\_OBT (total marks obtained), the results have been shown in different tables highlighted below. The results show that there is a significant relation among all predictable variables including General English (GE\_TOT) and dependent variable TOT\_OBT, Biology (BO\_TOT) and TOT\_OBT, Chemistry (CH\_TOT) and TOT\_OBT, and Zoology (ZO\_TOT) and TOT\_OBT. Among all the variables GE\_TOT and TOT\_OBT has highest  $r$ ,  $r$  squared ( $r^2$ ) and adjusted  $r$  square values of (.873), (.762) and (.749) respectively as can be clearly seen in the table 4. Therefore, GE\_TOT and TOT\_OBT have maximum degree of correlation and subsequently  $r$  value indicates that there is 87.3% prediction accuracy. The  $r^2$  value shows 76.2% variance in data which is highly significant in comparison to all other variables and can be visualized in the below mentioned tables viz. table 4, table 5 and table 6. Also, the table 3 shows high significance in  $f$ -test with  $F=1.595$  and with 1794 degree of freedom which suggests there is a considerable linear relationship among variables in our model.

The coefficient matrix shows that  $t$ -test for intercept and variable are also statistically significant with  $p=0.000$  which is less than 0.05. Hence, it is reasonable to conclude that GE\_TOT draws an impact on TOT\_OBT (dependent variable)

in contrast to other predictable variables. Additionally, it is reasonable to report that if a student is having good marks in GE\_TOT, then he is having good marks in TOT\_OBT. This signifies that if a student performs well in English then his score and grade also improves.

TABLE 3 shows results associated with General English and Total Marks Obtained

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.873 <sup>a</sup>	.762	.749	874.779	
a. Predictors: (Constant), GE_TOT					
ANOVA <sup>b</sup>					
Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	1381025.458	1	1381025.458	1.595	.000 <sup>a</sup>
Residual	1552276.701	1793	865.743		
Total	2933302.158	1794			
a. Predictors: (Constant), GE_TOT					
b. Dependent Variable: TOT_OBT					
Coefficients <sup>a</sup>					
Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	183.676	4.278		42.940	.000
GE_TOT	2.203	.055	.686	39.940	.000
a. Dependent: TOT_OBT					

As can be seen from Table 3 that R associated with the test statistic is 0.873 with 0.762 R Square which are highly satisfactory and indicates the reasonable linear relationship and degree of association between the two variables.

TABLE 4 shows results associated with Biology and Total Marks Obtained

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.842 <sup>a</sup>	.709	.709	21.820	
a. Predictors: (Constant), BO_TOT					
ANOVA <sup>b</sup>					
Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	2079658.083	1	2079658.083	4.368	.000 <sup>a</sup>
Residual	853644.075	1793	476.098		
Total	2933302.158	1794			
a. Predictors: (Constant), BO_TOT					
b. Dependent Variable: TOT_OBT					
Coefficients <sup>a</sup>					
Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	124.421	3.485	.842	35.697	.000
GE_TOT	2.436	.037		66.092	.000
a. Dependent: TOT_OBT					

a. Dependent: TOT\_OBT

Here R and R square associated with the test statistics are 0.842 and 0.709 which again demonstrate reasonable relationship between the variables under consideration (Table 4).

TABLE 5 shows results associated with Chemistry and Total Marks Obtained

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.820 <sup>a</sup>	.673	.673	23.124	
a. Predictors: (Constant), CH_TOT					
ANOVA <sup>b</sup>					
Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	1974581.311	1	1974581.311	3.693 E3	.000 <sup>a</sup>
Residual	958720.848	1793	534.702		
Total	2933302.158	1794			
a. Predictors: (Constant), CH_TOT					
b. Dependent Variable: TOT_OBT					
Coefficients <sup>a</sup>					
Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	103.516	4.129	.820	25.068	.000
GE_TOT	2.710	.045		60.769	.000
a. Dependent: TOT_OBT					

To make further roads in the direction of current problem, Linear Regression was again performed on Chemistry score and Total Marks Obtained and results are documented in Table 5. The findings linked with the test statistics are significant and reveal noteworthy linear association between the variables.

TABLE 6 shows results associated with Zoology and Total Marks Obtained

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.811 <sup>a</sup>	.658	.658	23.652	
a. Predictors: (Constant), ZO_TOT					
ANOVA <sup>b</sup>					
Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	1930234.883	1	1930234.883	3.450 E3	.000 <sup>a</sup>
Residual	1003067.275	1793	559.435		
Total	2933302.158	1794			
a. Predictors: (Constant), ZO_TOT					
b. Dependent Variable: TOT_OBT					
Coefficients <sup>a</sup>					
Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	117.769	4.031	.811	29.217	.000
GE_TOT	2.603	.044		58.739	.000

The application of Linear Regression on two attributes *visa-a-vis* Zoology and Total Marks Obtained have also revealed a significant linear association between the two factors and it amounts to 0.811 which is largely satisfactory (Table 6)

### A. Multiple Regressions

In this method, an attempt was made to find the multiple correlations among different predictable variables and as a consequence certify its impact on dependable variable. The results (table 7, table 8 and table 9) clearly indicate it that there is significant correlation among multiple attributes but GE\_TOT and BO\_TOT has maximum multiple correlation coefficient (r) of 0.923 which signifies 92.3% predication accuracy. The R Squared (r<sup>2</sup>) value is 0.852 which means that predictable variables impose 85.2% variability on dependent variables. The f-ratio in the ANOVA matrix implies that the overall model is significant and best fits the data. It also shows that predictable variables are statistically significant to predict dependent variable (TOT\_OBT) with F (2, 1792) =5.153E3 and p < 0.0005.

Furthermore, coefficient matrix explains that t-test for intercept and variable are also statistically significant with p=0.000 which is less than 0.05. Therefore, it can be concluded that if a student possesses maximum marks in GE\_TOT and BO\_TOT, his overall grade and his chances of success in passing the degree are maximized by 92.3%.

TABLE 7 shows relation between General English and Biology on overall result

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.923 <sup>a</sup>	.852	.852	15.571	
a. Predictors: (Constant), BO_TOT, GE_TOT					
ANOVA <sup>b</sup>					
Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	2498827.491	2	1249413.745	5.153 E3	.000 <sup>a</sup>
Residual	434474	1792	242.452		
Total	2933302.158	1794			
a. Predictors: (Constant), BO_TOT, GE_TOT					
b. Dependent Variable: TOT_OBT					
Coefficients <sup>a</sup>					
Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	68.090	2.832		24.040	.000
GE_TOT	1.327	.032	.413	41.580	.000
BO_TOT	1.953	.029	.675	67.900	.000
a. Dependent: TOT_OBT					

TABLE 8 shows relation between General English and Chemistry on overall result

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.914 <sup>a</sup>	.835	.835	16.418	
a. Predictors: (Constant), CH_TOT, GE_TOT					
ANOVA <sup>b</sup>					
Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	2450279.914	2	1225139.957	4.545 E3	.000 <sup>a</sup>
Residual	483022.244	1792	269.544		
Total	2933302.158	1794			
a. Predictors: (Constant), CH_TOT, GE_TOT					
b. Dependent Variable: TOT_OBT					
Coefficients <sup>a</sup>					
Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	47.017	3.226		14.576	.000
GE_TOT	1.399	.033	.436	42.010	.000
CH_TOT	2.159	.034	.654	62.983	.000
a. Dependent: TOT_OBT					

TABLE 9 shows relation between General English and Zoology on overall result

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	
1	.903	.816	.815	17.376	
a. Predictors: (Constant), ZO_TOT, GE_TOT					
ANOVA <sup>b</sup>					
Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	2392257.196	2	1196128.598	3.962 E3	.000 <sup>a</sup>
Residual	541044.962	1792	301.922		
Total	2933302.158	1794			
a. Predictors: (Constant), ZO_TOT, GE_TOT					
b. Dependent Variable: TOT_OBT					
Coefficients <sup>a</sup>					
Model	Unstandardised Coefficients		Standardised Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	61.132	3.296		18.546	.000
GE_TOT	1.388	.035	.432	39.119	.000
ZO_TOT	2.052	.035	.640	57.873	.000
a. Dependent: TOT_OBT					

After performing data analysis while taking different subjects together, it was examined that English and Biology collectively have significant r, r squared values and f-ratio. Therefore, after application of multiple regression it was analysed that collectively English and Biology play an imperative role on the outcome of a student.

## CONCLUSION

The main purpose of this study was to identify the factors which can ameliorate the student's performance and improve students' competence in academia. Based on the empirical results, it has been found that English as an academic subject in contrast to other subjects visa-a-vis Biology, Chemistry and Zoology is predominant and significant variable that largely shapes the dependent variable viz. overall result. Furthermore, results have also laid strong foundation for the relationship between predictor variables (individual subjects in this case) criterion variable (overall result) while controlling the effect of demographic elements and this could be for the reason that students with urban background have a significant record in English than other subjects. However, students that belong to rural areas have demonstrated significant score in Biology, Chemistry and Zoology than urban populace. Therefore, from the empirical results obtained from the dataset in the current study, it can be concluded that students in urban areas have a considerable potency in English rather than other subjects. Furthermore, statistical figures acquired from Regression, it can well be concluded that with one percent change in efficiency level in English Subject there would be 0.94 percent change in the overall result. Similarly, the results show that there is a change of 0.89, 0.84 and 0.85 in the efficiency of Biology, Chemistry and Zoology respectively. Moreover, from the results of the multiple regression, the results show significant transition as r value (0.923) associated in this case gets amplified with the inclusion of second variable that is performance in Biology. Therefore, the competence in the subject of Biology along with General English has larger influence on the performance of a student.

Similarly, inclusion of English with other subjects such as Chemistry and Zoology also shows a significant transition with r value 0.914 and 0.904 respectively. Therefore, our empirical results show that English as a subject plays a pivotal role in an academic system and needs due course of consideration for better academic performance of a student.

## REFERENCES

- [1] Barnes T, Desmarais M, Romero C, Ventura S. Educational data mining 2009: 2nd international conference on educational data mining. Proceedings. Cordoba, Spain; 2009.
- [2] Baker RSJD. Data mining for education. Int Encyclopedia Educ 2010;7:112-118.
- [3] Mostow J, Beck J. Some useful tactics to modify, map and mine data from intelligent tutors. Nat Lang Eng 2006;12(02):195-208.
- [4] C. Romero, S. Ventura, "Educational Data Mining: A Review of the State of the Art", IEEE Transactions on Systems Man and Cybernetics-Part C: Applications and Reviews, vol. 40, no. 6, November 2010.
- [5] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 40, pp. 601-618, 2010.
- [6] Ganesh, S. Hari, and A. Joy Christy. "Applications of Educational Data Mining: A Survey." Innovations in Information, Embedded and Communication Systems (ICIECS), 2015 International Conference on. IEEE, 2015.
- [7] Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, Sebastian Ventura, "Predicting students' final performance from participation in on-line discussion forums", Computers & Education, vol. 68, pp. 458-472, October 2013.

- [8] Sotiris B Kotsiantis, "Use of machine learning techniques for educational proposes: a decision support system for forecasting students grades", *Artificial Intelligence Review*, vol. 37, no. 4, pp. 331-344, 2012.
- [9] OTOBO Firstman, BAAH Barida Noah, Taylor Onate Egerton, "Evaluation of student performance using data mining over a given data space", *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 2, no. 4, pp. 2277-3878, September 2013.
- [10] R. S. Bichkar "Predicting Students Academic Perfonnance Using Education Data Mining", *World Journal of Computer Application and Technology* 2(2): 43-47, 2014.
- [11] J.K. Jothi, K. Venkatalakshmi, "Intellectual performance analysis of students by using data mining techniques", *International Journal of Innovative Research in Science Engineering and Technology*, vol. 3, no. 3, March 2014.
- [12] Shelke Nikitaben, Gadage Shriniwas, "A survey of data mining approaches in performance analysis and evaluation", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 5, 2015.
- [13] R. Junco, G. Heiberger, and E. Loken, "The effect of Twitter on college student engagement and grades", *Journal of Computer Assisted Learning*, Blackwell Publishing Ltd, vol. 27(2), pp. 119-132, 2011.
- [14] T. Stafford, H. Elgueta, and H. Cameron, "Students' engagement with a collaborative wiki tool predicts enhanced written exam performance", *Research in Learning Technology*, vol. 22, 2014.
- [15] C. Giovannella, E. Popescu, and F. Scaccia, "A PCA study of student performance indicators in a Web 2.0-based learning environment", *Proc. ICALT 2013 (13th IEEE International Conference on Advanced Learning Technologies)*, pp. 33-35, 2013.
- [16] Sayali Rajesh Suyal, Mohini Mukund Mohod, "Quality improvisation of student performance using data mining techniques", *International Journal of Scientific and Research Publications*, vol. 4, no. 4, April 2014.
- [17] Baradwaj Brijesh Kumar, Pal Saurabh, "Mining educational data to analyze Sstudents' performance", *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, 2011.
- [18] Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In *Data Mining and Advanced Computing (SAPIENCE)*, International Conference on (pp. 91-95). IEEE.
- [19] J.D.B. Baker Ryan, Yacef Kalina, "The state of educational data mining in 2009: A review and future revisions", *Journal of Educational Data Mining*, vol. 1, no. 1, February 2009.
- [20] Jiawei Han, Micheline Kamber, Jian Pei, "DATA MINING - Concepts and Techniques", Morgan Kaufmann.

# Automated Trust Evaluation Mechanism For Future Generation Internet

Arun Fera M

Assistant Professor  
Department of Information Technology  
Thiagarajar college of Engineering  
Madurai, India  
fera26@gmail.com

Umamaheswari S

PG Scholar  
Department of Information Technology  
Thiagarajar college of Engineering  
Madurai, India  
maheswariu95@gmail.com

**Abstract**—Cloud computing technology plays a major role in the business environment. One of the issues is to evaluate the trust of both cloud service provider and Internet service providers to obtain credible services. That is how to select the trustworthy services for cloud users remains a significant problem. In this paper, we propose a model, to evaluate the trust by filtering out the unfair ratings given by the witnesses. The ratings can be given by both cloud users and non-cloud users. The cloud consumer contains both functional and non-functional requirements. It includes VCPU, virtual processors, reliable services and safe, fast, high speed internet for CSP and ISP respectively. Then the effectiveness ratings are updated in the system and use of reinforcement learning algorithm to make the system automation process. Q learning is one of the subsets of reinforcement learning algorithms. This algorithm works based on the ratings given by the cloud consumer. After the completion of each interaction the cloud users allowed to give ratings for the service providers. The learning algorithm works on witnesses, that each witness is evaluated and updated into the learning algorithm. Thus, we present and evaluate an effective theoretical and experimental calculation of trust over the Cloud service providers and internet service providers.

**Keywords**—cloud computing, trust, reputation, feedback, Q-learning.

## I. INTRODUCTION

Cloud computing is the internet paradigm which provides services and resources such as hardware, software and network. It provides the services, the user can able to access the services, that is accessed via the network through anywhere from the environment. It is the complex computing technology, which eliminates the need of having own physical hardware, software and network. In cloud computing the user can place the request such as processors and memory requirements, etc., Internet service providers are the organization providing various services such as bandwidth, number of links, latency and turn around time. These are all the various services the user can place a request from the internet service providers. Internet service providers denote separation of both infrastructure provider and service providers.

Network virtualization- It is a key contribution to the future generation internet. In the virtualization-based future Internet,

an end-to-end network service delivery system is constructed by an SP through synthesizing resources from multiple network infrastructures. Network virtualization is the process decoupling between network service provisions from the underlying infrastructure. Network virtualization encapsulates the shares of resources and networking capabilities offered by the different infrastructure provider into a set of service components. The SP then assembles this set of service components to form an end-to-end network service that meets the application requirement. Therefore, an end-to-end service delivery system in network virtualization consists of a series of tandem service components, each of which is a logical abstraction of the infrastructure service provided by Infrastructure providers. i.e. the user request the service to service provider the service provider assigning the service based on the availability of the service. In trust evaluation the consumer request the service providers based on the degree of the reputation of both cloud service provider and internet service providers. The various sections involved in this paper are as follows: Section 2 describes related work in trust evaluation, Section 3 proposes basic Trust evaluation techniques, section 4 describes the proposed work section 5 and 6 describes the techniques used for service selection and trust evaluation respectively. Section 7 describes the experimental results of our proposed techniques.

## II. RELATED WORK

**Yu H, Shen Z, et al.,[1]** author proposed filtering of unfair ratings from the set of feedback given by the cloud consumer. For evaluating the trust they use a technique of the beta reputation system. They propose an algorithm called actor-critic trust model which is the adaptive trust aggregation model and dynamic updating of the trust parameter for the credible services they use the machine learning algorithm called reinforcement learning. **Josang A, Ismail R.et al., [2]** have proposed a method for uncertain probabilities. The uncertain probabilities are evaluated using Dempster-shafer theory. It computes all the possible situations such as belief, disbelief and uncertainty. Initially belief mass is computed based on the relative atomicity of the value is computed. The probability density function is evaluated using beta reputation system. **Taneja S, Rathi K.et al.,[3]** proposed method, the trust is evaluated by the third parties. For that, there are three

techniques are proposed in this paper. They are public recommendation system(based on the feedback of the users), self-recommendation system and third party recommendation system. They use SLA's is one of the factors for evaluating trust. In this paper the trust is evaluated over the period of time for that they use a technique called windowing technique. **Pan Y, Ding S et al.,[4]** proposes that Service provider can be chosen based on the QOS. In QOS both functional and non-functional parameters are considered. Functional and non-functional parameters such as accuracy and response time respectively. And also consider the location and usability of the services. Then the missing QOS values are filled with the traditional methods. Based on these parameters and the frequency of interaction between the users, rank the cloud service provider. **Wu Q, Zhang X [5]** proposes that to filter the un-faired ratings in the dynamic environment. The ratings are filtered based on the similarity between the feedbacks this paper the filtered feedback is combined with the historical user evaluation and preferences, to elevate the trust between the customers and providers. To filter the malicious feedback the filtering algorithms are used and it filters based on the similarities, and the similarities are evaluated based on the Euclidian distance. Then it is it is quantified based on some QOS values such as response time, cost and the reliability which is placed in the knowledge repository. **Fan WJ, Yang SL, et al.,[6]** proposes that integrates two different techniques, perception based and reputation based mechanisms which are also called direct and indirect trust respectively. The evidential reasoning approach is used to find the real time trust by aggregate the feedback ratings of the users and based on the result select the trustworthy service providers. The perception is based on the direct service interaction with the customer to the services and the reputation technique is based on the values assigned to the service based on the user interaction. This is the effective method for evaluating the trust of the cloud service providers. **Noor TH, Sheng QZ et al.,[7]** proposed that the proposed trustworthiness evaluation framework for cloud service selection. There are five blocks in the framework, the preprocessing block is to maintain a trustworthy record, and the next block is management block is used to categorize the trust factor into common trust factors and special trust factors. The trust factor processing block updates the common factor in the common factor pool. To enhance the efficiency the common factor attributes can be reused. The special factors are dealt with ontology. The trustworthiness decision making block that employs the trust factor in a multi-criteria analysis approach. **Ye X. et al.,[8]** proposes how to deal with unfair ratings, The trustworthiness of the customer is identified based on the similarity of the user, popular ratings given by the user and the experience of the customer to the cloud service provider. The similarity of the user is identified by adjusting the weight factor for the similar customer that is similar QOS ratings and the location similarities. In similarity based ratings the advisor group uses a collaborative filtering technique to identify the accuracy score of the advisors. To identify the similar users Pearson, co-efficient techniques will be used. **Raghebi Z, Hashemi MR et al.,[9]**, proposes that In this paper the trust is evaluated based on the reliability of the customer feedback. In this paper they considered both direct and indirect trust. Direct trust

states that the customers have direct experience with the service provider. Indirect trust states that the customer communicates with the people who have already shared their resources with them. Based on the opinion of the existing user, the upcoming users trust the service. This paper proposes the customer feedback reliability is based on the similarity between the customer feedbacks for the shared services. If the customer doesn't have any connection then the belief is based on the opinion of the majority of the existing customers who are all having direct contact with them. **Huang J, Xu C et al.,[10]** Network virtualization which is a key contribution to the future generation internet. In this paper they addressed the problem of path selection in a virtual environment and QOS service provisioning for multimedia services. In the virtualization-based future Internet, an end-to-end network service delivery system is constructed by an SP through synthesizing resources from multiple network infrastructures. Network virtualization is the process decoupling between network service provision from the underlying infrastructure. Network virtualization encapsulates the shares of resources and networking capabilities offered by different infrastructure providers into a set of service components. The SP then assembles this set of service components to form an end-to-end network service that meets the application requirement. Therefore, an end-to-end service delivery system in network virtualization consists of a series of tandem service components, each of which is a logical abstraction of the infrastructure service provided by an infrastructure provider.

### III. SYSTEM ARCHITECTURE

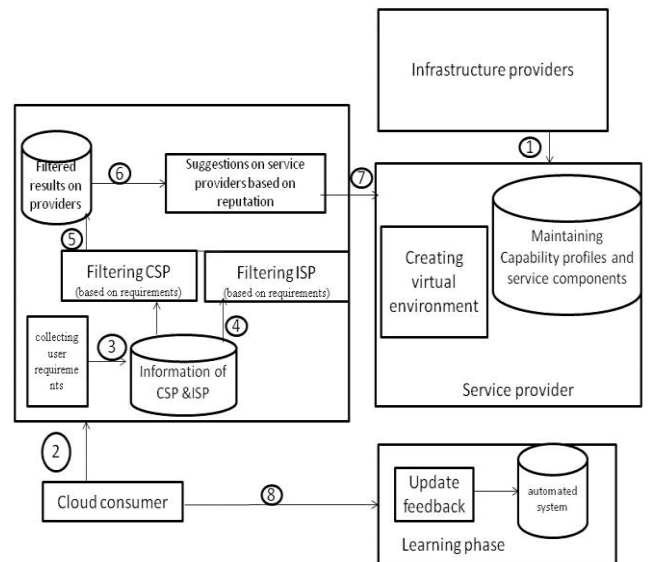


Fig. 1. Trust Evaluation Architecture

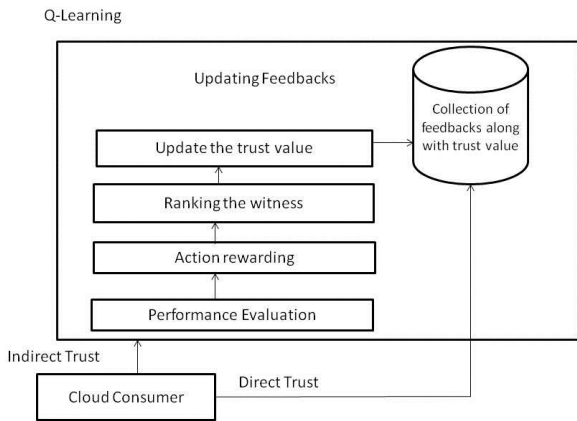


Fig. 2. Learning Phase

#### IV. TRUST EVALUATION FOUNDATIONS

Trust Evaluation is an important component in a business environment. Trust evaluation on the cloud service providers is done for investigating the features of the services provided by the service providers such as capability, applicability in a cloud environment, interoperability, security, personalization and privacy of data. Similarly Trust evaluation on the internet service providers helps to evaluate the reliability, security and latency on the services provided by the internet service providers. Basically, there are three types of trust, they are

- Direct Trust
- Indirect Trust
- Hybrid Trust

**Direct Trust-** It states that the users or the consumers have direct interaction with the service providers that is both cloud and internet service providers without having any interaction or suggestion of the third parties. These are the people they have directly requested the services by the service provider based on the own satisfaction level on the service providers or based on the own belief. Some people process it as trail and error model that is trust is evaluated based on the outcome or the satisfaction level of accessing services.

**Indirect Trust** – It is defined as that the consumer interacted with the cloud service provider and internet service providers is based on the suggestion of the services provided by the third parties. Based on the reputation value given by the third parties they deal with the cloud and internet service providers. Here the suggested peoples are called as witnesses who are all having prior interactions with the selected service providers. In this type, the trust level of the users is based on the people who are all already interacted with service providers.

**Hybrid Trust** – It is the combination of both direct and indirect trust values.

The problem with existing architecture is evaluating the trust of both cloud service provider and internet service providers. The trust is evaluated based on the ratings and feedback given by the users. But there is probability of giving false rating for the service providers. The false ratings for the service provider may be intentionally given by the competitors in the business environment or there may be the possibility of the fake users giving unfair ratings to the service providers, so that kind of ratings should be filtered out before evaluating the trust. Based on the similarity of the users the unfair ratings should be filtered out. The indirect trust is evaluated based on the suggestion of the credible user who is all already interacted with the service provider. The credibility of the user is verified based on the user who use their suggestion for the accessing the both cloud service providers and internet service providers. The suggested users are also called as a witness. The problem is ranking the witnesses for recommending the service provider for the cloud consumers. The witnesses ranking is based on their level of credibility of ratings. Then the reputation value of both cloud service providers and internet service providers should be updated. The updating of the reputation value of cloud service providers and internet service providers should consider the previous interaction values. Finally, based on the reputation of the cloud service provider and internet service providers are suggested to the user for selecting the best service providers.

#### V. PROPOSED WORK

The proposed system evaluates the reputation of both cloud service provider and the internet service provider. Initially the requirements are collected from the users for accessing both cloud and internet services. Based on the requirements of the users the set of service providers is filtered out. To filter out the service providers or selection of service for the user is implemented using bin-packing algorithm. In bin-packing algorithm, first fit allocation is implemented by selecting the service provider for both cloud and internet service providers. After selecting service providers the cloud consumer will interact with the service provider.

Based on the outcome or satisfactory level of the consumer they provide the ratings for both cloud service provider and internet service provider. The trust value is evaluated based on the ratings given by the users. The directors trust is evaluated using a beta reputation system. Beta reputation system computes trust based on number of successful and failure interaction for the service providers. The reputation value is directly updated to the knowledge based system. After the reputation value has updated the system starts the automation process for evaluating the reputation of both cloud service providers and the internet service providers. To make the calculation as system automation Q- learning algorithm is used for updating the feedback. The indirect trust state that the interaction is based on the suggestion the interacted users with the service providers. Here the user interact with the service provider based on the suggestion of the users have direct interaction with the service providers. In indirect trust, after the interaction with the service provider the credible value of the suggested users should be updated. The suggested users are also called as a witness. Through the correction of the ratings

of the user the un-faired ratings are filtered out. UN-faired ratings are given by the competitors for the service providers in the business environment.

In the process of filtering the unfair ratings the, the ratings of the users will be corrected based on the satisfactory and the similarity level of the user who are all having indirect trust with both cloud service providers and internet service providers. For the indirect trust also the top ranked witness are allowed to suggest the service provider in both cloud and internet service providers. The witness ranking is done based on the credibility level of the cloud consumers. Using the ratings and testimonies of the users the trust value will be calculated. After the computation the reputation value is updated to the system to suggest the best service in both cloud and internet service providers to the cloud consumers.

### System Parameters

<i>ISP</i>	<i>Set of Internet service providers</i>
<i>SP</i>	<i>Service providers</i>
<i>SPS</i>	<i>Service providers Specifications</i>
<i>CSP</i>	<i>Cloud service provider Specifications</i>
<i>CCR</i>	<i>Cloud consumer's requirements</i>
<i>CC</i>	<i>Cloud consumers</i>
<i>B</i>	<i>Bandwidth</i>
<i>L</i>	<i>Number of links</i>
<i>NP</i>	<i>Number of processors</i>
<i>SS</i>	<i>Number of Storage servers</i>
<i>c<sub>i</sub></i>	<i>Individual customer</i>
<i>CU</i>	<i>Cloud user requirements</i>
<i>IU</i>	<i>Internet users requirements</i>
<i>F</i>	<i>Feedback given by the customer</i>
<i>R</i>	<i>Reward</i>
<i>η</i>	<i>Learning rate</i>
<i>γ</i>	<i>Discount factor</i>
<i>D<sub>i</sub></i>	<i>Direct trust</i>
<i>ind<sub>i</sub></i>	<i>Indirect trust</i>
<i>w1 and w2</i>	<i>Weights for Evaluating trust</i>
<i>R(t)</i>	<i>Reputation value for the service providers</i>

### VI. SERVICE SELECTION

In the proposed method the cloud service selection and internet service selection are used bin-packing algorithm and it is working based on first fit bin-packing algorithm. Initially the requirements are collected from the users for the services. In case of cloud service provider the user requirements may be a number of the RAM and the RAM specifications, the number of storage servers and number of processor with them specifications. In case of internet service provider the requirements are bandwidth, latency, turnaround time for accessing the services by the users. Then it uses a bin packing algorithm for the user request to allocate the services for them.

### Algorithm

Read the input as *CU<sub>i</sub>, IU<sub>i</sub>, CSP, ISP*

**WHILE** (*CU is not empty*)

**For** (*any item a, b, c in userCU<sub>i</sub>*)

**IF** (*a, b, c in CU<sub>i</sub> is matched with an service in CSP*)

*Match the first selected provider in CSP*

**ELSE**

*Continue;*

**WHILE** (*IU is not empty*)

**For** (*any item a, b, c in userIU<sub>i</sub>*)

**IF** (*a, b, c in IU<sub>i</sub> is matched with an service in ISP*)

*Match the first selected provider in ISP*

**ELSE**

*Continue;*

ALGORITHM 1 SERVICE SELECTION ALGORITHM

It denotes if the requested service specification can be provided by the service provider, then the service provider is allocated for the consumers who request their services. Based on the result of the filtered service providers, the users select the trusted service provider among the cloud and the internet service provider.

### VII. EVALUATION OF TRUST

The trust is evaluated based on the outcome of the interaction of the requested users that is both direct and indirect trust users. If the service providers cannot able to provide service to the users then it consider as the failure outcome for the service providers. Additionally, if the user results the interaction results are 1 then it considers as the successful interaction for the service provider otherwise it considers as failure interaction for that service provider. The direct trust is evaluated using a Beta Reputation system.

$$D_i = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad (1)$$

Where  $\alpha$  denotes number of successful interactions for the service provider and  $\beta$  denotes the number of failure interaction in the service provider. In this method the direct trust value is evaluated for both cloud service provider and internet service providers.

The indirect trust is evaluated based on the previous interaction of the users with the service provider. The ratings given by direct trust users with the service provider is aggregated based on the value the indirect trust is evaluated. To filter the un-faired ratings given by the users Q-learning algorithm is used. The un-faired ratings are filtered by giving penalty to the user who gives the ratings. It is evaluated based on the outcome of the interaction of the user in the indirect trust. If the outcome of the interaction is 0 then give a penalty to the user who gives high ratings for the user already interacted with the service provider.



### Q-Learning algorithm

For all  $\langle c_i, f \rangle$  do  
 $Q^{\wedge}(c_i, f) \leftarrow 0$   
**End For**  
 Observe the current state of  $c_i$   
**Loop**  
 Select a Feedback  $f$  and collect it  
 Receive the immediate reward  $R$   
 Observe the new value  
 Update the trusted value of the customer  
 $Q^{\wedge}(c_i, f) \leftarrow (1-\eta) Q^{\wedge}(c_i, f) + \eta(r_{t+1} + \gamma \max_{f'} Q^{\wedge}(c_{i+1}', f_{i+1}'))$   
**End loop**

ALGORITHM 1 Q-LEARNING ALGORITHM

If the outcome of the interaction is 1 then given some reward to the users who are all suggest for that service provider to that user in the indirect trust. Based on the result of the ratings of the users the indirect trust is evaluated. The suggested users are the direct trust users who had a prior interaction with that service provider. They are also called as witnesses. The indirect trust users change the value of the ratings of the direct trust users by applying reward or penalty based on the outcome of their own interaction with the cloud service provider and the internet service providers. Based on the change of ratings the witnesses will be ranked. Based on the ranking value of the witness, the indirect trust user accesses the users for suggesting cloud and internet service providers. Then finally the reputation of the service providers is evaluated by giving equal weightage to both direct and indirect trust users.

$$R(t) = w_1 * D_i + w_2 * ind_i \quad (2)$$

Where  $R(t)$  is the reputation value of the each and every cloud and internet service providers. Based on the credibility and the experience of the users the different weightage is assigned for both direct and indirect trust value. After evaluating the reputation is updating to the learning system. Assign value for  $w_1$  and  $w_2$  based on the weighted for the direct and indirect trust respectively. Rank the Cloud and internet service provider. Based on the reputation of the service provider suggest the best service to the requested users.

### VIII. EXPERIMENTAL RESULT

In order to evaluate the performance of the proposed model under different possible conditions from the user requirements for both direct and indirect trust. For that we simulate the proposed model using MATLAB. 500 user requirements are passed as parameters for each round of the algorithm. There are 6 cloud service providers and 6 internet service providers. All the cloud service providers have 4 common services to provide the services for the consumers. Similarly, all the internet service providers have 4 common services to satisfy the requested users.

During simulation, initially 500 customers interacted with the 6 cloud service providers and with the internet service providers. In the first round of interaction all the customers are considered as the direct trust users. For the next round the

users are considered as direct and indirect trust users as multiple cases. For example, case 1) 10% are direct trust user remaining 90% are indirect trust users. Case2) 50% are direct trust users and 50% are indirect trust user case 3) 25% are direct trust users and 75% are indirect trust users. Case4) 75% are direct trust users and remaining 25% are indirect trust users.

TABLE I. SHOWS THE REPUTATION CALCULATION OF CLOUD SERVICE PROVIDER.

CSP id	Direct trust	Indirect trust	Reputation
1	0.997636	2.833363	0.600773
2	0.994536	2.924988	0.661476
3	0.999139	3.074889	0.767421
4	0.99802	2.749716	0.543679
5	0.99902	0	0.034115
6	0.999002	0.158895	0.12322

Based on the result of the reputation of the service provider the suggestion will be provided for the user. The user having high reputation value is suggested for the requested user. If that provider cannot satisfy the user requirements, then suggest the trusted provider who is next to him based on the reputation value.

TABLE II. SHOWS THE REPUTATION CALCULATION OF INTERNET SERVICE PROVIDERS.

ISP id	Direct trust	Indirect trust	Reputation
1	0.998476	0.8	0.116501
2	0.998969	2.5	0.265113
3	0.998054	3.142857	0.804991
4	0.998836	2.922787	0.620568
5	0.999046	3.175036	0.832887
6	0.998926	0.569432	0.135854

Similarly, based on the result of the direct and indirect trust value degree of the reputation for each service provider is evaluated. Based on the degree of the trust, the provider will be suggested to the requested consumers.

### Results and Discussion

Fig 3 shows the reputation value for the cloud service providers. According to the graph in Fig 1 the highest trust factor is 0.767421 obtained by the service provider 3. The degree of the trust is evaluated based on the number of successful and the unsuccessful interaction having each service provider for the requested users. Number of successful and failure interaction is represented in a graph. Fig 4 shows that the provider 3 has nearer successful and failure outcomes for the users while the provider 5 and provider 6 have a higher range of failure interaction compare to the other provider. This is evaluated by the interaction of 1000 users. So the provider 5 and provider 6 have the lowest degree of trust compare to other cloud service provider.

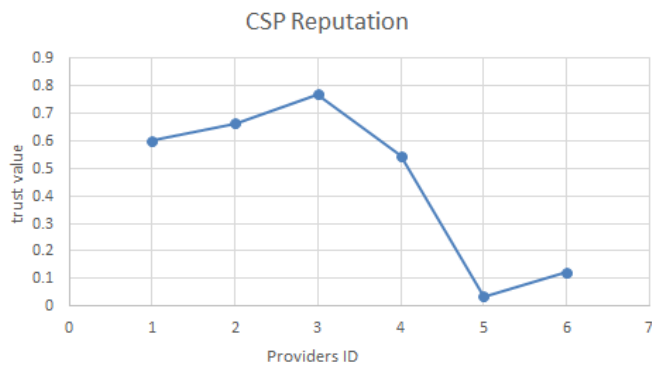


Fig. 3. Reputation Of The Cloud Service Provider

In case of a service provider they have lowest successful interaction compare to other cloud service providers, but compare to provider 5 and provider 6 it has the lowest failure rate in outcome.

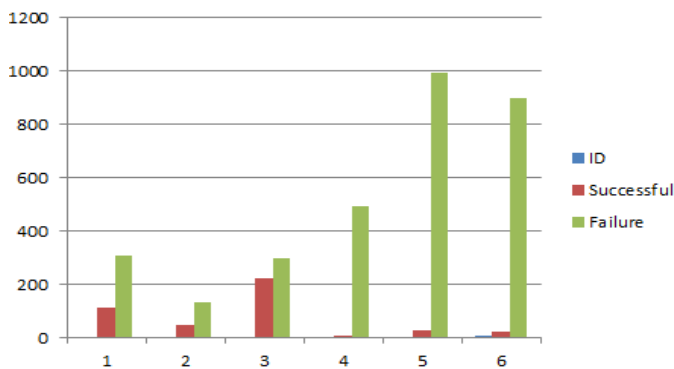


Fig. 4. Outcomes of the Interaction for the CSP's

Similarly for the internet service provider, the degree of the reputation of the service provider is represented in the graph which is shown in the fig5. It shows that the service provider 5 have the highest degree of trust 0.832887. The number of successful and failure interaction for the cloud service provider is shown in the fig 6.

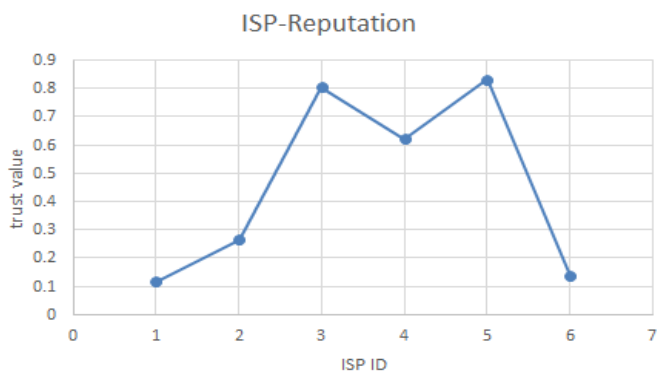


Fig. 5. Reputation of the Internet Service Providers

The number of successful and failure interaction is shown in the graph which is represented in the fig 4. The figure 4

shows that the service provider 5 has the high successful interaction among the request of 1000 providers. It results that the reputation is evaluated based on the outcome of the interaction.

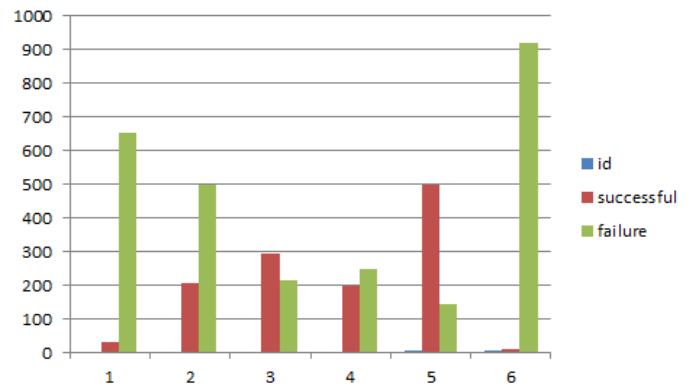


Fig. 6. Outcomes of the Interaction for the ISP's

Based on the reputation of both cloud and internet service provider, the best service provider is suggested for the users, requested for the service.

## IX. CONCLUSION

In this paper the trust value of both cloud service providers and the internet service providers are evaluated based on the ratings of both direct and indirect trust users. The un-faired ratings are also filtered out in the process of indirect trust. The witnesses are also ranked for providing a better suggestion for the indirect trust users. Using the evaluation, best service is obtained by combining both cloud and the internet service providers. The major issue in the evaluation of the reputation of the service providers is time complexity. This will be overcomes by evaluating the multiple possible choices in a parallel manner. The result is more effective and accurate.

## X. REFERENCES

- [1] Yu, H., Shen, Z., Miao, C., An, B., & Leung, C. (2014). Filtering trust opinions through reinforcement learning. *Decision Support Systems*, 66, 102–113. <https://doi.org/10.1016/j.dss.2014.06.006>
- [2] Jøsang, A. (2002). The Beta Reputation System, 1–14.
- [3] Taneja, S. (2015). A Trust Evaluation Model to Recommend a Service Provider to a Customer in Cloud Environment. *International Journal of Computer Applications*, 121(2), 975–8887.
- [4] Pan, Y., Ding, S., Fan, W., Li, J., & Yang, S. (2015). Trust-enhanced cloud service selection model based on QoS analysis. *PLoS ONE*, 10(11), 1–19. <https://doi.org/10.1371/journal.pone.0143448>
- [5] Wu, Q., Zhang, X., Zhang, M., Lou, Y., Zheng, R., & Wei, W. (2014). Reputation revision method for selecting cloud services based on prior knowledge and a market mechanism. *The Scientific World Journal*, 2014. <https://doi.org/10.1155/2014/617087>
- [6] Fan, W.-J., Yang, S.-L., Perros, H., & Pei, J. (2015). A multi-dimensional trust-aware cloud service, selection mechanism based on the evidential reasoning approach. *International Journal of Automation and Computing*, 12(2), 208–219. <https://doi.org/10.1007/s11633-014-0840-3>
- [7] Supriya, M., Sangeeta, K., & Patra, G. K. (2016). Trustworthy Cloud Service Provider Selection using Multi Criteria Decision Making Methods, (February)
- [8] Ye, X. (2013). Dealing with unfair ratings. *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013*, 2951–2956. <https://doi.org/10.1109/SMC.2013.503>

- [9] Raghebi, Z. (n.d.). A New Trust Evaluation Method based on Reliability of Customer Feedback for Cloud Computing.
- [10] Huang J, Xu C, Duan Q, Ma Y, Muntean GM. Novel end-to-end quality of service provisioning algorithms for multimedia services in

virtualization-based future internet. IEEE Transactions on Broadcasting. 2012 Dec;58(4):569-79.

# Design of Automated Data Extraction System for Medical Web Forums using Semantic Analysis

Umamageswari Kumaresan,  
Asst. Professor, Dept. of IT,  
New Prince Shri Bhavani College  
of Engineering & Technology,  
Chennai, India.  
umamage@gmail.com

Kalpna Ramanujam,  
Professor, Dept. of CSE,  
Pondicherry Engineering College,  
Puducherry.  
rkalpana@pec.edu

**Abstract**—Medical Web Forums contain discussions on various medical issues. The discussions are organized in the form of threads and clicking on a thread, leads to series of web pages containing posts related the thread. Each post is considered as a data object with attributes such as date of post, author, content, number of views etc. Information available on such forums have variety of applications including disease diagnosis based on symptoms, adverse drug reactions, treatments etc. This paper exploits the fact that posts are generated using single server-side template and therefore detection of templates makes extraction of data objects easier. First, DOM tree corresponding to webpage is generated and the semantic type of the leaf nodes are determined based on rules expressed using regular expression. Then, XPath of leaf nodes corresponding to each attribute-value pair of the data object is determined. XPaths are stored as template in the database which is used for extraction of data objects from similarly structured pages.

**Keywords:** Medical forum; XPath; DOM tree; wrapper induction; RDF graph

## I. INTRODUCTION

Automatic extraction of data from semi structured pages poses several challenges. The page structure is heterogeneous across various web sites. Another major challenge is the presence of non-informative data which has keywords related to informative content. Also, the extracted data should be represented using data structure which facilitates carrying out analysis task. The process of extracting informative content from semi-structured HTML pages is known as scraping[1]. Medical Forums act as a rich source of information in the field of medicine as patients freely express their views, symptoms, side effects of medications etc. The information can be used for various purposes including disease detection based on symptoms, determination of adverse drug reactions, determination of medical examinations that has to be carried out based on symptoms and many more. Organization of web pages in Medical forum site is shown in fig. 1.

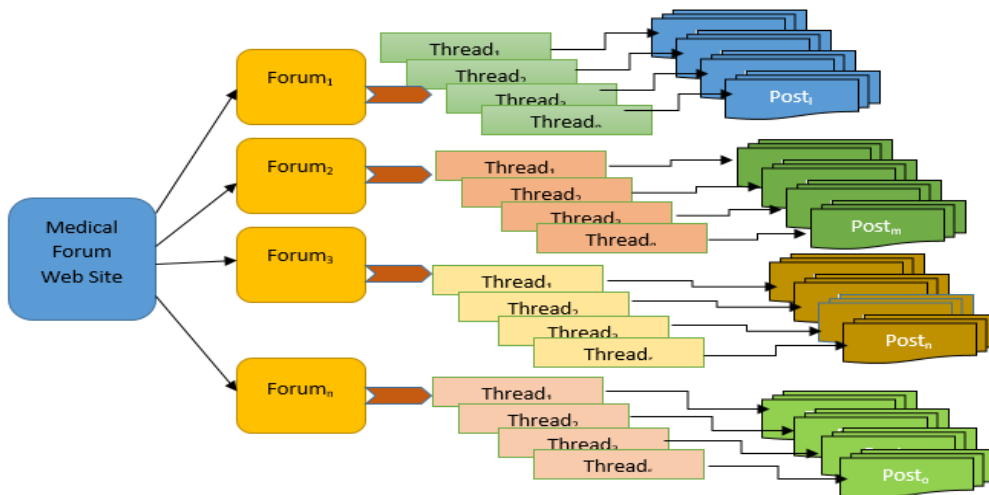


Fig. 1. Organization of posts in Medical Web Forum Websites

The Forum websites consist of collection of Forums arranged in chronological order. When we click on a link

corresponding to a Forum it leads to a web page containing collection of threads. As we click on a thread, it in turn

leads us to a web page containing collection of posts related to the thread, which is our target of extraction.

## II. RELATED WORKS

Based on the comprehensive review of web data extraction techniques [2], [3], [4] available so far in the literature, it can be classified into four types namely hand crafted, supervised, semi-supervised and fully automated techniques. Hand crafted wrapper induction techniques require extensive programming knowledge. Systems such as TSIMMIS[5], W4F[6] require extraction rules to be written using some declarative programming languages such as PERL, Python etc. Due to high level of human intervention needed in manually created wrapper induction systems, supervised techniques came into existence. Supervised techniques such as WIEN[7], SoftMealy[8], STALKER[9] automatically induces wrapper based on labeled training samples. The drawback of these techniques includes manually labeling training samples and inability to adapt to differently structured pages. Later, Semi-supervised techniques such as Thresher[10], IEPAD[11] and OLERA[12] came into existence. These techniques are capable of detecting templates automatically based on configuration file which specifies the target of extraction. These techniques face limitations such as manually labeling the data, post extraction. Finally many automated techniques such as EXALG[13], RoadRunner[14], DELA[15], DEPTA[16] and Stavies[17] came into existence. The challenges faced by automatic wrapper induction techniques include heterogeneous page structuring across sites and domains, which results in extraction of non-informative content or noise, in addition to the target of extraction and thereby reducing precision and recall values. In [18], a heuristic algorithm is designed for automatic extraction of data from scientific publishers' site which performs well for scientific publishing domain but it faces difficulty when applied to retailers' website. This is due to the dependency of extraction mechanism on domain keywords. Domain related keywords remain the same across web sites for scientific publishers' website, which is not the same for product websites'. In [19], a mechanism for extraction of data records from retailers' web sites based on semantic density and case based reasoning is proposed. Semantic density is used to discover nodes representing data rich region (DRR) and then, the data records within the DRR are extracted. It works well for extracting structured data records embedded in server-side templates. In [20], semi-supervised extraction of posts from Medical Forum Websites has been proposed in which the extraction is done based on configuration file specified by the user. This system faces the limitation of manually writing configuration file which requires programming knowledge and adaptability of technique for differently structured pages. These limitations are addressed in this paper by designing a fully automated framework for DRR detection

### 1) Data Objects Extraction Phase

A medical forum site consists of collection of forums. URLs of the forums from which posts have to be retrieved

and extraction of posts. Extraction of posts has a wide variety of applications in the medical field including prediction of disease based on symptoms, medical complications and measures for mitigation etc.

## III. DESCRIPTION OF PROPOSED FRAMEWORK

### A. Mathematical Model

A Medical Forum Site (S) consists of finite set of forums (F) where each Forum (F<sub>i</sub>) contains a set of threads (T<sub>j</sub>). Each thread (T<sub>j</sub>) consists of finite collection of posts (P<sub>k</sub>) which contains the data objects of concern. Each post (P<sub>k</sub>) contains set of attribute, value pairs (A,V). Therefore, a medical forum site can be represented as follows:

Thus,

F is a subset of S, Where S is the set of Medical Forum Sites

T is the subset of F

P is the subset of T,

Where  $F = \{f_1, f_2, f_3, \dots, f_n\}$  is the set of forums,

$T = \{t_1, t_2, t_3, \dots, t_m\}$  is the set of threads,

$P = \{ \langle a_1, v_1 \rangle, \langle a_2, v_2 \rangle, \dots, \langle a_w, v_w \rangle \}$  is the set of posts.

### B. Architecture

The proposed framework has two major phases: Data Objects (posts) Extraction phase and Data Objects (posts) Representation Phase. The architecture of the proposed framework is shown in Fig. 2.

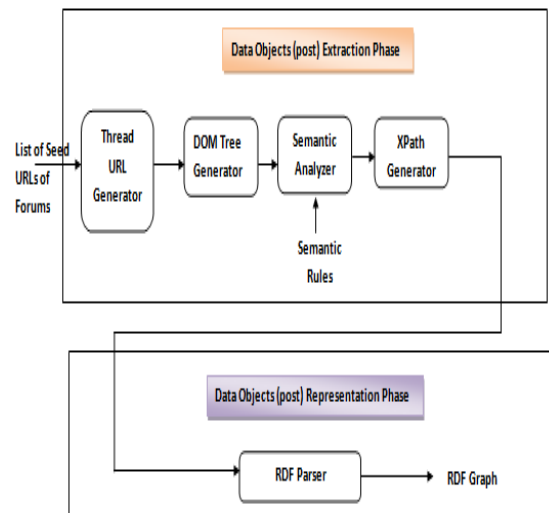


fig. 2. Architecture of WDE system for Medical Forum Website

are given as input. Navigate to forum URL to retrieve a list of thread URLs. Thread URLs are stored onto a file. Before extracting data from the thread, the URL is checked for uniqueness to avoid duplicates.

a) Thread URL Generation

The algorithm takes as input the seed URL obtained from the forum page and concatenates with the base\_url of the Forum site to get the absolute URL which is used during the extraction process.

```

Algorithm ThreadURLGenerator(SeedURL)
For each surl in SeedURL do
Thread_url.append(base_url+surl)
End for
    
```

Algorithm.1. ThreadURLGenerator

b) DOM tree Generator: Retrieve the HTML page corresponding to thread URL and convert it into DOM tree.

c) Semantic Analyzer: This module automatically identifies nodes in the DOM tree which contains the data of interest. In order to identify the semantic type of leaf nodes in the DOM tree semantic rules are used. For ex, if the node contents have more number of capitalized words compared total number of words then, its semantic\_type is title.

d) XPath Generator

Used to generate the XPaths of attribute value pairs whose semantic type matches with any one of the following: Ptitle, Pcontent, Pdate, Pauthor, Pcount.

```

Algorithm XPath_gen(DOM tree)
Identify the semantic type of leaf nodes and
determine the XPath
for each leaf_node whose semantic_type in
[Ptitle, Pcontent, Pdate, Pauthor, Pcount]
Determine XPath and store in file
end for
Determine XPath and store in file
end for
    
```

Algorithm.2. XPath\_gen

**Semantic Rules**  
 Semantic rules are expressed as regular expressions:  
 Ele: Identifies tags such as <img>, <a>, <b>, <p> etc.  
 W: identifies a word  
 CW: identifies capitalized word  
 (O)<sup>+</sup>: one or more occurrence  
 (O)<sup>\*</sup>: zero or more occurrence  
 (O)<sup>1</sup>: one occurrence  
 (O)?: zero or one occurrence  
 P<sub>title</sub>: Identifies title of post  
 P<sub>date</sub>: Identifies date of post  
 P<sub>content</sub>: Identifies description of post  
 P<sub>count</sub>: Identifies number of views of post  
 P<sub>author</sub>: Identifies author of post

**Rules for semantic types:**  
 Title: If content of the node has more than one word and number of capitalized words is greater than half the number of words in the content then, set node\_type as title.  
 P<sub>title</sub>: if ((W)<sup>+</sup> and count(CW) >= 1/2\*count(W)) then, assign node\_type = P<sub>title</sub>  
 Content: If content of the node has more than one word and number of capitalized words is lesser than half the number of words in the content then, set node\_type as content.  
 P<sub>content</sub>: if ((W)<sup>+</sup> and count(CW) < 1/2\*count(W)) then, assign node\_type = P<sub>content</sub>  
 If content of the node matches with the pattern corresponding to date then, assign node\_type as date.  
 P<sub>date</sub>: if O.matches(^([0-9]{1,2})/([0-9]{1,2})/([0-9]{1,2})) then, node\_type = P<sub>date</sub>  
 P<sub>author</sub>: if O.matches(<[A-Za-z][A-Za-z0-9]\*.\*? class=[""](byLineTag|byline|author|by)[""]>\*) then, node\_type = P<sub>author</sub>  
 Reply: If content of a node corresponds to number then, set node\_type as count.  
 P<sub>count</sub>: if O.matches(^([0-9]\*\$)) then, node\_type = P<sub>count</sub>

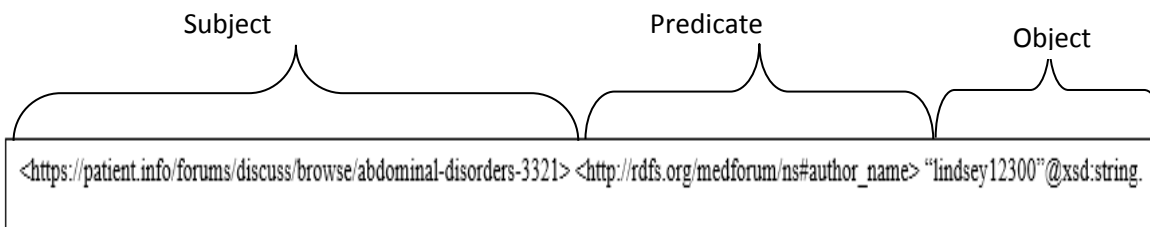


Fig. 3. N-triple format in RDF file

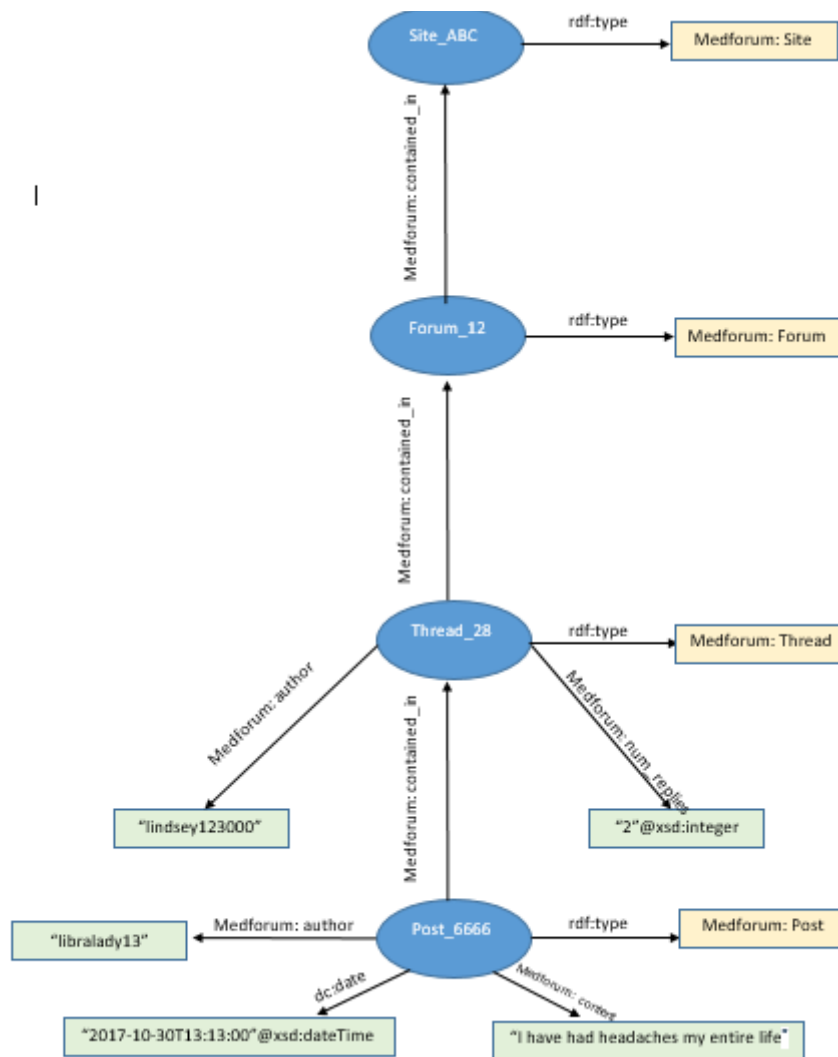


Fig. 4. Sample RDF sub-graph for a Medical Forum Website

## 2) Data Objects Representation Phase

RDF Graph[21] with N-triples syntax is exploited for representing the extracted data objects. Each line corresponding to N-triples file contains the triple: subject→predicate→object separated by whitespace and terminated by full stop (.) at the end of each triple.

Fig.3. shows sample N-triple in RDF file for author name and Fig.4. displays sample RDF sub graph for Medical Forum website. In this example, the post “Post\_6666” has the type “Medforum:Post, and it is associated with the author, text and date by the semantic relations “Medforum:author”, “Medforum:content”, and “dc:date”, respectively. The “Post\_6666” belongs to the thread “Thread\_28” (of type Medforum:Thread), which is associated with the author, and the number of replies via the semantic relations “Medforum:author”, “Medforum:num\_replies”, respectively. Finally, this thread appears in the forum “Forum\_12” (of type Medforum: Forum), which belongs to the website “Site\_ABC” (of type Medforum:Site).

## V. CONCLUSION AND FUTURE DIRECTIONS

In this paper, architecture for automatic extraction of posts from Medical web forums has been proposed. The use of semantic rules for automatic identification of data objects helps in application of technique to retrieve data from multiple heterogeneous web sites. Also, this system requires no user intervention. The extracted data is represented using RDF which facilitates carrying out data analysis task in an efficient manner. Anonymization of extracted posts needs to be addressed in future to ensure privacy.

## REFERENCES

- [1] Glez-Peña, D., Lourenco, A., Lopez Fernandez, H., Reboiro Jato, M., and Fdez Riverola., F. Web Scraping Technologies in an API World. Briefings in Bioinformatics, 15 (5):788-797, 2014. <http://doi.org/10.1093/bib/bbt026>
- [2] Schulz A, Lässig J, Gaedke M (2016) Practical web data extraction: are we there yet?—a short survey. In: 2016 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2016, Omaha, NE, USA, October 13–16 2016. IEEE Computer Society, pp 562–567

- [3] Umamageswari B, Kalpana R (2017) Web harvesting: web data extraction techniques for deep web pages. In: Kumar A (ed) *Web usage mining techniques and applications across industries*, pp 351–378
- [4] Varlamov MI, Turdakov DY (2016) A survey of methods for the extraction of information from webresources. *Program Comput Softw* 42(5):279–291
- [5] Hammer, J., McHugh, J., & Gracia-Molina, H. (1997). *Semistructured data: The TSIMISS experience*. Proceedings of the First East-European Symposium on Advances in Databases and Information Systems.
- [6] Sahuguet, A., & Azavant, F. (2001). Building Intelligent Web Applications using Lightweight Wrappers. *IEEE Transactions on Data and Knowledge Engineering*, 36(3), 283–316. doi:10.1016/S0169-023X(00)00051-3
- [7] Kushmerick, N., Weld, D., & Doorenbos, R. (1997). Wrapper Induction for Information Extraction. Proceedings of the Fifteenth International Conference on Artificial Intelligence.
- [8] Hsu, C.-N., & Dung, M. (1998). Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web. *Journal of Information Systems*, 23(8), 521–538. doi:10.1016/S0306-4379(98)00027-1
- [9] Muslea, I., Minton, S., & Knoblock, C. (1999). A Hierarchical Approach to Wrapper Induction. Proceedings of the Third International Conference on Autonomous Agents (AA-99). doi:10.1145/301136.301191
- [10] Hogue, A., & Karger, D. (2005). Thresher: Automating the Unwrapping of Semantic Content from the World Wide Web. Proceedings of the 14th International Conference on World Wide Web (WWW). doi:10.1145/1060745.1060762.
- [11] Chang, C.-H., & Lui, S.-C. (2001). IEPAD: Information Extraction based on Pattern Discovery. Proceedings of the Tenth International Conference on World Wide Web (WWW), Hong-Kong. doi:10.1145/371920.372182
- [12] Chang, C.-H., & Kuo, S.-C. (2004). OLERA: A Semi-Supervised Approach for Web Data Extraction with Visual Support. *IEEE Intelligent Systems*, 19(6), 56–64. doi:10.1109/MIS.2004.71
- [13] Arasu, A., & Garcia-Molina, H. (2003). Extracting structured data from Web pages. Proceedings of the ACM SIGMOD International Conference on Management of Data.
- [14] Crescenzi, V., Mecca, G., & Merialdo, P. (2002). Roadrunner: Automatic Data Extraction from DataIntensive Websites. SIGMOD. doi:10.1145/564691.564778
- [15] Wang, J., & Lochovsky, F.-H. (2003). Data extraction and Label Assignment for Web databases. Proceedings of the Twelfth International Conference on World Wide Web (WWW). doi:10.1145/775152.775179
- [16] Zhai, Y., & Liu, B. (2005). Web Data Extraction Based on Partial Tree Alignment. Proceedings of the 14th International Conference on World Wide Web (WWW). doi:10.1145/1060745.1060761
- [17] Papadakis, N., Skoutas, D., Topoulos, K.-R., & Varvarigou, T.-A. (2005). STAVIES: A System for Information Extraction from Unknown Web Data Sources through Automatic Web Wrapper Generation using Clustering Techniques. *IEEE Transactions on Knowledge and Data Engineering*, 17(12), 1638–1652. doi:10.1109/TKDE.2005.203
- [18] Umamageswari Kumaresan, Kalpana Ramanujam, "Web Data Extraction from Scientific Publishers' Website Using Heuristic Algorithm", *International Journal of Intelligent Systems and Applications(IJISA)*, Vol.9, No.10, pp.31-39, 2017. DOI: 10.5815/ijisa.2017.10.04
- [19] Umamageswari Kumaresan, Kalpana Ramanujam, "Web Data Extraction from Retailers' Site using Semantic Density and Case Based Reasoning", *ICIA-16: Proceedings of the International Conference on Informatics and Analytics*, Pondicherry Engineering College, April 2016. DOI: 10.1145/2980258.2980265
- [20] Audeh B, Beigbeder M, Zimmermann A, Jaillon P, Bousquet C (2017) Vigi4Med Scraper: A Framework for Web Forum Structured Data Extraction and Semantic Representation. *PLoS ONE* 12(1): e0169658. <https://doi.org/10.1371/journal.pone.0169658>.
- [21] Lassila O, Swick RR. Resource description framework (RDF) model and syntax specification. 1999. Available from <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222> (Linked accessed 11/2017)





# ICoAC 2017 Sponsors

**IEEE**

**CSIR, New Delhi**

**Centre for International Affairs  
Anna University**

**Centre for Technology Development and Transfer  
Anna University**

**Planning and Development  
Anna University**