

Volume 6, Issue 1 (VII)
January - March 2019

ISSN 2394 - 7780



ज्ञान-विज्ञान विमुक्तये
UGC
University Grants Commission
Journal No.: 63571

International Journal of
Advance and Innovative Research
(Conference Special)

Indian Academicians and Researchers Association
www.iaraedu.com



Hindusthan

College of Arts & Science

An Autonomous College - Affiliated to Bharathiar University
Approved by AICTE and Govt. of Tamilnadu
Accredited by NAAC-An ISO Certified Institution

5th International Conference on Computer Applications and Information Technology (CAIT 2019)

Organized by
PG and Research Department of Computer Applications
Hindusthan College of Arts and Science
Coimbatore

January 24, 2019



Publication Partner

Indian Academicians and Researcher's Association



Hindusthan College of Arts and Science

Coimbatore

Affiliated to Bharathiar University

Approved by AICTE and Govt of Tamilnadu

Accredited by NAAC

An ISO Certified Institution

Special Volume Editor

Dr. A. V. Senthil Kumar

Professor & Director

Department of MCA

Hindusthan College of Arts and Science

Coimbatore

ORGANIZING COMMITTEE

Chief Patron

Mr. T. S. R. Khannaiyann

Chairman,

Hindusthan Educational and Charitable Trust, Coimbatore

Chairman

Tmt. Sarasuwathi Khannaiyann

Managing Trustee

Hindusthan College and Charitable Trust, Coimbatore

Patrons

Thiru. K. Sakthivel

Trustee Administration

Hindusthan Educational and Charitable Trust, Coimbatore

Tmt. Priya Satishprabhu

Executive Trustee and Secretary,

Hindusthan Educational and Charitable Trust, Coimbatore

Co-Chairman

Dr. A. Ponnusamy

Principal,

Hindusthan College of Arts and Science, Coimbatore

Programme Chair

Dr. A. V. Senthil Kumar

Director, Department of MCA

Hindusthan College of Arts and Science, Coimbatore

Organizing Secretary

C. Daniel Nesa Kumar

Associate Professor, Department of MCA

Hindusthan College of Arts and Science, Coimbatore

Organizing Committee

A. Ganesan

Associate Professor, Hindusthan College of Arts and Science, Coimbatore

Dr. N. Revathy

Associate Professor, Hindusthan College of Arts and Science, Coimbatore

P. Hemashree

Assistant Professor, Hindusthan College of Arts and Science, Coimbatore

Dr. V. Kavitha

Associate Professor, Hindusthan College of Arts and Science, Coimbatore

Mrs. B. Sathyabama

Assistant Professor, Hindusthan College of Arts and Science, Coimbatore

TECHNICAL PROGRAM COMMITTEE

Dr. Ibrahiem M. M. El Emary, King Abdulaziz University, Jeddah
Dr. Adnan I. Al Rabea, Al Balqa Applied University, Al Salt, Jordan
Dr. Abdel-Badeeh M. Salem, Ain Shams University, Abbasia, Cairo-Egypt
Dr. SEDDIK BRI, Moulay Ismail University - Meknes, Morocco
Dr. Mohammad Israr, Dungarpur College of Engineering & Technology, Rajasthan, India
Dr. Monendra Grover, National Bureau of Plant Genetic Resources, New Delhi, India
Dr. Ehssan Mohamed Reda Nassef, Pharos University in Alexandria, Egypt
Dr. Shafqat Alauddin, Shibli National College, Azamgarh, India
Dr. R.V.Krishnaiah, DRK Institute of Science & Technology, Hyderabad, India
Dr. Nasib Singh Gill, Maharshi Dayanand University, Rohtan, India
Dr. Suvarna Nandyal, P.D.A. College of Engineering, Gulbarga, Karnataka, India
Dr. Mohammad Arif Kamal, Aligarh Muslim University, Aligarh, India
Dr. Boubker Sbihi, I- Shcool ESI, Morocco
Dr. M. Seetha, G. Narayanamma Institute of Technology and Sciences, Hyderabad, India
Dr. Mohammad Masdari, Urmia-Islamic Azad University, Iran
Dr. Madhumita Dash, Orissa Engineering College, India
Dr. Rahul M. Mane, Shivaji University, Kolhapur, India
Dr. Urmila Shrawankar, G H Raisoni College of Engineering, Nagpur, India
Dr. Sobhana N.V, Rajiv Gandhi Institute of Technology, Kottayam, India
Dr. Uma N. Dulhare, Muffham Jah College of Engg. & Tech. Andhra Pradesh, India
Dr. Abhishek Roy, St. Xavier's College, West Bengal, India.
Dr. Raman Kumar, DAV Institute of Engineering & Technology, Punjab, India
Dr. Muhammad Sarfraz, College of Computing Sciences & Engineering, Kuwait University, Kuwait
Dr. Anuj K. Gupta, Punjab Technical University, Punjab, India
Prof. T. V. Prasad, Chirala Engineering College, Andhra Pradesh, India
Dr. Shadi R. Masadeh, Isra University, Amman , 11622, Jordan
Dr. Rabinarayan Satpathy, HI-TECH Institute of Technology, Bhubaneswar, Odisha, India
Dr. Eng. Antonio Formisano, University of Naples 'Federico II', Italy
Dr. Ahmad Taher Azar, Benha University, Egypt
Dr. Amir J. Majid, Ajman University, UAE
Dr. Ang Kean Hua, Universiti Putra Malaysia, Malaysia
Dr. Warusia Yassin, Universiti Teknikal Malaysia Melaka
Dr. Marzanah A. Jabar, Universiti Putra Malaysia, Malaysia
Dr. Sasanko Sekhar Gantayat, GMR Institute Technology, Andhra Pradesh, India
Dr. Rahmita Wirza O.K. Rahmat., FSKTM, UPM, Malaysia
Dr. Ajay Goel, Baddi University of Emerging Sciences & Technology, Baddi, India
Dr. Shereen Aly Taie, Fayoum University, Egypt
Dr. Mahmood A. Mahmood, Cairo University, Egypt
Dr. H.R. Chennamma, Sri Jayachamarajendra College of Engineering, Mysore, India
Dr. Ismail Rakip Karas, Karabuk University, Demir Celik Campus, 78050 Karabük, Turkey
Dr. Biswajit Rath, GIT GITAM University, Andhra Pradesh, India
Dr. Mrinal Sarvagya, REVA University, Bengaluru, India
Dr. Venu Gopala Rao.Yadavalli, Vignan's University, Andhra Pradesh, India
Dr. Sadhana J. Kamatkar, University of Mumbai, India
Dr. Jairo R. Coronado-Hernandez, Universidad de la Costa, Colombia
Dr. Chaman Verma, Jan Nayak Ch. Devi Lal Vidyaeepth, Sirsa, India
Dr. K. Anitha, S. A. Engineering College, Anna University, India
Dr. D. Bala Krishna Kamesh, St. Martin's Engineering College, Hyderabad, India
Dr. M. Praveen Kumar, Malla Reddy Institute of Technology/JNTUH , Hyderabad, India
Dr. Garyfalia Charitaki, Faculty of Primary Education, University of Athens

ABOUT HINDUSTHAN EDUCATIONAL AND CHARITABLE TRUST

Hindusthan Educational and Charitable Trust, one of the finest in education and teaching is strategically placed in the heart of the city, and since 1992 has established itself firmly in the fields of Arts, Science, Education and Technical Education. The Trust aims at providing education that is world class and on the par with global standards. The Trust firmly believes in education from “Pre-KG to PhD” and is true to its motto – ‘Get the best in everything’. The Management has always stood by its commitment to the betterment of the student community and had at first established itself as a brand in the ‘Power Sector’ and today in the field of Education has reigned supreme with the ‘Life Time Educational Achievement Award’ for giving back to Society.

The Trust also believes in making education affordable to the underprivileged students and those from the weaker strata of the society are given educational scholarships. The various institutions run by Hindusthan Educational and Charitable Trust are:

Hindusthan College of Arts and Science

Hindusthan College of Engineering and Technology

Hindusthan Institute of Technology

Hindusthan College of Education

Hindusthan Polytechnic College

Hindusthan Matriculation and Higher Secondary School

Hindusthan Kangaroo Kids Pre School

ABOUT THE DEPARTMENT

The Programme Master of Computer Applications was started in the year 2002. The department is bestowed with an experienced team of Faculty members with excellent academic records and proficiency. Excellent infrastructure, State of art computers with the latest technology and comprehensive study materials make the department class apart. The department of MCA is fanatical towards the quality learning among students. The labs are well equipped with all software’s and students can develop extra potential beyond books for practical knowledge.

ABOUT IARA

Indian Academicians and Researchers Association (IARA) is an educational and scientific research organization of Academicians, Research Scholars and practitioners responsible for sharing information about research activities, projects, conferences to its members. IARA offers an excellent opportunity for networking with other members and exchange knowledge. It also takes immense pride in its services offerings to undergraduate and graduate students. Students are provided opportunities to develop and clarify their research interests and skills as part of their preparation to become faculty members and researcher. Visit our website www.iaaedu.com for more details.

ABOUT CONFERENCE

5th International Conference on Computer Applications and Information Technology (CAIT 2019) is an International Forum for the Presentation of Technological Advances and Research results in the fields of Computer Applications and Information Technology. The Conference will bring together leading Researchers, Engineers, Developers, Scientists and Practitioners from Academia and Industry working in all Interdisciplinary Areas of Computer Science and Information Technology to share their experience, and exchange their ideas. CAIT 2019 is organized by the PG and Research Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore-28, Tamilnadu, India.

International Journal of Advance and Innovative Research

Volume 6, Issue 1 (VII): January - March 2019

Editor- In-Chief

Dr. Tazyn Rahman

Members of Editorial Advisory Board

Mr. Nakibur Rahman

Ex. General Manager (Project)
Bongaigoan Refinery, IOC Ltd, Assam

Dr. Alka Agarwal

Director,
Mewar Institute of Management, Ghaziabad

Prof. (Dr.) Sudhansu Ranjan Mohapatra

Dean, Faculty of Law,
Sambalpur University, Sambalpur

Dr. P. Malyadri

Principal,
Government Degree College, Hyderabad

Prof.(Dr.) Shareef Hoque

Professor,
North South University, Bangladesh

Prof.(Dr.) Michael J. Riordan

Professor,
Sanda University, Jiashan, China

Prof.(Dr.) James Steve

Professor,
Fresno Pacific University, California, USA

Prof.(Dr.) Chris Wilson

Professor,
Curtin University, Singapore

Prof. (Dr.) Amer A. Taqa

Professor, DBS Department,
University of Mosul, Iraq

Dr. Nurul Fadly Habidin

Faculty of Management and Economics,
Universiti Pendidikan Sultan Idris, Malaysia

Dr. Neetu Singh

HOD, Department of Biotechnology,
Mewar Institute, Vasundhara, Ghaziabad

Dr. Mukesh Saxena

Pro Vice Chancellor,
University of Technology and Management, Shillong

Dr. Archana A. Ghatule

Director,
SKN Sinhgad Business School, Pandharpur

Prof. (Dr.) Monoj Kumar Chowdhury

Professor, Department of Business Administration,
Guahati University, Guwahati

Prof. (Dr.) Baljeet Singh Hothi

Professor,
Gitarattan International Business School, Delhi

Prof. (Dr.) Badiuddin Ahmed

Professor & Head, Department of Commerce,
Maulana Azad Nationl Urdu University, Hyderabad

Dr. Anindita Sharma

Dean & Associate Professor,
Jaipuria School of Business, Indirapuram, Ghaziabad

Prof. (Dr.) Jose Vargas Hernandez

Research Professor,
University of Guadalajara, Jalisco, México

Prof. (Dr.) P. Madhu Sudana Rao

Professor,
Mekelle University, Mekelle, Ethiopia

Prof. (Dr.) Himanshu Pandey

Professor, Department of Mathematics and Statistics
Gorakhpur University, Gorakhpur

Prof. (Dr.) Agbo Johnson Madaki

Faculty, Faculty of Law,
Catholic University of Eastern Africa, Nairobi, Kenya

Prof. (Dr.) D. Durga Bhavani

Professor,
CVR College of Engineering, Hyderabad, Telangana

Prof. (Dr.) Shashi Singhal

Professor,
Amity University, Jaipur

Prof. (Dr.) Alireza Heidari

Professor, Faculty of Chemistry,
California South University, California, USA

Prof. (Dr.) A. Mahadevan

Professor
S. G. School of Business Management, Salem

Prof. (Dr.) Hemant Sharma

Professor,
Amity University, Haryana

Dr. C. Shalini Kumar

Principal,
Vidhya Sagar Women's College, Chengalpet

Prof. (Dr.) Badar Alam Iqbal

Adjunct Professor,
Monarch University, Switzerland

Prof.(Dr.) D. Madan Mohan

Professor,
Indur PG College of MBA, Bodhan, Nizamabad

Dr. Sandeep Kumar Sahratia

Professor
Sreyas Institute of Engineering & Technology

Dr. S. Balamurugan

Director - Research & Development,
Mindnotix Technologies, Coimbatore

Dr. Dhananjay Prabhakar Awasarikar

Associate Professor,
Suryadutta Institute, Pune

Dr. Mohammad Younis

Associate Professor,
King Abdullah University, Saudi Arabia

Dr. Kavita Gidwani

Associate Professor,
Chanakya Technical Campus, Jaipur

Dr. Vijit Chaturvedi

Associate Professor,
Amity University, Noida

Dr. Marwan Mustafa Shamot

Associate Professor,
King Saud University, Saudi Arabia

Prof. (Dr.) Aradhna Yadav

Professor,
Krupanidhi School of Management, Bengaluru

Prof.(Dr.) Robert Allen

Professor
Carnegie Mellon University, Australia

Prof. (Dr.) S. Nallusamy

Professor & Dean,
Dr. M.G.R. Educational & Research Institute, Chennai

Prof. (Dr.) Ravi Kumar Bommiseti

Professor,
Amrita Sai Institute of Science & Technology, Paritala

Dr. Syed Mehartaj Begum

Professor,
Hamdard University, New Delhi

Dr. Darshana Narayanan

Head of Research,
Pymetrics, New York, USA

Dr. Rosemary Ekechukwu

Associate Dean,
University of Port Harcourt, Nigeria

Dr. P.V. Praveen Sundar

Director,
Shanmuga Industries Arts and Science College

Dr. Manoj P. K.

Associate Professor,
Cochin University of Science and Technology

Dr. Indu Santosh

Associate Professor,
Dr. C. V.Raman University, Chhattisgarh

Dr. Pranjal Sharma

Associate Professor, Department of Management
Mile Stone Institute of Higher Management, Ghaziabad

Dr. Lalata K Pani

Reader,
Bhadrak Autonomous College, Bhadrak, Odisha

Dr. Pradeepta Kishore Sahoo

Associate Professor,
B.S.A, Institute of Law, Faridabad

Dr. R. Navaneeth Krishnan

Associate Professor,
Bharathiyar College of Engg & Tech, Puducherry

Dr. Mahendra Daiya
Associate Professor,
JIET Group of Institutions, Jodhpur

Dr. Parbin Sultana
Associate Professor,
University of Science & Technology Meghalaya

Dr. Kalpesh T. Patel
Principal (In-charge)
Shree G. N. Patel Commerce College, Nanikadi

Dr. Juhab Hussain
Assistant Professor,
King Abdulaziz University, Saudi Arabia

Dr. V. Tulasi Das
Assistant Professor,
Acharya Nagarjuna University, Guntur, A.P.

Dr. Urmila Yadav
Assistant Professor,
Sharda University, Greater Noida

Dr. M. Kanagarathinam
Head, Department of Commerce
Nehru Arts and Science College, Coimbatore

Dr. V. Ananthaswamy
Assistant Professor
The Madura College (Autonomous), Madurai

Dr. S. R. Boselin Prabhu
Assistant Professor,
SVS College of Engineering, Coimbatore

Dr. A. Anbu
Assistant Professor,
Achariya College of Education, Puducherry

Dr. C. Sankar
Assistant Professor,
VLB Janakiammal College of Arts and Science

Dr. G. Valarmathi
Associate Professor,
Vidhya Sagar Women's College, Chengalpet

Dr. M. I. Qadir
Assistant Professor,
Bahauddin Zakariya University, Pakistan

Dr. Brijesh H. Joshi
Principal (In-charge)
B. L. Parikh College of BBA, Palanpur

Dr. Namita Dixit
Associate Professor,
ITS Institute of Management, Ghaziabad

Dr. Nidhi Agrawal
Assistant Professor,
Institute of Technology & Science, Ghaziabad

Dr. Ashutosh Pandey
Assistant Professor,
Lovely Professional University, Punjab

Dr. Subha Ganguly
Scientist (Food Microbiology)
West Bengal University of A. & F Sciences, Kolkata

Dr. R. Suresh
Assistant Professor, Department of Management
Mahatma Gandhi University

Dr. V. Subba Reddy
Assistant Professor,
RGM Group of Institutions, Kadapa

Dr. R. Jayanthi
Assistant Professor,
Vidhya Sagar Women's College, Chengalpattu

Dr. Manisha Gupta
Assistant Professor,
Jagannath International Management School

Copyright @ 2019 Indian Academicians and Researchers Association, Guwahati
All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, or stored in any retrieval system of any nature without prior written permission. Application for permission for other use of copyright material including permission to reproduce extracts in other published works shall be made to the publishers. Full acknowledgment of author, publishers and source must be given.

The views expressed in the articles are those of the contributors and not necessarily of the Editorial Board or the IARA. Although every care has been taken to avoid errors or omissions, this publication is being published on the condition and understanding that information given in this journal is merely for reference and must not be taken as having authority of or binding in any way on the authors, editors and publishers, who do not owe any responsibility for any damage or loss to any person, for the result of any action taken on the basis of this work. All disputes are subject to Guwahati jurisdiction only.



Journal - 63571

UGC Journal Details

Name of the Journal : International Journal of Advance & Innovative Research

ISSN Number :

e-ISSN Number : 23947780

Source: UNIV

Subject: Multidisciplinary

Publisher: Indian Academicians and Researchers Association

Country of Publication: India

Broad Subject Category: Multidisciplinary

CONTENTS

Research Papers

LUNG SEGMENTATION IN CHEST X-RAY IMAGES USING CANNY WITH MORPHOLOGY AND THRESHOLDING TECHNIQUES	1 – 7
V. Thamilarasi and Dr. R. Roselin	
ANALYSIS ON CLASSIFICATION OF PROTOCOL, ATTACKS & CHALLENGES IN MANET	8 – 13
N. Thangamani	
ANALYSIS ON DIGITAL IMAGE PROCESSING ENHANCEMENT TECHNIQUES AND THEIR ANALYSIS PROCESS	14 – 18
S. Saravanan	
ANALYSIS ON MOBILE CLOUD COMPUTING APPLICATION ADVANTAGES AND ISSUES	19 – 23
S. John Grasiyas	
TOWARDS INCORPORATING IOT AND BIG DATA IN HEALTH CARE INDUSTRY USING MACHINE LEARNING ALGORITHMS	24 – 26
Parimala. S and Dr. P. Senthil Vadivu	
IMPACTS OF IOT IN BIG DATA ANALYSIS	27 – 33
N. Suganthi and Dr. K. Sarojini	
NEW TRENDS IN OPINION MINING AND SENTIMENT ANALYSIS	34 – 39
R. Komalavalli and Dr. R. Vidyabanu	
A COMPREHENSIVE STUDY OF CURRENT TRENDS IN WEB MINING	40 – 46
Suhasini V and Dr. N. Vimala	
COMMUNICATION PROTOCOL FOR SECURED DATA TRANSFER ON IOT	47 – 49
Dr. P. Ananthi	
AN EFFECTIVE REPORTING SOLUTIONS FOR VALUE ADDED SERVICES	50 – 56
R. Sivakkolundu and Dr. V. Kavitha	
MANOEUVRING DATA MINING PERCEPTIONS FOR CRIME SCRUTINIZING THROUGH KNOWLEDGE REPRESENTATION USING AND-OR GRAPH	57 – 64
Bala Sneha. M, Saraswathi. S and Jayasri. G	
A STUDY ON VARIOUS LOAD BALANCING ALGORITHM IN CLOUD COMPUTING ENVIRONMENT	65 – 69
V. Malathi and Dr. V. Kavitha	
AN EFFICIENT WATERSHED TRANSFORMATION TECHNIQUE TO DETECT CANCER CELLS IN HUMAN ORGANS	70 – 74
P. Jayapriya	

INVENTIVE INQUIRY ON HEART DISEASE REPRESENTATION USING AND - OR GRAPH	75 – 82
Arulmozhi S, Saraswathi. S and Hemapriya. G	
STUDY ON LUNG CANCER DIAGNOSIS AND TECHNIQUES	83 – 89
K. S. Gowri laksshmi, Dr. R. Umagandhi	
INTELLIGENT ANALYSIS ON BREAST CANCER THROUGH KNOWLEDGE REPRESENTATION USING AND – OR GRAPH	90 – 96
Swetha. S, Saraswathi. S and Ayesha Begam. S	
USES OF ICT LEARNING TECHNIQUES TO IMPROVE STUDENTS LEARNING SATISFACTION	97 – 101
Nirmala Grace Rani S. and Dr. Joseph Catherine	
ANATOMIZATION OF MACHINE LEARNING TECHNIQUES AND ITS APPLICATIONS	102 – 120
R. Shinoj Robert and M. Maria Dominic	
COMPARATIVE STUDY ON CLUSTERING TECHNIQUES	121 – 124
Sathiyapriya. S and Dr. Kanagaraj. A	
INTERNET OF THINGS: SECURITY ISSUES, CHALLENGES AND COUNTER MEASURES	125 – 130
D. Sakthivel and R. Kumaramoorthy	
A SURVEY ON SOFTWARES FAILURE IN 2018	131 – 133
Ashin Thejas. J, Dinesh kumar. R, Sathyabama B and Dr. V. Kavitha	
CREDIT CARD FRAUD DETECTION TECHNIQUES AND EXISTING METHODS TO FRAUD DETECTION	134 – 137
S. Senthil Kumar, P. Balakrishnan and Sathyabama B	
PREDICTION OF OVARIAN CANCER CLASSIFICATION USING BACTERIAL FORAGING ALGORITHM WITH MULTIPLE SWARM OPTIMIZATION TECHNIQUES	138 – 141
C. J. Pavithra, G. S. Narmada and P. Prema and T. Guhan	
A SURVEY ON DATA MINING TECHNIQUES FOR HEART DISEASE PREDICTION	142 – 146
K. Dheenathayalan and Dr. K. K. Savitha	
WIRELESS ATTACKS	147 – 148
K. Baby and Dr. M. Shanthakumar	
A REVIEW ON SECURE AND DYNAMIC SEARCH METHODS OVER ENCRYPTED CLOUD DATA	149 – 153
P. Muthukumar, S. Sasikala and Renjith K V	
A STUDY AND ANALYSIS OF IMAGE MINING TECHNIQUES AND ITS APPLICATIONS	154 – 157
N. Sundaravalli and Dr. R. Vidya Banu	
A COMPARATIVE STUDY ABOUT OUTLIER DETECTION TECHNIQUES	158 – 163
R. Sinivasan, S. Sivasankar, Dr. V. Kavitha and B. Sathyabama	
HEARTBEAT PULSE DETECTION BY HEARTBEAT SENSING USING INTERNET OF THINGS: IOT	164 – 167
S. Subhasini, Dr. V. Kavitha, J. S. Anita lily and B. Satyabama	

BOUTIQUE MANAGEMENT SYSTEM	168 – 170
Semitha Esthar Rani S, Shanmuganathanj, Dr. V. Kavitha and G. Subashini	
CERTAIN INVESTIGATIONS ON INTEGRATED MATERIALIZED VIEW BASED APPROACH IN ETL WITH DSS TO ACHIEVE FAST DATA TRANSFORMATION	171 – 176
A. Uma Maheswari and Dr. N. Revathy	
STAGING IMPROVEMENT OF ENERGY OPTIMIZATION IN WSN USING BIO-INSPIRED ENVIRONMENT	177 – 182
B. Karthik and Dr. N. Revathy	
ADVANCEMENT OF CLOUD COMPUTING IN HEALTHCARE SECTOR	183 – 189
S. S. Shanthi	
DISCRIMINATIVE PATTERN WITH MULTIPLE SUPPORT VECTOR MACHINE (MSVM) PREDICTION FRAMEWORK (DPMSvmPF)	190 – 196
J. Jeyaboopathiraja and Dr. G. Maria Priscilla	
DIGITAL MARKETING IN INTERNATIONAL ARENA	197 – 202
P. Chandradevi, A. Priya and M. Subhashini	
AN INTRODUCTION OF EDGE DETECTORS FOR BRAIN TUMOR	203 – 205
B. Chitradevi ¹ , N. Thinaharan and G. Suganya	
AN INTRODUCTORY STUDY ON WEB BASED SUPPORT SYSTEM	206 – 209
V. Rengaraj and V. Thiagarasu	
AN EFFICIENT FUZZY TECHNIQUE INCORPORATED WITH MAMMOGRAM IMAGE ENHANCEMENT	210 – 214
Dr. T. A. Sangeetha and Dr. N. Revathy	
SURVEY ON MACHINE LEARNING WITH BIG DATA SOFTWARE DEVELOPMENT	215 – 217
Dr. M. Thangamani, K. PavithKumar, T. Preethi	
E-COMMERCE	218 – 222
Subha S, Blessy Ebanazar. D and Carolin Joyce Paiva	
DATA MINING APPLICATION IN CREDIT CARD FRAUD DETECTION	223 – 226
J. Karunanithi and S. Jagadesh Kumar and V.Ponmagal	
WEB SERVICES	227 – 231
Sindhu T	
CLOUD COMPUTING SECURITY ISSUES & CHALLENGES..,	232 – 237
Chandra Devi P, Afritha A. Nazar and Anjana A	
DIABETES MELLITUS (DM) PREDICTION AND CLASSIFICATION USING BAT OPTIMIZATION BASED ON ADAPTIVE NEURO-FUZZY	238 – 244
R. S. Padma Priya and Dr.P.Senthilvadivu	
CLASSIFICATION PROBLEMS IN SUPERVISED MACHINE LEARNING	245 – 249
Karpagadeepa S, Theinmozhi M. and Vasuki S	
DATA MINING AND ITS IMPLICATIONS	250 – 254
B. Saravana Kumar and Dr. P. Senthil Vadivu	

ESSENTIALS OF CLOUD COMPUTING WITH ITS CHALLENGES	255 – 260
S. Kowsalya, B. Reethu and S. Suprraja	
AUTOMATIC TIMETABLE GENERATOR	261 – 263
Aiswarya Lekshmi and Dr. A. V. Senthil Kumar	
CREDIT CARD FRAUD DETECTION SYSTEM	264 – 266
Abidh Ali A and Dr. A. V. Senthil Kumar	
PUBLIC DISTRIBUTION SYSTEM	267 – 269
Viswapal Venugopal and Dr. A. V. Senthil Kumar	
PERFORMANCE EVALUATION OF MOBILE ADHOC NETWORK USING DYNAMIC SOURCE ROUTING PROTOCOL	270 – 278
S. Sasikala	
CRIME RECORDS MANAGMENT SYSTEM	279 - 285
Aathinamarun, Dr. A. V. Senthilkumar	

LUNG SEGMENTATION IN CHEST X-RAY IMAGES USING CANNY WITH MORPHOLOGY AND THRESHOLDING TECHNIQUES

V. Thamilarasi¹ and Dr. R. Roselin²Research Scholar¹ and Associate Professor², Sri Sarada College for Women (Autonomous), Salem

ABSTRACT

Segmentation of lung region in image is a tedious and time consuming process. Every Chest X-ray image contains all the regions with irrelevant information. There is a lot of CXR based segmentation methods but some of them concentrate on noise removal, remaining aims at segmentation only. Lung Cancer is one of the fearful and painful challenging diseases in both human and medical world. At the Same time, rising complication of environmental pollution created by vehicle, smoking and wrong diet collapse respiratory system of human body. Hence Lungs are mostly affected part at any age group which results in cancer. In olden days cancer identification and treatment takes more time and maximum results to failure. So identification of early stages of cancer increases life rate of humans and proper segmentation of lungs and there by identify the nodule and non nodule portion of lungs leads hope to success of cure from cancer. Now a days medical field uses many computer aided technologies to identify cancers. This paper discusses some techniques likes Binarization edge detection, Morphology and Thresholding method to identify and segment the lung region. The results are compared using Jaccard Similarity coefficient to identify best technique for Segmentation of lung region. Hence this paper considers both methods and compare ground truth image with segmented image and estimate which gives better result.

Keywords: Chest X-ray images, Canny edge detection, Morphology, Thresholding, Jaccard Similarity Coefficient.

I. INTRODUCTION

Lung region segmentation is the basic and important task in identification of lung cancer. Mostly detection of cancer needs proper region of lungs from other regions in human body. Because chest bone covers most of the organs in our body and separate each organs from one another is a challenging and cumbersome process in medical field. Now a days many Computer aided techniques (CAD) are used to detect the lung cancer in Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Sputum Cytology and Chest Radiography (X- ray). In 2016 the lung cancer patients range increased from 3170 to 4788 and so it becomes the emerging research area. In 2020 there will be red alert in lung cancer [11]. Most of the patients are still in non-identifiable stage of cancer. Sometimes it leads to wrong treatment and it increases death rate of cancerous patients. Computer aided tools give better result for these problems. In medical results Chest Radiography (X-ray) needs more help from image processing techniques for enhancement, to give perfect perception of image. There are variety of methods available for cancer detection particularly lung segmentation and cancer in the literature. At the same time they are unable to produce best result. So in this paper Canny, Morphology, Thresholding methods used to segment the lung region and accuracy of those results were compared with Jaccard Similarity coefficient method and give best solution for lung segmentation. JSRT database is used for experiment. It is the standard digital image database with and without chest lung nodules. It contains 247 Chest X-ray images and also contains information about patient age, gender, diagnosis (malignant or benign), X and Y coordinates of nodule, simple diagram of nodule location, degree of subtlety in visual detection of nodules. In this experimental setup 100 images were taken for segmentation, only 10 image results are presented visually in this paper. Section II summarizes the related work; Section III explains the proposed system. Implementation platform explained in Section-IV. Results and discussions are given in Section V, Conclusion about this paper explained In section VI.

II. RELATED WORKS

This paper mainly focuses on lung region segmentation from chest X-ray images. There are lot of works in the literature in lung segmentation, some of them are Aroop Mukherjee et al.,[2010] [1] used binarization to enhance the image resolution by the combination of threshold with binarization and also present algorithms to execute it . A. Dawoud., [2010] [2] first computed lung shape information like size, orientation, eccentricity, minor and major ellipse lengths, and left and right lung centroid locations and compared tested results of statistical model with iterative binarization method and similarity between these methods were measured by using mahalanobis distance, finally proposed method to compare with ASM and AAM method. Fatemeh Shahsavari Alavijeh et al.,[2015] [3] enhance the chest X-ray images by using morphological operations . M. Gomathi et al., [2010] [5] analyse different C-means clustering algorithm in noisy images and proved Fuzzy

Possibilistic C-means Algorithm gives good result for Lung Segmentation. Hui Luo et al., [2000] [6] used object –oriented knowledge model to detect lung portion. M.Jannathl Firdouse et al., [2017] [7] survey about seven recent lung segmentation papers and discuss about algorithms and drawbacks of methods , S.Juhasz et al., [2017] [8] develop a rip detection algorithm to find accurate rips there by identify the clear portion of lung . Mercy Theresa M et al., [2017] [9] used watershed for segmentation and feature extraction done by using five different transforms on all 247 images in JSRT database. The result of this method is analysed using different statistical feature like Energy, Entropy, Mean and Median. Mohd Nizam Saad et al., [2014] [10] used canny and morphological method to segment the lung region and they used Jaccard similarity coefficient method to compare and find the accuracy of proposed method. Otsu , N., [1979] [11] discusses about automatic threshold selection for image segmentation . P.M.Parikh et al., [2016] [12] discussed the need for CAD techniques in Lung cancer, and mention some people biological status in Lung cancer. Reddi, S., et al., [1984] [13] discussed about finding single and multiple threshold and also implement with some examples . Sema Candemir et al., [2014] [14] used different chest x-ray database sets and detect the lung boundaries using SIFT-Flow Non-Rigid Registration, Graph Cut Segmentation and fast partial Randon profile similarity selection and predict the success rate of datasets. R.Senthilkumar et al ., [2014] [15] used different enhancement techniques like HE, AHE, CLAHE in Chest X-ray images and estimated that CLAHE gives best result than other two. Senthil kumaran, N., et al., [2016] [16] used best thresholding algorithm and also compared with Niblack and sauvola local thresholding algorithm and segmented image measure by using Jaccard similarity coefficient method and PSNR. Tsai.D.M.,[1995] [17] Tao Xu et al., [2012] [18] developed new technique with a combination of global edge and force field information for lung segmentation and compared with ASM and hybrid LSSP methods and declared that new technique gives better result. B. Van Ginneken et al [2000][19] compared different segmentation techniques and proposed new method called hybrid scheme, it's a combination of the rule-based segmentation algorithm and pixel classification approach which gives good result for segmentation. Vanita Naronha et al.,[2016][20] discussed about cancer creating elements and recent therapy of lung cancer. Weszka, J.S., et al [1978] [21] discussed about various threshold evaluation techniques . Zhao Yu-qian et al .,[2005] [22] used different noise removal techniques and edge detectors and mathematical morphological operations like dilation, erosion, opening, closing , structuring elements and concluded mathematical morphological operation is suitable for edge detection. S. Zulaikha Beevi et al.,[2010] [23] used improved FCM algorithm to avoid the sensitivity of noise in conventional FCM algorithm.

III. PROPOSED SYSTEM

3.1 The Proposed Work

The proposed work aims at segmentation of lung region with less noise and also reduces the processing steps to get the exact segmented lung portion. Two techniques are used to segment the lung region; First method is based on Canny edge detection with Morphology. Second method based on Thresholding. The resultant image of these techniques compared with ground truth image of JSRT database. This comparison complete with the help of Jaccard Similarity coefficient method and provides best technique based on higher percentage accuracy of result .

3.2 Image Dataset Acquisition

The standard digital image database with and without chest lung nodules was created by the Japanese Society of Radiological Technology (JSRT) in cooperation with the Japanese Radiological Society (JRS) in 1998. The JSRT database has been used by a number of researchers in the world for various research purposes such as image processing, image compression, evaluation of image display, computer-aided diagnosis (CAD), picture archiving and communication system (PACS), and for training and testing. The Chest X-ray images were taken from Japan Society of Radiological Technology (JSRT) public at <http://www.jsrt.or.jp/jsrt-db/eng.php>. The database contains 247 images and it is categorised based on nodule and non-nodule .These images were at size of 2048×2048 pixels . JSRT also provide special software called ImageJ, which is used to get images from database . By using this software we import image as a raw image and save it as jpeg format for further processing. In this paper 1024×1024 pixel image used for processing.

3.3 Segmentation Using Edge Detection and Morphology

The proposed method of segmentation of the lung region is based on shape of the lung. To segment the lungs, many techniques used like binarization, canny edge detection, morphological operations. First the gray scale image is converted into binary image which gives basic segmented portion of the lungs, by manually setting threshold values. Second binary image is processed using canny edge detection. Third morphological operations erosion and dilation techniques applied on edge detected image . After applying these operations we got visible edge of segmented lungs. Finally segmented lung regions filled to get the segmented lungs. Segmented portion

is compared with ground truth segmentation which is taken from JSRT database, Jaccard similarity coefficient method is used to measure the accuracy of segmentation.

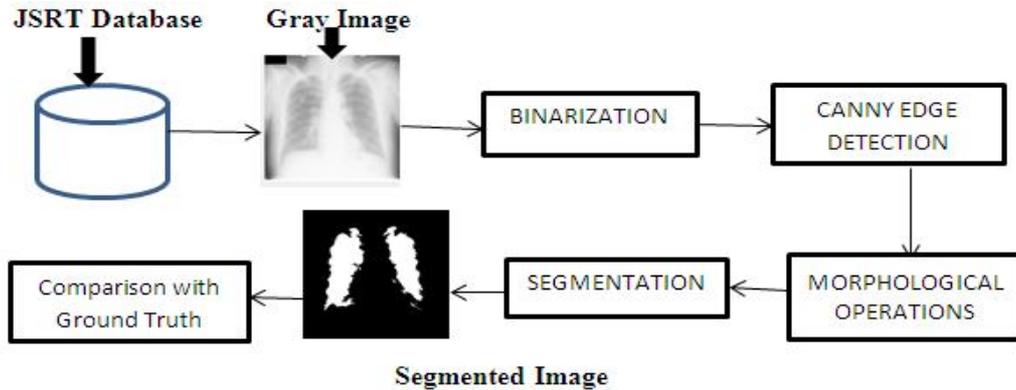


Figure-1: Edge Detection and Morphology

3.3.1 Binarization

The enhanced image is converted to binary image to identify background and foreground from the image. So the image is converted into binary image based on threshold values which set manually by every image. It gives the basic segmented image.

3.3.2 Canny Edge Detection

Edge detection is a basic tool for image segmentation. It transforms original images into edge images like Gray tones. Basically it detects the object and its boundaries and its background in the image. Edge detection implements set of mathematical methods which aim at identifying points in a digital image at where the image brightness changes sharply or, more formally, has discontinuities. The points at which image brightness changes sharply are typically organized into a set of curved line segments termed edges. So edge detection is a fundamental part for machine vision and computer vision. Edge detected image is find by using canny edge detection method.

3.3.3 Morphological Erosion & Dilation

Mathematical morphology is based on the concept of mathematical set theory. It is a powerful tool for extracting feature from an image and examining shapes in an image. Dilation, Erosion, Opening, Closing are different methods of morphology. Dilation operation thickens the image. Erosion performs thinning the image. Erosion followed by dilation is named as opening. Dilation followed by erosion is named as closing. Here Dilation and Erosion operations are applied to canny image to improve the edges. Dilation used to compute, every image pixel, and the maximum value of its neighboring pixels. The structuring element used to define neighbourhood. Erosion used to compute minimum value of its neighboring pixels. Finally the lung image with boundary of both left and right lungs are obtained. After morphological operation perfect boundary of image is identified. Finally lung region is segmented after fill the regions successfully.

3.3.4 Jaccard Similarity Coefficient

It is used to compare segmented image with ground truth image. Thereby identify the accuracy of the computed result. The function used to compute the Jaccard similarity coefficient is $Jaccard(bw1, bw2)$. Here bw1 is ground truth image and bw2 is segmented image. The accuracy value is high means the segmentation gives good result.

Following figure shows the segmentation result of canny & morphological method.

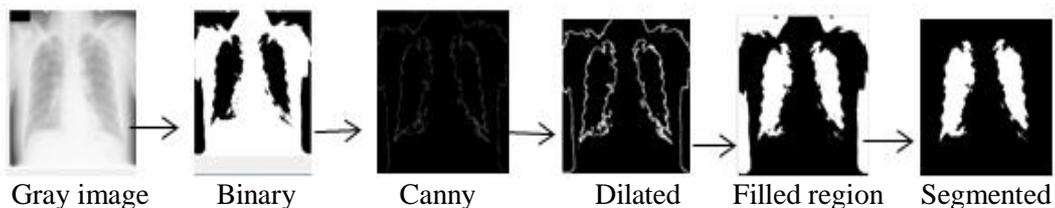


Figure-2: Canny and Morphological method Based Segmentation

3.4 SEGMENTATION USING THRESHOLDING

Based on several research work, the aim of Thresholding segmentation is to get fast and noiseless segmented images. Image segmentation is the method of partitioning the image into several segments. The aim of segmentation is to present the image portions in accurate and identifiable manner. First the gray image is taken

from JSRT database by ImageJ tool. Second the grayscale image contrast is enhanced by using histogram equalization. Third contrast enhanced image is converted to binary image by fixing appropriate threshold values. Fourth segmented image background and connected lines are removed to get exact portion of lung region, the holes are filled. Finally appropriate lung portion is segmented. To evaluate the accuracy of segmentation Jaccard Similarity coefficient is use. Figure 3 shows the working method of Threshold based segmentation.

JSRT Database Gray image

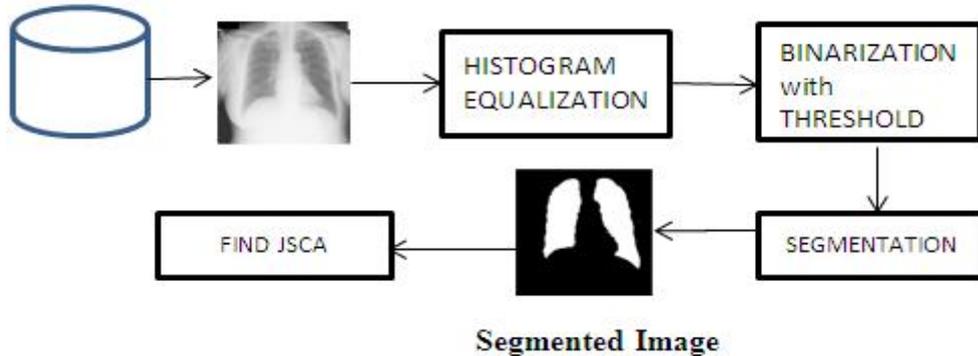


Figure-3: Threshold based Segmentation

3.4.1 Histogram Equalization

It is used to enhance the low contrast image into high contrast one by frequent intensity values. It is more efficient when compared with other image segmentation methods because it requires only one pass through the pixels. In medical field, the digital X-rays are displayed in different colour, based on observation of light, patient and surroundings. Therefore bone and water represented by white structure and air represented by black structure. So for pathologists the contrast variation or similarity leads difficulty in achieving accurate diagnosis. Thus contrast enhancement is achieved by using histogram equalization. This method is very useful in close contrast value images because there only the brightness of both foreground and background are same. Histogram equalization is a point operator, so the histogram of output image is constant. Often it used to correct for varying illumination conditions.

3.4.2 Thresholding and Binarization

Thresholding is basic and important method for image segmentation. Binary image is always the output of thresholding. It is an effective method of partitioning an image into back ground and fore ground. In digital image processing book Gonzalez and Woods defined T as threshold and $f(x,y)$ as represent background and foreground points if $f(x,y) > T$ the point is foreground and if $f(x,y) < T$ the point is background. The thresholded image $g(x,y)$ is represented by

$$g(x,y) = \begin{cases} a & \text{if } f(x,y) > T \\ b & \text{if } f(x,y) \leq T \end{cases}$$

a = foreground and b = background

ie a = 1(White) and b = 0 (Black) [24].

Insufficient contrast, bad illumination, gray level variations reduce the performance of thresholding. To overcome this problem thresholding is done with binarization.

Binarization is another principle segmentation method of image processing. It also convert grayscale image into black and white image, pixel intensity range from 0 to 255. It is very important to fix correct threshold value to binarize the image. It is achieved by using global thresholding. Thresholded binary image is segmented by using available matlab functions and reduce the structures lighter and which are connected to the image border and fills holes in the input binary image. Finally the segmented image is compared with ground truth image by using Jaccard Similarity coefficient method. Figure 4 shows the Segmentation using Thresholding

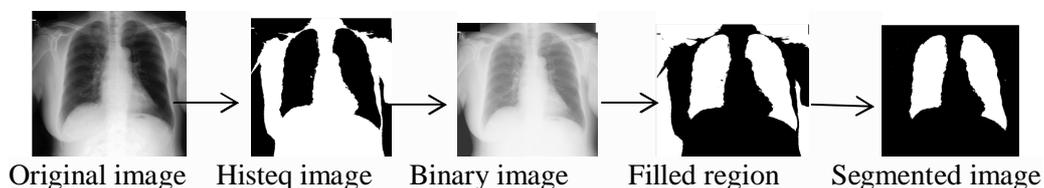


Figure -4: Segmentation using Thresholding

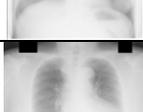
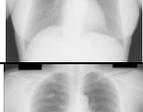
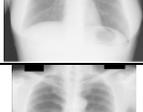
IV. IMPLEMENTATION

MATLAB is used for processing. It is a powerful tool for technical computing and also it can be used for algorithm development, data analysis, math computations, visualization and graphics, modelling and simulations. MATLAB environment gives required functions and tools from toolboxes.

V. RESULT AND DISCUSSION

To analyse the performance of segmentation, set of 100 images were tested. For all the images only results of 10 images presented in the following table. Both results verified by using Jaccard Similarity Coefficient method. The performance of canny with morphological segmentation gives average of 76% accuracy whereas thresholding method gives nearly 80% accuracy. So compared by this accuracy value of both methods, Experimental result reveals that that Thresholding method achieves higher performance rate than another one. The following Table 1 shows the original image, ground truth image and the similarity co-efficient for both methods

Table-1: The Jaccard Similarity Coefficient of Segmented and Ground Truth Images

Original Gray Image	Ground Truth Image	Segment I	Segment II	JSCA I Segment	JSCA II Segment
				0.7571	0.7840
				0.7337	0.7569
				0.7686	0.7923
				0.7867	0.7997
				0.8467	0.8613
				0.7248	0.7182
				0.72	0.7466
				0.7588	0.7879
				0.7450	0.7717
				0.7875	0.7986

Segment I - Canny & Morphological Segmented Image

Segment II - Threshold based Segmented Image

JSCA - Jaccard Similarity Coefficient Accuracy

VI. CONCLUSION

Today's environment Lung cancer is a destroyer of human life. So there is a need for good procedure to identify the early stage of lung cancer. Still now most of the research work ended with some limitation. In this paper two segmentation methods are taken for research and based on accuracy it is concluded that Thresholding is the best method for lung segmentation and success rate of this work is 80%. This work is a basic one, because it focussed only lung regions and its result also not exceeded previous related research work results. In future this technique can be experimented by all 247 images and apply different techniques for segmentation and identify which gives better performance in lung segmentation.

REFERENCES

1. Aroop Mukherjee, Soumen Kanrar, (2010), 'Enhancement of image Resolution by Binarization', *International Journal of Computer Applications*, 10(10) .
2. Dawoud, A, (2011), 'Lung Segmentation in Chest Radiographs by fusing shape Information iterative thresholding', *IET Computer Vision*. 5:185-190 .
3. Fatemeh Shahsavari Alavijeh and Homayoun Mahdavi Nasab,(2015), 'Multiscale Morphological image Enhancement of Chest Radiographs by a Hybrid Scheme', *J Med Signals Sens*, 5(1) : 59–68.
4. Gonzalez and Woods, *Digital image processing*, 2nd Edition, prentice hall, pp. 511, 2002.
5. Gomathi, M., Thangara, Dr.P., (2010), 'A New Approach to Lung Segmentation using Fuzzy Possibilistic C-Means Algorithm', *International Journal of Computer Science and Information Security*, 7(3) .
6. Hui Luo, Gaborski, R., Acharya, R., (2000), 'Automatic Segmentation of Lung Region in Chest Radiographs: A Model Guided Approach', *proceedings 2000 International conference on image processing*, ISBN: 0-7803-6297-7., (IEEE Explore):10-13,
7. Jannathl Firdouse, M., Balasubramanian, Dr.M. ,(2017), 'A Survey on Lung Segmentation Methods', *Advances in Computational Science and Technology*, 10(9) : 2875-2885 .
8. Juhasz , S., Horvath, A., Nikhazy, L.,Horvath, Horvath, A., 'Segmentation of anatomical structures on chest radiographs' , ARON_Sanyi_ver2_corr.doc .
9. Mercy Theresa, M .,Subbiah Bharathi, V., (2017) 'Computer aided diagnostic (CAD) for Feature extraction of lungs in chest radiograph using different transform features', *Biomedical Research - India special issue* S208 – S214 .
10. Mohd Nizam Saad, Zurina Muda, Noraidah Sahari@Ashaari, Hamzaini Abdul Hamid, (2014), ' Image Segmentation for Lung Region in Chest X-ray Images using Edge Detection and Morphology', *IEEE International Conference on Control System, Computing and Engineering* : 28-30 .
11. Otsu, N., (1979), 'A Threshold Selection Method from Gray-Level Histogram', *IEEE Transactions on System, Man and Cybernetics*, 9(1) : 62-63.
12. Parikh, P.M., Ranada,A.A., Babu Govind, Ghadyalpatil, N., Singh,R.,Bharath, R., Bhattacharyya, G.S. , Koyande, S., Singhal, M., Vora, A., Verma, A., and Hingmire, S., (2016), 'Lung Cancer in India : Current status and promising Strategies' *South Asian Journal of Cancer*, 5(3) : 93-95.
13. Reddi, S., Rudin, S., and Keshavan, H.,(1984) 'An optimal multiple threshold Scheme for image segmentation', *IEEE Trans. Syst, Man Cybern.*, 14 : 661-665, 1984.
14. Sema candemir, Stefan Jacger, Kannappan Palaniappan, Jonathan P.Musco, Rahul K.Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Zhoma, and Clement J.McDonald, (2014) , 'Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration", *IEEE transactions on medical imaging*, 33(2) .
15. Senthilkumar, R., Senthilmurugan, Dr.M. ,(2014), 'Traid Histogram to Enhance Chest X-ray image", *International journal of Advanced Research in Computer and Communication Engineering*, 3(11).
16. Senthil kumaran, N., Vaithegi, S.,(2016), 'Image Segmentation by Using Thresholding Techniques For Medical Images', *Computer Science & Engineering: An International Journal (CSEIJ)*, 6(1) .
17. Tsai.D.M.,(1995), 'A fast thresholding selection procedure for multimodal and Unimodal histograms', *Pattern Recognit. Lett*, 16(6) : 653-666 .

-
18. Tao Xu, Mrinal Mandal, Richard Long, Irene Cheng, Anup Basu, (2012), 'An edge region force guided active shape approach for automatic lung field detection in chest radiographs', *Computerized Medical Imaging and Graphics*, 36(6) : 452-463 .
 19. Van Ginneken, B., ter Romeny, B. M., (2000), 'Automatic segmentation of lung fields in chest radiographs', *Medical physics*, 27(10) : 2445-2455 .
 20. Vanita Naronha, Rakesh Pinnini, Vijay M. Patil, Amit Joshi, and Kumar Prabhash, (2016), 'Lung Cancer in the Indian subcontinent', *South Asian Journal of Cancer*, 5(3) : 95-103 .
 21. Weszka, J.S., Rosenfeld, A., (1978) , Threshold evaluation techniques, *IEEE Trans. Systems, Man and Cybernetics* , SMC-8(8) : 627-629 .
 22. Zhao Yu-qian, Gui Wei-hux, Chen Zhen-cheng, Tang jing-tian, Li Lingyun, (2005), 'Medical Images Edge Detection Based on Mathematical Morphology', *Engineering in Medicine and Biology 27th Annual Conference (Proceedings of IEEE)*
 23. Zulaikha Beevi, S., Mohammed Sathik, M., Senthamaraikannan, K., (2010), 'A Robust Fuzzy Clustering Technique with Spatial Neighborhood Information for Effective Medical Image Segmentation', *International Journal of Computer Science and Information Security*, 7(3) .

ANALYSIS ON CLASSIFICATION OF PROTOCOL, ATTACKS & CHALLENGES IN MANET

N. Thangamani

Assistant Professor, Department of Computer Science, AJK College of Arts and Science, Coimbatore

ABSTRACT

Headway in the field of internet because of wireless networking technologies offers ascend to numerous new applications. Mobile specially appointed network (MANET) is a standout amongst the most encouraging fields for research and advancement of wireless network. As the ubiquity of mobile gadget and wireless networks essentially expanded over the previous years, wireless specially appointed networks has now turned out to be a standout amongst the most dynamic and dynamic field of communication and networks. A mobile specially appointed network is an independent gathering of mobile gadgets (PCs, smart phones, sensors, etc.) that speak with each other over wireless links and coordinate in a circulated way so as to give the important network functionality without a settled infrastructure. This paper analyzed to kinds of attacks, security challenges and different protocol classifications.

Keywords: Protocol, Classification, MANET, Security, Routing.

1. INTRODUCTION

Specially appointed wireless network is an accumulation of wireless mobile nodes that self-design to build a network without the requirement for any established infrastructure or backbone. Specially appointed networks are another wireless networking worldview for mobile hosts. Unlike traditional mobile wireless networks, specially appointed networks don't depend on any settled infrastructure. Impromptu networks utilize mobile nodes to empower communication outside wireless transmission run. Because of the nonappearance of any settled infrastructure, it ends up hard to make utilization of the current routing techniques for network services, and this represents a number of challenges in guaranteeing the security of the communication. A considerable lot of the specially appointed routing protocols that address security issues depend on understood trust relationships to route packets among taking part nodes. The general security goals like authentication, confidentiality, respectability, accessibility and non-denial, the impromptu routing protocols should also address location confidentiality, participation decency and nonattendance of traffic preoccupation. In opposition to the traditional wireless network, MANET has a decentralized network infrastructure. MANET does not require a settled infrastructure; thus, all nodes are allowed to move arbitrarily. MANET is equipped for making a self-designing and self-keeping up network without the help of a centralized infrastructure, which is regularly infeasible in critical mission applications like military clash or crisis recuperation. Minimal configuration and quick organization make MANET prepared to be utilized in crisis conditions where an infrastructure is inaccessible or unfeasible to install in situations like natural or human-incited catastrophes, military clashes, and medical crisis circumstances. Attributable to these one of a kind characteristics, MANET is winding up increasingly more broadly executed in the business. However, considering the way that MANET is prevalent among critical mission applications, network security is of vital significance.



Figure-1: MANET Architecture

Shockingly, the open medium and remote appropriation of MANET make it defenseless against different sorts of attacks. For example, because of the nodes' lack of physical assurance, malicious attackers can without much of a stretch catch and trade off nodes to achieve attacks. Specifically, considering the way that most routing protocols in MANETs expect that each node in the network behaves agreeably with other nodes and apparently not malicious attackers can without much of a stretch trade off MANETs by embeddings malicious or non-

helpful nodes into the network. Furthermore, as a result of MANET's conveyed architecture and changing topology, a traditional centralized observing technique is never again plausible in MANETs. In such case, it is crucial to build up an interruption detection system (IDS) specially designed for MANETs. In mobile specially appointed networks, security relies upon several parameters (authentication, Confidentiality, integrity, non-repudiation and availability). Without one of these parameters, security won't be complete. Without authentication, an attacker could disguise a node, thus having the capacity to have unauthorized access to the resources and to delicate data.

2. LITERATURE SURVEY

Wei Li recommended that utilized the Genetic Algorithm (GA) to network Intrusion Detection Systems. Here they are depicting about the Parameters and advancement process for GA. And afterward it's valuable for distinguishing proof of complex anomalous behaviors amid the procedure. There are two kinds of intrusion detection techniques utilized that is abuse detection and anomaly detection. Abuse detection alludes to techniques that characterize known methods to penetrate a system. System reactions depend on distinguished penetrations. Anomaly detection Refers to techniques that characterize and characterize normal or satisfactory behaviors of the system. Genetic algorithm thinks about the both temporal and spatial data of network association amid encoding problems. This is utilized to locate the anomalous behaviors. Utilizing this genetic algorithm with such parameters to discover the intrusion. **A. Kartit, A. Saidi et al**, suggested that article about the security enhanced network. Numerous mechanisms have been created to enhance the security of PC networks. In this paper, proposed an approach dependent on security arrangement at three dimensions for complex PC systems. That is Strategies for External Protection, Functional Security Policies and Operational Security Policies. First dimension is a classical so it will be set in firewall to keep network attacks from outside by declining malicious association endeavors by unauthorized third gatherings outside And then second dimension depends on the tasks doled out to clients in the organization by the division of the network to VLAN "Virtual Local Area Network" and the utilization of ACL "Access Control List".Atlast third dimension control will keep character usurpation from inside or from outside to the internal PC network. **Kiran Dhangar Prof. Deepak Kulhare Arif Khan** proposed OSI layer based Intrusion Detection System [IDS].here two kinds of attribute find in security occasion log file one is login-logoff time and another is unauthorized getting to of the host. Numerous methods have been produced by associations and assume essential jobs to anchor network infrastructure and communications by means of the Internet such as through the utilization of firewalls, against infection software packages and intrusion detection systems. Intrusion Detection System[IDS] are discovering layer attack or abnormality in the caught packets which is pursues: in application layer attacks are discovering —Backl, —Buffer overflowl and —Port Scanl. In Transport layer discovering —TCP SYN FLOAD Attackl —Landl and —Smurfl. Finally network layer attack is future work of this research. —Host based Intrusion Detection Systeml is proposed security analyzer to check or discover attack in local host then it will detect security attack in security occasion log file. In the wake of completing this it is create results. On the off chance that any illegal movement find in this log file like unauthorized getting to or login flopped then it will go to alarm system for data that this system is experiencing attack. In future have to work on network layers protocol and endeavor to discover attack on network layers.

3. ATTACKS ON MANET

3.1 Active Attacks

Performed by attackers for repeating, changing and deletion of exchanged data. They attempt to change the behavior of the protocol. These attacks are intended to corrupt or avert message stream among the nodes. Such attacks by and large can be called as DOS attacks that either corrupt or completely block the communication between the nodes. Another sort of attack includes inclusion of unessential packets in the network to cause clog. Obsolete routing data might be replayed back to the nodes in the network. Dynamic attacks can be detected sometimes and this reason makes dynamic attack less utilized by an attacker.

3.2 Passive Attacks

As talked about in this sort of attack includes unauthorized tuning in of the routing packets. Attacker may listen in on all the routing refreshes. For this situation an attacker does not upset the task of a routing protocol rather it just tunes in to it to find the valuable data about the routing. Such attacks are hard to be detected. From the routing packets an attacker may comprehend about a node which is essential in the network and route to that node is being asked for all the time by each other node. So an attacker endeavors to debilitate.

3.3 Physical Layer Attacks

1) Eavesdropping: In eavesdropping attack, attacker attempts to get the secret data amid communication. **2) Jamming:** Jamming attack will be actualized by knowing the recurrence malicious nodes sends stick signal to

aggravate the communication. **3) Active Interference:** An active interference is a kind of denial of service attack which misshapes the communications.

3.4 Link Layer Attacks

The data link layer can be classified about what impact it has on the condition of the network as a whole.

1) Selfish Misbehavior of Nodes: In the selfish misbehavior nodes will go about as selfish and won't partake in sending process. **2) DOS Attack:** This attack forestalls authorized access of resources to the authentic node. **3) Resource Exhaustion:** Malicious nodes make a rehashed impact to deplete the battery control. **4) Malicious Behavior of nodes:** The primary task of malicious node is to disturb normal activity of routing protocol. The effect of such attack is expanded when the communication takes place between neighboring nodes..

3.5 Network Layer Attacks

Black Hole Attack: In black hole attack the attacker node publicizes other node that it has shortest route to reach goal. In the event that this answer reaches before the actual answer a manufactured route will be established including the malicious node. Presently the malicious node can drop packet or perform DOS attack or Man in the center attack.

Wormhole attack : includes the participation between two attacking nodes. One attacker catches the packet and passages it to the other attacker. The link between the attackers is high speed communication link. These two attackers make the topology under their control.

Routing Table Poisoning Attack: In routing table harming attack attacker harms the routing table by changing the routes in the routing table. Other route is to infuse RREQ packet with high succession number. The packet with low grouping number will be deleted. This prompts determination of wrong routes.

3.6 Transport Layer Attacks

Session Hijacking: In session hijacking attacker hijacks the session after its set up. Here the attacker spoofs the IP address and launches the various attacks using the right sequence number.

3.7 Application Layer Attacks

Malicious code attacks: Malicious code attacks include, Viruses, Worms can attack both operating system and user application. **Multilayer Attacks:** The DoS attacks, impersonation attacks, man-in-the-middle attacks, and many other attacks can target multiple layers.

4. SECURITY CHALLENGES

A central helplessness of MANET originates from Peer-to-Peer architecture in which each node demonstrates like a router to forward packets to other nodes. Additionally, these nodes on network share the equivalent open environment that gives open door for malicious attackers. In the challenges for MANET security can be outlined as follows:

1. Lacking of central focuses: in view of characteristics of MANET such as lacking portals, routers, etc, the mobile nodes simply know a few neighbors in its range. This presents new troubles for security designs such as looking with the change of network topology, resource requirement .
2. Mobility: MANET nodes can leave, join, and meander in the network without anyone else will, so the topology of network is changed as often as possible. Therefore, some proposed security answers for adjusted with the change of topology. However, this also raises new problems for these systems.
3. Wireless link: In wireless environment, a lot of crash happened when nodes send and receive the packets. The wireless channel is also subject to interferences and errors, exhibiting unstable characteristics as far as bandwidth and postponement. Likewise, a few services such as routing protocols, communication services have to speak with others in real-time, this can surge the network traffic.
4. Limited resources: The mobile nodes like workstation, PDA are generally requirement in battery control, processing pace, storage, and memory limit. Therefore, the activity of security arrangements can be decreased the precision, proficiency such as dropping packets, various time for calculation.
5. 5)Cooperativeness: MANET is a mobility network, so nodes have to speak with others by utilizing routing protocol such as AODV, DSR... Therefore, this can make these protocols to end up a target of the attacks.

5. CLASSIFICATION OF PROTOCOLS

MANET or routing protocols can be extensively classified into three classifications such as Table Driven Protocols or Proactive Protocols, On-Demand Protocols or Reactive Protocols, Hybrid Protocol.

A. Table Driven or Proactive Routing Protocols

Proactive routing protocols are also called as table driven routing protocols. In this each node keep up routing table which contains data about the network topology even without requiring it. The routing tables are refreshed periodically whenever the network topology changes. In the proactive protocol all the nodes keeps up the data about the following node. All the nodes of any protocol have to hand-off it's whole to its adjoining nodes. The nodes send the packet data from one node to the other node after mutual understanding therefore the whole node always refresh their position. Proactive protocols are not proper for huge networks as they have to keep up node passages for each and every node in the routing table of each node. Network mobility is another factor that can corrupt the execution of specific protocols. When the network is moderately static, proactive routing protocols can be utilized, as putting away the topology data is increasingly efficient. There are different proactive routing protocols. Example: DSDV (Dynamic Destination-Sequence Distance-Vector Routing Protocol), OLSR (Optimized Link State Routing), WRP (Wireless Routing Protocol) etc.

B. On Demand Protocol or Reactive Routing Protocols

Reactive protocol also called as on interest routing protocol. Reactive protocol depends on a type of question answer dialog. Reactive protocol is better than the proactive protocol .Reactive routing is best adjusted to the most challenging manifestations of the adhoc networks. Their significant goal is to limit the network traffic overhead. A large portion of time everybody can utilize the reactive protocol since it is an on interest routing protocol. .As the mobility of nodes in the network increments, reactive protocols perform better. The mobility and traffic pattern of the network must assume the key job for choosing a fitting routing methodology for a specific network. In this kind of protocol, route is found whenever it is required. Nodes start route discovery when requested. A route is procured by the inception of a route discovery process by the source node. Examples: DSR (Dynamic Source Routing), AODV (Adhoc on Demand Distance Vector), TORA (Temporary Ordered Routing Protocol).

C. Hybrid Routing Protocols

This kind of protocol is an exchange off between proactive and reactive protocols. Proactive protocols have all the more overhead and less dormancy while reactive protocols have less overhead and more idleness. Thus a Hybrid protocol is expected to conquer the shortcomings of both proactive and reactive routing protocols. This protocol is a blend of both proactive and reactive routing protocol. It utilizes the on interest mechanism of reactive protocol and the table upkeep mechanism of proactive protocol in order to dodge inactivity and overhead problems in the network. Hybrid protocol depends on distance vector protocol however contain numerous highlights and preferred standpoint of link state protocol. Hybrid protocol enhances inside entryway routing protocol. Hybrid protocol is suitable for extensive networks where huge numbers of nodes are available. In this, expansive network is partitioned into a set of zones where routing inside the zone is finished by utilizing proactive approach and outside the zone routing is finished utilizing reactive approach. It is increasingly suitable to apply a hybrid protocol rather than an entirely proactive or reactive protocol as hybrid protocols frequently have the benefits of both sorts of protocols. There are different hybrid routing protocols for MANET like ZRP (Zone Routing Protocol), (Sharp Hybrid Adaptive Routing Protocol) DHAR (Dual-Hybrid Adaptive Routing), ADV (Adaptive Distance Vector Routing) etc.

6. TECHNIQUES TO DETECT ATTACKS**A. Intrusion detection System**

A system to give completes security arrangements and furthermore used to detect and report about malicious exercises in the network. MANET is also design for route traffic mechanism when there is blockage in the network, flawed nodes and topology changes because of its dynamic behavior. IDS use to detect critical nodes and afterward analyze its data traffic, critical node also debase network execution. There are distinctive IDS systems which have some explicit highlights, some of them are given underneath: 1. Cluster based casting a ballot, 2. Neighbor-checking, 3. Trust building.

B. Multi-layer Intrusion Detection Technique

These kind of attacks predominantly attack at cross layer which are all the more alarming and frightening as contrast with single layer attack and they can without much of a stretch be gotten away. Although these kind of attacks can be detected by a multiple layer rebellion detector, where as for all network layer's info are use to join and inspect by the cross-layer detector in a detailed fashion. There is also another approach to detect these kinds of attacks by working together with RTS/CTS and network layer detection concerning dropped packets.

C. Watchdog

Watchdog and Path rater are two vital parts that endeavor to enhance execution of specially appointed networks with the problematic nodes in network. Watchdog protects and distinguish the misbehavior by duplicating

packets into the support which should be sent and screen adjoining nodes behavior to these packets. The sending of neighboring nodes with or without change is chosen by snooping of Watchdog. The snooped packets that match with the cradle of watching node are disposed of; whereas the packets stayed in support past the timeout without any match found are hailed as dropped or changed. The packet sending node is noted as suspicious on the off chance that it disregards in excess of a predetermined threshold, the node is marked as malicious. Data about malicious nodes is passed to the Path rater part for consideration in path rating evaluation.

D. Path rater

The unwavering quality of all the known nodes in a specific network is evaluated. As needs be with the nodes point of view evaluations are made, and refreshed. Neutral rating nodes are altered after some time amid packet routing dependent on solid or questionable behavior. The Nodes are appraised - 100 that are misbehaved and seen by watchdog. Misbehavior is distinguished and detected as packet mishandling or change, whereas link breakage is detected as problematic behavior.

CONCLUSION

The advancement in the field of mobile computing is driving another alternative manner by which mobile gadgets shape a self-making, self-sorting out and self directing wireless network, called a mobile specially appointed network. Mobile Ad hoc networks are generally more helpless against physical security threats than settled or hardwired networks. This paper throws a light on various ideas of MANETS that can help researchers to the maximum. Its natural adaptability, lack of infrastructure, simplicity of arrangement, auto-configuration, minimal effort and potential applications make it an essential piece of future inescapable computing environments. As the association goes on, especially the need of thick organization such as combat zone and sensor networks, the nodes in impromptu networks will be smaller, cheaper, progressively able, and come in all forms. Taking all things together, although the far reaching sending of specially appointed networks is still year away, the research in this field will keep being extremely active and creative.

REFERENCES

1. Bing Wu, Jianmin Chen, Jie Wu, Mihaela Cardei, A Survey on Attacks and Countermeasures in Mobile Ad-Hoc Networks, Springer.
2. Satyam Shrivastava, Sonali Jain, A Brief Introduction of Different type of Security Attacks found in Mobile Ad-hoc Network, International Journal of Computer Science & Engineering Technology (IJCSSET), ISSN : 2229-3345, Vol. 4, No. 03, Mar 2013, pp:218-222.
3. Manjeet Singh , Gaganpreet Kaur A Surveys of Attacks in MANET, International Journal of Advanced Research in Computer Science and Software Engineering(IJARCSSE), Volume 3, Issue 6, June 2013 ISSN: 2277 128X, pp: 1631- 1636.
4. Khaleel Ahmad, Shikha Verma, Nitesh Kumar and Jayant Shekhar, Classification of Internet Security Attacks, Proceedings of the 5th National Conference; INDIACom-2011, ISSN 0973-7529, ISBN 978-93-80544-00-7. International Journal On Advanced Computer Theory And Engineering (IJACTE) ISSN (Print): 2319-2526, Volume -4, Issue -5, 2015 28
5. Pankaj Chaudhary, Sachin Chaudhary, Classification of Security Attacks in MANET, Asian Journal Of Computer Science And Information Technology, 3 : 5, 2013, pp: 77 – 80.
6. Aarti, Dr. S. S. Tyagi, Study of MANET: Characteristics, Challenges, Application and Security Attacks, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 3, Issue 5, May 2013 ISSN: 2277 128X, pp: 252- 257.
7. A.Rajaram, Dr. S. Palaniswami, Malicious Node Detection System for Mobile Ad-hoc Networks, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2), 2010, 77-85.
8. CH.V. Raghavendran, G. Naga Satish, P. Suresh Varma, Security Challenges and Attacks in Mobile Ad-Hoc Networks, Information Engineering and Electronic Business, 2013, 3, 49-58, DOI: 10.5815/ijieeb.2013.03.06.
9. Jiahong Weng, Security Issues in Mobile Ad Hoc Networks - A Survey.
10. Wenjia Li and Anupam Joshi, Security Issues in Mobile Ad Hoc Networks - A Survey.
11. Sevil Şen, John A. Clark, Juan E. Tapiador, and Security Threats in Mobile Ad Hoc Networks.

-
-
12. Rajni Sharma¹, Alisha saini², A Study of Various Security Attacks and their Countermeasures in MANET, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 1, Issue 1, December-2011, ISSN: 2277 128X, pp: 1-5.
 13. Sathish Alampalayam Kumar, Classification and Review of Security Schemes in Mobile Computing, scientific research, Wireless Sensor Network, 2010, 2, 419-440, DOI: 10.4236/wsn.2010.26054.

ANALYSIS ON DIGITAL IMAGE PROCESSING ENHANCEMENT TECHNIQUES AND THEIR ANALYSIS PROCESS

S. Saravanan

Assistant Professor, AJK College of Arts and Science, Coimbatore

ABSTRACT

The digital image processing (DIP) has been utilized in a number of areas, especially for highlight extraction and to get patterns of digital images. Acknowledgment of characters is a novel problem, and although, presently there are generally accessible digital image processing algorithms and implementations that can detect characters from images, choice of a suitable technique that can straightforwardly adjust to different kinds of images, that are unmistakable or complex is imperative. This paper exhibits a short diagram of digital image processing techniques such as image reclamation, image enhancements, and highlight extraction, a framework for processing images and goes for displaying a versatile digital image processing method for acknowledgment of characters in digital images.

Keywords: Image Enhancement, Noise, Filtering, Histogram, Contrast, Compression.

1. INTRODUCTION

The image processing is an analyzed and control of a digitalized image, especially so as to enhance the quality of image processing. Dunk technique can be connected in variety of various fields such as Diagnostic image analysis, Surgical arranging, Object detection and Matching, Background subtraction in video, Localization of tumors, Measuring tissue volumes, Locate questions in satellite images (streets, woodlands, etc.) ,Traffic control systems, Locating objects in face acknowledgment, iris acknowledgment, agricultural imaging, and medical imaging. DIP addresses challenges and issues like that loss of image quality, to enhance corrupted image. In this paper the audit of writing identified with DIP is talked about. In image enhancement, the goal is to complement certain image highlights for ensuing analysis or for image show. Examples incorporate contrast and edge enhancement, pseudo shading, noise filtering, sharpening, and amplifying. Image enhancement is valuable in highlight extraction, image analysis, and visual information show. The enhancement procedure itself does not expand the inherent information content in the data. It essentially emphasizes certain predetermined image characteristics. Enhancement algorithms are generally interactive and application-subordinate. Image enhancement techniques, such as contrast stretching, delineate dark dimension into another dim dimension by a predetermined change. An example is the histogram equalization method, where the information dim dimensions are mapped with the goal that the yield dark dimension dispersion is uniform. This has been observed to be a ground-breaking method of enhancement of low contrast image. Image enhancement is among the least complex and most appealing areas of digital image processing. Basically, the thought behind enhancement techniques is to bring out detail that is clouded, or just to highlight certain highlights of enthusiasm for an image

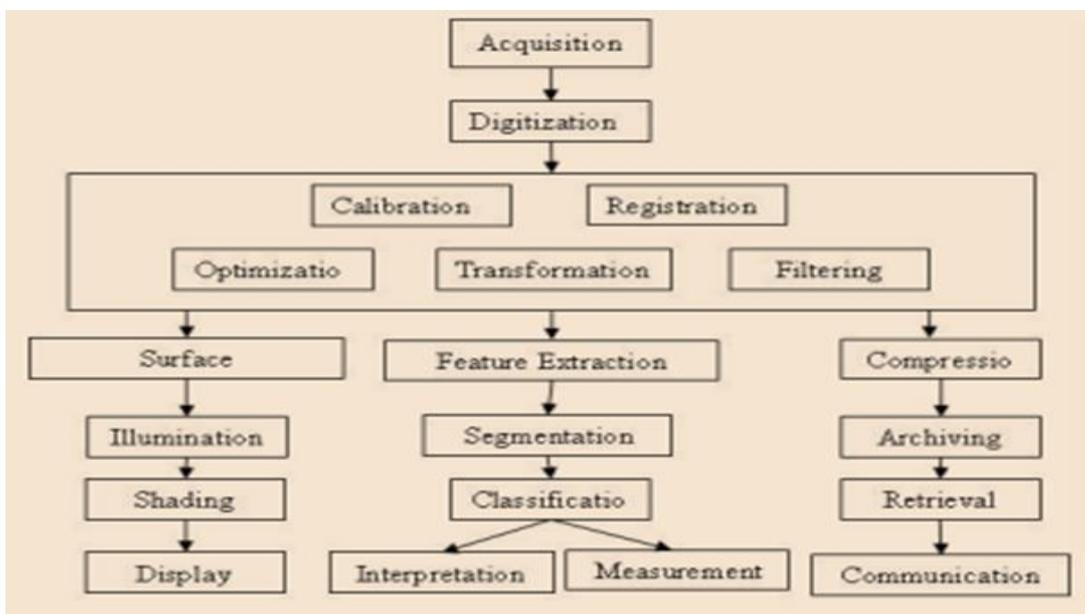


Figure-1: Modules of Digital Image Processing

In imaging science, Image Processing will be processing of images utilizing mathematical activities by utilizing any type of signal processing for which the information is an image, a progression of images, or a video, such as a photograph or video outline; the yield of image processing might be either an image or a set of characteristics or parameters identified with the image. Most image-processing techniques include regarding the image as a two dimensional signal and applying standard signal-processing techniques to it. Images are also prepared as three dimensional signals where the third-measurement being time or the z-hub. Image processing usually alludes to digital image processing, yet optical and analog image processing also are conceivable. This article is about general techniques that apply to all of them. The procurement of images (delivering the information image in any case) is alluded to as imaging. Firmly identified with image processing are PC graphics and PC vision. In PC graphics, images are manually produced using physical models of articles, environments, and lighting, rather than being obtained (through imaging gadgets such as cameras) from natural scenes, as in most energized motion pictures. PC vision, then again, is regularly viewed as high-level image processing out of which a machine/PC/software means to decipher the physical substance of an image or a succession of images (e.g., videos or 3D full-body magnetic reverberation filters). In present day sciences and technologies, images also increase much more extensive degrees because of the regularly developing significance of logical visualization (of frequently substantial scale complex logical/experimental data). Examples incorporate microarray data in genetic research, or real-time multi-asset portfolio exchanging money. It is among rapidly developing technologies today, with its applications in different parts of a business. Image Processing forms center research area within designing and software engineering disciplines as well. Image processing basically incorporates the accompanying three steps. Bringing in the image with optical scanner or by digital photography. Analyzing and controlling the image which incorporates data compression and image enhancement and spotting patterns that are not to human eyes like satellite photographs.

2. LITERATURE SURVEY

SuezouNakadate et al examined the utilization of digital image processing techniques for electronic speckle pattern interferometry. A digital TV-image processing system with a vast edge memory allows them to perform exact and adaptable tasks such as subtraction, summation, and level cutting. Digital image processing techniques made it simple contrasted with analog techniques with create high contrast borders. **Satoshi Kawata et al** talked about the characteristics of the iterative image-reclamation method altered by the reblurring methodology through an analysis in recurrence space. An iterative method for tackling synchronous direct conditions for image rebuilding has an inherent problem of union. The presentation of the strategy called "reblur" tackled this union problem. This reblurring technique also served to smother noise intensification. Two-dimensional recreations utilizing this method demonstrated that a boisterous image debased by straight movement can be all around reestablished without detectable noise enhancement. **William H** highlighted the advancement in the image processing and analysis of digital images amid the previous ten years. The subjects included digitization and coding, filtering, enhancement, and rebuilding, recreation from projections, hardware and software, highlight detection, matching, division, surface and shape analysis, and pattern acknowledgment and scene analysis. **David W. Robinson** exhibited the application of a general-reason image-processing PC system to programmed periphery analysis. Three areas of application were analyzed where the utilization of a system dependent on an arbitrary access outline store has empowered a processing algorithm to be produced to suit an explicit problem. Furthermore, it empowered programmed analysis to be performed with unpredictable and boisterous data. The applications considered were strain measurement by speckle interferometry, position location in three tomahawks, and blame detection in holographic nondestructive testing. A short portrayal of each problem is introduced, trailed by a depiction of the processing algorithm, results, and timings. **S V Ahmed** talked about the work arranged by concentrating upon the reenactment and image processing viewpoints in the transmission of data over the supporter lines for the improvement of an image processing system for eye measurements from eye. **P K Sahoo et al** exhibited an overview of thresholding techniques and refreshed the before review work. An endeavor was made to evaluate the execution of some programmed global thresholding methods utilizing the paradigm functions such as consistency and shape measures. The evaluation depended on some real world images.

3. IMAGE ENHANCEMENT TECHNIQUES

Sometimes images acquired from satellites and conventional and digital cameras lack in contrast and brightness due to the constraints of imaging sub systems and enlightenment conditions while catching image. Images may have distinctive kinds of noise. In image enhancement, the goal is to complement certain image highlights for consequent analysis or for image show. Examples incorporate contrast and edge enhancement, pseudo shading, noise filtering, sharpening, and amplifying. Image enhancement is helpful in highlight extraction, image analysis and an image show. The enhancement procedure itself does not build the inherent information content

in the data. It basically emphasizes certain predetermined image characteristics. Enhancement algorithms are generally interactive and application-subordinate. A portion of the enhancement techniques are: Contrast Stretching, Noise Filtering, Histogram change

A) Contrast Stretching: Some images (eg. over water bodies, deserts, thick woods, snow, clouds and under hazy conditions over heterogeneous locales) are homogeneous i.e., they don't have much change in their dimensions. Regarding histogram portrayal, they are characterized as the event of extremely thin peaks. The homogeneity can also be because of the off base enlightenment of the scene.

B) Histogram Modification: Histogram has a great deal of significance in image enhancement. It mirrors the characteristics of image. By changing the histogram, image characteristics can be adjusted. One such example is Histogram Equalization. Histogram equalization is a nonlinear stretch that redistributes pixel values so that there is roughly a similar number of pixels with each value within a range. The outcome approximates a level histogram. Therefore, contrast is expanded at the peaks and lessened at the tails.

C) Noise Filtering: Noise filtering is utilized to channel the superfluous information from an image. It is also used to expel different sorts of noises from the images. For the most part this element is interactive. Different channels like low pass, high pass, mean, middle etc.,

4. IMAGE ANALYSIS

Image analysis is worried about making quantitative measurements from an image to deliver a depiction of it. In the most straightforward frame, this task could be perusing a name on a basic supply item, arranging diverse parts on a mechanical production system, or estimating the size and introduction of platelets in a medical image. Further developed image analysis systems measure quantitative information and use it to make a sophisticated choice, such as controlling the arm of a robot to move an item in the wake of distinguishing it or exploring an airplane with the guide of images gained along its direction. Image analysis techniques require extraction of specific highlights that guide in the distinguishing proof of the item. Division techniques are utilized to separate the ideal item from the scene with the goal that measurements can be made on it along these lines. Quantitative measurements of item includes allow classification and portrayal of the image.

5. IMAGE SEGMENTATION TECHNIQUES

A. Threshold division: In this background and closer view can be isolated utilizing histogram. In this we utilize bimodal kind of images. We separate images in two sections. We can manually discover the values which separate the force values of these images. This works for the images which have high contrast and clear background is there so we simply need to fragment thresholding method. This method depends on a clasp level (or a threshold value) to turn a dark scale that background. The easiest method of image division is called the image into a parallel image. The key of this method is to choose the threshold value (or values when multiple-levels are chosen). Several mainstream methods are utilized in industry including the maximum entropy method, Otsu's method (maximum difference), and k-implies clustering. As of late, methods have been produced for thresholding figured tomography (CT) images. The key thought is that, unlike Otsu's method, the thresholds are gotten from the radiographs rather than the (recreated) image. New methods recommended the utilization of multidimensional fluffy standard based non-straight thresholds. In these works choice over each pixel's membership to a portion depends on multi-dimensional guidelines got from fluffy rationale and developmental algorithms dependent on image lighting environment and application.

B. Point Transformation Logarithm administrator: diminishing contrast of brighter districts. Histogram equalization: general method of altering force conveyance.

1) Logarithm administrator: When changing the brightness of an image, the pixel values are mapped to an alternate range, with the outcome that the brightness is enhanced or diminished. The main parameter of this function is a factor in the range. In the event that this factor is smaller than 0, the brightness is diminished, in the event that it is bigger than 0, the brightness is expanded. This activity can't enhance the quality of an image, so normally it won't be utilized in a division procedure. Be that as it may, it very well may be utilized to make an image 'looking' better. When changing the contrast of an image, the pixel values are mapped to an alternate range, with the outcome that the contrast is enhanced or diminished. The main parameter of this function is a factor in the range. In the event that this factor is smaller than 0, the contrast is diminished, on the off chance that it is bigger than 0, the contrast is expanded. This activity can't enhance the quality of an image, so normally it won't be utilized in a division procedure. However, it tends to be utilized to better visualize image details for the eye. Logical activities are point tasks which must be executed on parallel images. These tasks comprise of the application of logical operators on each pixel of the image.

2) Histogram equalization: Given a greyscale image, its histogram comprises of the histogram of its dim dimensions; that is, a graph demonstrating the number of times each dark dimension happens in the image. We can construe a lot about the presence of an image from its histogram. In a dark image, the dim dimensions would be clustered at the lower end. In a consistently bright image, the dim dimensions would be clustered at the upper end. In a very much contrasted image, the dark dimensions would be well spread out over much of the range. Problem is given an ineffectively contrasted image; we might want to enhance its contrast, by spreading out its histogram.

C. Watershed Transformation The watershed change considers the angle greatness of an image as a topographic surface. Pixels having the highest inclination extent powers (GMIs) relate to watershed lines, which speak to the locale limits. Water put on any pixel encased by a typical watershed line streams downhill to a typical local power least (LIM). Pixels depleting to a typical least shape a catch bowl, which speaks to a section.

D. Wavelet Transform Wavelets (little waves) are functions that are moved in time and in addition in recurrence around a specific point. Fourier change has the drawback of dealing with simply the recurrence segment in the signal. The temporal details are not accessible. Due to Heisenberg's vulnerability rule we know that we can have either high recurrence goals or poor recurrence goals and great temporal goals. This change is most suitable for no stationary signals. The premise function fluctuates both in recurrence extend and in spatial range. The wavelet change is designed such that we get great recurrence goals for low recurrence segments and high temporal goals for high recurrence segments.

CONCLUSION

Obviously, digital image processing is a vital part of photography thinking about that technology keeps changing. There are a host of digital image processing techniques that gives a wide application variety in highlight extraction and classification. Artificial neural networks are frequently used to undertake character acknowledgment in light of their high resilience to noise. The systems have the ability to realize immaculate results. The different image division techniques are talked about and analyzed in this paper. The MATLAB software is utilized to analyze the distinctive techniques and yield is contemplated. The algorithm administrator method shows that an image can be compacted by supplanting each pixel value with its logarithm. The yield from watershed change technique shows that we can isolate two blended items and can analyze it in the best way. The wavelet change technique help us to analzse horizontal, diagonal, and vertical two dimensional perspectives of an image.

REFERENCES

- [1] ShubhangiMeshram, AnuradhaThakare, Santwana Gudadhe, "An Automated Medical Support System based on Medical Palmistry and Nail Color Analysis", International Journal Of Engineering And Computer Science, Volume 5, Issue 6, June 2016, ISSN: 2319-7242, pp.17006-17009.
- [2] Nityash Bajpai, Rohit Alawadhi, Anuradha Thakare, Swati Avhad, Sneha Gandhat, "Automated Prediction System For Various Health Conditions By Analysing Human Palms And Nails Using Image Matching Technique", International Journal of Scientific & Engineering Research, Volume 6, Issue 10, October 2015, ISSN: 2229-5518, pp.609-613.
- [3] G.Chitra, "Chiromancy in the field of Medicinal science based Human health care using Digital Image Processing", International Journal for Research in Applied Science & Engineering Technology, Volume 4, Issue 2, February 2016, ISSN: 2321-9653, pp.21-26.
- [4] Anuradha Thakare, Shubangi Mashram, H.M. Baradkar, "An Automated Medical Support System for detecting human health conditions based on Noval Bicluster Method", IJCTA, Volume 10, Issue 8, 2017, ISSN: 0974-5572, pp. 223-228.
- [5] Dr. Hardik B. Pandit, Prof. Dipti Shah, "A system for palm color analysis in healthcare", International Journal of Advanced Engineering Technology, Volume 5, Issue 1, January-March 2014, E-ISSN: 0976-3945, pp. 30-32.
- [6] Hardik Pandit, Dr. D M Shah, "Application of Digital Image Processing and Analysis in Healthcare Based on Medical Palmistry", International Conference on Intelligent Systems and Data Processing, 2011, pp. 56-59.
- [7] Sneha Gandhat, A.D. Thakare, Swati Avhad, Nityash Bajpai, Rohit Alawadhi, "Study and Analysis of Nail Images of Patients", International Journal of Computer Applications, Volume 143, Issue 13, June 2016, ISSN: 0975-8887, pp. 38-41.

-
-
- [8] Vipra Sharma, Manoj Ramaiya, “Nail Color and Texture Analysis for Disease Detection”, International Journal of Bio-Science and Bio-Technology, Volume 7, Issue 5, 2015, ISSN: 2233-7849, pp. 351-358.
 - [9] Vipra Sharma, Aparajit Shrivastava, “System for Disease detection by analyzing finger nails Color and Texture”, International Journal of Advanced Engineering Research and Science, Volume 2, Issue 10, October 2015, ISSN: 2349-6495.
 - [10] Zainab Othman, Sarmad Saleem, “Diseases Diagnosis Using Medical Palmistry Fuzzy Model”, MATEC Web of Conferences 76, 2016.
 - [11] Shahzad, A., Sharif, M., Raza, M., and Hussain, K., “Enhanced Watershed Image Processing Segmentation”, Journal of Information & Communication Technology, vol 2, pp: 01-09, 2008.
 - [12] Ramadevi, Y., Sridevi, T., Pootnima, B., and Kalyani, B., “Segmentation and Object Recognition using Edge Detection Techniques”, International Journal of Computer Science & Information Technology, vol 2, pp: 153- 161, 2010.
 - [13] Nagabhushana Rao, M., Venkateswara Rao, M., and Bhagavi, Y, K., “Applications of Edge Based Segmentation in Bio-Metric Security System”, International Journal of Advanced Engineering & Application, pp:175-178, (jan) 2011.
 - [14] Wayalun, P., Laopracha, N., Songrum, P., and Wanchanthuek, P., “Quality Evaluation for Edge Detection of Chromosome G-band Images for Segmentation”, Applied Medical Informatics, vol 32, pp: 25-32, 2013.
 - [15] Acharjya, P.P., and Ghoshal, D., “An Approach to reduce Over Segmentation in Watershed Ridge Line Observation”, International Journal of Advancements in Research & Technology, vol 2, 2013.
 - [16] Vijayran, S., and Paramjeet singh., “A Watershed Based Morphological Operator Approach for Image Segmentation”, International Journal of Advanced Research in Computer and Communication Engineering, vol 2, 2013.

ANALYSIS ON MOBILE CLOUD COMPUTING APPLICATION ADVANTAGES AND ISSUES

S. John Gracias

Assistant Professor, Department of Computer Applications, AJK College of Arts and Science, Coimbatore

ABSTRACT

Mobile Cloud computing is an accumulation of substantial gathering of interconnected networks. This incorporates personal PCs, network servers; mobile computing can be classified in to two sorts as public and private cloud. Mobile cloud computing depict both a stage and sort of application. A cloud computing stage requires arrangements, designs, reconfigures and depravations servers. The security issues start to develop and raised there are number of loopholes and challenges that still exist in the security of mobile cloud computing. The security threats have got a hurdle in the rapid flexibility of the mobile computing worldview. Noteworthy endeavors have been submitted in research associations and the scholarly community to assemble secure mobile cloud computing environments and infrastructures. In cloud computing, top cloud services challenges are security, availability and execution. The cloud computing security issue is always the key factor and it is ranked one. This paper exhibits a review about the mobile cloud computing security issues and challenges concentrating on the cloud computing.

Keywords: Mobile Cloud Computing, Security, Applications, Issues.

1. INTRODUCTION

By expanding utilization of mobile computing, mobile gadgets assume an essential job in human life as best communication devices are prepared in each place and time. Mobile clients use different services from mobile applications such as Google applications which keep running on mobile gadgets locally or are offloaded to remote servers for remote execution. Mobile gadgets are confronting numerous problems about their resources (e.g. battery life, storage and bandwidth) and communications (e.g. mobility and security), Hence QoS isn't fulfilling. Mobile gadgets can't run resource-concentrated applications because of lack of adequate processing force and storage, hence mobile clients like to use all the more incredible gadgets like PCs and laptops about resource shortage problems. To conquer these obstructions we can bolster mobile computing by cloud computing. Cloud computing can be characterized as "the collection of computing as an utility and software as a service", also called pay-as-yougo-computing. Cloud service suppliers offer Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). Offloading heavy computing can help mobile gadgets to spare energy and mobile clients to quicker and increasingly efficient execution. There are some resource-serious applications that mobile gadgets never can run however with help of cloud computing, running them is effectively, on the grounds that cloud service suppliers have numerous incredible resources. Accordingly, with supporting of cloud computing for mobile clients, mobile cloud-computing is presented as mix of cloud computing by mobile computing and network technology, hence resource-escalated applications can be executed on resourcerequirement mobile gadgets.

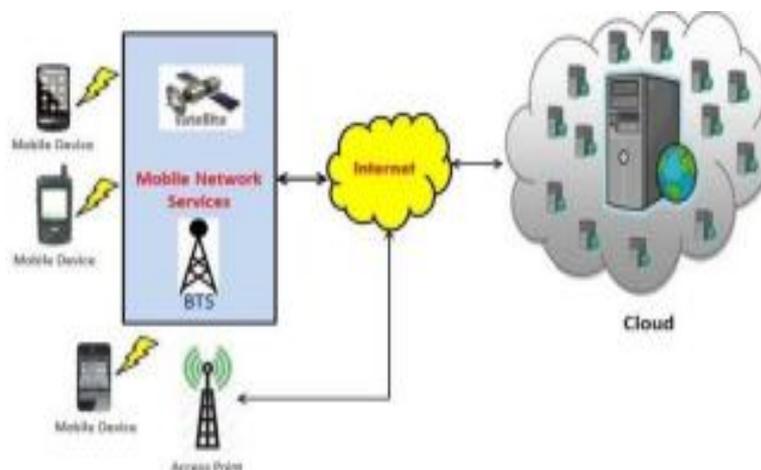


Figure-1: Mobile Cloud Computing Architecture

With the proliferation of smart mobile gadgets and cloud computing technologies, mobile cloud computing (MCC) has developed as another computing worldview for building the cutting edge MCC applications. MCC guarantees to bring new energizing MCC applications past mobile cloud computing (MCC) applications by consolidating cloud computing, mobile computing, and data analytics at the fingertip of a human administrator.

In this paper we overview existing and theorize future age MCC applications. We limit our review to infrastructure based MCC applications where the hardware infrastructure stays static and gives services to the mobile clients. We give insights to the empowering technologies and challenges that lie ahead for us to push ahead from Mobile Computing for building the cutting edge Mobile Computing applications. For each of the challenges, we give an overview of existing arrangements, distinguish research holes, and propose future research areas.

2. LITERATURE SURVEY

Le Guan et. al. addresses the challenges in Mobile Cloud Computing design such as network latency, limited bandwidth and availability. In order to analyze Mobile Cloud Computing technology, a concept model is proposed which includes context management, resource scheduling, client and transmission channel. A Cloud architecture of Mobile Cloud Computing is described for organization of Mobile Cloud Computing systems. Application partition and offloading and various context aware services are explained briefly. **Dejan et. al.** addresses several mobile cloud approaches. An overview of various possibilities of Mobile Cloud Computing is given. Native and web applications are too extremes of mobile applications. The cost model of elastic mobile cloud applications is described. **Han Qi et. al.** discuss Mobile cloud computing (MCC) as a development and extension of mobile computing (MC) and cloud computing (CC) which has inherited high mobility and scalability. The proposed system in the paper explains the principle of MCC, characteristics, recent research work, and future research trends. Proposed system analyzes the features and infrastructure of mobile cloud computing and also analyzes the challenges of mobile cloud computing. **Vinod et. al.** discuss about the cloud computing which enables the work anywhere anytime by allowing application execution and data storage on remote servers. This is useful for mobile computing and communication devices that are constrained in terms of computation power and storage. The goal of the paper is to characterize under what scenarios cloud-based applications would be relatively more energy-efficient for users of mobile devices. **Hung et. al.** analyzes the performance of many mobile applications which are weak due to lack of computation resources, storage, and bandwidth and battery capacity. To overcome this, application is rebuilt using the cloud services. The proposed system explains a framework to execute the mobile application in cloud based virtualized environment with encryption, and isolation to protect against unauthenticated cloud providers. Results show the execution of mobile application by offloading the workload with efficient application level migration method via mobile networks. The migration of application form one device to another is easy and quick in the proposed system. **Ricky et. al.** discuss that mobile cloud computing allows mobile applications to use the large resources in the clouds. In order to utilize the resources, migration of the computation among mobile nodes and cloud nodes is necessary. Therefore, a highly portable and transparent migration approach is needed. The paper uses a Java byte code transformation technique for task migration without effecting normal execution. Asynchronous migration technique is used to allow migrations to take place virtually anywhere in the user codes. The proposed Twin Method Hierarchy minimizes the overhead from state restoration codes in normal execution.

3. MCC APPLICATIONS

Mobile applications increase expanding share in a global mobile market. Different mobile applications have taken the upsides of MCC. We have seen a number of MCC applications as of late, including mobile business, multimedia sharing, mobile learning, mobile detecting, mobile healthcare, mobile gaming, mobile social networking, location-based mobile service, and expanded reality. Mobile business, such as e-banking, e-publicizing and e-shopping, utilizes scalable processing force and security measures to oblige a high volume of traffic because of synchronous client access and data transaction processing multimedia sharing gives secure review and sharing of multimedia information put away on smart phones while giving regulatory controls to oversee client benefits and access rights important to guarantee security. Mobile learning allows a thin terminal to get to learning materials on the cloud whenever and wherever. Mobile detecting using sensor prepared smart phones to gather data will alter numerous MCC applications including healthcare, social networking, and environment/health checking. Mobile healthcare allows a colossal measure of patient data to be put away on the cloud promptly. Mobile gaming achieves scalability by utilizing scalable calculation and momentary data refresh on the cloud side and screen refresh at the mobile gadget side. Mobile social networking allows a gathering of mobile clients to transfer sound/video/multimedia data for real-time sharing, with cloud computing giving storage to data, as well as security to ensure mystery and integrity of data.

4. ADVANTAGES OF MCC

Broadening battery lifetime: Computation offloading moves expansive calculations and complex processing from resource-constrained gadgets (i.e., mobile gadgets) to resourceful machines (i.e., servers in clouds). Numerous mobile applications take focal points from task relocation and remote processing, thereby allowing remote application execution to spare energy essentially. ‡

Enhancing data storage limit and processing power: MCC empowers mobile clients to store/get to expansive data on the cloud. It helps decrease the running expense for calculation escalated applications by not being obliged by storage limit on the gadgets in light of the fact that their data presently is put away on the cloud therefore enhancing dependability and availability: with data and services in the clouds, then are always (almost) accessible notwithstanding when the clients are moving. ‡

Dynamic provisioning: Dynamic on-request provisioning of resources on a fine-grained, self service premise therefore there is no requirement for timely booking ‡

Scalability: Mobile applications can be performed and scaled to meet the erratic client requests and Service suppliers can without much of a stretch include and extend a service ‡

Multi-occupancy: Service suppliers can share the resources and expenses to help a variety of applications and substantial no. of clients. ‡

Simplicity of Integration: Multiple services from various suppliers can be incorporated effortlessly through the cloud and internet to meet the clients end.

5. MCC ISSUES

Mobile communication issues

Low bandwidth: One of the biggest issues, on the grounds that the radio resource for wireless networks is much more rare than wired networks.

Service availability: Mobile clients will most likely be unable to interface with the cloud to get a service because of traffic blockage, network disappointments, mobile signal strength problems

Heterogeneity: Handling wireless availability with highly heterogeneous networks to fulfill MCC necessities (always-on availability, ondemand scalability, energy proficiency) is a troublesome problem ‡

Calculation offloading: One of the fundamental highlights of MCC ‡ Offloading isn't always viable in sparing energy ‡ It is critical to determine whether to offload and which bits of the service codes to offload.

6. CHALLENGES AND SOLUTION FOR MOBILE CLOUD COMPUTING

A. Challenges Regarding Mobile Communication

1) Low Bandwidth Problem: In communication network Bandwidth is one of the vital thing as the radio resource for wireless networks are transmitted over networks as indicated by the measure of bandwidth is available for moving the substance in the network. As the bandwidth is restricted sharing the constrained bandwidth among various mobile clients situated in a similar area or workstation and likely associated with a similar substance to be exchanged. This results in the enhancement of the quality and this arrangement is connected fundamentally for the situation when the clients in a specific area are keen on similar substance. It gathers client profiles that are utilizing the network periodically and makes choice tables, Based on which the clients choose whether or not to help other clients download substance that can't receive by them because of the bandwidth constraint.

2) Lack of Resource of Mobile Devices: Comparing mobile gadget with more established desktop PC shows that how the cost highlight of mobility is being achieved. As there is lack of resources makes it hard for the reception of mobile cloud computing in general conditions. For beating this impediment of mobile gadgets and there resources, they are added to the cloud infrastructure with the goal that they can be utilized on whenever on anywhere premise makes it simple for the vast majority of cutting edge applications. As the mobile gadget exhibitions, and the resource imperatives of mobile gadgets continuing expanding and settled gadgets will remain and should be represented the sorts of application chose for mobile cloud computing.

B. Challenges of Network

1) Challenges of Wireless Network and Access Control Policies: Wireless network is base for doing cloud computing and it has its own characteristic nature and imperatives. For better execution the reliable network bandwidth is critical however factor data rates, longer dormancy and availability with holes in inclusion are the principle problems related with network in the MCC. Some wild factors are also mindful like weather for fluctuating bandwidth limit and inclusion. For executing MCC, getting to the network with heterogeneous access situation and diverse access technologies like WiMAX, WLAN, 4G, etc, having their very own approaches and confinements. As the wireless network is something imperative to help MCC functioning there should be the best possible mechanism for limiting the idleness, expanding the bandwidth and diminishing the availability hole. We should keep diverse access schemes for maintaining a strategic distance from connection disappointment and connection re-establishment. So as to give quicker access for mobile gadgets, most

suppliers are putting forth 4G/Long Term Evolution (LTE) services. These services based on data storage limit, attachment and play highlights, low dormancy, etc. This gives download peak rates up to 100 Mbps and transfer up to 50 Mbps.

2) Seamless Connection Handover: Currently executing application is ended or it returns error message when one move starting with one passageway of network then onto the next point or one move from Wi-Fi network to 3G-based cell network. Since this makes the circumstance of communication disappointment and connection reestablishment. Along these lines, for giving data communication utilizing cell network mobile operators are attempting to set up WiFi Aps on street. This system is helpful to offload traffic of Wi-Fi systems can be diminished, and is to give seamless in decreased cell traffic clog.

C. Challenges Related To Mobile Applications

1) Interoperability: There are bunches of mobile gadgets running on various stage including iPhone, Android phones, BlackBerry and others also. This variety of gadgets are utilized by individuals in a similar association or a gathering of individuals sharing single network. Also, in such circumstance interoperability issue turns into a noteworthy challenge in pulling/pushing data over multiple gadgets. An application that are kept running on mobile cloud infrastructure should be bolstered by certain mobile cloud infrastructure that can without much of a stretch be judged perhaps based on its prerequisites against the cloud infrastructure characteristics. Along with the gadget, network bandwidth and inactivity vectors should perform calculation force, network bandwidth, and network inertness appropriately.

2) Mobile Cloud Convergence: Data dispersion is a critical issue for achieving preferred standpoint of mobility by making combination with cloud computing with mobile world. Concerning utilizing this cloud computing application services with mobile gadgets there a few issues with calculation of data, battery life and execution of this gadgets in conveyed stage. Mobile cloud intermingling is the technique that gives execution enhancement and answer for the calculation control problem. For this there is a parcel of application takes place such that parts that require more calculation keep running on the cloud and keep running on the mobile gadget. Wireless technologies, propelled hardware and internet are imperative to achieve inescapable and omnipresent computing.

D. Challenges Regarding Security

1) Information Security Devices Privacy: As cloud computing basically deals with giving all kind of services, data storage and processing. As all this is done remotely, so security is a vital worry for all who are utilizing these services. We are worried here with Mobile Cloud Computing hence its important to check the security identified with mobile gadgets along with cloud computing stage, which is the key worry in this area. This is on the grounds that there is plausibility of gadget stolen or lost, which prompts crucial data to be imperiled. Presently days as different security threats are conceived, cloud platforms also offers numerous hearty implicit security measures like SSL and digital endorsements gives as to empower external security. Abuse of data from stolen/lost mobile gadgets can be stayed away from by wiping of these mobile gadget remotely. For detecting security threats on any mobile gadget is finished by installing and running security software's developers called "Antiviruses" which are promptly accessible in the market.

2) Security Attacks and Hacking: All networking enacts are powerless to one or other sort of malicious attacks. As there is more utilization of Web sites that are sometimes getting to malicious code sites, for getting to the network and operational data of that specific individual or association. There are some occasion around then in the wake of executing best measures for giving the best security strategies to data and information prepared attackers with best surfing May makes occurrences that normally unpreventable as: There are different approaches and schemes are currently days accessible such as Fair Information Practice Principles (FIPP) which require thorough controls and methodology to ensure the privacy of individual people data and in addition associations information. Encryption is technique that is best to give best approach to keep up integrity and confidentiality of information. Along with other internal servers Web 2.0 servers may further alleviate the threat of unauthorized getting to of information through social media, Web sites and other internet sources.

CONCLUSION

This paper exhibits a study on MCC and clarifies the patterns and challenges in MCC. Mobile applications are developing step by step with the expanding utilization of mobile phones. Mobile utilization has been expanded so clients need to do all the functionalities on the mobile gadget. With the help of cloud computing new open doors are emerging in this field and this is the hot theme in research area. Calculations are expanding step by step in fields like trade, science and technology. In the ongoing years MCC is concentrating on enhancement of mobile restrictions and make it all the more amazing utilizing virtualization techniques. As talked about above,

various MCC models have been introduced; one thing basic in all is that they are lacking privacy of the application.

REFERENCES

- [1] S. O. Kuyoro, F. Ibikunle and O. Awodele, "Cloud Computing Security Issues and Challenges", *International Journal of Computer Networks (IJCN)*, vol. 3, Issue 5, (2011).
- [2] Rajesh Piplode, Umesh Kumar Singh " An Overview and Study of Security Issues & Challenges in Cloud Computing ", *International Journal of Advanced Research in Computer Science and Software Engineering* , Vol 2, Issue 9, September 2012 ISSN: 2277 128X.
- [3] Deyan Chen, Hong Zhao, "Data Security and Privacy Protection Issues inCloud Computing", (2012) *International Conference on Computer Science and Electronics Engineering*.
- [4] Abid Shahzad and Mureed Hussain, "Security Issues and Challenges of Mobile Cloud Computing ", *International Journal of Grid and Distributed Computing* Vol.6, No.6 (2013),
- [5] Soeung-Kon, J. -H. Lee and S. W. Kim, "Mobile Cloud Computing Security Considerations", *Journal of Security Engineering*, no. 9, (2012) April.
- [6] Shih-Hao Hung, Chi-Sheng Shih, Jeng-Peng Shieh, Chen-Pang Lee, Yi-Hsiang Huang, " Executing mobile applications on the cloud: Framework and issues ", *Computers and Mathematics with Applications* 63 (2012) 573–587. Elsevier Ltd.
- [7] Swarnpreet Singh, Ritu Bagga, Devinder Singh, Tarun Jangwal "Architecture Of Mobile Application, Security Issues And Services Involved In Mobile Cloud Computing Environment ", *International Journal O Computer Science And Electronics Research*.Vol.Issues.Agu(2012).
- [8] Abdul Nasir Khan, M.L. Mat Kiah, Samee U. Khan, Sajjad A. Madani " A review on interworking and mobility techniques for seamless connectivity in mobile cloud computing: *Journal of Network and Computer Applications* 43 (2014)84–102 (2014) Elsevier Ltd.
- [9] Hoang T. Dinh, Chonho Lee, Dusit Niyato, and Ping Wang, " A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches", *Department of Computer Science and Computer Engineering, La Trobe University, Australia* 31 Accepted 30 May 2012, Available online 6 June 2012.
- [10] Sriram N. Premnatha, Zygmunt J. Haas, " A Practical, Secure, and Verifiable Cloud Computing for Mobile Systems", *The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC-2014)*.
- [11] M.Rajendra Prasad, Jayadev Gyani, P.R.K.Murti, " Mobile Cloud Computing: Implications and Challenges ", *Journal of Information Engineering and Applications* ISSN 2224-5782 (print) ISSN 2225-0506 (online) Vol 2, No.7, (2012).
- [12] Huajian Mao, Nong Xiao, Weisong Shi, Yutong Lu, "A cloud-oriented file service for mobile Internet devices ", *J. Parallel Distrib. Compute.* 72 (2012) 171–184 31 October 2011, Available online 11 November (2011) Elsevier Ltd.
- [13] Nazanin Aminzadeh, Zohreh Sanaei, Siti Hafizah Ab Hamid, " Mobile storage augmentation in mobile cloud computing: Taxonomy, approaches, and open issues ", *Simulation Modeling Practice and Theory* (2014) Elsevier Ltd.
- [14] S. Subashini, V.Kavitha, "A survey on security issues in service delivery models of cloud computing ", *Journal of Network and Computer Applications* 34 (2011) 1–11, Elsevier Ltd.
- [15] Zaheer Ahmad, Keith E. Mayes, Song Dong, Kostas Markantonakis, " Considerations for mobile authentication in the Cloud information security technical report 1 6 (2011) 123to1 3 0 Elsevier Ltd.

TOWARDS INCORPORATING IOT AND BIG DATA IN HEALTH CARE INDUSTRY USING MACHINE LEARNING ALGORITHMS**Parimala. S¹ and Dr. P. Senthil Vadivu²**Assistant Professor¹, Department of Computer Science, SRM Institute of Science and Technology, Chennai
Head², Department of Computer Applications, Hindustan Arts and Science College, Coimbatore

ABSTRACT

Big data analytics and the Internet of Things (IoT) can aid resolve some of the most thought-provoking medical difficulties that we are presently facing as a society. Like any industry, the medical field can get advantage from the competence of technologies available in Big Data and IoT. The embryonic trend in the expansion of the internet, the Internet of Things (IoT) is a network of physical substances that can be retrieved via the Internet. These substances are everyday items like washing machines and vehicles, which contain some embedded technologies that can interrelate with an exterior background or standardize internal situations. This paper highlights how the technology of IoT and Big Data paves the way for the Medical Field in which Machine Learning Algorithms are implemented.

Keywords: IoT, Big Data Analytics, Internet of Things, Machine Learning Algorithms

I. INTRODUCTION

Internet of Things (IoT) and Big Data Analytics are the two challenging phases in the healthcare industry. Like any industry, the medical field can benefit from more efficiency introduced by big data analytics and IoT [11]. For the general physicians, these dual technologies would guarantee that patients are given treatment faster and with more precision.

The 4 Vs. of big data relevant to health care [5] are volume, variety, veracity and value. In healthcare, big data refers to electronic health data sets which are so large and complex EHR [Electronic Health Record] is created from the database which takes the patient details like name and address from a source system and matches with the scheduled appointments from another system and integrate it. After integrating with the multiple system patching multiple sources of information together into a centralized databank accessed by reporting or a query system can deliver a more comprehensive and actionable portrait of past details of the patient, diagnoses, treatments, and other risk profiles. It improves the healthcare systems are great repository of big data (like patient records, test reports, medical images etc.) that can be utilized to reduce the cost in healthcare and to get better trustworthiness and efficacy. The rapid change in healthcare is to the step of prevention to curing of diseases.

A new archetype, branded as the Internet of Things (IoT), has a widespread application in numerous industries, including healthcare. The full application of this prototype in healthcare area is a joint anticipation for the reason that it allows medical industries to perform more competently and patients can obtain better treatment. With the use of this technology-based healthcare technique, there is a consummate panorama to improve the excellence and effectiveness of treatments and which consequently improve the health of the patients.

II. REAL TIME DIAGNOSTIC

Some expediency problems can be easily solved by IoT in medical fields since it allows the medical practitioners to concentrate on other demanding issues. IoT can create and withstand acquaintance networks among interconnected devices with fully-functional feedback and looping systems with minimum input from the physicians. In a juncture, IoT can be used to monitor a patient's condition if the necessary devices are interrelated and proficient of allocating accurate real time data and triggering the corrective measures. IoT could possibly allow a general practitioner to pay for more time to give attention to other patients.

The big data analytics can be used for decision making in health care with the aid of machine learning algorithms [6]. The different approaches of machine learning algorithms and data mining can afford better cure to disease, build up personalized medicines and even prevent disease or epidemics.

There are different approaches Traditional Machine Learning Algorithms works on centralized databases. A centralized database is governed by the different approaches in Traditional Machine Learning Algorithms. There is an inevitability to modify the conventional algorithms or produce with some common mixed approaches to accomplish the large dataset in distributed environment. Apache Hadoop [4] offer Hadoop Distributed File System storage which takes care of distributed storage and fault tolerance.

The increasing digitization of information, management, and retrieval systems, together with the rapid improvement of devices and sensors, has controlled the strategy, expansion, and wide-ranging use of actual mobile health apps and accompanying estimate and wellness systems. Indeed, big data and the Internet of Things (IoT) play a vital role in health-related applications, particularly in resolving disease biology at multiple scales, ranging from intracellular networks to cellular, tissue, organ, and whole-body systems. By integrating large, diverse data from IoT-based devices and sensors and employing novel data mining, machine learning, and predictive analytics methods, it is quickly becoming possible to build effective prediction systems for automatic implication and endorsement of disease diagnosis and treatment. At the same time, these methods give rise to an amount of research challenges, including security and applications as well as the expansion of standards and the competence to constantly recover organized tools and skills.

Figure-I

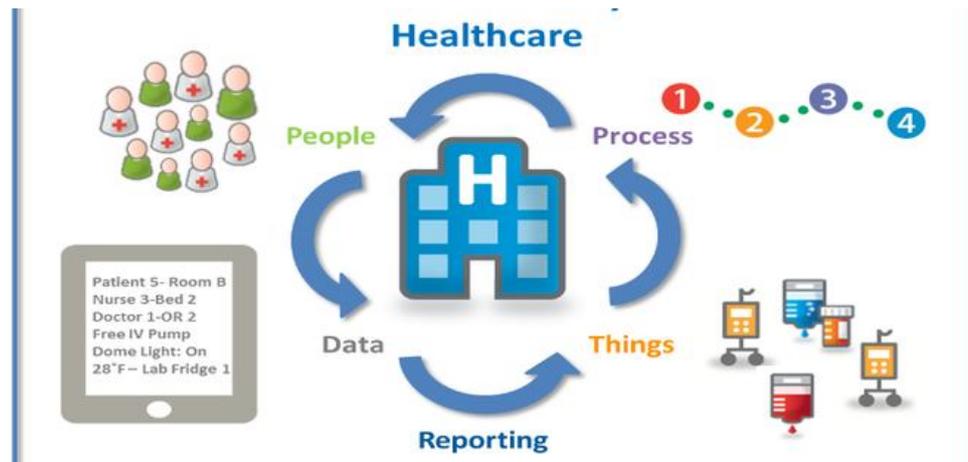


Fig-1: Health Care in IoT

A. The Significance of IoT in an Emergency Situation

In a circumstance of a deadly road accident, hospitals can save valuable time if they install ECG machines and transmission capabilities to share the patients’ vital stats with ER physicians/surgeons. As and when the ambulance reaches the hospital premise, everyone in the ER would be aware about how to stabilize and treat the patient immediately. When the hospitals have installed ECG and other equipment can be shared along with the quantitative data.

In the example mentioned above, the patients cannot do anything since they are in an emergency situation. But, in case of prolonging diseases, patients can try towards improving their healthcare by frequently recording and share their statistics with their doctors through smart sensors & other scientific instruments. Specifically, in population health, regular monitoring can bring a more widespread improvement in chronic condition management. For instance, to control diabetes in a particular area, physicians can monitor a group of high-risk people and request their daily Blood-sugar level on tablets or mobile phones. Sudden spikes in sugar level would send alert signals to the physicians who can immediately warn the population as well as suggest corrective measures to avoid any aggravation. According to research ‘There is almost 500 petabytes of healthcare data in being today and that number is predictable to skyrocket to greater than 25,000 petabytes within the following seven years’. [1].

B. IoT devices data

The Healthcare industry is one of the quickest businesses that stays among that quickly holds the technologies and from the Internet of Things. The main objective of incorporating the feature of IoT into the medical equipment improves the quality as well as the feasibility of the administrating the medical industry enormously. 'As indicated by a few assessments, spending on the Healthcare IoT arrangements will achieve an amazing \$1 trillion by 2025 and, ideally, will set the phase for exceedingly customized, available, and on-time Healthcare administrations for everybody'. [7]

IOT contraptions which can be aided for the patients

- Any wearable or any portable device can be integrated to the Cloud and we can collect and analyse the patients’ data in real-time.
- We can monitor the patient’s health condition using the live streaming using audio /video.
- We can also monitor vital health signs which are collected by portable devices like smartphones

- Emergency notifications sent to a doctor or family using intelligent devices.
- From health monitoring devices the data can be collected with the help of Diagrammatic visualizations.
- Data can be collected for analysis by the patient's itself with their own personal smart watches and other fitness bands which can be a favour done for the health care industry. These wearable devices sometimes give the best answer and give the conclusions which may depend on the patient's habits and life style.

C. Taxonomy of Machine learning Algorithm in Health Care and IoT

Machine learning is a prototype or a structure which quickly identifies and analyse the collected data and convey the results, also it influences the conventional as well as the real-time data. With machine learning, healthcare service providers can make better decisions on patient's diagnoses and treatment options, which lead to overall improvement of healthcare services.

A learning algorithm acquires a set of samples as an input named which is known as the training set. In general, there exist three main categories of learning: supervised, unsupervised, and reinforcement [2], [3], [9]. Prediction using traditional disease risk models usually involves a machine learning algorithm (e.g., logistic regression and regression analysis, etc.), and especially a supervised learning algorithm by the use of training data with labels to train the model [8], [9], [10].

Machine learning algorithms are supervised learning with supported vector machine can be used in finding the rheumatoid stiffness with the amalgamation of medicine records which are stored in EHR of the patients to progress the precision of prognostic prototypes of disease. For instance, implementation of hypoglycaemic agents from the treatment can propose the warning of the diabetes.

III. CONCLUSION AND FUTURE WORK

IoT and Big data are the two technologies that allow the devices to admit precarious data, wrapping it and decipher its meaning. IoT may directly adjust various occurrences based on the investigated data. On the other hand, big data may require some more widespread human interference.

In medical industry, the portrayal of the two technologies presents a unique occasion to save on costs incurred in tough medical problems, intercept indications early, and endorse conduct based on the data composed by the devices. Still there are many number of algorithms based on Machine Learning Concepts like Deep Learning, Supervised Learning to implement the Big Data Analytics in Healthcare Industry.

REFERENCES

1. Anon, (2017). [Online] Available at: http://www.ijarcce.com/upload/2013/october/8-h-mohammed_khaleel__locally.pdf [Accessed 15 Jan. 2017].
2. D. Barber Bayesian Reasoning and Machine Learning, Cambridge University Press (2012) Google Scholar
3. C.M. Bishop Pattern Recognition and Machine Learning Springer (2006) Google Scholar
4. Dharavath Ramesh, Member, IEEE, Pranshu Suraj, and Lokendra Saini. (2016). Big data Analytics in Healthcare: A Survey Approach .Presented at International conference on Microelectronics, Computing and Communications (MicroCom). [Online].Available: ieeexplore.ieee.org/document/7522520/
5. Jasleen Kaur Bains. (2016, 4). Big Data Analytics in Healthcare- Its Benefits, Phases and Challenges. [Online].6(4), pp- 430-435.Available: https://www.ijarcse.com/docs/papers/Volume_6/4_April2016/V6I4-0143.pdf
6. Prof. Jigna Ashish Patel, Dr. Priyanka Sharma. (2014). Big data/or Better Health Planning .Presented at IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014). [Online].Available: ieeexplore.ieee.org/document/7012828/
7. Kaa. (2017). IoT Healthcare Solutions - Medical Internet of Things for Healthcare Devices and Hospitals. [Online] Available at: <https://www.kaaproject.org/healthcare> [Accessed 15 Jan. 2017].
8. K. Lin, M. Chen, J. Deng, M. M. Hassan, G. Fortino, "Enhanced fingerprinting and trajectory prediction for IoT localization in smart buildings", IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294-1307, Jul. 2016
9. K. Lin, J. Luo, L. Hu, M. S. Hossain, A. Ghoneim, "Localization based on social big data analysis in the vehicular networks", IEEE Trans. Ind. Informat..
10. K.P. Murphy Machine Learning: a Probabilistic Perspective, MIT press (2012), Google Scholar
11. <http://www.iotevolutionworld.com/iot/articles/434509-iot-healthcare-big-data-iot-solve-some-difficult.htm>

IMPACTS OF IOT IN BIG DATA ANALYSIS**N. Suganthi¹ and Dr. K. Sarojini²**Research Scholar¹ and Assistant Professor², L. R. G. Govt Arts College for Women, Tirupur

ABSTRACT

In today's growing technological revolutions and advanced networking system Internet of Things (IoT) is the most popular and evolving term in technological world. Explosive growth of the devices connected to IoT and exponential growth of information conception makes it possible to overlap with the recent advancements of Big Data Analysis. IoT is used for variety of applications. The applications are interacting with text books using QR code printed on it, smart meters, home router, TV, light control, A/C control. The convergence of these areas creates several flourishing opportunities in Big Data and IoT systems. The advantages of IoT for the user there are many security aspects which need to be taken care by the user of IoT enabled devices. In an effort to understand the development of IoT and Bigdata. This paper reviews the current research of IoT, key enabling technologies, IoT applications, and identifies research trends and challenges.

Keywords: IoT, Big Data, Data Analysis, Taxonomy

1. INTRODUCTION

The Internet of things (IoT) is the network of physical devices, vehicles, home appliances, and other items embedded with electronics, software, sensors, actuators, and network connectivity which enable these objects to connect and exchange data. It refers to a network of objects, each of which has a unique IP address & can connect to the internet. The network can be a combination of people-things, things-things, and people-people. In other words, it is a system of interrelated computing devices, digital machines, object, animals or people that are provided with a unique identifier and the ability to transfer data over a network without requiring human-to-human or human-to-machine interaction. Convergence between wireless communications, Digital electronic devices, and Micro-electro-mechanical systems (MEMS) technologies led to the rise of the Internet of Things. Internet-connected objects like computers, smart phones, tablets and Wi-Fi de-vices, sensors, wearable devices and household appliances are all the objects of the IoT components [1]. Since there is a massive growth in number of devices day by day, the amount of data generated would also be enormous. Here is where Big Data and IoT go hand in hand. Big Data manages the enormous amount of data generated using its technologies. The Internet of Things (IoT) and big data are two vital subjects in commercial, industrial, and many other applications [2].

Given that sensors are used in nearly all industries, the IoT is expected to produce a huge amount of data. The data generated from IoT devices can be used in finding potential research trends and investigating the impact of certain events or decisions. These data are processed using various analytic tools. In this context, leveraging a big data plat-form that can assist in consuming and reading diverse data sources as well as in accelerating the data integration process becomes vital [3]. Data integration and analytics allow organizations to revolutionize their business process. Specifically, these enterprises can use data analytics tools to transform a huge volume of sensor-collected data into valuable insights. Given the overlapping research trends in these areas, this paper focuses on the recent advances in management of big data and analytics in the IoT paradigm.

II. TAXONOMY

Figure 1 shows the thematic taxonomy of big data and analytics solutions that are designed for IoT systems. These solutions are categorized based on the following attributes: a) big data sources, b) system components, c) big data enabling technologies, d) functional elements, and e) analytics type [4].

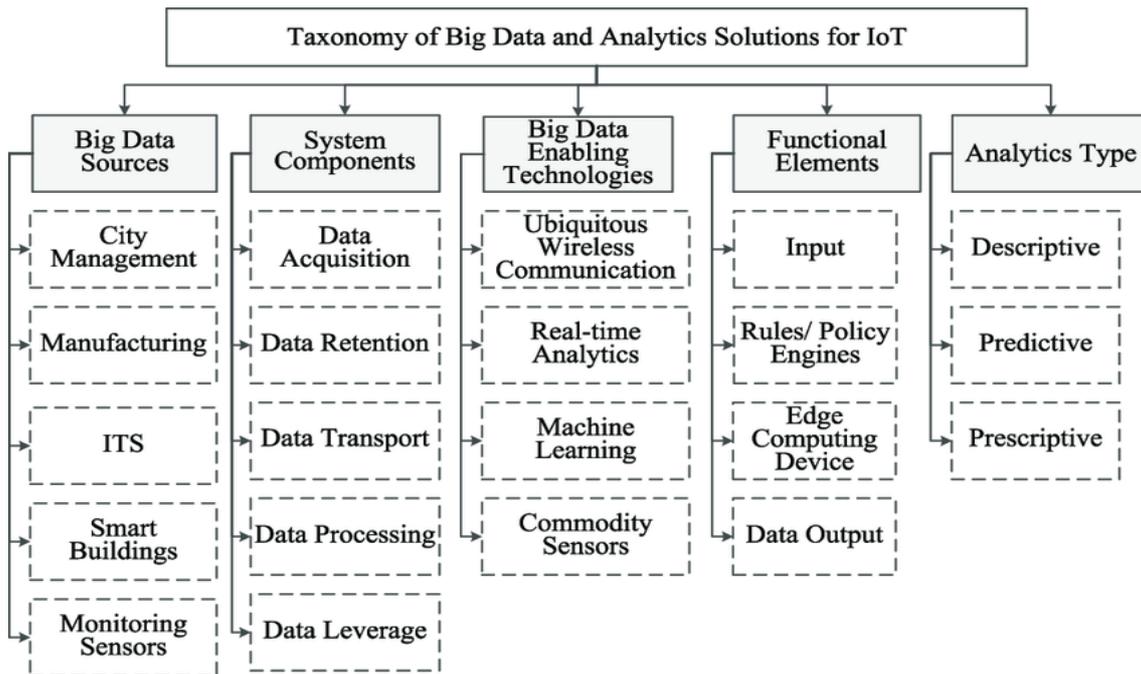


Figure-1: Taxonomy of Big Data and Analytics Solutions for IoT

BIG DATA SOURCES

Big data are generated by an infrastructure that is deployed to run various IoT applications, including city management, manufacturing, intelligent transport systems (ITS), smart building, and monitoring sensors. The city management uses connected cameras, sensors, and actuators to make the lives of citizens secure and convenient. However, these devices generate a bulk of data that must be managed and analyzed in real time to obtain relevant insights. Similarly, the manufacturing industry has deployed IoT devices that continuously generate a huge amount of data to maximize the productivity and efficiency of its operations. To obtain insights from these data, big data and analytics solutions have been used in designing and testing new products, optimizing services and marketing, minimizing defects, and improving yields.

Along with big data and analytics, the proliferation of sensors [5], connected vehicle technologies [6], [7], and IoT [8] have resulted in the creation of intelligent transportation systems, thereby significantly increasing the amount of real-time big data that must be communicated, aggregated, analyzed, and managed. The ITS can take advantage of big data and analytics to enhance the decision-making capabilities of its users.

SYSTEM COMPONENTS

Big data and analytics solutions usually comprise five system components, namely, data acquisition, data retention, data transport, data processing, and data leverage. Big data acquisition involves collecting, filtering, and cleaning the data before they are transferred into the data warehouse. This component is commonly governed by four attributes, namely, volume, variety, velocity, and value. Various big data retention policies involve privacy and legal concerns against economics to identify archival rules, retention time, data formats, and encryption methods. The big data must be transported across different data sites to guarantee load balancing, business continuity, and replication.

BIG DATA ENABLING TECHNOLOGIES

The big data enabling technologies in the IoT context are related to ubiquitous wireless communication, real-time analytics, machine learning, and data capturing elements, such as commodity sensors and embedded systems. The key ubiquitous wireless communication technologies that are used for transporting big data in IoT include IEEE 802.15.4, IEEE 802.11, IEEE 802.15.1, and IEEE 802.16.

KEY ELEMENTS

The big data and analytics solutions for IoT comprise four key elements, namely, input, rules or policy engines, edge computing devices, and data output. The raw data are collected from different resources and transferred to edge analytics systems. These systems are based on a rules/policy engine that defines and applies rules to the input data in order to obtain insights. The edge computing device is another key player in the operation of the analytics system. Processing the data that are generated by IoT devices on the edge devices can bring several advantages, such as low latency, minimal bandwidth consumption, data integrity, security, and low cost [9], [10], [11]. These data are also made avail-able to the consumer in real time.

ANALYTICS TYPE

Analytics can be divided into three types, namely, descriptive analytics, predictive analytics, and perspective analytics. Descriptive analytics, which defines “what has happened or what is happening,” helps find new business opportunities and challenges. Predictive analytics, which defines “what will happen and why it will happen,” is enabled by using various technologies, such as text/web/data mining, to accurately predict future conditions and states. Prescriptive analytics, which defines “what should I do and why should I do it,” utilizes simulation, expertise, and decision support systems to investigate various choices and provide suggestions to decision makers.

III. LAYERED ARCHITECTURE OF IOT

Layered architecture of IoT plays an important role in the reduction of the amount of data to be transmitted. The layers in the IoT make it possible to process data at each layer so that the complexity of processing big data at the final layer will be reduced. Chen [12] discusses three layers architecture for IoT while Shuo et al. divide the IoT architecture into four layers [13]. Zhang et al. proposed six layer architecture [14] shown in Figure 2

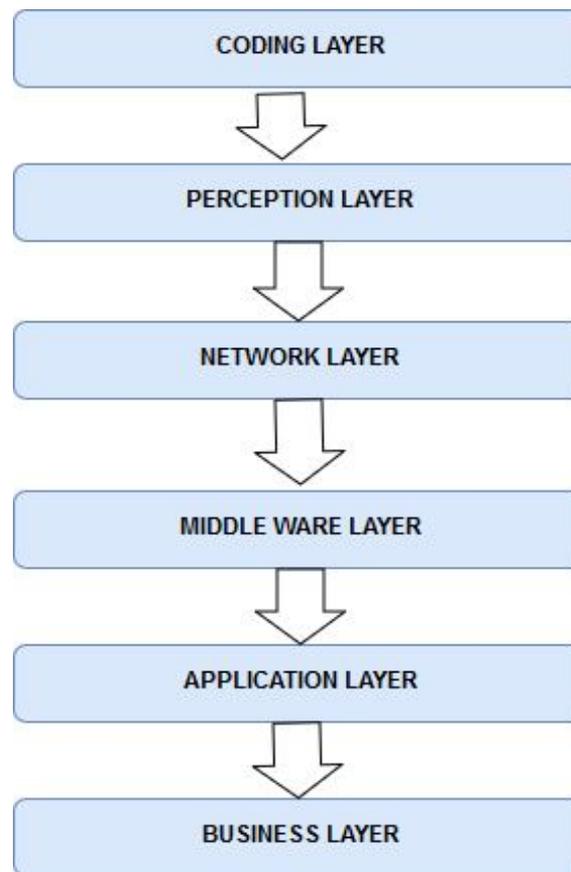


Figure-2: Layers of IoT

This architecture is suitable for IoT and services offered by various layers are discussed below:

1. Coding layer uniquely identifies the data rate requirement of each IoT network device [15]. It is used for selection of algorithm and route, which in turn maximizes data rate.
2. The perception layer works same as physical layer of the network. In addition to this, it has the capability of processing data due to which only processed data will be sent to above layers which results in reduction of big data.
3. The network layer is used as a bridge between perception layer and middleware layer.
4. The middleware layer is used for big data analysis. Hadoop can be one of the options to be used for efficient data analysis in this layer.
5. The application layer further generates responses on the basis of processed data which it receives from middleware layer.
6. Finally, the business layer handles the applications and services of a device which is a very necessary aspect in the overall effectiveness of the network.

IV. SURVEY ON BIG DATA ANALYTICS AND IOT

C. Lee et al. [16] propose an IoT-based cyber physical system that supports information analysis and knowledge acquisition methods to improve productivity in various industries. This system, which focuses on industrial big data analytics, integrates various data analytics components in the form of reconfigurable and interchangeable modules to meet different business needs.

P. Rizwan et al. [17] study the strengths and weaknesses of various traffic management systems. They propose a low cost, real-time traffic management system that deploys IoT devices and sensors to capture real-time traffic information.

Q. Zhang et al. [18] propose Firework, a new computing paradigm that allows distributed data processing and sharing in an IoT-based, collaborative edge environment. Firework combines physically distributed data by providing virtual data views to end users using predefined interfaces.

M. M. Rathore et al. [19] propose a smart city management system based on IoT that exploits big data and analytics. The data are collected by deploying different sensors, including weather and water sensors, vehicular networking sensors, surveillance objects, smart home sensors, and smart parking sensors.

B. Ahlgren et al. [20] discuss the significance of using IoT to deliver services for improving the lives of citizens, including transportation, air quality, and energy efficiency.

O. B. Sezer et al. [21] propose an augmented framework that integrates semantic web technologies, big data, and IoT. The key requirements for the proposed framework are analyzed, and the conceptual design of the envisioned IoT system is proposed based on the analysis results.

B. Cheng et al. [22] design GeeLytics, an edge analytics platform that performs real-time data processing at the network edges and in the cloud.

A. J. Jara et al. [23] conduct a survey to highlight the existing solutions and challenge to big data that are posed by cyber-physical systems.

C. Vuppalapati et al. [24] examine the role of big data in healthcare and find that body sensors generate massive amounts of health related data.

A. Ahmad et al. [25] analyze human behavior by using big data and analytics in the social IoT paradigm [26].

D. Arora et al. [27] utilize big data and analytics techniques to classify network-enabled devices. They also analyze the performance of four machine learning algorithms, such as k-nearest neighbor (KNN), NaveBayes (NB), support vector machines (SVM), and random forest.

This is exploration of evolving methods and technologies in IoT and Big Data Analysis. These are the only some of the previous works done in the mentioned technologies.

V. CHALLENGES IN IOT WITH BIG DATA

Major challenges that can fetch momentous rewards when they are solved

1. Huge data volumes
2. Difficulty in data collection
3. Incompatible standards
4. New security threats
5. No reliability in the data
6. Fundamental shifts in business models
7. Huge amount of data to analyze
8. A rapidly evolving privacy landscape

The above points are some of the challenges that IoT big data faces. The rate in data growth is expanding every second, storage is a big challenge, processing and maintaining is even more tedious. The tools that are developed to manage the both technologies are day by day changing as per the requirements. No doubt, both technologies are going to play a major role in the information technology field.

VI. IMPACT OF IOT IN DATA ANALYSIS

IoT is the next big thing impacting our lives in major ways and number of factors. Technologies like Column oriented databases, SQL in Hadoop, Hive, Wibidata, PLATFORA, SkyTree, Storage Technologies, Schema less databases, or NoSQL databases, Streaming Big Data analytics, Big Data Lambda Architecture, Map reduce, PIG, etc., helps in dealing with the enormous amount of data generated by IoT and other sources.

Big Data Storage: At basis, the key Necessities of big data storage are that it can handle very huge amounts of data and continuous balancing to keep up with expansion and that it can provide the input/output operations per second (IOPS) necessary to deliver data to analytics tools. The data is of different form and format and thus, a datacenter for storing such data must be able to handle the load in changeable forms. Obviously IoT has a direct impact on the storage infrastructure of big data. Collection of IoT Big Data is a challenging task because filtering redundant data is mandatorily required. After Collection, the data has to transfer over a network to a data center and maintained. Many companies started to use Platform as a Service (PaaS) to handle their infrastructure based on IT. It helps in developing and running web applications. By this way, Big data can be managed efficiently without the need of expanding their infrastructural facilities to some extent. IoT Big Data Storage is certainly a challenging task as the data grows in a faster rate than expected.

Data Security Issues: The IoT has given new security challenges that cannot be controlled by traditional security methods. Facing IoT security issues require a shift. For instance, how do you deal with a situation when the television and security camera at your home are fitted with unknown Wi-Fi access. Few security problems are

1. Secure computations in distributed environment
2. Secure data centers
3. Secure transactions
4. Secure filtering of redundant data
5. Scalable and secure data mining and analytics
6. Access control
7. Imposing real time security, etc.,

Big data analytics: Big data analytics will help you understand the business value it brings and how different industries are applying it to deal with their sole business necessities. According to the Gartner IT dictionary, Big Data is variety of information assets, high-volume, and high-velocity and, innovative forms of information processing for enhanced approach and decision making.

Volume refers to the size of data. Data sources can be social media, sensor and machine - generated data, structured and unstructured networks, and much more. Enterprises are flooded with terabytes of big data.

Variety refers to the number of forms of data. Big data deals with numbers, 3D data and log files, dates, strings, text, video, audio, click streams.

Velocity refers to the speed of data processing. The rate at which data streams in from sources such as mobile devices, click streams, machine-to-machine processes is massive and continuously fast moving. Big data mining and analytics helps to reveal hidden patterns, unidentified correlations, and other business information.

Impact on day to day life: IoT Big Data is slowly redefining our lives. Let us consider a few examples of our lives. At work, the cctv camera in the canteen estimating the time you spend there. The class room sensors can find out how much time you spend in writing on the board. This can be just to measure the productivity of an employee. At home, the home theatre playing the favorite movie of ours as soon as you switch on the television, smart devices could save a lot of power and money by automatically switching off electrical devices when you leave home. A smart wrist band tied to the elder people at home intimates the nearby hospital if they fall sick. The above said is going to happen in a very short time because of the rapid development in IoT and Big Data technologies

VII. CONCLUSION

A wide range of IoT applications have been developed and deployed in recent years. In an effort to understand the development of IoT and Big data, this paper reviews the current research of IoT, key enabling technologies, IoT applications, and identifies research trends and challenges. It discusses about the basic necessity of Big data with IoT. It explores the areas like the taxonomy of the Big data with IoT structure, real

time application and advancements of the IoT with Big Data Analysis techniques, challenges in the Big Data analysis with IoT. Finally it concludes with the impact of the IoT technology in big data analysis.

REFERENCE

1. Belgaum, M. R., Soomro, S. U., Alansari, Z., & Alam, M. (2018). Treatment of Reactive Routing Protocols Using Second Chance Based on Malicious behavior of Nodes in MANETS. *Pakistan Journal of Engineering, Technology & Science*, 6(2).
2. Alansari, Z., Soomro, S., Belgaum, M. R., & Shamshirband, S. (2018). The Rise of Internet of Things (IoT) in Big Healthcare Data: Review and Open Research Issues. In *Progress in Advanced Computing and Intelligent Engineering* (pp. 675-685). Springer, Singapore.
3. R. Bashir and A. Q. Gill, "Towards an iot big data analytics framework: Smart buildings systems," in *High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2016 IEEE 18th International Conference on. IEEE, 2016, pp.1325–1332.
4. Ahmed, Ejaz & Yaqoob, Ibrar & Hashem, Ibrahim & Khan, Imran & Ibrahim Abdalla Ahmed, Abdelmutilib & Imran, Muhammad & Vasilakos, Athanasios. (2017). The role of big data analytics in Internet of Things. *Computer Networks*. 10.1016/j.comnet.2017.06.013.
5. E. Ahmed, S. Ali, A. Akheenzada, and I. Yaqoob, "Cognitive radio sensor networks: Bridging the gap for network," *Cognitive Radio Sensor Networks: Applications, Architectures, and Challenges: Applications, Architectures, and Challenges*. IGI Global, p. 160, 2014.
6. S. A. A. Shah, E. Ahmed, F. Xia, A. Karim, M. A. Qureshi, I. Ali, and R. M. Noor, "Coverage differentiation based adaptive tx-power for congestion and awareness control in vanets," *Mobile Networks and Applications*, pp. 1–12.
7. I. Yaqoob, I. Ahmad, E. Ahmed, A. Gani, M. Imran, and N. Guizani, "Overcoming the key challenges to establishing vehicular communication: Is sdn the answer?" *IEEE Communications Magazine*, 2017.
8. Yaqoob, E. Ahmed, I. A. T. Hashem, A. Ahmed, A. Gani, M. Imran, and M. Guizani, "Internet of things architecture: Recent advances, taxonomy, requirements, and open challenges," *IEEE Wireless Communications*, 2017
9. E. Ahmed and M. H. Rehmani, "Mobile edge computing: Opportunities, solutions, and challenges," pp. 59–63.
10. A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Intelligent Systems and Control (ISCO)*, 2016 10th International Conference on. IEEE, 2016, pp. 1–8.
11. Y. Jararweh, A. Doulat, O. AlQudah, E. Ahmed, M. Al Ayyoub, and E. Benkhelifa, "The future of mobile cloud computing: integrating cloudlets and mobile edge computing," in *Telecommunications (ICT)*, 2016 23rd International Conference on. IEEE, 2016, pp. 1–5.
12. W. Chen, "An ibe-based security scheme on internet of things," in *Cloud Computing and Intelligent Systems (CCIS)*, 2012 IEEE 2nd International Conference on, vol. 3.IEEE, 2012, pp. 1046–1049.
13. H. Suo, J. Wan, C. Zou, and J. Liu, "Security in the internet of things: a review," in *Computer Science and Electronics Engineering (ICCSEE)*,2012 International Conference on, vol. 3. IEEE, 2012, pp. 648–651.
14. M. Zhang, F. Sun, and X. Cheng, "Architecture of internet of things and its key technology integration based-on rfid," in *Computational Intelligence and Design (ISCID)*, 2012 Fifth International Symposium on, vol. 1. IEEE, 2012, pp. 294–297.
15. M. Farooq, M. Waseem, S. Mazhar, A. Khairi, and T. Kamal, "A review on internet of things (iot)," *International Journal of Computer Applications*, vol. 113, no. 1, 2015.
16. C. Lee, C. Yeung, and M. Cheng, "Research on iot based cyber physical system for industrial big data analytics," in *Industrial Engineering and Engineering Management (IEEM)*, 2015 IEEE International Conference on. IEEE, 2015, pp.1855–1859.
17. Rizwan, K. Suresh, and M. R. Babu, "Real-time smart traffic management system for smart cities by using internet of things and big data," in *Emerging Technological Trends (ICETT)*, International Conference on. IEEE, 2016,pp. 1–7.

18. Q. Zhang, X. Zhang, Q. Zhang, W. Shi, and H. Zhong, "Firework: Big data sharing and processing in collaborative edge environment," in *Hot Topics in Web Systems and Technologies (HotWeb)*, 2016 Fourth IEEE Workshop on. IEEE, 2016, pp. 20–25.
19. M. M. Rathore, A. Ahmad, and A. Paul, "Iot-based smart city development using big data analytical approach," in *Automatica (ICA-ACCA)*, IEEE International Conference on IEEE, 2016, pp. 1–8.
20. B. Ahlgren, M. Hidell, and E. C.-H. Ngai, "Internet of things for smart cities: Interoperability and open data," *IEEE Internet Computing*, vol. 20, no. 6, pp. 52–56, 2016.
21. O. B. Sezer, E. Dogdu, M. Ozbayoglu, and A. Onal, "An extended iot framework with semantics, big data, and analytics," in *Big Data (Big Data)*, 2016 IEEE International Conference on. IEEE, 2016, pp. 1849–1856.
22. B. Cheng, A. Papageorgiou, F. Cirillo, and E. Kovacs, "Geelytics: Geo-distributed edge analytics for large scale iot systems based on dynamic topology," in *Internet of Things (WF-IoT)*, 2015 IEEE 2nd World Forum on. IEEE, 2015, pp. 565–570.
23. A. J. Jara, D. Genoud, and Y. Bocchi, "Big data for cyber physical systems: an analysis of challenges, solutions and opportunities," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, 2014 Eighth International Conference on. IEEE, 2014, pp. 376–380.
24. C. Vuppalapati, A. Ilapakurti, and S. Kedari, "The role of big data in creating sense ehr, an integrated approach to create next generation mobile sensor and wearable data driven electronic health record (ehr)," in *Big Data Computing Service and Applications (BigDataService)*, 2016 IEEE Second International Conference on . IEEE, 2016, pp.293–296.
25. A. Ahmad, M. M. Rathore, A. Paul, and S. Rho, "Defining human behaviors using big data analytics in social internet of things," in *Advanced Information Networking and Applications (AINA)*, 2016 IEEE 30th International Conference on. IEEE, 2016, pp. 1101–1107.
26. E. Ahmed and M. H. Rehmani, "Introduction to the special section on social collaborative internet of things," p. 382384, 2017.
27. D. Arora, K. F. Li, and A. Loffler, "Big data analytics for classification of network enabled devices," in *Advanced Information Networking and Applications Workshops (WAINA)*, 2016 30th International Conference on. IEEE, 2016, pp. 708–713.

NEW TRENDS IN OPINION MINING AND SENTIMENT ANALYSIS

R. Komalavalli¹ and Dr. R. Vidyabanu²Research Scholar¹ and Assistant Professor², L. R. G. Government Arts College for women, Tirupur**ABSTRACT**

World is evolving around the World Wide Web and information sharing technologies. Nowadays, people express their reactions to various public issues, events or products in social media applications. An organization can analyze such reactions of people to take an action on the event. In the current scenario, at the crossroad of computational linguistics and data retrieval opinions and emotions are more valuable than the subject of the document. Linguistic resources are used to retrieve sentiments and also to classify it. It is essential to develop methods to automatically classify and gauge them to identify the underlying sentiment about the product. Analyzing the polarity of sentiment expressed in data is Opinion Mining (OM). It is a system that identifies and classifies opinion/sentiment as represented in electronic text. Economic and marketing researches depend heavily on accurate method to predict sentiments of opinions extracted from internet and predict online customer's preferences. OM has many steps and techniques for each step

Keywords: Emotion Mining, Machine Learning, Opinion Mining, Sentiment Analysis, Semantics.

I. INTRODUCTION

The electronic world change the way of expressing feelings of particular. Major leading industries uses written customer reviews in comments for the business intelligent [1]. Different products from numerous amounts of retailers are available for E-Shopping. Different shopping sites like amazon, flipkart, snapdeal, myntra, offers customers to write their opinion about different features of the product [2]. This enormous corpus of review is playing very important role in competitive intelligence and it gives right direction to consumer as well as retailer. Not only the consumer uses this information for smart purchasing but also retailer uses this information to find out the pitfalls in their product and improve the quality, for finding out the current requirement of the market and to adapt the change of marketed [3].

Before purchasing any goods user find out the response of others about that product. for example a particular Smartphone contains various features like mp3, Bluetooth, calendar ,alarm ,browser ,wifi etc. as shown in figure1. People can debate some of the aspects play more important role than others, this may have a major weightage for general users for taking decision and to retailer for their future development plans[4]. for example consider some of the aspects of smart phone like "Battery", —Browsing speed, have more importance by majority of users that other aspects like "alarm", "calendar". The term sentiment mining is also referred as opinion mining which will help us for the process of sentiment analysis and for classifying sentiments. Some of the challenges faced in Sentiment analysis are that an opinion word which is positive in one situation can be negative in another situation and opinions are not expressed similarly by different people. Most reviews have positive and negative comments and are analyzed sentence by sentence. But, in more informal media like twitter or blogs, people are more likely to combine different opinions in same sentence which may or may not be easy to comprehend, but difficult for an algorithm to analyze [5].

SENTIMENT ANALYSIS AND OPINION MINING

Opinion mining (sometimes known as **sentiment analysis** or **emotion AI**) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Not surprisingly, there has been some confusion among practitioners, students and even researchers about the difference between sentiment and opinion and whether the field should be called sentiment analysis or opinion mining. In Merriam-Webster's dictionary, sentiment is defined as an attitude, thought, or judgment prompted by feeling, whereas opinion is defined as a view, judgment formed in the mind about a particular matter. The difference is quite little, and each contains some elements of the other. The definitions indicate that an opinion is more of a person's view about something, whereas a sentiment is more of a feeling. For example, the sentence "I am concerned about the current state of the economy" expresses a sentiment, whereas the sentence "I think the economy is not doing well" expresses an opinion. In a conversation, if someone says the first sentence, we can respond by saying, "I share your sentiment," but for the second sentence, we would normally say, "I agree/disagree with you." However, the underlying meanings of the two sentences are related because

the sentiment depicted in the first sentence is likely to be a feeling caused by the opinion in the second sentence. We can also say that the first sentiment sentence implies a negative opinion about the economy, which is what the second sentence is saying. Although in most cases opinions imply positive or negative sentiments.

Sentiment analysis and opinion mining is almost same thing however there is minor difference between them that is opinion mining extracts and analyze people's opinion about an entity while Sentiment analysis search for the sentiment words/expression in a text and then analyze it.

EMOTION MINING

Emotions are human beings subjective feelings and thoughts expressed by them. The significant study has been done on the emotions and studied in many fields like biology, social science, psychology, philosophy, etc. Still it is found that researchers are not agreed on basic set of emotions. Mainly, there are six types of emotions, namely joy, love, surprise, anger, fear and sadness. These emotions further can be classified into their own subtypes. Ekman identified six basic emotions: happiness, sadness, anger, fear, disgust and surprise [6]. Other approaches do not search to categorize emotions in specific categories, but rather identify them on two scales: the valence of the emotion indicating if the feeling is positive or negative and the arousal level indicating the energy level associated with the emotion [7].

In fact, Thelwall et. al. performed emotion mining from texts retrieved from the online social network MySpace. They argued that studying emotions based on a 2-dimensional scale (i.e. valence and arousal) is more reliable and provides more accurate results than studying emotions on a finer grain [7]. Although emotions are universal, there are huge differences between cultures and between individuals in the way and the extent in which these emotions are expressed. In general, women are more likely to share their emotions and their feelings than men; and this observation was also verified in online social networks [7]. Furthermore, personality is an important factor influencing emotions. Social factors have also an effect on emotions where the expression of emotions is not limited to a person's internal feelings but influenced by the society, a person's strategic goals and previous experiences [7].

Emotion mining can be divided into three categories depending on the purpose for mining emotions. The first category aims at extracting the valence of the text, indicating if the text has positive or negative emotions associated with it. The second category aims at identifying whether the text is subjective or factual (i.e. objective), thus the purpose is to find if the text is emotionally rich or not. The third category aims at recognizing not just the emotion but also its strength or arousal.

Though emotions are influencing the strength of opinion and are exhibiting close relationship with opinion, there is no equivalence between emotion and opinion. So when the impact of emotions or opinions is discussed, it is important to differentiate between these two things i.e. feelings and language expressions used to describe feelings. As stated earlier, there are only six types of primary emotions; but on other side there are many expressions of the language that can be used to convey emotions. In the same sense, there are also many opinion expressions used to explain positive sentiments or negative sentiments by users. Therefore, mainly sentiment analysis attempts to conclude individual sentiments based on expressions provided with the help of any language to portray their emotions.

II. TAXONOMY

Sentiment analysis can be achieved at various levels, the levels are: Phrase Level, Aspect Level, Sentence Level, Document Level, Natural Language Processing. Depending upon nature of use level of Sentiment analysis is selected.

1] Phrase Level: It works with phrases of a document. The words which are closer to each other are called as phrases. Phrase level classification is performed on the phrases which contain opinion words. Here we cannot find the long range dependency. Usability is depends upon situation [8]. In a sentence if opinion words are too far then phrase level analysis is not useful [9].

2] Aspect Level: Aspect level analysis is also called as feature based classification. Instead on studying structure of the sentence it directly deal with the aspects in the text[10]. For example —Nokia internet speed is good but its battery drain out early —it consist two aspects, internet speed and battery life of the product Nokia. The opinion on the Nokia internet speed is positive, but on its battery life is negative. so internet speed and battery life are targets. Based on this analysis, a structured summary of sentiments about products and their features can be produced [11].

3] Sentence Level: It works with finding out polarity of each sentence. It considers subjectivity and objectivity of the sentence. Subjectivity is related to domain, one sentence is opinion about single domain. If the sentence is too complex then sentence level classification fails [12].

4] Document Level: It works with the documents which contain E-text. It finds out polarity of the document, if a positive phrase is present it does not mean that user like everything vice versa for the negative phrase [13]. Document Level analysis classify whole document here sentiments of the single document is considered so it will not work well with news portal data ,blogs .It is used with both supervised and unsupervised learning algorithms[14].

5]Natural Language Processing: Natural Language Processing works with the grammar of the sentence , on the basis of grammar it searches the nouns , adjectives, verbs, etc. for aspect level classification it is best suited[15]. Consider example,"The mobile has excellent battery backup". here we can identify the camera is a feature but for machine understanding the grammatical structure takes place by grammar if we consider then the noun term is the feature of the product. The demerit of NLP is it fails if the sentences contain grammatically incorrect words, but we can overcome it by preprocessing the sentences [16].

III. TECHNIQUES USED IN SENTIMENT ANALYSIS

Sentiment Analyzing and Classification uses various techniques which are mainly divided in to two approaches they are Machine Learning and Lexicon based approach. Each technique is again sub divided. They are explained as below:

1] Machine Learning: Machine learning methodology consist of supervised ,unsupervised and semi supervised categories .each category is again sub divided as shown in figure1

a] Supervised Learning: It predicts attribute classes on the basis of given set of training values. It contains training and testing dataset. Training dataset is smaller which contain same

b]Unsupervised Learning: Unlike supervised learning classification training data is not needed for unsupervised learning, here for data classification it uses clustering algorithms like K-Mean clustering, Hierarchical Clustering etc. for determining threshold values of words neural network may be used and then classify them depending upon thresholds[17]. Below Table consist different models which use unsupervised learning. Table 2: Different Models Unsupervised Learning Method

c) Semi supervised Learning: It is the combination of supervised learning and unsupervised learning .It aggregates the accuracy of supervised learning and word stability and readability of unsupervised learning [18].

2] Lexicon Based Techniques

Opinions words are categories into positive and negative sense. Sentiment phrases and idioms are also there which are called as sentiment lexicon. Sentiment lexicon is mainly categorized in to three types Manual, dictionary based and corpus based [19]. Manual approach is more time consuming than other two automated approaches. in Dictionary based category tiny set of opinion words are gathered and then they increased by looking their word finder for their similar words and opposite words[20]. If new words found then it will append to list otherwise iteration starts again until no new word found. It will not work for domain and context specific words. Corpus based category searches the opinion word in domain specific approach. It will categorized in to

1: Semantic Approach: It is use to search co-occurrence sequence and seed opinion words 2: Statistical Approach: This approach give sentiment values directly relied on various standards for computing the synonym words [21].

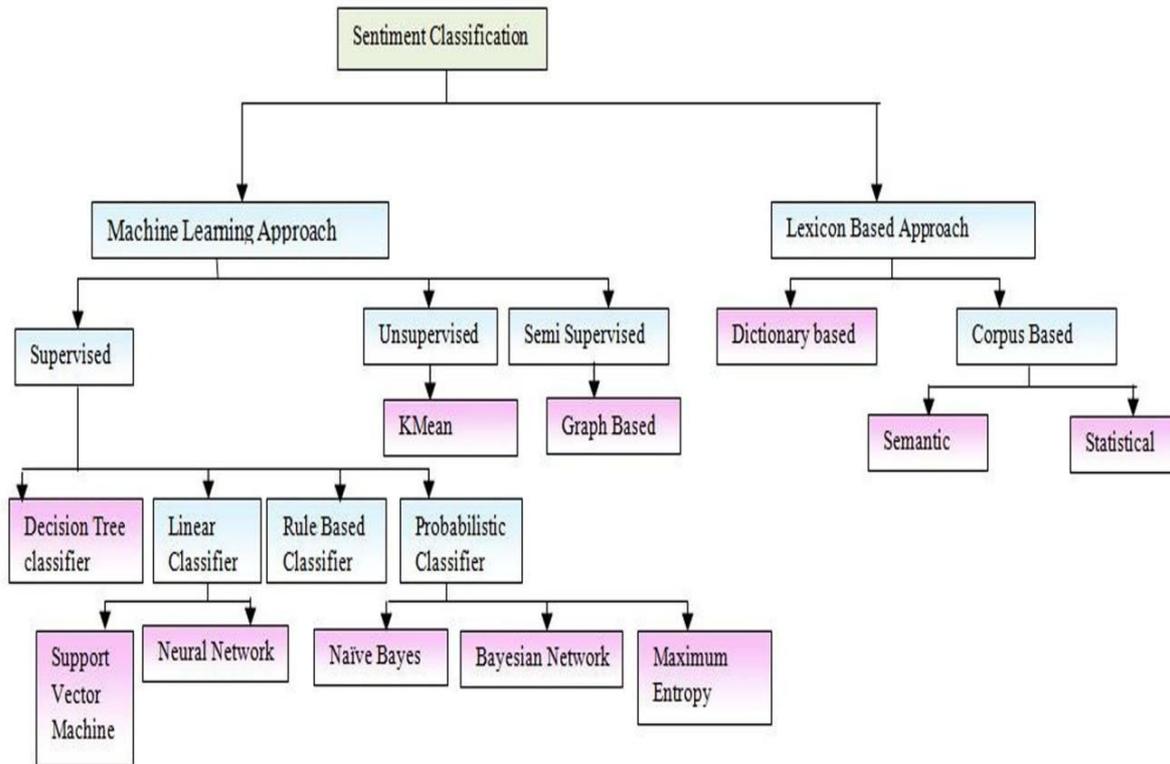


Figure-1: Different Techniques used in Sentiment Analysis

IV. TOOLS USED IN SENTIMENT ANALYSIS

The tools which are used to track the opinion or polarity from the user generated contents are:

- Review Seer tool [22]–This tool is used to automates the work done by aggregation sites. The Naive Bayes classifier approach is used to collect positive and negative opinions for assigning a score to the extracted feature terms.
- Web Fountain [23]-It uses the beginning definite Base Noun Phrase (bBNP) heuristic approach for extracting the product features.
- Red Opal [24]–It is a tool that enables the users to determine the opinion orientations of products based on their features. It assign the scores to each product based on features extracted from the customer reviews’
- Opinion observer[25] - This is an opinion mining system for analyzing and comparing opinions on the Internet using user generated contents. This system shows the results in a graph format showing opinion of the product feature by feature. Along with these automated tools, there are various online tools like Twitrratr, Twendz, Social mention, and Sentimetrics are available to track the opinions in the web.
- Konstanz Information Miner (KNIME) [26] - Konstanz Information Miner (KNIME) which is a user-accommodating graphical workbench capacities of the whole analysis process. KNIME uses six distinctive strides to process writings: perusing and parsing archives, named substance acknowledgment, shifting and manipulation, word numbering and catchphrase extraction, and transformation and perception. Taking after work processes and assignments are created and executed utilizing KNIME
- Retrieving data from a database
- Dictionary improvements and use
- Review scores

.V. CONCLUSION

Sentiment analysis and opinion miming researches are indeed good advancement in science and engineering fields. In this study general introduction on sentiment analysis and opinion mining. This paper discusses about the necessity of the sentiment analysis and opinion mining in real time operations. It gives the basic classification and techniques used for the sentiment analysis. Finally it discusses the tools used for the sentiment analysis in real time. With these details given it is hoped that researchers will engage in sentiment analysis and opinion mining researches to attain more successes.

REFERENCE

1. Zheng-Jun Zha, Member, IEEE, Jianxing Yu, Jinhui Tang, Member, IEEE, Meng Wang, Member, IEEE, And Tat-Seng Chua, "Product Aspect Ranking And Its Applications", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014
2. Singh and Vivek Kumar, —A clustering and opinion mining approach to socio-political analysis of the blogosphere. Computational Intelligence and Computing Research (ICCIC), 2010 IEEE
3. M Fan, G WU —Opinion Summarization of Customer comments. International conference on Applied Physics and Industrial Engineering in 2012.
4. B. B. Khairullah Khan, Aurangzeb Khan, —Sentence based sentiment classification from online customer reviews. ACM, 2010.
5. Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6).
6. P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, Vol. 6, No. 3, pp. 169-200, 1992.
7. M. Thelwall, D. Wilkinson and S. Uppal. "Data mining emotion in social network communication: Gender differences in MySpace". In *Journal of the American Society for Information Science and Technology*, pp. 190-199, 2010.
8. Liu, B. (2010), — Sentiment Analysis and Subjectivity. Appeared in Handbook of Natural Language Processing, Indurkha, N. & Damerau, F.J. [Eds.].
9. Bing Liu, Mining Hu, Junsheng Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web", WWW 2005, May 10-14, 2005, Chiba, Japan, ACM 1-59593-046-9/05/0005.
10. Zhang, Z. and B. Varadarajan. Utility scoring of product reviews. In Proceedings of ACM International Conference on Information and Knowledge Management (CIKM-2006), 2006.
11. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using Machine Learning techniques. In: Proc. of CoRR (2002)
12. Nasukawa, T. and Yi, J. 2003. Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the 2nd Intl. Conf. on Knowledge Capture (K-CA'2003).
13. Camelin, N., Damnati, G., Béchet, F. and De Mori, R., —Opinion Mining in a Telephone Survey Corpus. International Conference on Spoken Language Processing in 2006.
14. Mohammad S, Dunne C, Dorr B. Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'09); 2009.
15. Bing Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, May 2012. p.18-19, 27-28, 44-45, 47, 90-101.
16. Kim S, Hovy E. Determining the sentiment of opinions. In: Proceedings of international conference on Computational Linguistics (COLING'04); 2004.
17. Pang, B. & Lee, L. (2008), —Opinion Mining and Sentiment Analysis. In Foundations and Trends in Information Retrieval 2 (1-2), p. 1–135.
18. Ayesha Rashid, Naveed Anwer, Dr. Muddaser Iqbal, Dr. Muhammad Sher, "A Survey Paper: Areas, Techniques and Challenges of Opinion Mining", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013
19. Daille, B. 1996. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. The Balancing Act: Combining Symbolic and Statistical Approaches to Language. MIT Press, Cambridge
20. Somasundaran S, Wiebe J. Recognizing stances in online debates. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP; 2009. p. 226–34
21. Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis" book.

-
-
22. G.Vinodhini and RM. Chandrasekaran 2012. Sentiment analysis and Opinion Mining: A survey International Journal of advanced Research in Computer Science and Software Engineering vol. 2 Issue 6.
 23. Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng and Chun Jin 2007. Red Opal: product-feature scoring from reviews, Proceedings of 8th ACM Conference on Electronic Commerce, pp. 182-191, New York.
 24. Yi and Niblack 2005. Sentiment Mining in Web Fountain” ,Proceedings of 21st international Conference on Data Engineering, pp. 1073-1083, Washington DC.
 25. M. Hu and B. Liu 2004. Mining and summarizing customer reviews, Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 168–177.
 26. Banan Awrahman¹, Bilal Alatas² Sentiment Analysis and Opinion Mining within Social Networks using Konstanz Information Miner, ISSN: 2180-1843 e-ISSN: 2289-8131 Vol. 9 No. 1

A COMPREHENSIVE STUDY OF CURRENT TRENDS IN WEB MINING

Suhasini V¹ and Dr. N. Vimala²Research Scholar¹ and Assistant Professor², L. R. G. Government Arts College, Tirupur**ABSTRACT**

The World Wide Web is a huge, information center for a variety of applications. Web contains a dynamic and rich collection of hyperlink information. It allows Web page access, usage of information and provides numerous sources for data mining. The goal of Web mining is to discover the pattern of access and hidden information from huge collections of documents. The present era is engulfed in data and it is quite difficult to churn the emergence of unstructured data in order to mine relevant information. The purpose of this paper is to provide the current evaluation and update of web mining research and techniques available. Current advances in each of the three different types of web mining are studied in the categories of web content mining, web usage mining, and web structure mining.

Keywords: Data Mining, Web Mining, Web Content Mining, Web content mining tools, Web Structure Mining, Web Usage Mining.

I. INTRODUCTION

In Today's world, there is a rapid increase in the usage of Internet applications in day to day life and it grows significantly and steadily day by day, thereby distressing the lives of people in almost all the sectors like health, education, business etc. The web applications are gaining more popularity in the present scenario due to the contributing factors like expediency and flexibility of services provided by web applications. Web applications could able to work with a huge data which consistently consists of various user operations, transactions and user activity logs. In order to enhance the decision making process, the framework of Knowledge Discovery from Databases (KDD) [1] have been used and various people have conducted many experiments to discover the various ways of retrieving possibly useful information which is embedded in large databases. The main process of KDD, is called as data mining, and its main work is to retrieve the frequent patterns which includes association rules and sequential patterns mining. Web mining is one of the applications of data mining and it specified about the web data [2].

Many of the researchers are involved in mining the data due to the tremendous increase in the growth of the information sources available on the web and also ecommerce. According to the authors Madria, et al. [3] and Borges and Levene [3], Web mining have been categorized into three broad areas of interest namely: Web content mining, Web structure mining, and Web usage mining. The above three mining tasks can be used in isolation or it can combined with other tasks since they might contain the links of the web document.

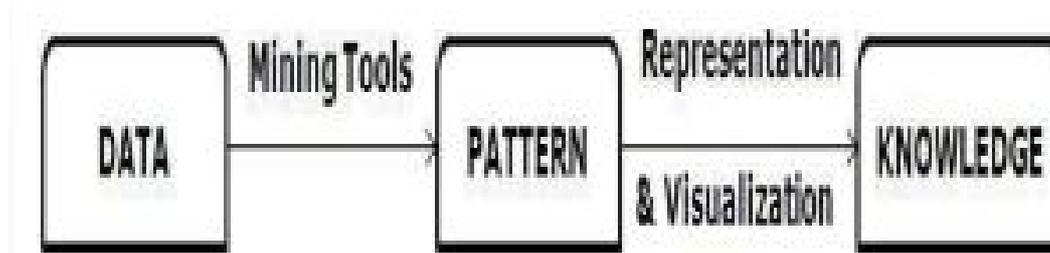


Fig-1: Route of Web Mining

Web mining is categorized under data mining technique due to which kind of information is to be extracted and retrieved process from large number of documents and web services mechanically. The main aim of data mining is to the retrieve the necessary and exciting patterns from a collection of enormous data sets in the current trend as well as used in the typical data mining. Web mining uses big data as the data set from which it tries to retrieve the data. Web data typically consists of various profile, structure, documents, information etc.

Web mining is broadly based on two major concepts namely process-based and data-driven. In Web mining mainly try to extract knowledge from the web [4]. The steps involved in web mining as: collection of data, data selection before processing, knowledge discovery and analysis of data [5].

II. TAXONOMY

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Figure 2 explains the web mining taxonomy.

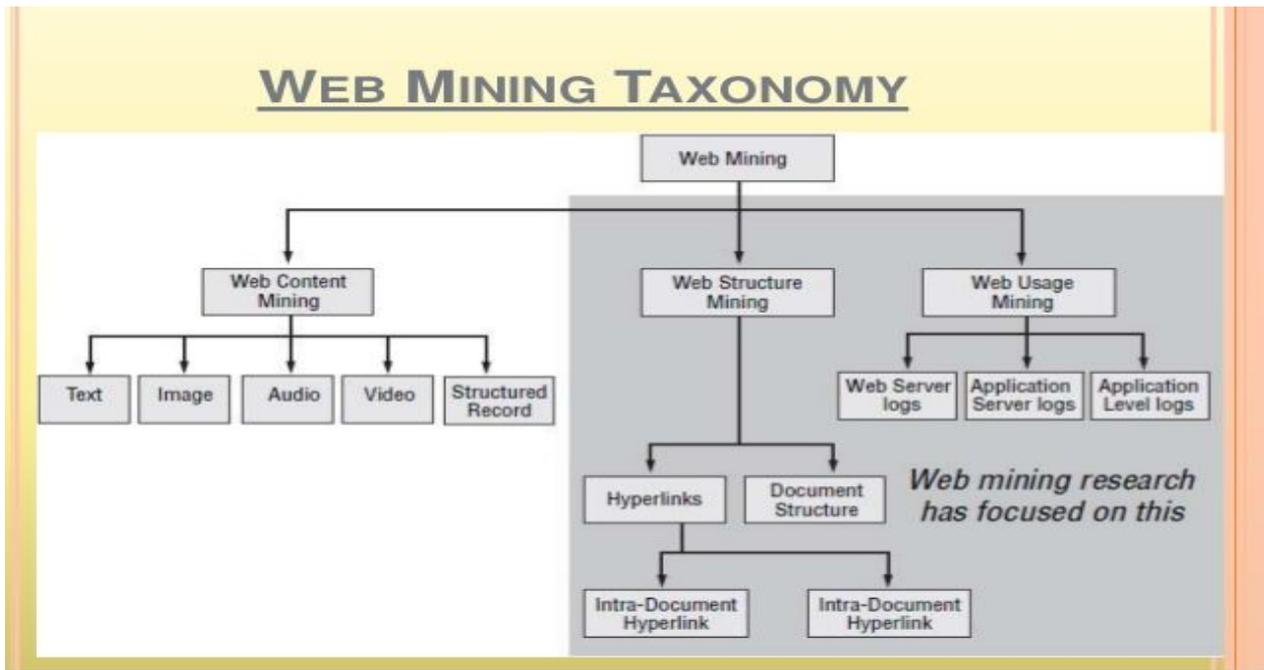


Fig-2: Taxonomy of Web Mining

WEB CONTENT MINING

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain.

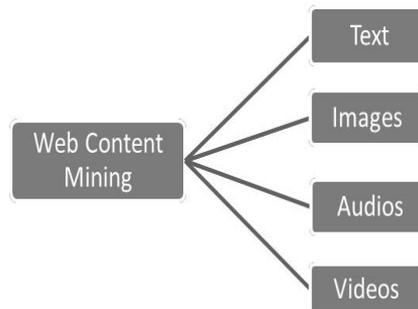


Fig-3: Web Content mining

It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited. Web content mining can be performed on images, videos, audio and texts. The mining techniques for each are different. Here is the brief detail of each mining techniques are

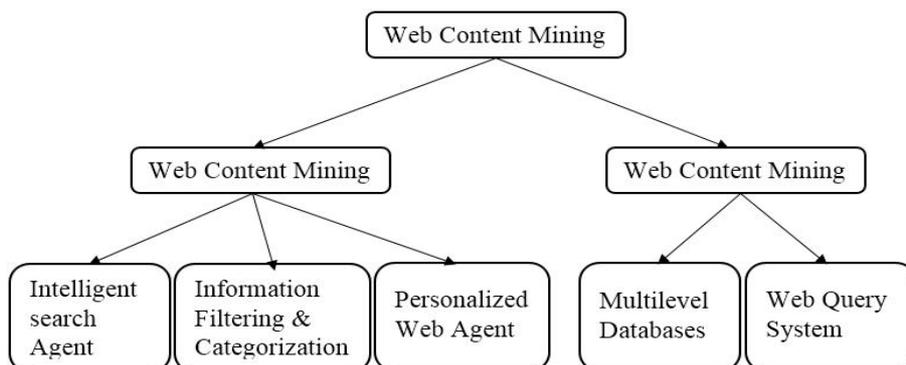


Fig-4: Type of Web Content Mining

Image Mining: Image Mining is the technique which is used to detect unusual patterns and extract important and useful information from the images stored on the web and large database. Thus image mining mainly deals with defining relationships between different mages from the web and large databases [5]. Image mining is used in various fields like medical diagnosis, space research, remote sensing, agriculture and industries. The images include maps, geological structures and even it is used in the field of education.

Video Mining: Mining video is more complicated than mining image data. Video is the collection of moving images like animation. There are 3 types of videos: 1) the produced (movies, news videos etc), 2) the raw (traffic, surveillance etc) and 3) the medical video (x-ray, cardiogram etc). The information from the video can be a) detecting trigger events (movement of vehicle and people), b) determining typical patterns of activity, generating person or object centric views of activity, and c) classifying activities into named categories (walking, sleeping etc), clustering and determination of interactions between two entities [6]. Video mining can also be classified into pixel based, statistics based, feature based and histogram based.

Audio Mining: As audio is the continuous media like video, the techniques and tools for audio processing and mining information is similar to video mining. Audio data can be in the form of radio, speech or spoken languages. Also television news has audio which are integrated with the videos. Mining audio data requires conversion of audio to speech for better processing. Audio data can be directly mined by using audio information processing methods and extracting selected audio data. Very few works has been carried out in the field of audio mining.

Text Mining: The most trending research in the field of web content mining is text mining. The text mining refers to the text representation, classification, clustering, information extraction and search for hidden patterns. Text mining is the process of extracting useful information from the text and converting to automated discovery of knowledge [7]. It is natural extension of data mining or applying data mining techniques on a specific domain. Text categorization is one of the technique of text mining which classifies the texts in a particular domain [8]. It works on keywords extraction from the document. A lot of work has been carried out in the field of text mining.

Structured records: Structured records are most important technique in web content mining. Structured records are classified as: semi-structure, unstructured and structure formats. 1) Semi-structure data is a form structure data that does not with formal structure data modules associated data with relational database or other forms of data tables. 2) Unstructured data mining as the practice of looking at relatively unstructured data and trying to get refined datasets out of it. It consists of extracting data from source not traditionally data mining activities. 3) Structure mining can be divided into two kinds extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects webpage to a different location.

WEB STURCTURE MINING

This type of mining focuses on the data which describes the structure of the content of the web page. It is classified into two types: 1) intra-page structure: existence of links within the page, 2) inter-page structure: the connection of one web page with other web pages. This can be classified into two types based on structure of information.

Hyperlinks: A hyperlink is a structural unit that connects one web page with other web page either within same location or different location. A hyperlink connecting web pages in the same location is called intra-document hyperlink and a hyperlink connecting web pages at different locations is called inter-document hyperlink.

Document Structure: The content within a web page can also be organized in a tree structure based on HTML and XML Tags used to create a web page. Mining can be done to identify document object model (DOM) structures automatically from the documents.



Fig-5: Web Structure Mining

WEB USAGE MINING

Web usage mining is one techniques of data mining to extract interesting and useful patterns from the web usage logs. Usage logs stores the identity or the origin of web users along with the browsing behavior on the web site. Web mining can be grouped based on the type of usage logs:

Web Server Data: User logs are collected by the web server which includes IP address, page references and the time accessed by the user.

Application Server Data: Application servers are used to track various types of business events which can be used to improve the performance for any business firms. For e.g. E-commerce websites uses such servers to know the events, business policies developed by their competitors.

User Level Data: User level data is the software developed using the information available from the web server and application server data. It is an end user application which is used by different users for different purposes.

II. DIFFERENCE BETWEEN CLASSIC DATA MINING AND WEB MINING**Web mining**

- The web is a collection of interrelated files, even though it is not a relation.
- Web mining is the discovery of knowledge from the web.
- Usage data is huge and growing rapidly.
- Ability to react in real-time usage of patterns.

Data mining

- Textual information and linkage structure.
- Google's usage logs are bigger than their web crawl.
- Data generated per day is comparable to largest conventional data warehouse.
- No human in the loop.

III. TECHNIQUES AND STAGES IN WEB MINING**Techniques**

The Web Usage Mining comprises of different steps which help in overall cleaning and finding relevant pattern from the data. Different techniques are used in different steps of the web mining process. the techniques used for web mining are listed below.

Preprocessing: It is the first step used for data which is stored in web logs. It comprises of data cleaning, session reconstruction, content and structure information retrieval and data abstraction. These stages deals with cleaning of data, finding use session and reconstruction, enriching web log data and development of abstraction inform of session and page views. This step is further categorized into usage preprocessing which is difficult task as it has incomplete server log data, content preprocessing which deal with converting unstructured and semi -structured documents into suitable forms by using vector space model and structure preprocessing which handle hyperlinks.

Pattern discovery: This phase focuses on some algorithms which are quite useful in extracting pattern and understandable knowledge by using statistics, machine learning and data mining techniques. Various algorithms are - Statistical Analysis which helps in improving system performance by focusing on data. Association Rules uses Apriori algorithm which help in finding correlations among web pages. Clustering which is used for grouping similar session and helps to find user having similar behavior. Classification deals with categorization of documents which help in detecting interesting patterns. Sequential Patterns help to find sessions that are placed in a particular order.

Pattern Analysis: This step helps in system performance by separating interesting patterns from uninteresting ones. Various techniques like Knowledge Query Mechanism, Visualization tools and Intelligent Agents are used for analyzing discovered patterns. This phase help in redesigning of the websites, improving access time of fetching the pages and helps in enhancing the browsing of desired pages.

Stages

Web Usage Mining helps in recovering user access patterns from web servers. It comprises of four processing stages.

Data Collection: This stage helps the web managers for improving the management, performance and controlling of web servers by collecting the hidden information and usage patterns.

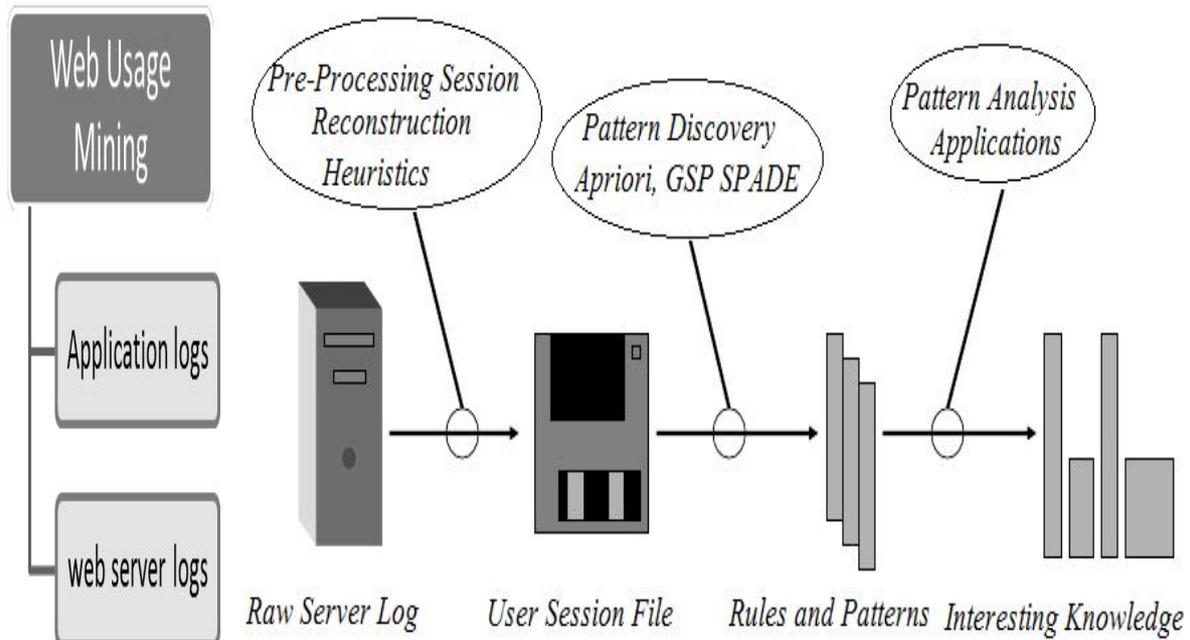


Fig-6: Phases of Web Usage Mining

Data Preprocessing: This stage focuses on selecting useful data by eliminating irrelevant and noisy data. It tries to arrange most recently accessed data with higher index with respect to least used by providing lower index value. It is regarded as a critical step as it helps in the analysis to obtain precise result.

Data Clustering: Clustering algorithm are most frequently used in mining web pages and forming cluster objects. This method is helpful in finding usage patterns and user profiles.

Pattern Discovery & Analysis: With help of data analysis and graph relevant and useful information can be easily predicted.

APPLICATION AND TOOLS OF WEB MINING

Application

Application areas of Web Mining

- E-Commerce
- Search Engines
- Personalization
- Website Design

The various fields where web mining is applied are:

- E-Commerce
- Information filtering
- Fraud detection
- Education and research

E-Commerce: In e-commerce, web mining helps in generating user profiles by customizing the choice of users. For example, web mining enables a user to search for an advertisement and information regarding a product of his interest. Internet advertising is one of the major fields in e-commerce, where web mining is widely used. Advertising in a specific domain of an e-commerce web site or a general web site and is considered as one of the major application area of web mining.

Information filtering: Information filtering is the method to identify the most important results from a list of discovered frequent set of data items for which you can make use of web mining.

Fraud detection: Fraud detection can be performed using web mining by maintaining a list of signatures of all the users. Web mining is also applied for plagiarism detection and research works.

TOOLS OF WEB MINING

Web mining tools are the software or the applications which helps the users to download essential information from the web. It collects the exact and required information for the user which can be helpful in mining. Number of sources which help in accessing data in Web Usage Mining includes cookies, web access log, login information, client or server side scripts etc. Different tools are used which help in recovering and performing data preprocessing, pattern discovery and pattern analysis are Data Preparator, SumatraTT, Lisp Miner, SpeedTracer, SEWEB AR-CMS, i-Miner, Argonaut, Webalizer, WebViz and WebMiner. These tools performs the task of cleaning, preprocessing, reconstructing, pattern discovery, rules selection, clustering, analyzing patterns and forming visualization of graphical patterns. Some of the tools are explained below.

Automation Anywhere: It is a tool which is used to extract web data very easily, screen shots of web data which can be used in web mining. It is unique Intelligent and Smart Automation Application used for quick automation of any complex tasks [10].

Web Info Extractor: This tools is used to collect web content, constantly updating data and analyzing data. This tool can be used to extract information like images, vidoes and texts [11].

Screen-Scraper: It is a tool to extract information from the websites. It can be used in searching databases and document structure. It provides a graphical interface allowing the user to navigate through URLs, data elements and hyperlinks and extract useful information from it [12].

Mozenda: This tool is used to extract and manage web data. User is allowed to setup tools at different places which can store and publish data at a regular interval of time [13] . This data can be used with other applications for mining purposes.

Web Content Extractor: It is a tool for extracting web scraping, data mining and information retrieval from the internet. This tool is used to extract information from various websites like online auctions, online shopping, business directories, financial sites etc. The data can be represented in the form of excel, HTML, XML or any other script [14].

VI. CONCLUSION

This paper characterizes about web structure mining, web content mining, and web usage mining including its structure. Internet and websites provides opulent platform for searching the information. Most of the websites are complex and larger in their size and structure. It gives the basic introduction to web mining. It includes the taxonomy of the web mining. And it also elaborates the difference between web mining and basic data mining. After then it includes techniques and stages in the web mining process. Finally it describes application and tools used for web mining.

REFERENCES

1. Kamika Chaudhary and Santosh Kumar Gupta, "Web Usage Mining Tools & Techniques: A Survey", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013.
2. Shadab Irfan, Subhajit Ghosh, " Optimization of Information Retrieval Using Evolutionary Computation: A Survey", International Conference on Computing, Communication and Automation (ICCCA2017).
3. Nasrin JOKAR, Ali Reza HONARVAR, Shima AgHAMIRZADEH, Khadijeh ESFANDIARI, "Web mining and Web usage mining techniques", Bulletin de la Société des Sciences de Liège, Vol. 85, 2016.
4. Shadab Irfan, Dr. B V Babu, "Information Retrieval in Big Data Using Evolutionary Computation: A Survey", International Conference on Computing, Communication and Automation (ICCCA2016).
5. Gadamsetty Vasavi and Dr. T. Sudha, "Information Extraction from Online Shopping Sites using Web Content Mining Methods and Techniques", International Conference on Innovative Applications In Engineering and Information Technology (ICIAEIT-2017).
6. Ji Zhang, Wynne Hsu and Mong Li Lee "An Information -Driven Framework for Image Mining" Database and Expert Systems Applications in Computer Science, 2001, Volume 2113/2001, 232-242, DOI: 10.1007/3-540-44759-8_24
7. Boreczky J. S. and L. A. Rowe, "A Comparison of Video Shot Boundary Detection Techniques", Storage & Retrieval for Image and Video Databases IV, Proc. SPIE 2670, 1996, pp.170-179

8. A. Czyzewski, "Mining Knowledge in Noisy Audio Data", in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, pages 220-225, 1996.
9. Ahmad Tasnim Siddiqui and Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications", International Journal of Computer Applications Volume 69–No.8, May 2013.
10. Automation Anywhere Manual. AA, <http://www.automationanywhere.com>
11. Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques, and Software, International Journal of Information Technology & Decision Making. Vol.7, No.4, pp. 683-720. World Scientific Publishing Company (2008)
12. Mozendo, <http://www.Mozendo.com>
13. www.screen-scraper.com
14. Web Content Extractor help. WCE, <http://www.newprosoft/web-content-extractor.htm>
15. Strategy and Implementation of Web Mining
16. TOOLS ,Danish Ahamad College of Science and Arts, Sajir, Shaqra University, KSA danish@su.edu.sa, Md Mobin Akhtar, College of Computing and Information Technology, Shaqra, Shaqra University, KSA jmi.mobin@su.edu.sa, Dr. Adil Mahmoud Mohamed Mahmoud College of Science and Arts, Sajir, Shaqra University, KSA adilmahmoud@su.

COMMUNICATION PROTOCOL FOR SECURED DATA TRANSFER ON IOT

Dr. P. AnanthiAssistant Professor, Kongu Engineering College, Tamilnadu

ABSTRACT

In Internet of Things (IoT), there is a vast number of connected devices that exist. These devices are collecting and transmitting great volumes of data from device to device, device to enterprise systems, and occasionally from device to humans. Due to the billions of connected devices, there is a great risk of identity and data theft, device manipulation, data falsification and server/network manipulation. While the number of these interconnected devices continues to grow every day, so does the number of security threats and vulnerabilities posed to these devices. Security is one of the most paramount technological research problems that exist today for IoT. Security has many facets - security built within the device, security of data transmission, and data storage within the systems and its applications. This paper focuses on secure IoT device communication. The multi hop routing protocol enables the IoT devices to authenticate before forming a new network or joining an existing network. The authentication uses multi-layer parameters to enhance the security of the communication. The proposed routing protocol embeds the multi-layer parameters into the routing algorithm, thus combining the authentication and routing processes without incurring significant overheads. The multi-layer parameters include a unique User-Controllable Identification, users' pre agreed application(s), and a list of permitted devices, thus saving resources by maintaining smaller routing information. Experimental and field tests were conducted with results showing that our secure multi-hop routing is suitable to be deployed for IoT communication.

I. INTRODUCTION

The emerging technical space is growing with the Internet of Things (IoT). IoT is bringing about a paradigm shift in services, infrastructure, and consumer industries [1]. While this paradigm shift is happening, trust and security are necessary requirements to tackle different kinds of attacks, threats, malfunctions, and devastating impacts to society. The responsibility of securing IoT lies with device manufacturing companies and companies that use the devices. Having a complete set of security terms is a priority to organize the threat and overcome all security challenges in IoT. With increased commercialization of IoT devices, society is becoming more and more connected with the IoT infrastructure - making society more susceptible to the vulnerabilities of the current IoT environment. IoT will increasingly touch our lives in more ways than before. Hence, research community must tackle and resolve security aspects of IoT. Compromised IoT devices present the risk of misusing personal information, compromising other connected systems, and safety risks.

The Multi-Hop Routing Protocol (MRP) that requires an insignificant overhead for discovering IoT devices. The MRP embeds the security information of the IoT devices in the routing protocol. The MRP is adopted from the User-Controllable Multi-Layer Secure Algorithm for mobile ad-hoc networks. The efficient and effective authentication method for IoT devices that uses three notifications such as matching User-Controllable Identification (UCID), similar users' pre-agreed application(s) and Permitted devices list based on the network and data-link addresses. Using this methodology, the MRP disregards those IoT devices that do not conform to any of the three recognitions.

II. MULTI HOP ROUTING PROTOCOL

This protocol has the capability to isolate the IoT devices based on the UCID, the application(s), the network address and the data-link address. The adaptive nature of the architecture is capable of dynamically re-defining logical networks within a network formed by IoT devices. The network formation and the authentication process occur simultaneously, thus providing a secured IoT communication. It is resilient to a hacker's attack by able to change the embedded secure information in the routing protocol frequently.

The protocol stack of the MRP composes of the following four layers

- 1) Application Layer, where the devices' applications are executed.
- 2) Transport Layer, where end-to-end communication services such as connection-oriented data stream support, reliability, flow control, and multiplexing are provided.
- 3) User-Controllable Multi-Layer Layer (UML), where a routing agent that provides network addressing and routing, an Application(s), network address and Data-link address Lookup (ANDL) module to verify the users' pre-agreed application(s) and addresses, a mechanism to enter UCID and a Unique Code Generator (CG) and

4) Data Link Layer, that provides the functional means to transfer data between devices and the means to detect and possibly correct errors that may occur physically.

In this routing protocol HELLO message is forwarded in regular intervals by the IoT devices to discover its neighbors. This message is generated and forwarding to sense the link and neighbor. During the arrival of IoT HELLO message from device1, the device2 verifies the incoming message header with its own message header. If the header matches two devices form a network for communication.

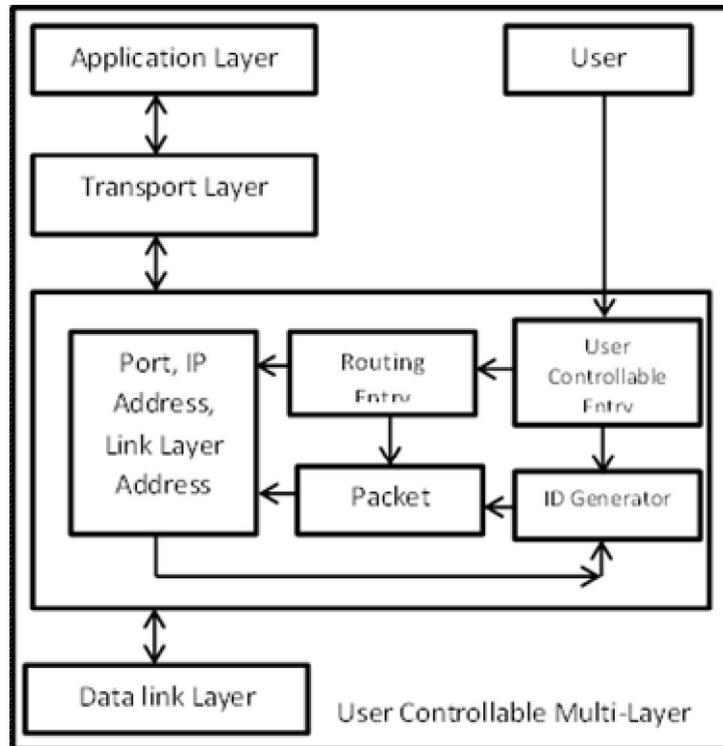


Figure-1: MRP Layers

Consider a device3, that intends to form a new IoT network. A user enters a UCID using an Application Programming Interface (API) into the UML. The ANDL module in the UML examines the application(s) executing in device3 and keeps a record of it. When a neighboring IoT device, device4 requests to join device3 the network address and data-link address of device4 are checked against the permitted devices in the EF stored in device3. If device4 is permitted and the application(s) running in device4 matches the application(s) in device3, a signal is sent to GC to create a unique code together with the UCID.

Table-1: HELLO message format

0	16	31
Reserved		TTL
Link Identifier	Message Length	Type
Neighbor Link Interface Address		
Neighbor Link Interface Address		

The ID generated by CG is then embedded into the “Reserved” bits of the HELLO packet before passing the packet to the Data Link Layer. This process is executed simultaneously in device3 and device4. Device4 is then able to communicate with device3. The cryptographic algorithm presented in this paper is by no means extensive; however it is sufficient to serve as a proof-of-concept to embed the authentication information into the routing algorithm. The UCID entered by the user and the information of the application(s), network and data-link addresses of the IoT device obtained by ANDL are fed into an encryption algorithm. The information obtained from the encryption algorithm is stored in the header of the HELLO message. The Timer module keeps a synchronized time among all the IoT devices in the network. Once the Timer reaches a pre-assigned value, it will trigger the Scrambler module to change the sequence of the “Reserved” bits of the HELLO message to enhance the security of the embedded secure information in MRP.

III. EXPERIMENTS

The equipment used in our experiments include a Windows XP laptop, an iPad and 3 IoT devices using PhidgetSBC 1072 single board computers. The specification of the laptop was Compaq 6510b Intel Core2Duo CPU T7100@1.8GHz with 2GB RAM and Broadcom 802.11a/b/g WLAN. It was installed with Ubuntu 11.04. The iPad was running on iOS 4.3.3, had 16GB of storage and WiFi capabilities. The IoT devices were running on Debian 6.0 with 64MB of SDRAM and 512 MB flash.

IV CONCLUSION

In this paper, a Multi-Hop Routing Protocol (MRP) for IoT communication is presented. The proposed MRP merges the routing and authentication processes for forming a secured IoT network without incurring significant overheads. Our experimental and field tests results showed that the MRP produces a secured multi-hop IoT communication network without performance degradation as compared to the well-known OLSR protocol.

REFERENCES

- [1] Atzori L, et al., "The internet of things: A survey," *Computer Networks*, vol. 54, pp. 2787-2805, 2010.
- [2] Babar S, Mahalle P, Stango A, Prasad N, and Prasad R, "Proposed security model and threat taxonomy for the internet of things (IoT)," in *Anonymous Berlin, Heidelberg: Springer Berlin Heidelberg*, 2010, pp. 420-429.
- [3] Iqbal A, Saleem R, and Suryani M, "Internet of Things (IOT): on going Security Challenges and Risks," *International Journal of Computer Science and Information Security*, vol. 14, pp. 671, 2016.
- [4] Kortuem G, et al., "Smart objects as building blocks for the internet of things," *Internet Computing, IEEE*, vol. 14, pp. 44-51, 2010.
- [5] Kranz M, et al., "Embedded interaction: Interacting with the internet of things," *Internet Computing, IEEE*, vol. 14, pp. 46-53, 2010
- [6] Miorandi D, et al., "Internet of things: Vision, applications and research challenges," *Ad Hoc Networks*, vol. 10, pp. 1497-1516, 2012.
- [7] Weber R.H, "Internet of Things–New security and privacy challenges," *Computer Law & Security Review*, vol. 26, pp. 23-30, 2010.
- [8] Welbourne E, et al., "Building the internet of things using RFID: the RFID ecosystem experience," *Internet Computing, IEEE*, vol. 13, pp. 48-55, 2009.
- [9] Yan Z, Zhang P and Vasilakos A, "A survey on trust management for Internet of Things," *Journal of Network and Computer Applications*, vol. 42, pp. 120-134, 2014.

AN EFFECTIVE REPORTING SOLUTIONS FOR VALUE ADDED SERVICES

R. Sivakkolundu¹ and Dr. V. Kavitha²

Research Scholar¹ and Associate Professor², Hindusthan College of Arts and Science, Coimbatore

ABSTRACT

Value added services are services beyond the standard services provided by any service industry, for services available at low cost or no cost, to promote their primary business. Value added services are supplied either in house by the mobile network operator or by a third party value added service provider. Value added services are provided via SMS and MMS, in case of a mobile network. VAS applications which send and receive these messages are connected to the mobile network using messaging gateways which connects to the SMSC/MMSC. Data Warehouse is a database specifically designed to store historic snapshots of various operational system data, normally in an aggregated form which is used by data analysts and other end users for analysing, reporting, tracking, and supporting strategic decisions. Business intelligence data required by VASP and mobile operators to efficiently manage and develop their service need to be generated from the messaging gateways. Since the VAS transactions handled by the messaging gateways are hundreds of millions in number, using a reporting tool to query the database and viewing the transactions will require considerable amount of time and efforts. This system provides an efficient way of getting the business intelligence data required by the VASP and mobile operators using a data ware house.

I. INTRODUCTION

In any business, it is crucial that information about the business and operations is available for decision making. In small companies, it may be easy to see the current status and direction of the business, but large corporates need tools to achieve the same level of clarity. In particular, large corporates need to build specialized systems to collect data about their operations, and to convert that data into information and finally to knowledge about the business [12]. The process of turning data into information and then into knowledge is generally known as business intelligence (BI). In the present day, an example of large business that needs data about their operation is the value added service provided using SMS and MMS by the mobile operators. A mobile operator may have a customer base of millions of subscribers. If an average subscriber uses the operator’s value added services tens of times in a month, the total number of monthly transactions served by the operator can reach hundreds of millions. In the case of the value added service provided by the mobile network operator, challenges are set for BI by the large transaction volumes that are related to serving the subscribers. A report about the weekly transaction volume of mobile messaging services over the last three months could cover hundreds of millions of transactions. Traditionally, reporting on this kind of transaction volumes has been done by summarizing the detailed transaction data before storing it for later analysis. However, the problem with data summarization is that it severely restricts the analysis. The only way to remove these restrictions is to store and make available transaction data at detailed, single transaction level with as many attributes as possible [5][10]. The foremost intention of this system is to design and implement a prototype of a data warehouse based reporting solution for the messaging gateway described in the problem statement. The new solution should be capable of storing large number (in the range of hundreds of millions) of detailed messaging transaction data, and yet enable efficient reporting on the data. The notion of detailed transaction data means that the data must not be summed up, but stored as individual transactions, and as many details as possible should be stored about individual transactions.

II. ATTRIBUTES OF THE TRANSACTION DOMAIN MODEL

In order to reach the target of defining the dimensions for the data mart being designed, Table 1 lists all attributes of the transaction domain model as well as classifies the attributes to dimension and fact attributes. The classification between dimension and fact attributes is done by considering the cardinality of the attribute – that is, the size of the set of possible attribute values is considered. Attributes having large cardinality are considered as fact attributes (also known as measures), and attributes with low cardinality are candidates for dimension attributes. The dimension attributes are grouped into dimensions by considering which of the attributes do have either a hierarchical relation with each other (clearly belong together) or are correlated with each other (weaker than hierarchical relation).

Table-1 : Classification of the transaction attributes

Attribute	Description	Cardinality	Dimension or Fact
Message ID	Messaging gateway transaction ID	Large (unique ID)	Fact
Type	The message type	Small	Message type dim.

	(SMS or MMS)		
Priority	Messaging gateway transaction priority	Small	Message type dim.
Direction	The message direction (MO or MT)	Small	Message type dim.
Message size	The message size in octets	Large	Fact
Status	The status of the message	Small	Status dimension
Status reason	Detail reason of the status	Small	Status dimension
Resend count	Number resends done for the message	Small	Status dimension
DR requested	Has a delivery report been requested?	Small (boolean)	Message type dim.
Ack received	Has an acknowledgement been received? Note, the value of this attribute is derived from the status attribute.	Small (boolean)	Status dimension
DR received	Has a delivery report been received? Note, the value of this attribute is derived from the status attribute.	Small (boolean)	Status dimension
RR received	Has a read-reply report been received? Note, the value of this attribute is derived from the status attribute.	Small (boolean)	Status dimension
Receive time	The message receive timestamp (in millisecond precision)	Large	Fact
Send time	The message send timestamp (in millisecond precision)	Large	Fact
Expiration time	Expiration time of the message. (in millisecond precision)	Large	Fact
Ack time	The acknowledgement receive timestamp (in millisecond precision)	Large	Fact
DR time	The delivery report receive timestamp (in millisecond precision)	Large	Fact
RR time	The read-reply report receive timestamp (in millisecond precision)	Large	Fact
Subscriber address	The address of the subscriber (Sender in MO / recipient in MT)	Large	Fact
VASP name	The name of the VASP	Small	VAS dimension
VAS name	The name of the VAS	Small	VAS dimension
Shortcode	The shortcode to which the message was sent to (MO) or from (MT)	Small	VAS dimension
Keyword	Keyword used in VAS identification (if multiple VASs share a shortcode)	Small	VAS dimension
Mobile network connection name	Name of the mobile network connection towards an SMSC or MMSC	Small	Network connection
Mobile network connection protocol	The name of the protocol used in the connection towards an SMSC or MMSC	Small	Network connection
Application connection name	The name of the network connection towards the VAS	Small	VAS dimension
Application connection protocol	The name of the protocol used in the connection towards the VAS	Small	VAS dimension
RR requested	Has a read-reply report been requested?	Small (boolean)	Message type dim.

As a result of the above classification, five dimensions are identified

- **VAS:** The VAS dimension provides information about the VAS application to which the message was sent (MO), or that sent the message (MT). The basis of the VAS dimension corresponds to the four objects (and their four attributes) related to VAS management: VAS, VASP, short code and keyword. These attributes are used by the messaging gateway in VAS identification. The VAS dimension is augmented with the information about the connection name and protocol between the VAS and the messaging gateway.
 - **Message type:** Like the name implies, the message type dimension describes aspects of the message type. This dimension includes a collection of low cardinality attributes that are actually independent of each other, but can all be considered to describe the message type (related by category).
 - **Message status:** The message status dimension gathers the attributes describing different aspects of the message status.
 - **Network connection:** This dimension describes the connection between the messaging gateway and the mobile network. At this point, the only attributes are the network connection name and the protocol. A separate dimension is created for the network connection because the attributes do not correlate with any other dimensions, and they are of too low cardinality to justify their storage in the fact table.
 - **Date:** The date dimension exists to make the complex date attributes easily available for querying, and to make the design independent of any particular database management system. The dimension contains various data attributes and their combinations.
1. **Identify the facts** - The last step is the identification of the facts. This step is about the identification of the measures. The previous step identified already the dimensions of the new data mart, and as a side-product, also the measures. Table 1 identifies these measures with the word “Fact” in the last column. To complete the design, the measures that shall be included to the final design are selected. Table 2 lists the measures identified in the previous section, as well as describes which of the measures are included to the fact table as measures.

Table-2 : Summary of the fact measures

Measure	Description	Include?	Basis for the decision
Message ID	Messaging gateway transaction ID	Yes	Needed for status updates by the ETL process
Subscriber Addr	The address of the subscriber (sender in MO / recipient in MT)	Yes	Relevant information for reporting (e.g. number of distinct subscribers)
Message size	The message size in octets	Yes	Relevant, additive measure
Receive time	The message receive timestamp (in millisecond precision)	Yes	Basis for second, minute, and hour granularity of reporting
Send time	The message send timestamp (in millisecond precision)	Yes	Allows analysis of the messaging gateway processing delay
Expiration time	Expiration time of the message. (in millisecond precision)	No	Mostly irrelevant information from reporting point of view (usually derived from receive time by adding a constant validity period)
Ack time	The acknowledgement receive timestamp (in millisecond precision)	Yes	Allows analysis of the SMSC/MMSC delay
DR time	The delivery report receive timestamp (in millisecond precision)	Yes	Indicates the message receive time at the final recipient; limited use, but included not to restrict analysis possibilities.
RR time	The read-reply report receive timestamp (in millisecond precision)	Yes	Indicates the time the recipient read the message (MMS only); limited use, but included not to restrict analysis possibilities.

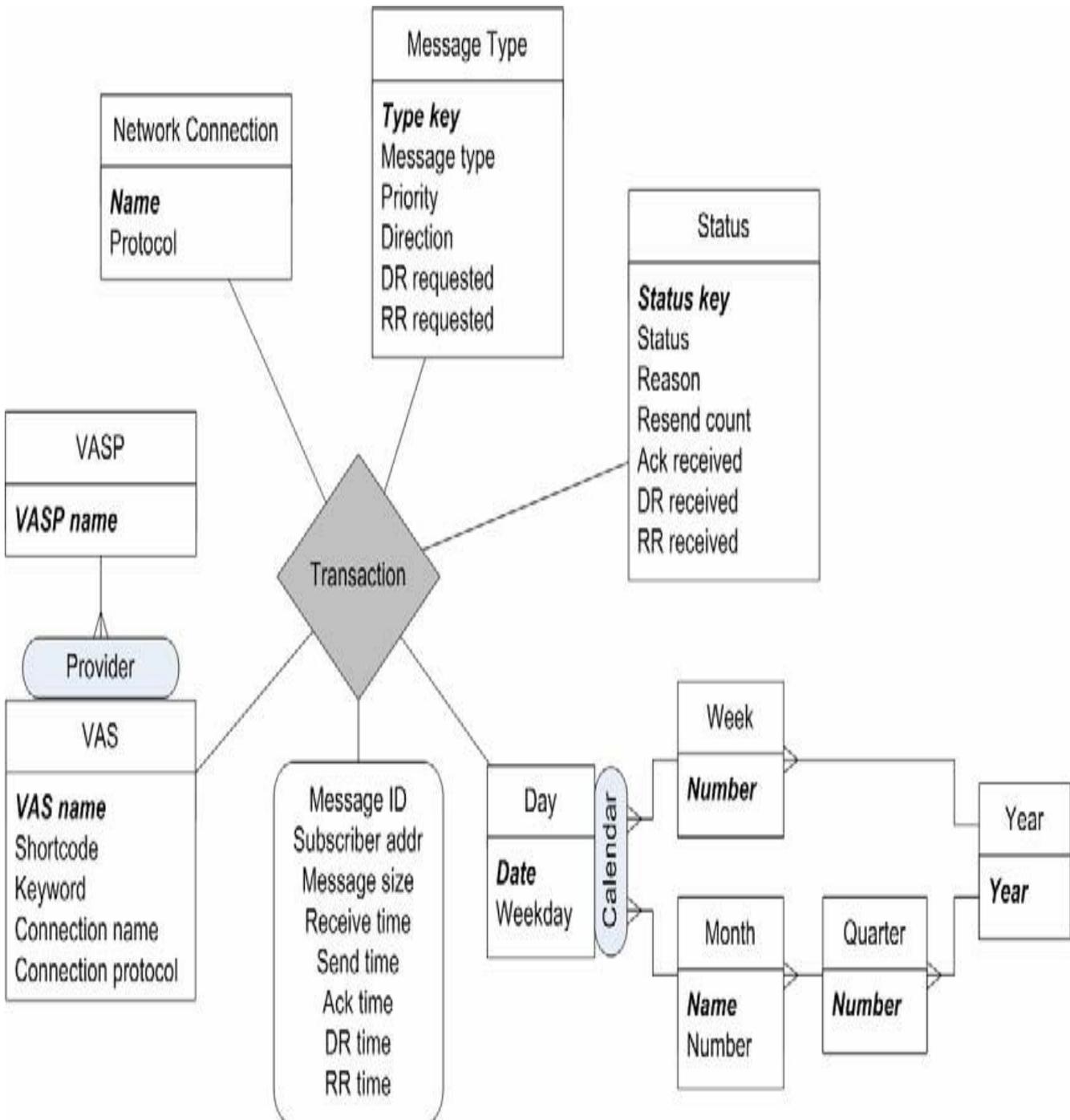


Figure-1: VAS DataMart

The fact table consists of the measures, and the dimension keys. The fact table matches the grain definition of the Mobile Messaging Data Mart – each fact table row identifies one transaction meaning a message delivered to a single recipient. Thus, the fact table rows are identified by the message ID and the subscriber address. The message ID is actually so called degenerate dimension: it contains just a unique ID that replaces the original message ID. This reduces the size of the fact table row compared to storing the original message ID string for each transaction. The smaller the fact table row size, the faster the database is to query [10]. The ETL process of the Mobile Messaging Data Mart is responsible for maintaining the mapping between the messaging gateway message IDs and the generated message ID numbers for the duration that the transactions may be updated. In practice this period is defined by the validity period of the messages. Figure 1 illustrates the VAS DataMart

The subscriber address is stored as a string as it can be either an MSISDN as specified by the E.164 numbering or an Internet email address. The subscriber address could also be converted to a degenerate dimension key to improve the performance on of the database, and hide the subscriber address information from the data mart users while still allowing to query, for example, how many distinct subscribers have used a particular VAS application monthly.

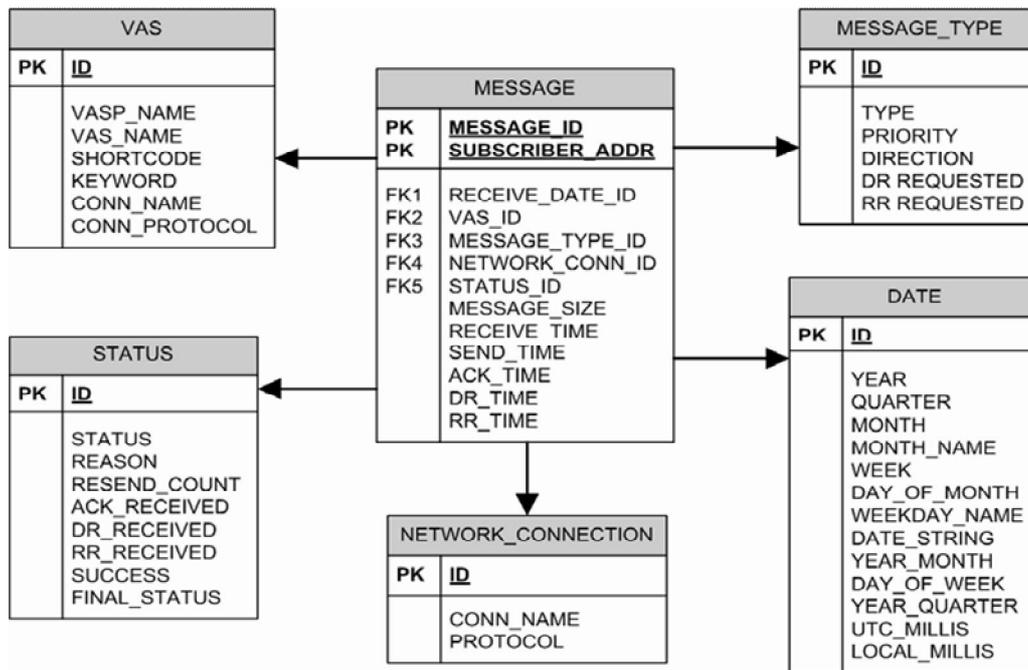


Figure-2: Logical Schema of the data storage of the new reporting solution

Figure 2 illustrates Logical Schema of the data storage of the new reporting solution. The last distinctive feature to mention is the use of surrogate keys in the dimension tables. In other words, the keys to the dimension tables are not natural keys, but generated and meaningless ID numbers. This has several benefits – the most important being performance (small keys, quick and easy to join for the database), maintenance and ease of use (easy to create joins in queries) [10].

ETL PROCESS AND METADATA

This section describes the logical design of the ETL and metadata components. The design is based on the conceptual design as well as on the logical design of the data storage component. The operation of the ETL process shall be based on the following steps:

1. Query the database of the existing reporting solution to get new transactions at the data granularity chosen for the new reporting solution.
2. Transform the data, one transaction at a time, to the format of the new data storage by resolving the dimension IDs for the new data (populating the dimension tables with new entries if corresponding entries do not exist already).
3. Clean the data by ensuring there are no duplicates, and that attributes that may have variations in their presentation are normalized to a uniform format.
4. Write the fact table data, one transaction at a time, to the fact table of the new data mart (with SQL inserts).
5. Identify transactions whose status has been updated after the last execution of the ETL process, and update the content of the new data mart accordingly (with SQL updates).

To support the above process the ETL process shall maintain a mapping between the message IDs of the messaging gateway and their corresponding numeric IDs in the new data mart. The unique numeric IDs shall be generated by using an ID generation mechanism provided by the chosen RDBMS. For each entry in the mapping, there shall be a data field indicating the day during which the transaction corresponding the mapping was received. This information is used by the ETL process to purge mappings that are not needed anymore. The validity of the mappings is configurable and shall be based on transaction validity period on the messaging gateway system.

The message ID mapping is also used to prevent the creation of duplicate copies of transaction data. While copying the transaction data, the ETL process checks for each message whether a mapping exists already. Hence, if a mapping exists, the ETL process can act accordingly and not write a new copy of a transaction to the data storage. Another data cleaning operation is the normalization of the subscriber and VAS addresses which may be either MSISDN or email address. The MSISDNs are converted to international format and the email addresses are converted to lower case letters. Figure 3 illustrates the logical design of the ETL process and the metadata components.

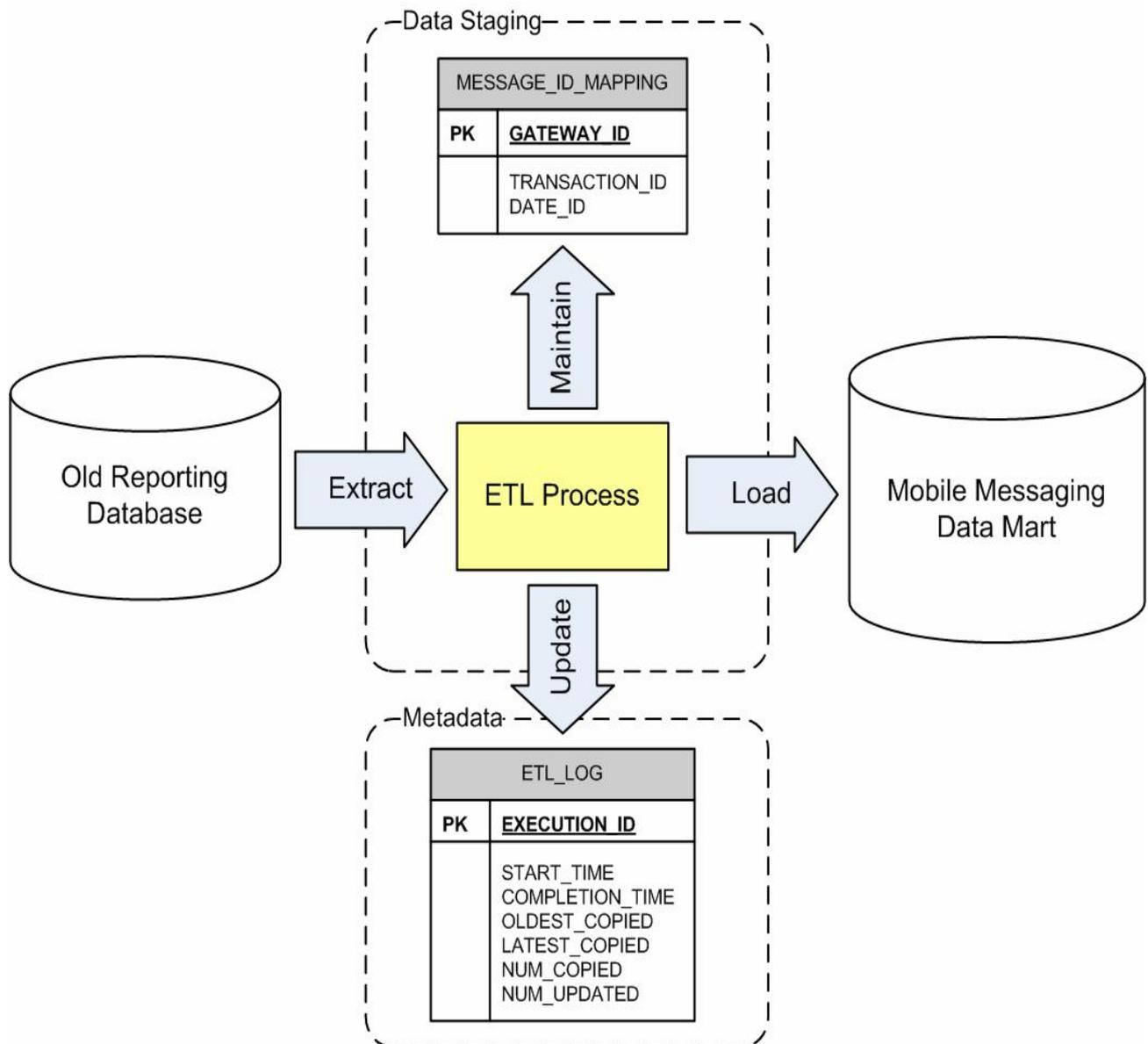


Figure-3: Logical design of the ETL process and the metadata components

III. CONCLUSION

Companies need to process and analyse an increasing amount of information about their business environment as well as their own operations in order to make informed decisions. This is the domain of business intelligence (BI), which aims to leverage organization’s internal and external information assets for making better business decisions [10]. Starting from the 1990s data warehousing has established its position in supporting corporations in their BI activities. Especially gains in processing power and the ever increasing data storage capacities have made the data warehousing technology more accessible to companies, and thus helped the data warehousing to develop to the point where it is today[12]. This System studied data warehousing on a general level as well as the existing data warehousing approaches and solutions. The main objective of the system is to design and implement a prototype of a new data warehouse based reporting solution for the messaging gateway of a mobile network operator. The new reporting solution is developed to replace an existing solution whose performance did not meet the performance requirements of the mobile network operator. The study on data warehousing revealed that there are two main architectural approaches to data warehousing. Most of the methods covered only a part of the complete design process that can be considered to start from requirements analysis, and to continue through conceptual and logical design all the way to physical design. Based on the study, dimensional modelling was selected as the architectural approach to be used as the basis of the architecture of the new reporting solution. The design of the new reporting solution was done by dividing the work to conceptual, logical and physical design phases. Both the study and the experience gained from the prototype implementation showed that the data warehousing paradigm provides a good framework for the design and implementation of an information system.

REFERENCES

- Breslin, M., 2004. *Data warehousing battle of the giants: Comparing the basics of the Kimball and Inmon models*. Business Intelligence Journal. Vol. 9:1, p. 6—20. Available at <http://www.bi-bestpractices.com/view-articles/4768>.
- Chaudhuri, S. & Dayal, U. 1997. *An Overview of Data Warehousing and OLAP Technology*. ACM SIGMOD Record. Vol. 26:1. 1997.
- Chaudhuri, S. & Dayal, U. & Venkatesh, G. 2001. *Database Technology for Decision Support Systems*. Computer. Vol. 32:12. December 2001.
- EMI, 2003. Short Message Service Centre 4.6, EMI – UCP Interface, Specification. CMG Wireless Data Solutions.
- Gardner, S., 1998. *Building the Data Warehouse*. *Communications of the ACM*, Vol. 41, No.9, 1998.
- Imhoff, C. & Galemno, N. & Jonathan, G. 2003. *Mastering Data Warehouse Design*. Indianapolis, Indiana, USA. Wiley Publishing Inc. 438 p. ISBN 0-471-32421-3.
- Inmon, W.H. 1999. *Data Mart Does Not Equal Data Warehouse*. DM Direct, November 1999. http://www.dmreview.com/article_sub.cfm?articleId=1675.
- Inmon, W.H. 2002. *Building the Data Warehouse*. 3rd ed. USA. John Wiley Sons Inc. 412 p. ISBN 0-471-08130-2.
- Jacobson, I. & Booch, G. & Rumbaugh, J., 1999. *The Unified Software Development Process*. USA. Addison-Wesley. 463 p. ISBN 0-201-57169-2.
- Kimball, R. & Ross, M. 2002. *The Data Warehouse Toolkit*. 2nd ed. USA. John Wiley & Sons Inc. 436 p. ISBN 0-471-20024-7.
- Peersman et al. 2000. *The Global System for Mobile Communications Short Message Service*. IEEE Personal Communications, June 2000.
- Ponniah, P. 2001. *Data Warehousing Fundamentals*. USA. John Wiley & Sons Inc. 516 p. ISBN 0-471-41254-6.
- Sen, A. & Sinha, A.P. 2005. *A Comparison of Data Warehousing Methodologies*.
- Silfvast, D. 2004. *A Web Services Interface for Managing Mobile Messaging Services*. Master's thesis. TKK, Helsinki University of Technology, Department of Computer Science and Engineering. Espoo.
- SMPP, 2003. *Short Message Peer to Peer Protocol Specification v5.0*. SMS Forum. 166 p. Available at <http://www.smsforum.net/>
- Ullman, J.D., Widom, J. 1997. *A First Course in Database Systems*. USA: Prentice-Hall Inc. 470 p. ISBN 0-13-861337-0.
- W3C, 2000. *Simple Object Access Protocol (SOAP) 1.1*. W3C Note, May 2000. Available at <http://www.w3.org/TR/2000/NOTE-SOAP-20000508/>.
- Wikipedia, 2007. *Online Computer Dictionary for Computer and Internet Terms and Definitions*. Available at <http://www.webopedia.com/>.

MANOEUVRING DATA MINING PERCEPTIONS FOR CRIME SCRUTINIZING THROUGH KNOWLEDGE REPRESENTATION USING AND-OR GRAPH

Bala Sneha. M, Saraswathi. S and Jayasri. GDepartment of Computer Science & Application, Sri Krishna Arts and Science College, Coimbatore

ABSTRACT

Crime rate is being escalated at an alarming rate. Scrutinizing these crimes and providing the inhibitor ways are the prominent goals of the proposed work. Crime scrutinizing is an emerging field through which the user could identify the crimes and scrutinize the patterns associated with them. Here the impacts of data mining techniques were utilized to enhance the scrutiny so that scrutinizing becomes a lot easier. With the help of knowledge representation large unorganized sets of crime information are collected and interpreted with formulating facts that resembles the generality which exists in the application. Predictions are done in a facile way by generating assertions with the derived fact. This effort results in formulating the precautionary measures to be taken in order to control crime.

Keywords: Crime rate, datamining, knowledge representation, And or graph

1.1. INTRODUCTION

The term crime is derived from the Latin word **cernere** which means “to judge”. It is also derived from **crime** which means judgment or an offense committed. In today’s world scenario crime rate is escalated at an enormous rate. Both the crime and the criminals are becoming advanced day by day. Hence it is very important to scrutinize the crime and provide inhibitor ways. Crime prone areas could be easily found using data mining strategies [1].

Initially, a clear idea about the terminologies associated with crime is analyzed. Crime is defined as an act which is punishable by law. Accused or the suspect is a person who is charged of committing a crime but not yet legally confirmed. Victim is the person who is affected by the crime. He is considered as the target of the crime. In most of the cases victim is the person who files a complaint about the crime. Accomplice is a person who helps the accused to commit a crime. Allegation is the claim of the illegal activity which one has made without any proof. Cause of action is the reason why the crime has occurred. Evidence or witness is the proof to find whether the crime has occurred or not.

Considering the crime rate is an important factor for the people residing in a particular area. Their safety mainly depends upon the protection which is given against the crime [2]. This acts as a main factor for taking important decision as to move in a particular area or not. Whether travelling through a particular area is safe or not? Another important effect which crime leaves on people is that they are put forth to a fear of occurrence of crime. It also indirectly affects the economy of a region.

To maintain the safety of the residents, the crime should be controlled by analyzing them. But the main disadvantage associated with them is the crime data sets are huge. They are being increased day after another. In order to analyze these data we would need a lot of time and effort [3]. Retrieving the information from them is very interesting but a challenging one. In order to reduce the challenges, the concept of data mining is used. Data mining is defined as the process of extracting valuable information from the existing large sets of data. It is now considered as one of the emerging fields across the globe. It is also known as knowledge discovery process.

Crime analysis plays a major role in finding solution for the crime by identifying the patterns and trends associated with them [4]. Crime analysis or criminology is one of the applications of data mining. The main aim of the criminology is to suppress the activity of crime. Initially the crime data set are collected and organized by eliminating the unessential attributes. Then we interpret them to identify the common patterns of the crime incidents and the frequent occurrences of crime. Finally the areas where high crime rate are suspected is found and awareness is given to the people of those areas about the crime scene and how to inhibitor them.

1.2. STUDY OF BASESYSTEM’S ARCHITECTONICS

To build an intelligent predictive system a depth analysis of the problem plays an initial role. Then challenges associated for analyzing a particular problem are taken into account. Considering those problems in mind, conclusion is provided by using knowledge engineering process. Knowledge engineering is one of the technologies used by many of the experts for finding the simple solutions for big complicated problems. It employs many machine learning concepts for deriving equitable results.

Initially the problem for which solutions must be derived is identified. Here the main problem is to scrutinize the crime and provide the inhibitor ways. The problem i.e. the crime is taken as **Cis** considered. Crime is

composed of many minute details which aid us in finding the solution in a better way. These minute details are known as **facts** .this provides the conclusion that our problem crime is completely dependent on the facts

C →d (F)

The facts which are available are generally inconsistent in nature. Some data which are available are even noisy in nature. Concluding decisions from the available facts or data becomes a tedious process. Hence we are employing the technique of data pre- processing to remove all the noisy data .Analyzing becomes a lot easier by converting the facts into **information I**.

F → I

But all the data which is present is not necessarily important for finding the solution S. Now we need to analyze what data is necessary for arriving at a solution.The important attributes are selected by a selecting process. Selecting process starts by questioning ourselves whether this data is useful or not? After selecting the required data we make sure whether the available data is accurate or not. Maintaining accuracy is very important becomes our problem crime is a sensitive one. The information available is refined and processed (RI).

RI → I

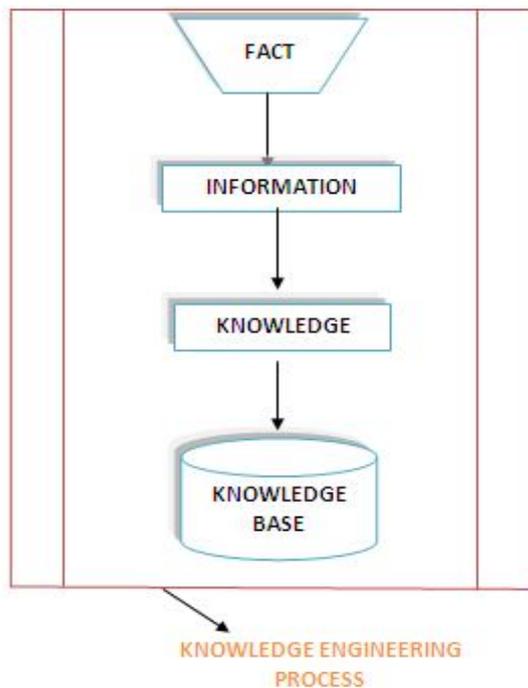


Fig-1.1: Knowledge engineering process

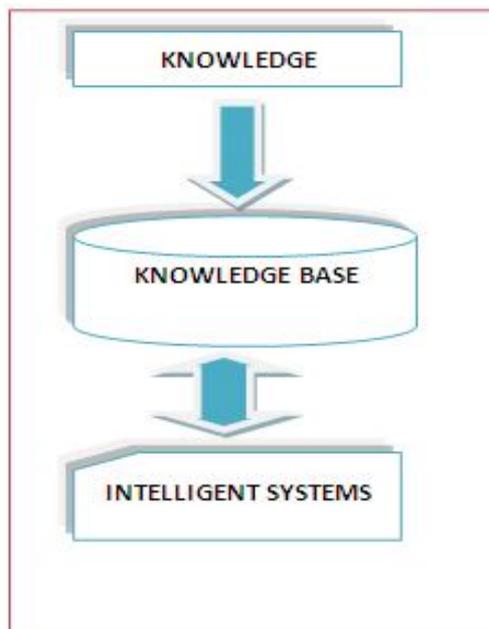


Fig-1.2: Knowledge retrieval process

Investigation of a fact is very important to arrive at a particular required information which is now available is unstructured, unorganized and even misplaced. Making analyses becomes a tedious one. Hence the information is organized and structured. Then they are interpreted in a specific way using many algorithms and data mining perceptions to gain knowledge from this processed information. The knowledge which is gained benefits and supports us in recommending a proper definite **solution**.

S → RI

The knowledge which is available or the knowledge which is gained from this processed information must be stored safely. Only then they could be used for further purposes. This knowledge and the data are stored in a warehouse called as knowledge base. This knowledge base could be accessed at any time for the retrieval of information. There are many techniques available for investigating a particular fact or information. The knowledge which is acquired could be represented in many ways. Some of them are Propositional calculus, First Order Predicate logic, Probabilistic logic, Fuzzy Logic, Decision trees, Data Mining Techniques, etc., These methods which are specified are used to represent the knowledge in a formal way.

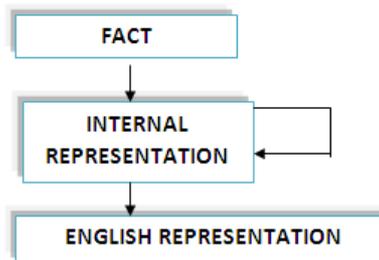


Fig-1.3: Fact representation

The natural language in which the facts are present is very difficult for machines to understand. First the facts are converted into a representation, then that internal representation is converted to English format. It is called as symbol level because the representation is in the form of symbol

Another method which is available is fact acquisition. It is one of the formal procedures used by many researchers and technicians. This process is also known as fact finding or fact analysis. It also helps us in finding answers and arrives to a conclusion for a particular question from the available facts. There are many perceptions available for fact analyzing. One of the most prominent methods used for fact analysis is forward mapping and backward mapping. Mapping is the process of creating mapping elements between the available data models. It is one of the most primary steps for data integration and data transformation.

In the technique of forward mapping a fact which is related to crime is found. This initial fact is given as input. Then the fact is converted to a logical representation. In backward mapping the formal logical representation is converted to normal English representation form. The answer which is given by the backward representation is called as the final fact because it acts as a solution. Some of the facts which are analyzed are

FORWARD MAPPING

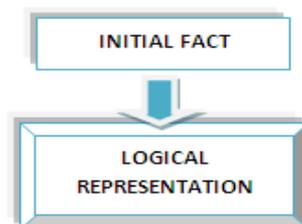


Fig-1.4: Forward mapping

BACKWARD MAPPING

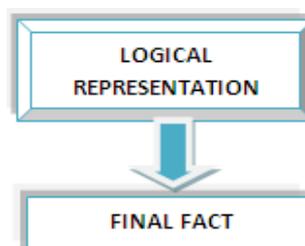


Fig-1.5: backward mapping

FACT	REPRESENTATION
Emotional torture is one of the crime	English representation
Crime (emotionaltorture) Crime(x)	Logical representation
X:punishable(crime)	Formal logical representation All crimes are punishable
Punishable(emotional torture)	Deductive mechanism
Emotional torture is punishable by law	English representation

Table-1.1: Fact Representation1

FACT	REPRESENTATION
Every crime has an id	English representation
Crime(crime id) Crime(x)	Logical representation
X:Unique(crime)	Formal logical representation Crimes which occur are unique
Unique(x) Unique (crime id)	Deductive mechanism
Crime id is unique in nature	English representation (background representation)

Table-1.2: Fact Representation2

FACT	REPRESENTATION
Burglary is one of the crime	English representation
Crime(burglary) Crime(x)	Logical representation
X:punishable(crime)	Formal logical representation All crimes are punishable
Punishable(burglary)	Deductive mechanism
Burglary is punishable by law	English representation (background representation)

Table-1.3: Fact Representation3

FACT	REPRESENTATION
Delhi is one of the city in India	English representation
City(Delhi) city(x)	Logical representation
X:Increased crime(city)	Formal logical representation Crime rate is being increased cities
Increased crime(city) Increased crime(Delhi)	Deductive mechanism
Crime rate is increased in Delhi	English representation (background representation)

Table-1.4: Fact Representation4

1.4. TYPES OF CRIME ANALYSIS

Analyzing crime could be done in many ways. But mainly there are three types of approaches used for finding solutions in a facile way. They are tactical, strategic and administrative approaches.

I. TACTICAL APPROACH

The main aim of tactical approach is to find the trends and patterns associated with the specific current crime. This approach deals with crimes which are currently occurring. This aids in providing clue to the police for their further investigation. They give a very quick solution. It also aids in identifying the hot spots where maximum crime occurs. Hence extra care would be given for those areas. [5] The patterns of the crime are found by including the fields like suspect, victim, evidence, time and area. With this type of analysis the crime which has repeated occurrences are found.

II. ADMINISTRATIVE APPROACH

This type of crime analysis is done mainly for administrative reasons especially for police departments, detective agencies and for government agencies. It makes use of many statistics and data for performing the analysis. It is one of the cost benefit analysis. They are used for supplying useful information about the crime scenes. This is a valid approach which police utilizes for planning and proceeding accordingly. It is mainly used for enforcing laws in a certain community.

III. STRATEGIC APPROACH

This type of analysis provides a strategic way of providing solutions to the occurrence of crime. With the help of strategic approach police could perform their duties in an efficient and a facile way. These type of approaches deals with a very longrange of facts and makes comparisons based on it. This approach help us in finding the crime pattern based on the long term of crime occurrence. The Crime pattern and crime trend would help us to determine the increase or the decrease of crime rate in a particular region or a community.

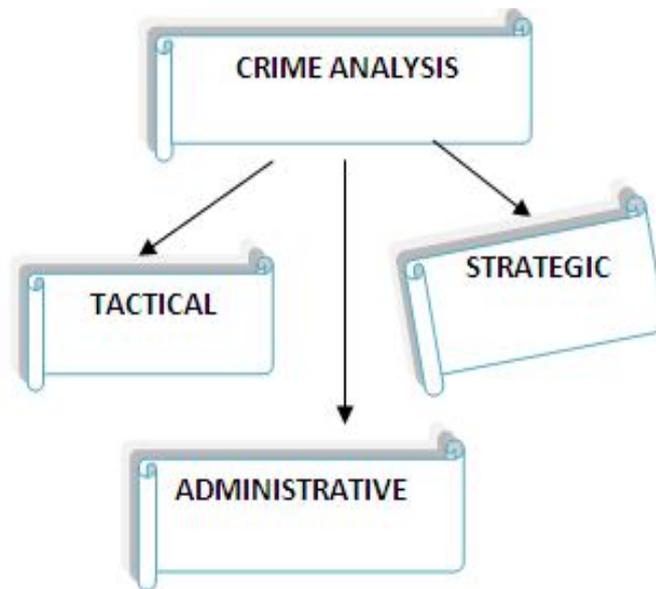


Fig-1.6: Types of crime analysis

1.5. Manipulating Knowledge

To generate a proper solution or to predict a result it becomes very important to manipulate knowledge. It helps us in making decisions and take actions according to the results produced. This is one of the key factors for making important decisions For example: If the crime rate > 75 there are high chances for the occurrences of the crime. Hence proper awareness and protection must be given in those areas. But one thing to note is that a correct decision could be taken by manipulating the information in a correct proper way. There are many techniques available for manipulating knowledge. Choosing the correct method plays a prominent role.

AND OR GRAPHS

One of the most important methods for manipulating knowledge is by using and or graphs. It is very useful to represent a solution to a specific problem. It provides a solution to a problem by decomposing the problem into smaller parts of problem where each problem is solved separately

Operation	Representation
And - Or	$\wedge - \vee$

Fig-1.7: Or graph representation

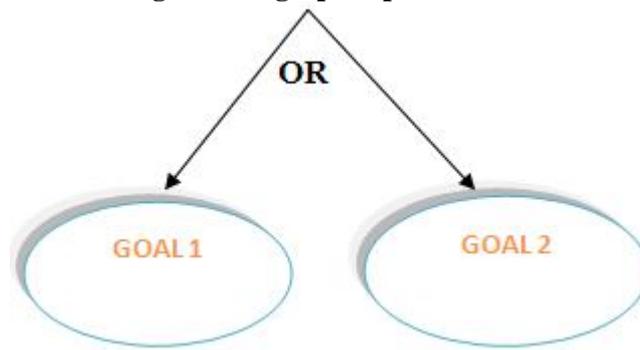
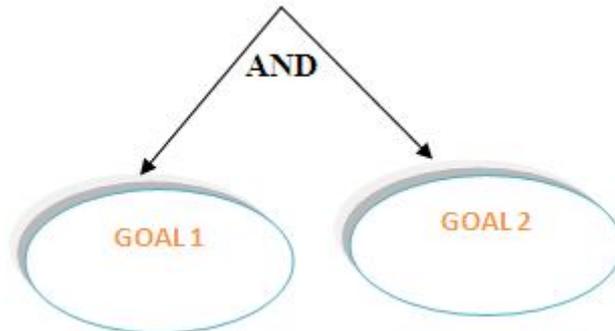


Fig-1.8: and graph representation



PROBLEM

The following statement below mentioned is the main problem of the proposed work. When the crime rate >65 and ≤ 95 there is a high chance for the occurrence of the crime in the particular region. To control the crime rate by providing awareness are the prominent goals.

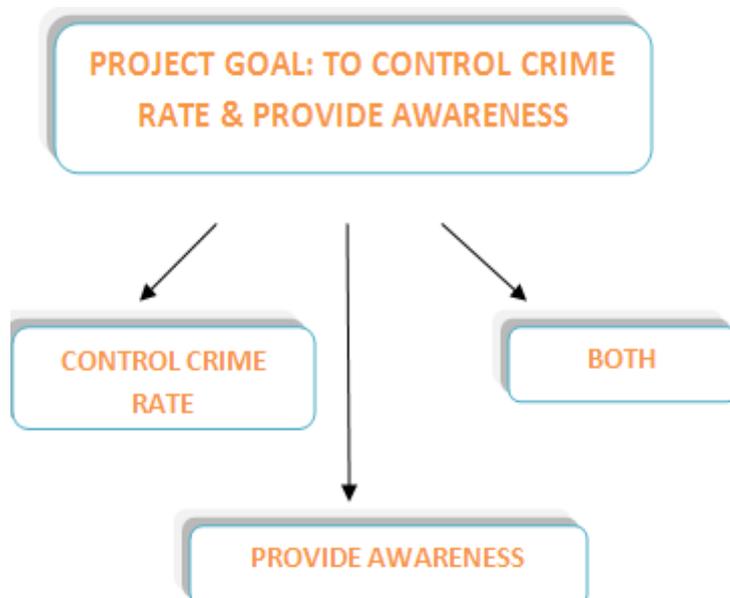
ASSERTION

Control crime rate and provide awareness

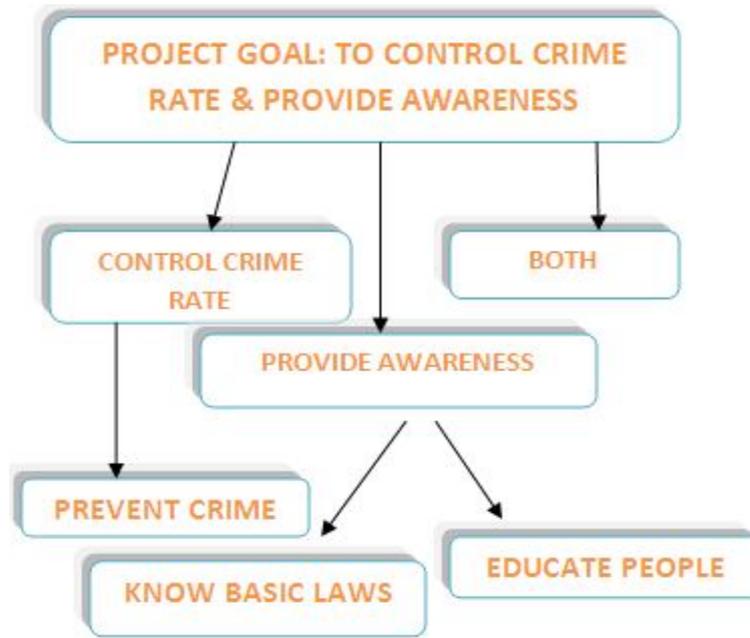
STEP-1: CREATE A NODE AT THE TOP



STEP-2: IT SPLITS INTO THREE NODES



STEP-3: IT IS AGAIN SPLITTED AS



CASE 1: It is only true, in order to control crime rate we need to prevent the occurrences of the crime

CASE 2: It is true that in order to provide awareness we need to educate people about the type of crime and make them understand the basic laws that are available for a citizen.

CASE 3: It is a mixture

1.5. TYPES OF KNOWLEDGE REPRESENTATION

In general knowledge means gaining awareness or understanding of a particular concept. The study of knowledge is called as epistemology. There are many ways available for representing knowledge. The following are the approaches for a good knowledge representation system.

1.5.1 RELATIONAL KNOWLEDGE

One of the simplest ways to store the facts or information is through the relational knowledge concept. Here the facts about the objects are present in sequential columns.

Crime Rate	Chances of Occurrence
≤ 25	very low
≤ 50	low
$50 < \leq 75$	moderate
$75 < \leq 95$	high

Table-1.6: Relational knowledge example

1.5.2. INHERITABLE KNOWLEDGE: It consists of objects and attributes that inherit the new objects from the existing objects. It makes use of property inheritance where data is organized as a hierarchy of class.

1.5.3. INFERENCE KNOWLEDGE

Here the knowledge is represented through formal logic. It uses a set of protocols for deriving new facts

1.6 CONCLUSION

In the proposed work data sets were analyzed and the crucial attributes that have more contribution are identified and organized. Through pre-processing steps the ignorable attributes are eliminated. The facts are derived from the real data. The extracted general knowledge were evaluated and interpreted through And or graph that results in formulating assertions. These assertions will enhance to prevent and control crime.

REFERENCES

1. Devan M.S., ShijuSathyadevan, Crime analysis and prediction using data mining. In First International Conference on network and computing
2. EmmanuelAhishakiye crime.prediction using decision tree. In International Journal of computer and Information technology.

3. SujaRadhatitle A surveyto analyze crime using data mining techniques. , International Journal of Pharmacy and Technology.
4. Dr.ZakairaSulimanZubi, AymanAltaherMahmmud.Crime data analysis using data mining techniques to improve crime prevention. International Journal of computers
5. Jessicab.LeBlanc and Christopher W.Bruce.definitionand types of crime analysis.In International association of crime analysis.
6. ShyamVaranNath. Crime Pattern Detection Using Data Mining.In Oracle corporation
7. TusharSonawanev, ShirinShaikh, ShaistaShaikh, Rahul Shinde, AsifSayyad. Crime Pattern Analysis, Visualization And Prediction Using Data Mining.in IJARIE-ISSN(O)-2395-4396
8. Malathi. A, Dr. S. SanthoshBaboo. An Enhanced Algorithm to Predict a Future Crime using Data Mining.In International Journal of Computer,Applications (0975 – 8887) Volume 21– No.1, May 2011
9. RasoulKiani ,SiamakMarvdasht, Amin Keshavarzi. Analysis and Prediction of Crimes by Clustering and Classification.International Journal of Advanced Research in Artificial Intelligence.
10. Dr.M.Sreedevi[1] , A.HarshaVardhan Reddy[2] , Ch.VenakataSai Krishna Reddy[3] Review on Crime Analysis and Prediction Using Data Mining Techniques International Journal of Innovative Research in Science, Engineering and Technology.

A STUDY ON VARIOUS LOAD BALANCING ALGORITHM IN CLOUD COMPUTING ENVIRONMENT

V. Malathi¹ and Dr. V. Kavitha²

Assistant Professor¹ and Associate Professor², Hindusthan College of Arts and Science, Coimbatore

ABSTRACT

Cloud computing is emerging as a new paradigm for manipulating, configuring, and accessing large scale distributed computing applications over the network. Cloud users may access the resources through computer notebook, smart phone or other devices, so they can store data on servers and can access data through Internet. Load Balancing is one of the most significant concepts in distributed environments. The load on the cloud is increasing tremendously with the development of new applications. Load balancing aims at high user satisfaction and usage of resource ratio by guaranteeing a proficient and reasonable allocation of each computing resource. Load is a measure of the amount of work that a computation system performs which can be classified as CPU load, network load, memory capacity and storage capacity. There are numerous difficulties in load balancing techniques such as security, fault tolerance etc which are prevalent in modern cloud computing environments. Proper load balancing aids in implementing fail-over, enabling scalability, over-provisioning, minimizing resource consumption and avoiding bottlenecks etc. In this paper, we explore the diverse types of algorithms that are proposed by a number of researchers to solve the problem of load balancing in cloud computing.

Keywords: Cloud computing, Load balancing, Load balancing algorithm

1. INTRODUCTION

The term "cloud" originates from the world of telecommunications when providers began using virtual private network services for data communications. Cloud computing is an on demand service in which shared resources, information, software and other devices are provided according to the client’s requirement at specific time. Cloud computing is a virtualization of cloud program through internet connection, in which there is no need of installing application on everywhere. In cloud computing applications are provided and managed by the cloud provider. It offers the facility of pay per use to their customers.

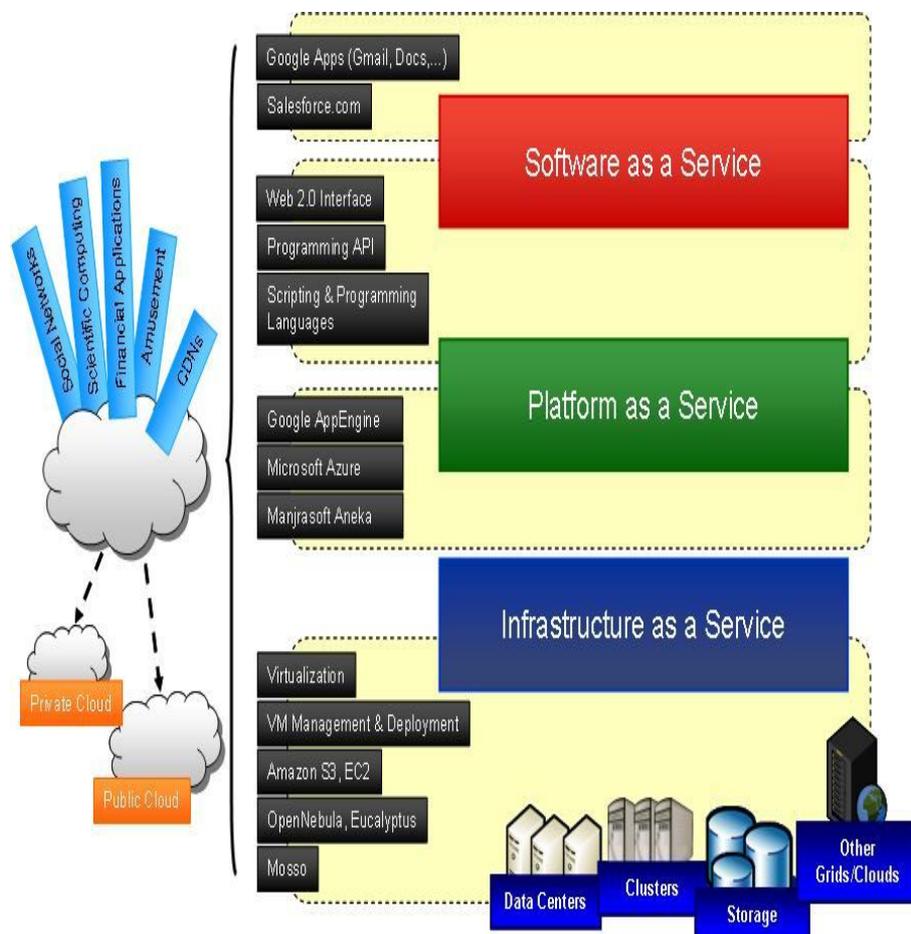


Figure-1: Cloud Computing Architecture

1.1 DEPLOYMENT MODEL

There are four types of cloud Deployment Model:

1.1.1 Private Cloud

A private cloud is owned by a single organization. Private clouds enable an organization to use of centralizing access to IT resources by different parts, locations, or departments of the organization.

1.1.2 Public Cloud

A public cloud is a publicly accessible cloud environment owned by a third-party cloud provider. The cloud provider is responsible for the creation and on-going maintenance of the public cloud and its IT resources.

1.1.3 Community Cloud

A community cloud is similar to a public cloud except that its access is limited to a specific community of cloud consumers. The community cloud may be jointly owned by the community members or by a third-party cloud provider that provisions a public cloud with limited access.

1.1.4 Hybrid Cloud

Hybrid cloud is a combination of two or more models, private cloud, public cloud or community cloud. Though these models maintain their separate entities they are amalgamated through a standard technology that enables the portability of data and applications.

1.2 CLOUD SERVICE MODELS

Cloud service models can be broadly defined in three categories

1.2.1 SaaS (Software as a Service)

SaaS is a software service provided over the internet and requires no prior installation. These services can be availed from any part of the world at a minimal per month fee.

1.2.2 PaaS (Platform as a Service)

PaaS runs on physical servers, database software and web servers. All these are basically known as platforms, and cloud computing firms provide platforms as a service allowing you to use the service without having to build it from the scratch.

1.2.3 IaaS (Infrastructure as a Service)

IaaS on demand allows organizations to make use of operating systems and associated software without having to pay for hefty license fees. IaaS includes infrastructure such as servers, private networks, disk drives, long term storage solutions, email servers, domain name servers etc.

2. LOAD BALANCING IN A CLOUD COMPUTING ENVIRONMENT

Load Balancing is the method that allows us to create a proper balance of the amount of work that is being performed on various pieces of device or hardware equipment. Usually, what happens is that loads of the devices are balanced amongst different servers or cloud server between CPUs and hard drives. Load Balancing was introduced for better the speed and performance of every single device and to keep saving individual devices from hitting their threshold by dropping down their performance. Cloud computing can have either static or dynamic environment based upon how developer configures the cloud demanded by the cloud supplier.

2.1 Static Algorithm

In static algorithm the traffic is separated evenly among the servers. This algorithm requires a previous knowledge of system resources, so that the choice of shifting of the load does not depend on the present state of system. The entire traffic is equally divided amongst the servers. This algorithm needs an in-depth knowledge of server resources for better performance of the processors which is determined at the beginning of the implementation. There is one major drawback of Static Load Balancing Algorithm that is load balancing tasks work only after they are created, it could not be implemented to other devices for load balancing.

2.2 Dynamic Algorithm

Dynamic algorithms are decision concerning load balancing based upon the current state of the system i.e. any prior knowledge about the system is not required. It is a complex algorithm, but they can provide better performance than static algorithm and fault tolerance. This Algorithm first searches the lightest server in the entire network and gives it the preference for load balancing. It requires real time communication with the network that can help in increasing the traffic of the system. The specialty of Dynamic Algorithm is to make load transfer decisions to the actual current system state.

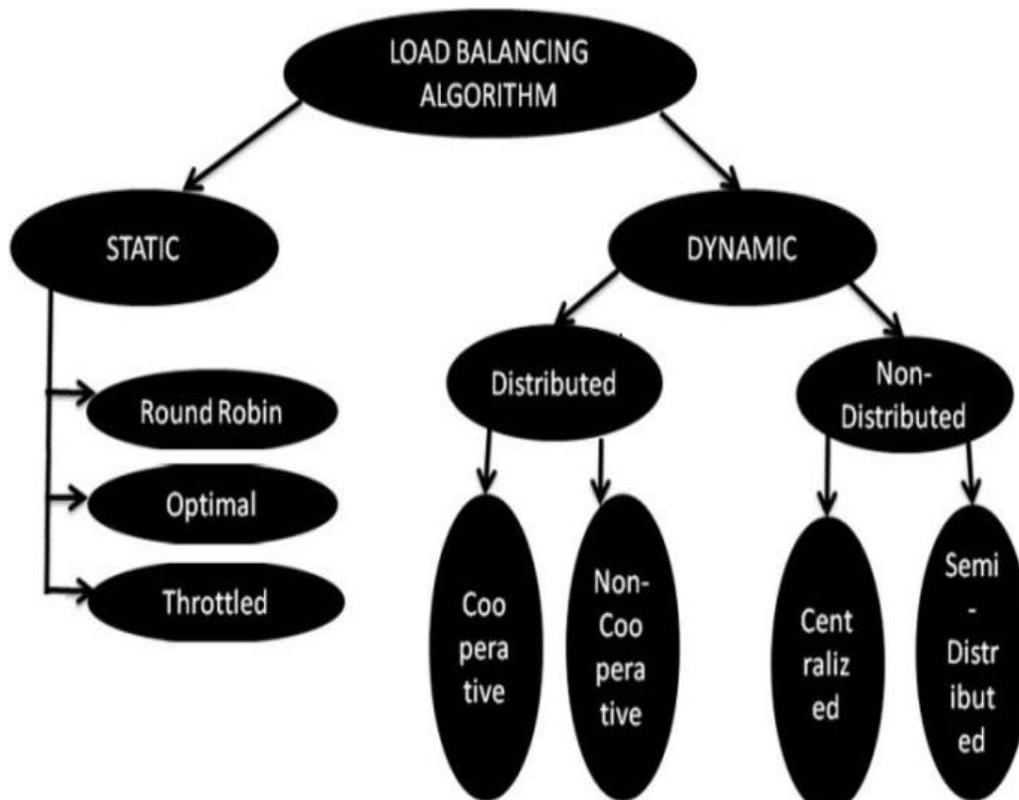


Figure-2: Classification-of-load-balancing-algorithms-in-cloud

3. RELATED WORK TO LOAD BALANCING ALGORITHMS

The following load balancing algorithms are currently prevalent in clouds

3.1 Round Robin Load Balancing

It is the static load balancing algorithm which uses the round robin scheme for allocating job. The round robin mechanism follows a time slice while processing the data. Each process is going to execute in the time slice and then switch to other process and follow on ring manner. The process is going to occur in round robin until the all processes complete their task such that a balance technique is implemented in order to balance the process in a group. All the processes have different loading time. Therefore, some nodes get heavily loaded and some nodes get lightly loaded because the running time of any process is not known in advance. This limitation is overcome in the weighted round-robin algorithm. In the weighted round-robin algorithm there are equal assignment of weight, each node receive some traffic.

3.2 Shortest Job Scheduling Algorithm

In this algorithm shortest executable job is selected first. The approach follows to perform the complete execution of short jobs to utilize the resources in completion of heavy jobs. Shortest job had an advantage that the waiting time for the processes is less which makes it a powerful approach.

3.3 Min-Min Load Balancing

This is static load balancing algorithm so the parameters related to the job are known in advance. Min-Min algorithm the cloud manager identifies the execution and completion time of the unassigned tasks waiting in a queue. The job with minimum completion time is selected. Finally, the selected node and the selected job are mapped. The ready time of the node is updated. This process is repeated until all unassigned jobs are assigned. This algorithm performs better when the numbers of jobs having small execution time is more than the jobs having large execution time. The advantage of this algorithm is that job with the smallest execution time is executed. The drawback of this algorithm is that some jobs may experience starvation.

3.4 Max-Min Load Balancing

Maximum to Minimum Load Balancing Algorithm is almost same as the Min-Min algorithm where the maximum value is consider to execute first. Here maximum value is selected after finding out minimum completion time of jobs. After that, as per the maximum period, the task is scheduled on the machine. The execution time of all the tasks are updated and the assigned task is removed from the list. Then the ready time of the node is updated by adding the execution time of the assigned task. A maximum time taken process is shifted one by one.

3.5 Throttled Load balancer

This load balancing technique ensures that only as per-defined number of internet cloudlets are allocated to a single virtual machine at any point of time. If more groups are presents in the data center than then number of available virtual machine than some of the requests have to be queued until the next virtual machine is available.

3.6 Two-phase (OLB + LBMM) load balancing algorithm

It is proposed by S.-c. Wang et al. it is combination of Opportunistic Load Balancing (OLB) and Load Balance Min-Min (LBMM) scheduling algorithms to utilize better executing efficiency and maintain the load balancing of the system. Working principle of OLB scheduling algorithm keeps every node in working state to achieve the goal of load balance. On the other hand LBMM scheduling algorithm is used for minimizing the execution time of the tasks on node which reduce of overall completion time. Combining these two algorithms help achieving proper utilization of all resources and enhances the work efficiency in the network of multiple processor.

3.7 Ant Colony based Algorithm

The ant colony optimization to search an optimal path between the source of food and colony of ant on the basis of their behavior. The Ants while seeking a path from their colony in search of food secrete a chemical called pheromone on the ground thus leaving a trail for other ants to follow the path. This approach aims efficient distribution of work load among the node. When request is initialized the ant starts movement towards the source of food from the head node. Ants keep records the every node they visits ant record their data for future decision making. Ant deposits the pheromones during their movement for other ants to select next node. The intensity of pheromones can vary on the bases of certain factors like distance of food, quality of food etc. When the job gets successful the pheromones is updated. Each ant build their own individual result set and it is later on built into a complete solution. The ant continuously updates a single result set rather than updating their own result set. By the ant pheromones trials, the solution set is continuously updated

3.8 HBB-LB Algorithm

This algorithm has similar principle in balance the work of the virtual machine. The HBB algorithm calculates the virtual machine work load then it decides whether it is overloaded, light weighted or balanced. The high priority of the task is off from the overload virtual machine and tasks are waiting for the light weight virtual machine. These tasks are known as scout bee in the next step. Honey Bee Behaviour inspired Load Balancing technique reduces the response time of virtual machine and also reduces the waiting time of task.

3.9 Opportunistic Load Balancing (OLB)

This is the static load balancing algorithm and it does not consider the current workload of the virtual machine. It attempts to keep each node busy. It never considers the current workload of each system. This algorithm schedules the new task in randomly available virtual machine without checking workload of that machine. It provides load balance schedule without good result. The task will process in slow in manner because it does not calculate the current execution time of the node.

4. CONCLUSION

Cloud computing provides everything to the user as a service over network. The major issues of cloud computing is Load Balancing. Overloading of a system may lead to poor performance which can make the technology unsuccessful, for the efficient utilization of resources, the efficient load balancing algorithm is required. In this paper, we discussed the already proposed various load balancing schemes by various researchers, each having some pros and cons. The static load balancing scheme provide easiest simulation and monitoring of environment but fail to model heterogeneous nature of cloud. The dynamic load balancing algorithm are difficult to simulate but are best suited in heterogeneous environment of cloud computing. The key function of load balancing is to satisfy the customer requirement by distributing load dynamically along with the nodes as well as to make highest resource utilization by reassigning the total load to individual node.

REFERENCES

1. A. Bhadani and S. Chaudhary, "performance evaluation of web servers using central load balancing policy over virtual machine on cloud", proceedings of third Annual ACM.
2. Che-Lun Hung¹, Hsiao-hsi Wang² and Yu-Chen Hu², —Efficient Load Balancing Algorithm for Cloud Computing Network. IEEE Vol. 9, pp: 70-78, 2012
3. Dharmesh Kashyap, Jaydeep Viradiya "A Survey of Various Load Balancing Algorithms In Cloud Computing" International Journal Of Scientific & Technology Research, Volume 3, Issue 11, November 2014, (ISSN 2277-8616).

4. Karanpreet Kaur, Ashima Narang, Kuldeep Kaur, "Load Balancing Techniques of Cloud Computing", International Journal of Mathematics and Computer Research, April 2013.
5. Kundan Pagar, Sachin Patil, "A Survey on Load Balancing Techniques in Cloud Computing", International Journal of Science and Research (IJSR), Volume 3 Issue 12, December 2014.
6. Nusrat Pasha, Dr. Amit Agarwal and Dr.Ravi Rastogi,"Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4, Issue 5, May 2014.
7. Poonam Devi,Mr. Trilok "GabImplementation of cloud computing by using short job scheduling" International Journal of Advanced Research in Computer Science and Software Engineering.
8. Pooja Samal, Pranati Mishra, |Analysis of variants in Round Robin Algorithms for load balancing in Cloud Computing| (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (3), 2013, 416-419.
9. R. Rajeshkannan, M. Aramudhan, "Comparative Study of Load Balancing Algorithms in Cloud Computing Environment", Indian Journal of Science and Technology, Volume 9, May 2016.
10. Shahrzad Aslanzadeh, Venkatesh Mahadevan, Christopher Mcdermid," Availability and Load Balancing in Cloud Computing" International Conference on Computer and Software Modeling IPCSIT vol.14 IACSIT Press, Singapore 2011.
11. S. Wang, K. Van, W. Liao, and S. Wang, "Towards a Load Balancing in a Three-level Cloud Computing Network", Proceedings of the 3rd IEEE International Conference on Computer Science and Information Technology (ICC SIT), Chengdu, China, pp.108-113, September 2010.
12. T. Kokilavani and Dr. D.I. George Amalarethinam, "Load Balanced Min-Min Algorithm for Static MetaTask Scheduling in Grid Computing"International Journal of Computer Applications Volume 20–No.2, pp.0975-8887, April 2011.
13. Upendra Bhoi, Purvi N. Ramanuj, "Enhanced Max-min Task Scheduling Algorithm in Cloud Computing" International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 4, April 2013.

**AN EFFICIENT WATERSHED TRANSFORMATION TECHNIQUE TO DETECT CANCER CELLS
IN HUMAN ORGANS****P. Jayapriya**

Assistant Professor, Department of Computer Science, NGM College, Pollachi

ABSTRACT

Cancer is a significant health problem throughout the world. It is very important to detect such types of cancer at an earlier stage than the later stage where the treatment becomes unsuccessful. Early detection helps surgeons to provide necessary therapeutic measures which also benefit the patients. In this paper, a technique is proposed to detect cancer cells present in organs like Brain, Breast, Lung, Liver by MRI, 3D, Image Segmentation, Watershed and Morphological Operators. A novel mathematical morphological watershed algorithm is proposed to preserve these edge details as well as prominent ones to identify tumors in dental radiographs. Applying watershed on images leads to segmentation hence it is preprocessed.

Keywords: Image processing, Watershed Segmentation, Brain cancer, Breast Cancer, Lung cancer, Liver cancer, Mammogram Detection.

INTRODUCTION

Image processing techniques is widely use in the medical image currently. Ultrasound and X-ray medical images play an important role in the detection of lung, liver, brain, breast tumors. Different methods were employed to detect the cancerous cell in image pre processing such as Gabor filter, image segmentation using watershed segmentation and feature extraction by using MATLAB. This paper is an attempt for comparison of cancer detection in various organs which mentioned above. This paper deals with tumors that start within the brain, breast, lung, liver using morphological watershed algorithm on Magnetic Reasoning Images (MRI) to identify, locate and segment the tumor. The basic idea in this paper is to design software and use the proposed system to help the physician reading and classifying these types of tumors into benign or malignant. Segmentation method based on morphological watershed transforms to extract watershed lines from a topographic representation of the input image. The diagnosis of cancer is achieved by algorithms based on morphological operation and segmentation using watershed transformation. The basic idea in this paper is to design software and use the proposed system to help the physician reading and classifying the type of tumors into benign or malignant.

MEDICAL ASPECTS

Cancer disease begins in the cells of the human body, which is generated by abnormal division of those cells. There are two types of cancer, benign tumors are not cancerous and malignant tumors are cancerous [7]. Image segmentation by mathematical morphology is a methodology based upon the notions of watershed transformation. Watershed transform is a powerful tool for image segmentation [2].

IMAGE SEGMENTATION

The goal of image segmentation is to cluster pixels into salient image regions, i.e., regions corresponding to individual surfaces, objects, or natural parts of objects. Using this process most of the image analyzing task can be done subsequently. In specific, many methods depend highly on the segmentation result for image description and recognition. But in this paper a study on Watershed segmentation and thresholding method used is analyzed. Obtained image after segmentation from thresholding had much significance like fast processing speed less storage space and simply by manipulation of 256 levels of grey level image. Thresholding is the most dominant tool for image segmentation by replacing original pixel values by black pixel values (converts grey image into binary image). Thresholding selects a threshold value T and it assigns two levels to the image that is above value and below value for original threshold value.

THE WATERSHED TRANSFORMS

The watershed transform proposed by Vicent and Soille [9] is a well-known segmentation technique, which is based on immersion simulation, and allows the generation of an initial image partition into regions. It is based on visualizing a gray level image as topographic surface with three dimensions: two special coordinates versus intensity. In such a topographic surface view, the gray level of a pixel is interpreted as its altitude. Suppose a water source is placed in each regional minimum (also called 'catchment basins') and the entire topography structure is flooded from below. When water from two sources (regional minima) are about to meet, a dam is constructed to prevent the merging as in Fig.1. The flooding and dam construction process continues until only the dams are visible from above. These dams (watershed lines or 'watersheds') effectively segment the image

into regions. Image segmentation is an essential process for most subsequent image analysis process. In particular, many of the existing techniques for image description and recognition, image visualization, and object based image compression highly depend on the segmentation results [8].The steps shown in figure Fig.2.

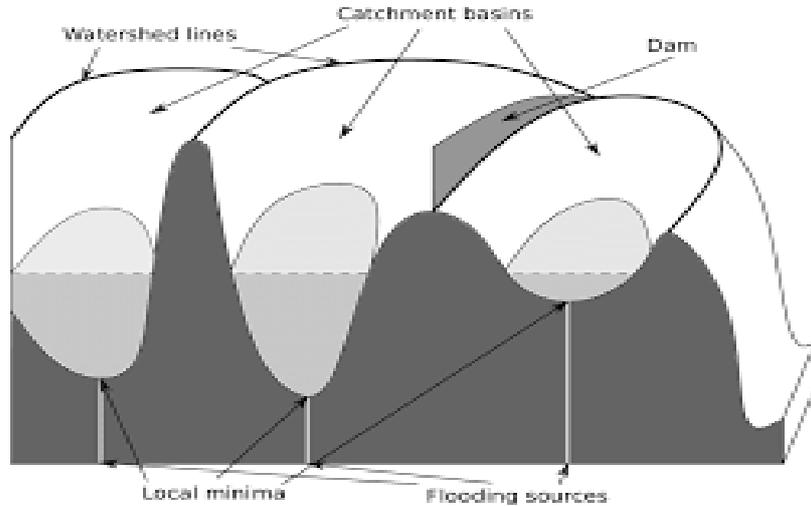


Fig-1

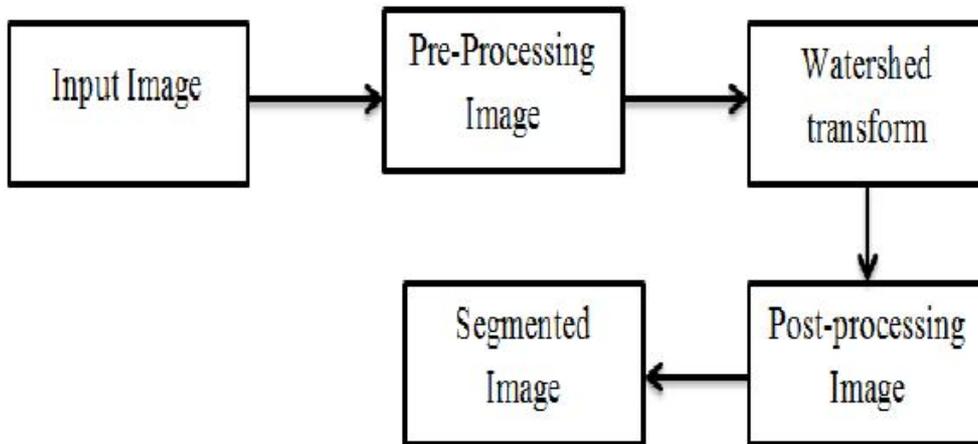


Fig-2: Block Diagram for Applying Watershed Method

BRAIN CANCER

One of the main causes for increasing mortality among children and adults is brain tumor. It has been concluded from the research of most of the developed countries that number of peoples suffering and dying from brain tumors has been increased to 300 per year during past few decades. The National Brain Tumor Foundation (NBTF) for research in United States estimates the death of 13000 patients while 29,000 undergo primary brain tumor diagnosis. This high mortality rate of brain tumor greatly increases the importance of Brain Tumor detection. Hence the MRI, 3D, Image Segmentation, Watershed & Morphological Operators are the fundamental problem of Tumor Detection [6]. Input image and its Segmentation image is as shown in Fig.3 and Fig 3.1

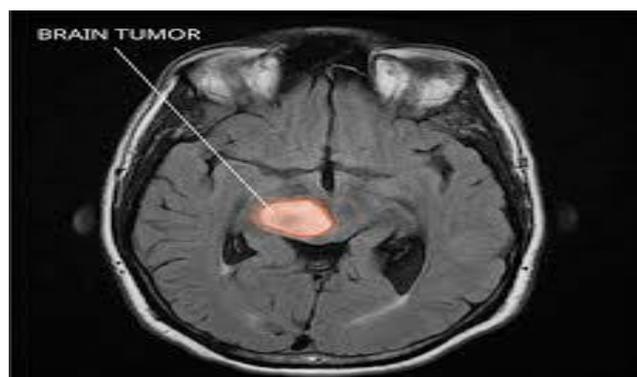


Fig-3

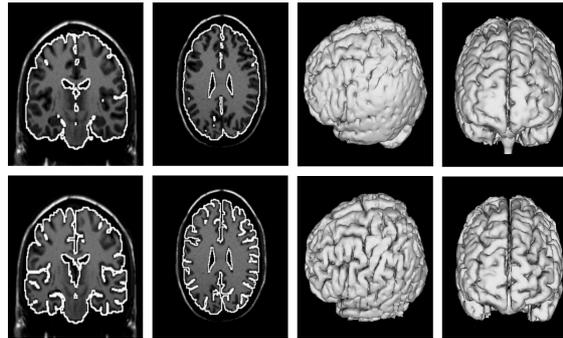


Fig-3.1

BREAST CANCER

Breast cancer is a malignant tumor. The malignant tissue begins to grow in the breast. The symptoms of breast cancer include breast mass, change in shape and dimension of breast, differences in the color of breast skin, breast aches and gene changes. The early detection and diagnosis of breast cancer is the key to decrease rate and to provide prompt. In recent year, a variety of imaging techniques used to study breast tumor such as: magnetic resonance imaging (MRI), Computed Tomography (CT), Ultrasound, X-ray Ultrasound and X-ray mammogram are the most widely used techniques, because their ability to produce resolution images of normal pathological tissues. Mammogram is a low dose x-ray procedure for the visualization of internal structure of breast. Mammography has been proven to be the most reliable method and it is the key screening tool for the early detection of breast cancer. Ultrasound imaging is non- invasive, real time, low cost, and convenient for patients [4]. Segmentation is an important way to extract information from medical image. In Segmentation the inputs are images and, outputs are the attributes extracted from those images. Segmentation divides image into its constituent regions or objects as shown in Fig.4. Segmentation based on morphological operation, and watershed transform applied to grey level images is a fast, robust and widely used in image processing and analysis, but it suffers from over-segmentation [5].

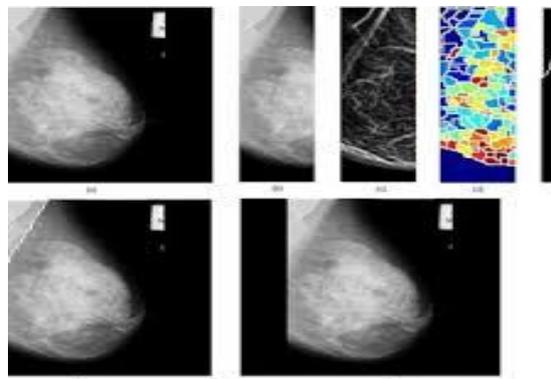


Fig-4

LUNG CANCER

Lung cancer cell can be identified using image pre processing, feature extraction of cancer image and finally the classification process. Lung cancer images are collected from a private hospital Chennai. Various stages are involved in the cell identification of lung cancer. Lung cancer cell identification processes involves convolution filter for smoothening the cancer cell images. To enhance the image contrast and color, then the nucleases were segmented by using thresholding process in the images. It's a simple image processing techniques on lung cancer detection system which can extract the feature from the MRI lung image [1]. Fig.5 shows the lung image and Fig 5.1 shows segmentation image.

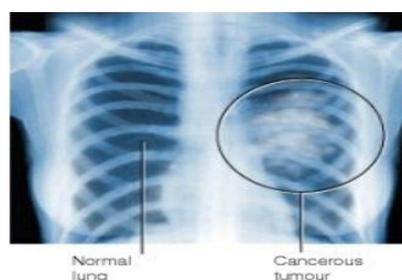


Fig-5



Fig-5.1

LIVER CANCER

Segmentation method of liver cancer CT image based on the watershed algorithm by the methods of edge detection, the watershed algorithm and region merging, and has obtained good results. MRI scanned liver image Fig.6 and segmentation applied image as shown in Fig.6.1. This method can better solve the over-segmentation of the traditional algorithm, and get closed, continuous, more accurate lesion area contour curve. However, in this method, the thresholds in region merging is the experience values gained through many tests, because the selection of the threshold directly affects the segmentation results, how to select the threshold adaptively will become a study emphasis in research environment[3].

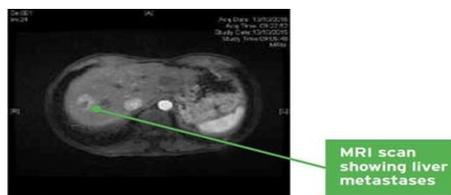


Fig-6

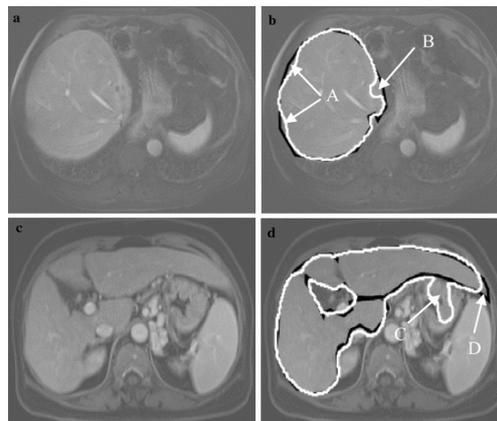


Fig-6.1

CONCLUSION

In this paper, use image processing techniques in the medical image. The proposed approach algorithm helps in detection of cancer in brain, breast, lung and liver cancer using medical image processing such as MRI, Ultrasound and x-ray. Since medical image are complex, requirement preprocessing aids in gray scale image use some operation. Then applied the watershed transform. The areas in the image are highlighted and that could be under analyzed to detect cancerous and non-cancerous. The proposed algorithm has been tested on standard digital image techniques. Through the work the value of neighbor adopted has been reached a good results ratio. In future, this work can be detection the tumors using any of techniques such as combined the watershed transform algorithm and clustering algorithm, Neural network system, Fuzzy logic and also control problem of over-segmentation using watershed transform.

REFERENCES

1. Al Tarawneh MS. "Lung cancer detection using image processing techniques", Leonardo Electronic Journal of Practices and Technologies. 2012; (20):147–58.
2. Beucher.S, "The Watershed Transformation Applied to Image segmentation", Center De Morphologic Mathematique, Ecole Desmines Paris,77305 Fontainebleau Cedex.
3. Jianhua Liu, Zhongyi Wang, Rui Zhang "Liver Cancer CT Image Segmentation Methods based on Watershed Algorithm" IEEE pp.4244-4507 ,2009.

4. Kadhim D.A , "Development Algorithm Computer Program of Digital Mammograms Segmentation for Detection of Masses Breast Using Marker-Controlled Watershed in LAB Environment",First Scientific Conference the Collage of Education for Pure Sciences, 2012,Kerbala University.
5. Ponraj.D.N , Jenifer.M.E, Poongodi.P , and Manoharan.J.S , "A Survey On The Preprocessing Techniques Of Mammogram For The Detection Of Breast Cancer", Journal of Emerging Trends in Computing and Information Sciences ,Vol. 2, No. 12, 2011,ISSN 2079-8407.
6. Roopali R Ladhake,Dr. Siddharth A Ladhake, "Brain Tumor Detection using Morphological and watershed Operators" IJIEM, Vol. 3, No. 3, pp. 383-387, 2014.
7. Sharma.J, and Sharma.S, "Mammogram Image Segmentation Using Watershed", International Journal of Information Technology and Knowledge Management, Vol. 4,No. 2, 2011,pp.423-425.
8. Tulsani.H, Saxena.S and Yaday.N, "Segmentation Using Morphological Watershed Transformation For Counting Blood Cells", International Journal of Computer Applications and Information Technology ,Vol. 2, Issue III,2013,ISSN:2278-7720.
9. Vincent.L, Soille.P, "Watersheds in Digital Spaces: An Efficient Algorithm based on Immersion Simulation", IEEE Trans. PAMI, Vol. 13, No. 6, pp. 583-598, 1991.

INVENTIVE INQUIRY ON HEART DISEASE REPRESENTATION USING AND - OR GRAPH

Arulmozhi S, Saraswathi. S and Hemapriya. G

Department of Computer Science & Applications, Sri Krishna Arts and Science College, Coimbatore

ABSTRACT

Data mining is the widely accepted and implemented technique for analyzing huge set of database in order to derive an new information. Data mining techniques are almost used in every field in order to reduce complexity. In this proposed effort the issues involved in healthcare applications are analyzed. To have effective knowledge representation, raw facts are derived from healthcare dataset. Out of 76 raw attributes, only 14 of them are having high dominant impact in predicting the heart disease. Using the preprocessing technique 8 crucial attributes are derived and its associations where analyzed. Five different categories of problem facts were identified and tabulated with required representations. The represented knowledge is interpreted to derive optimal assertions using AND-OR-Graphs. The assertions enhance the clarity in understanding the prediction of heart disease under different age group of people.

Keywords: Facts, Knowledge, Assertions

1. INTRODUCTION

Data extraction technique plays a vital role in day to day life due to the increase in the object behavior. In order to derive or extract new meaningful information from the huge amount of data, the data mining techniques are used. There are many techniques like Classification, Clustering, Association rules etc. It plays a major role in Healthcare industry for multiple purposes like data entry, storage, maintenance, and analysis. It is used to retrieve the hidden data from the huge storage of data. There are many data mining applications that is helpful in identifying the diseases with the help of symptoms to identify the causes and giving the best treatment to the people. The major impact of Data mining is low cost and improved quality it is helpful in reducing the data loss and fraud mechanisms. [1]Heart disease is the major cause for death occurring now-a-days. Heart problems can cause many other disease, disorders, stroke etc. With the help of chosen attributes many clustering algorithms are implemented and results are found out which one gives the highest accuracy in prediction of heart disease. [2][3]There are many factors that causes heart diseases such as family history, obesity, smoking, BP, cholesterol etc. with the help of 14 attributes by using the attribute selection data reduction techniques they reduced the irrelevant data in the data set and using the classification algorithms they found that Naïve Bayes and ANN results in highest accuracy.[4] Knowledge discovery in database [KDD] of Data mining provides a user friendly nature to the data by improving the service quality. It helps in analyzing and working with millions of data. [5] [6] The aim of data mining is to extract hidden data from large amount of data. In Data mining classification algorithm uses two techniques to categorize the attributes. It uses CRISP methodology and other algorithms. [8][9]Data mining applications in healthcare are very useful in policy making precautionary measures of diseases and death in the hospitals. It says that for every 34second one person died because of heart disease there are many other diseases such as cardiovascular, cardiomyopathy, coronary heart diseases. The main attributes for prediction of heart diseases are age, sex, obesity, BP, Chest pain, Cholesterol, ECG results etc.

2. KNOWLEDGE REPRESENTATION

Knowledge representation lies in different forms like English, logical, formal, deductive etc. for understanding the real world complex applications.

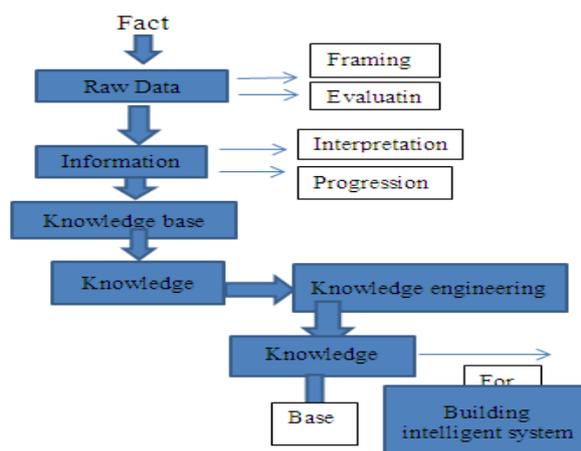


Fig-1: Schematics of Knowledge Engineering

The crucial data are filtered from the Healthcare dataset. The selected data are organized and evaluated based on the objective of the proposed work. The facts were identified and represented using different techniques through interpretation and progression, the extracted initial information are recorded in knowledge base. Then the knowledge engineering process were applied to extract the hidden knowledge. That shows high values in serving as a base to build intelligent heart disease predicting system.

Various techniques are available to represent the knowledge such as,

- Propositional calculus.
- 1st Order Predicate logic.
- Probabilistic logic.
- Fuzzy Logic.
- Decision trees.
- Data Mining Techniques etc.,

Since Natural language is burdensome to have machine interpretation. The healthcare dataset were analyzed to frame “Facts and Symbols”.

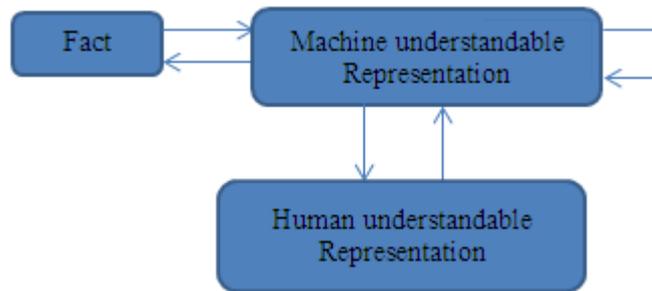


Fig-2: English Understanding

Facts which are in human readable format will not be understood by the machine. The user need to represent the data in (0's and 1's) that makes machine to read and act. So, the facts are converted into internal representation i.e. machine language(0's and 1's) it can interpret easily and it will decrease burdensome of the machine.

2.1 Forward Mapping and Backward Mapping

Mapping is a process of converting natural language into formal language, machine can easily interpret by this conversion. It is used to map the entities by using this method. The user will give the inputs in the form of human readable representation (i.e English representation) that will be converted into machine language (i.e formal language). By this conversion machine can be able to read easily and interpret the data. Another method is used to convert the machine language into human readable format(i.e English language).

3. DERIVING KNOWLEDGE FROM FACTS

The facts are used to derive initial knowledge about the domain which provides clarity in understanding the problem. The formal known language will support and then converted into logical language. Then the next stage of conversion will takes place to represent it in machine readable representation for further interpretation.

Fact	Representation
X Peoples are affected by heart disease	English Representation
People(x)	Logical Representation
X.peoples (x) -> are affected by heart disease(x)	Formal Logical Representation
Has affected (x)	Deductive Mechanism
Peoples are affected by heart disease	English Representation (background representation)

Table-1: Fact denoting normal category

Fact	Representation
People (X) age Below(30) has affected by heart disease	English Representation
People(x)	Logical Representation
X:People's(x) below (30) affected by heart disease	Formal Logical Representation
Below(30)has affected (people)	Deductive Mechanism
People who are below (30) affected by heart disease	English Representation (background representation)

Table-2: Fact representing below 30 categories

Fact	Representation
People (X) age above(30+) has affected by heart disease	English Representation
People(x)	Logical Representation
X:People's(x) above (30+) affected by heart disease	Formal Logical Representation
Above(30+)has affected (people)	Deductive Mechanism
People who are above (40+) affected by heart disease	English Representation (background representation)

Table-3: Fact representing above 30 category

Fact	Representation
Children (X) age Below(15) has affected by heart disease	English Representation
Children(x)	Logical Representation
X:Children (x) below (15) affected by heart disease	Formal Logical Representation
Below(15)has affected (children)	Deductive Mechanism
Children who are below (15) affected by heart disease	English Representation (background representation)

Table-4: Fact representing below 15 categories

Fact	Representation
2 Million (X) people has affected by heart disease in India	English Representation
People(x)	Logical Representation
X:People 2 Million (x) affected by heart disease(x)	Formal Logical Representation
2Million has affected (people)	Deductive Mechanism
2 Million people are below affected by heart disease	English Representation (background representation)

Table-5: Fact representing in India

4. PROPERTIES OF GOOD KNOWLEDGE BASE SYSTEM

The approach which is used to encode knowledge in intelligent system knowledge base is Knowledge representation. It is dependency of one variable from other variables. To represent a knowledge system in intelligent system knowledge base four properties need to be considered.

1. Representational adequacy

By analyzing the datasets and attributes, it is possible to find out which attribute is more important and functional. Dependency from one to other variables is analyzed. Also possible to eliminate the non-important attributes.

2. Inferential adequacy

The skill to represent an knowledge in a new form from extracting it from older form.

3. Inferential efficiency

The skill of extracting new facts from the old one.

4. Acquisitional knowledge

The skill to gain a new knowledge easily from automatic mechanisms are possible. Admirably the agent should not able to control its own knowledge acquisition, but direct assertion of information by a “Knowledge Engineer” would be suitable.

5. Knowledge Acquisition

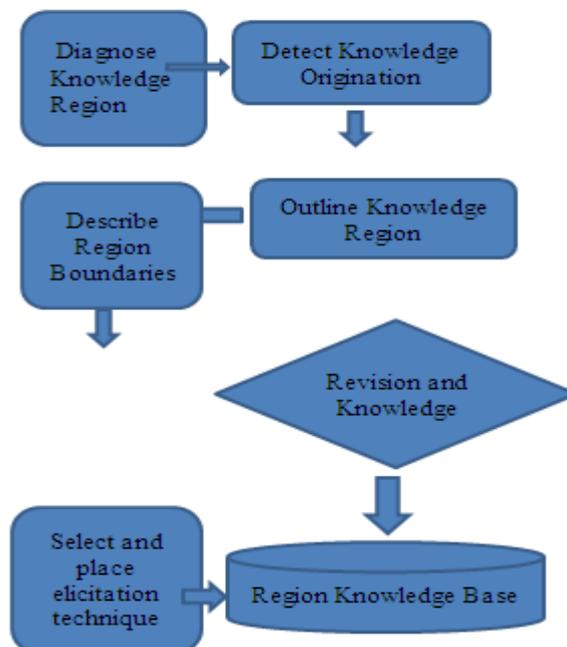


Fig-3: Knowledge Acquisition

6. Knowledge Manipulation

Knowledge manipulation are denoted as, if independent then corollary, it has two primary proper prediction and description.

For example:

If age < 18 to >=80 there are many chances of getting affected by heart disease both for men and women. The factors of affecting heart disease are BMI, Cholesterol, smoking, physical inactivity etc.

7. AND OR GRAPHS

AND – OR Graph shows structured representation and knowledge, this graph supports problem disintegration like divide and over throw, finding solution to small parts and ploy to upper node.

AND operation can be represented as (^) – It relies True if both the operands are true.

OR operation can be represented as (v) - It relies True if any one of the operands are true.

7.1 Combination of AND node & OR node

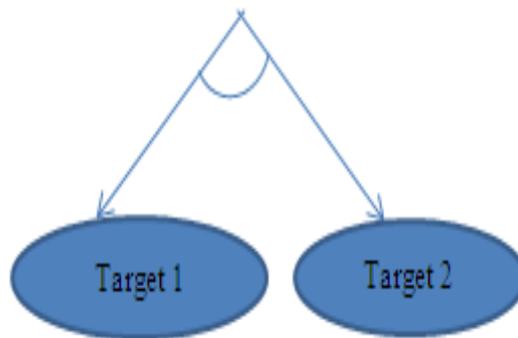


Fig-4: OR Node

The outcome of OR operation results in “Positive” when any one of the targets is achieved.

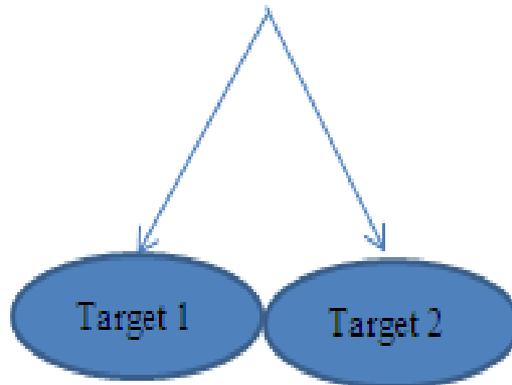


Fig-5: AND Node

The outcome of AND operation results in “Positive” when both targets are achieved.

Problem-1: Here the below mentioned problems are described with people’s age and advice given by the doctor. These two are the some of the causes and how to prevent it from heart disease.

1. People with the age >40 <=70 are affected by heart disease. To control this assertions are formulated with the criteria or having doctor advised them to control sugar and BP level, do daily walking.

Assertion: Possibility exist is to have physical activity (walking, exercise etc) controlling measures.

Step-1

Doctors advice :X To control Sugar and BP

Step-2: It splits the nodes into 3 sub nodes.

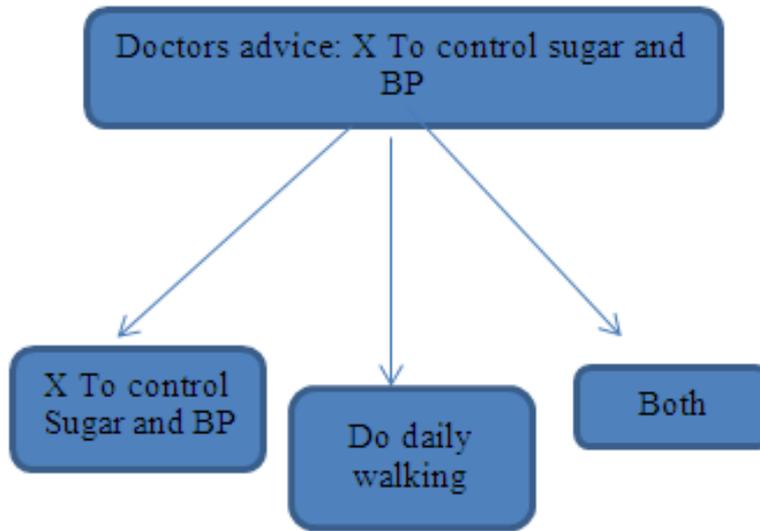


Fig-6: Doctors advice

Step-3: Again the nodes are splitted up into

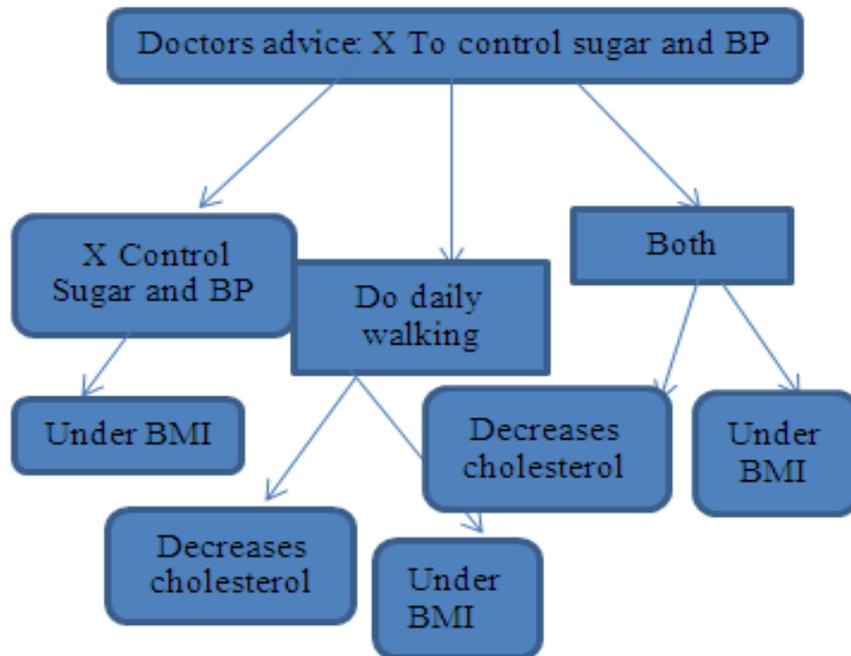


Fig-7: Problem using AND OR graph

Case-1: It is only “TRUE” that X to control sugar and BP “if we have proper healthy diet that will be splitted up under “X control sugar and BP”.

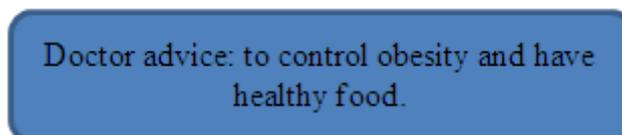
Case-2: If daily walking decreases cholesterol and weight is under BMI then it is only true “X is both”.

Case-3: “X is mixture”

Problem-2: Peoples with age $<10 \geq 30$ are affected by heart disease. Some of the reasons are due to family history, bad habits, Unhealthy foods, obesity.

Assertion: Possibility exist is to have Healthy food.

Step-1



Step-2

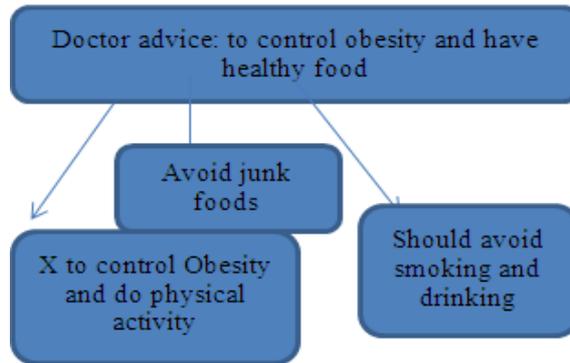


Fig-8: Doctors advice

Step-3: Nodes are split into 4 sub nodes

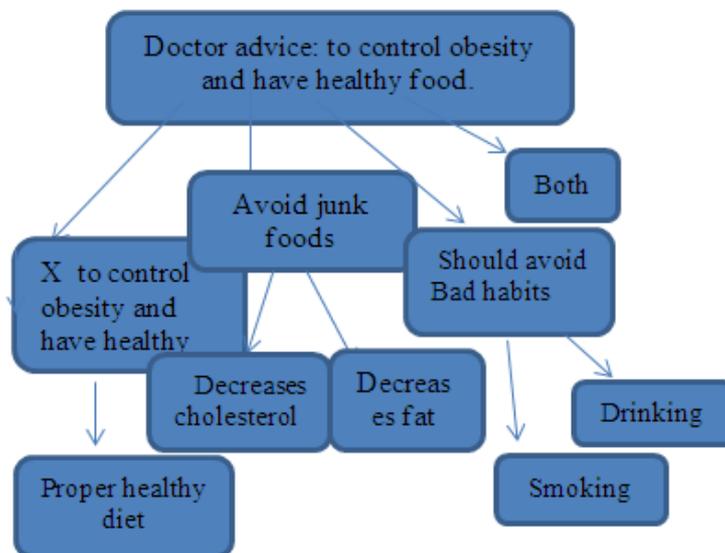


Fig-9: Problem solving using AND OR Graph

Case-1: It is only “TRUE” that X to control obesity “if we have proper physical activity and have healthy food that will be splitted up under “X control obesity and have healthy food”.

Case-2: If people avoid junk foods it decreases fat and cholesterol then it is only true “X is both”.

Case-3: Bad habits are splitted up into smoking and drinking. It is true “X is both”.

Case-4: “X is a mixture”.

8. TYPES OF KNOWLEDGE

Study of knowledge is called as **epistemology**

➤ **Relational Knowledge**

Relational Knowledge compares two object based on attribute

Age	Factors	Percentage
≥18	Use of tobacco	15.2%(16.2% males,13.7% females)
≥18	Physical Inactivity	49.8%
Both young and Adults	Overweight and Obesity	50%
≥20	Family History and Genetics	12.2%

Table-6: Derived knowledge

➤ **Inheritable Knowledge**

Inheritable Knowledge consists of objects and attributes which inherits the new object from existing object.

➤ **Inferential Knowledge**

Inferential Knowledge consist of AND – OR Graphs.

Predicate Logic	Representation
Heart disease is a name of disease	disease(Heart disease)
All Heart problems belong to the class of heart diseases.	$E:C(v) \rightarrow diseases(v)$
All heart problems either disease or heart disease	$\forall E: heart\ problems(E) \rightarrow Diseases(E, Heart\ Disease)$ $\forall Disease(X, Heart\ disease)$

Table-10: Representation of Factors

9. CONCLUSION

Heart disease is one of the fatal diseases. The WHO (World Health Organization) has estimated that 2 million people died every year because of heart disease. This is due to many factors. This study shows, In early days only the persons age >50 will get the heart disease because of obesity and medical issues. But now a days, Even a new born baby get a heart problem due to family history and genetics. Young people get heart problems because of this modern unhealthy lifestyle and their bad habits like smoking, drinking etc. In this paper we have represented the factors and problems that leads heart disease by using the AND OR graphs with the help of knowledge engineering. The assertions are formulated to control the occurrence of heart disease among different group of people.

10. REFERENCES

1. Atul Kumar Pandey, Prabhat pandey, K.L.Jaiswal, Ashish Kumar Sen “Data mining clustering techniques in the prediction of Heart Disease using Attribute selection method” International journal of science, Engineering and Technology Research, Volume(2) ISSN: 2278-7798.
2. Umair Shafique, Fiaz Majeed, Mr.Haseeb Qaiser and Irfan UI Mustafa “Data mining in Healthcare for Heart diseases” International Jornal of Innovation and Applied studies, Volume (10), ISSN: 2028-9324.
3. Sulabha s.Apte, Chaitrali S. Danare “Improved study of Heart prediction system using Data Mining classification techinques” International Journal of Computer Applications, Volume(47), ISSN: 0975-888.
4. Harleen kaur and Siri Krishnan Wasan “Empirical study on Applications of Data mining techniques in Healthcare” Jornal of Computer Science, Volume (2), ISSN: 1549-3636.
5. Sellappan Palaniappan,Rafiah Awang “Intelligent Heart Disease Prediction System using Data mining Techniques” International Journal of Computer Science and Network Security, Volume(8).
6. Syeda Farha Shazmeen,Mirza Mustafa Ali Baig,M. Reena Pawar “Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis” IOSR Journal Of Computer Engineering, Volume(10), ISSN:2278-8727.
7. N.Aditya Sundar, P.Pushpa Latha, Ramachandra “Performance Analysis Of Classification Data Mining Techniques Over Heart Disease Database” International journal of Engineering science and Advanced Technology, Volume (2), ISSN: 2250-3676.

STUDY ON LUNG CANCER DIAGNOSIS AND TECHNIQUES

K. S. Gowri laksshmi¹, Dr. R. Umagandhi²Research Scholar¹ and Associate Professor², Kongunadu Arts and Science College, Coimbatore**ABSTRACT**

Image processing methods plays vital role in medical field to detect the earliest stage of diseases and treatment stage. It is importance and necessity of digital Image processing because it provides a clear enhancement image and restoration are used to process degraded or blurred image. As per medical, imaging is concerned most of the images may be used in detection of tumors. It solves the problem of machine vision so as to attain the good result. By using this Image processing method serious disease of Lung cancer can be detect in the earlier stage. By obtaining Chest X-ray radiographs or CT scan image, the image processing techniques perform. The main aim of this paper is to present a survey about lung cancer detection at primary stage by using various techniques image mining techniques, applications. Disease detection is a significant problem. To fight against these diseases, the accurate prediction must be needed. The increase in the number of different types of cancer in the last few years has forced governments to take precautions. Therefore, there is a big need for predicting the behavior and stages of cancer. In this paper various techniques of Image mining and its applications and different algorithms used to analyze a lung cancer were discussed.

Keywords: CT, MRI, PET, classification, feature extraction, pre-processing, segmentation

INTRODUCTION

A cancer that begins in the lungs and most often occurs in people who smoke and Also occurs in people who never smoke. In this case they may be no clear cause of lung cancer.

There are two types of lung cancer, which grow and spread differently are small cell lung cancer (SCLC) and non-small cell lung cancer cell (NSCLC). About 10% to 15% stage called SCLC and about 80% to 85% are (NSCLC).

Lung cancer typically doesn't causes signs and symptoms in its earlier stage. Signs and symptoms of lung cancer typically occurs only the disease in advance, some symptoms:

- Shortness of breath
- Chest pain
- Headache
- Hoarseness
- Losing weight without any reason.

If any persistent signs and symptoms occur make a doctor appointment. There are many advanced technologies in medical field to detect the cancer in the initial stage. Image processing techniques are widely used in medical areas for image improvement in earlier detection.

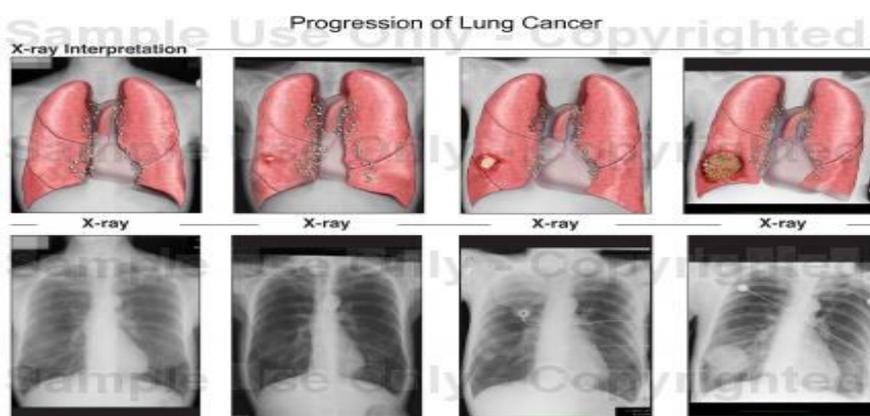


Fig-1: Sample images of lung cancer

The main intent of the image mining is to remove the data loss and obtain the meaningful information which is expected need of human. Image Mining has become an important research issue in the area of Multimedia Mining. Research in Image mining can be broadly classified in two main directions:

- Domain specific applications
- General applications

Image Mining is different from computer vision, and image processing techniques. This is because the focus of image mining is in the extraction of patterns from a large collection of images, whereas the focus of computer vision and image processing techniques is in understanding and/or extracting specific features from a single image [15]. Variety of image data for instance digital photographs, medical images and satellite TV images are generally generated each day. The process starts with Preprocessing, Feature Extraction and Segmentation.

The image Pre-processing stage starts with image enhancement; the aim of image enhancement is to improve the interpretability or perception of information included in the image for human viewers, or to provide better input for other automated image processing techniques. Image segmentation is an essential process for most image analysis subsequent tasks. Thresholding is one of the most powerful tools for image segmentation. The segmented image obtained from Thresholding has the advantages of smaller storage space, fast processing speed and ease in manipulation, compared with gray level image which usually contains 256 levels. Image features Extraction stage is an important stage that uses algorithms and techniques to detect and isolate various desired portions or shapes (features) of a given image[2012]

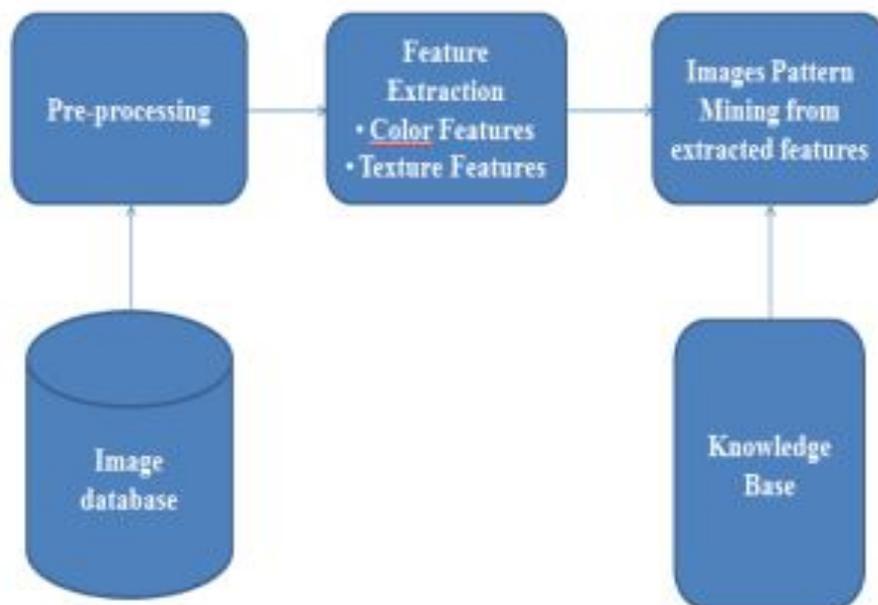


Fig-2: Image mining process

Figure: 2 shows a common structure model for image mining System. The system considers a particular sample of images as an input, whose image features are extracted to correspond to succinctly the image content for detecting the cancer Chest X-ray and CT-scan image can be taken. Most of the researchers used CT-scan image for their research work to get clear visibility. The studies shows that, however 30% of nodules in CXRs were missed by radiologist and missed nodules are partly obscured by overlying bones. Therefore Computer-aided detection scheme gives clear visibility for research. The Fig-3 is a computer scan image, its show with cigarette packet which was patient kept in his pocket.



Fig-3: Chest X-ray



Fig-4: CT-scan

RELATED WORKS

Image mining facilitates the extraction of hidden information, image data association, or other patterns not clearly accumulated in the images [16]. Image mining is an interdisciplinary effort that provides significant application in the domain of machine learning, image processing, image retrieval, data mining, database, computer vision, and artificial intelligence. Even though the growth of several applications and techniques in the individual research domain mentioned above, research in image mining has to be explored in investigated the research problems in image mining, modern growth in image mining, predominantly, image mining frameworks, modern techniques and systems

National cancer Institute taken the Lung cancer as serious problem and went through many research to detect the disease. Many researchers did their research works in Image processing techniques to predict the lung cancer. Some of the related works are:

[1] Xuechen Li, Linlin Shen, and Suhuai Luo [2018], the aim of this research is to pin point the exact cancerous lung nodule. This work is carried through solitary feature-based lung nodule detection method. Using Chest X-ray radiographs images the work will start. Stationary wavelet transform and convergence index filter used to extract the texture feature and Ada boost classifier is used to classify the nodule.

The following steps are included to detect the lung nodule

- First segmenting the lung field and rib suppression done for visibility enhancement.
- For feature extraction stationary wavelet and convergence index filter is used.
- Pixel based detecting the white nodule likeliness map done using Laplace Gaussian blob detection.
- By using Adaboost classifier classifies and ranking according to the combination and texture and measured the solitary degree using proposed solitary feature.
- Finally detecting the nodules. This proposed method shows promising performance more than 80% in lung nodule detection

[2] Jayshree Talukdar, Dr. Parismita Sarma [2018], the main goal of the research is to diagnosis the lung cancer in the primary stage. Artificial neural network is used for classification, Gaussian noise is used to remove the noise reduction, Otsu’s Thresholding used for automatic segmentation and computed Aided Diagnosing used to detect the cancer cell in earliest stage. The dataset used is QIN LUNG CT from The Cancer Imaging Archive (TCIA) .The Computed tomography (CT) Image data was obtained on patients diagnosed with Non-Small Cell Lung Cancer (NSCLC) with mixed stage. The dataset comprises of 67 images. So according to this research it is found that since real data are not easily available, so we should focus on processing of image with better techniques in the limited set of data and then only move to our next phase

Table-1: Result of Diagnosis

Patient ID	4302
Name	XXX
Sex	Male
Diagnosis Date	21-8-2015
Result	Malignant tumor indicated
Suggestion	Go to hospital on or before Dec 22-12-2015

[3] Alakwaa, W, Nassef, M. and Badr, A (2017). The objective of this research work is carried through CAD system to classify CT image of the lung cancer. The image processing step included segmentation (Thresholding) followed by normalization down sampling and lastly zero centering. The 3D CNN produces an accuracy of 86.6 % which was tested in Kaggle’s CT scans. The future work of the researchers aims to detect the exact location of the cancerous nodules with the help of the current model. Also, they have thought of use of watershed algorithm for segmentation and making the network deeper.

[4] Rodiah, Emy Haryatmi, Fitrianinghsih, Herio Susanto, [2016], the motive of this research is to diagnosis a lung cancer with three assumptions: (i) No tumor is indicated, (ii) tumors is indicated but benign (iii) malignant tumors’ are indicated.

This research is carried with sugeno Fuzzy logic, PHP, MYSQL. Using PHP the website is constructed with four pages: (I) main page, (ii) information page regarding the sickness,

(iii) Diagnosis page (IV) the page related to the result of diagnosis. The first page of the website is main menu, which contains login form, which is used by listed user whether they may be doctors or patients. The information is collected as Dataset on the basis of interview methods and variable information; there are five input variables for lung cancer detection. They are: (I) anamnesis total (information regarding cough, headache, and bone pain symptoms), (ii) the degree of smokers (information about smokers), (iii) age, (IV) performance status, (v) doubling time (time needed for tumor to grow doubled the size).

Fuzzification is a step to transform the variables explained above in to the form of variable fuzzy. The variable status performance of Fuzzification consists of three categories: (I) very weak, (ii) weak (iii) normal. As result lung cancer detection with Fuzzy logic method using web can be used as support medical tool.

[5] The work published by Gruetzemacher, R and Gupta, a (2016), proposed a methodology which mainly focused on image recognition algorithm and deep learning for detection of malignant pulmonary nodules. The Caffe, an open source deep neural network solver was used to train the DNN model. This paper performed 4 architecture of convolution layer with convolution layer of 3,4,5,6 and the architecture with five convolution layers gave the best performance with an accuracy of 82%. Sensitivity of this method is 78.2% and specificity of 86.3%.

[6] Nastaran Emaminejad, Wei Qian, Yubao Guan, Maxine Tan, Yuchen Qiu, Hong Liu, and Bin Zhen [2016], the aim of this research is to prognosis early stage of lung cancer patients. They used a technique of new quantitative images feature analysis scheme and find out its role with two genomic biomarkers. Two biomarkers namely excision repairs cross complementing [ERCCI] gene product and regulatory subunit of ribonucleotide reeducates [RRM1] and reported prognostic biomarkers of Non-Small lung cancer. CT image used and it is segmented lung tumor by computer-aided detection scheme and also called related tumor features.

Trained Naïve Bayesian network based classifier used for image features and multilayer perceptron classifier is used to predict lung cancer. In this research work quantitative image feature is compared with genomic biomarkers ERCCI and RRM1. [2013] Quantitative image feature technique carry this work with following steps :(I) Standard deviation of tumor pixel value (CT number) (ii) The maximum tumor diameter computed from the targeted CT slice, (iii) A tumor shape factor, (IV) the maximum CT number within the tumor volume, (v) Standard deviation of pixel values in central volume of tumor.

[7] Bhagyarekha U.Dhaware, Anjali C.Pise [2016], the motive of the research is to detect the lung cancer using Bayasein classifier and FCM segmentation. The Lung Database is collected from PRISM Diagnostic centre, Maharashtra. First CT scan image collected and categorized in to normal and abnormal. The taken image is preprocessed for noise reduction for image enhancement and it can be done through "Contrast limited Adaptive Histogram Equalization" (CLAHE). To extract features gray level co-occurrence matrix used. They extract 12 texture features and 7 shapes. Bayasein classifier is used to classify the input CT lung image whether it is normal or abnormal condition. Then Fuzzy c-means is a technique applied when clustering is needed. This technique will help to get more than 90% accurate result.

[8] Manasee KurKure, Anuradha Thakare [2016], the proposed work of this paper is to classify the image whether cancer or non cancerous. The image taken may be X-ray, CT, or PET .The input images are preprocessed using canny detection method to detect the strong edges. Genetic algorithm and Naïve Bayes classifier is used to classify the different stages of cancer images fast and accurately. It obtained 80% of accuracy.

[9] Suman Shrestha, Jeff Petermann, Tannaz Farrahi, Aditi Deshpande, and George C. Giakos, this research work is carried through liquid crystal optical polarimeter instrument by testing the tissue diagnostic method and classified the images on histopathological samples of lung cancer cell. [2005], [2012] it is an automated polar metric imaging system developed for efficient and reliable diagnostic method. [2010] the experimental results provide useful information in discriminating healthy and cancer cells.

[10] Mr.Vijay A.Gajdhane, Prof.Deshpande L.M [2014], the main of this research work is to detect the lung cancer using Image Processing techniques. The images acquired from CT scan for clear visibility and started the first stage of Image processing method called Preprocessing. For Preprocessing Gabour Filter and Watershed segmentation algorithm used, then the features are extracted on the basis of area, perimeter and eccentricity. For classification support vector machines are used. This result gives different dimension of lung tumors. If it is cancer tumor then it is little bit bigger than the other.

[11] Hashemi et.al. (2013) his aim has to improve the efficiency of the lung cancer diagnosis system, with the help of proposing a region growing segmentation method to segment CT scan lung images. First of all, for noise

removing linear-filtering and contrast enhancement technique was used as pre-processing step, to prepare the image for segmentation. After that for differentiating among malignant, benign and advance lung nodules, the cancer recognition was presented by fuzzy inference system. The authors also compare the diagnosis performance of the proposed method with the artificial neural network.

[12] Mokhled S. AL-TARAWNEH [2012], the aim of this research was to detect features for accurate images comparison as pixels percentage and mask-labeling. The work is divided in to three stages: Image Enhancement, Image Segmentation, and Feature Extraction. The image Pre-processing stage starts with image enhancement is to improve the interpretability or perception of information included or to provide better input for other automated image processing techniques. The following three methods are used for this purpose: Gabor filter (has the best results), Auto enhancement algorithm, and FFT Fast Fourier Transform (shows the worst results for image

Table-3: Results of three techniques used for image enhancement

Subject	Auto enhancement	Gabour filter	FFT filter
Sub1	37.95	80.95	27.075
Sub2	47.725	80	36.825

Image Segmentation stage: to divide and segment the enhanced images, the used algorithms on the ROI of the image (just two lungs, the methods used are: Thresholding approach and Marker-Controlled Watershed Segmentation approach (this approach has better results than Thresholding).

Table-4: Image segmentation Experimental Result

Subject	Thresholding	Watershed Filter
Sub1	81.625	85.375
Sub2	82.2	85.25
Sub3	82.125	85.55

[13] Daw, Chung, Wen, presented an extension of neural network based fuzzy model for the detection of lung nodule .After the Thresholding stage, some part of the blood vessels or the large airways may also be removed. So, in order to fill these areas, morphological closing and labeling was done. In order to make distinction between the nodules and other structure in lung region, three main features area, brightness and circularity were calculated. This neural network based fuzzy model consists of four layers: input layer, Fuzzification layer, rule inference layer, Defuzzification layer. With this system, the classification accuracy of 89.3% was achieved. The false positive value was 0.21. The main advantage of this system, it was faster, no prior knowledge was required, the fuzzy rules were defined using learning procedure and Detection rate was high

[14] The lung cancer detection on CT image using image processing achieves a higher detection rate of 90.1% while Neuro - Fuzzy Model achieves 89.3% and Region Growing Method achieves 88.5% detection rate. In this paper [2], Fuzzy C-Means (FCM) and different extensions of FCM algorithm are discussed. The exclusive FCM algorithm yields better results for segmenting noise free images, but it fails to segment images downgraded by noise, outliers and other imaging artifacts. This paper presents an image segmentation approach using Modified Fuzzy Possibility C-Means algorithm (MFPCM), which is a generalized adaptation of the standard Fuzzy C-Means Clustering algorithm and Fuzzy Possibility C-Means algorithm. The flaw of the conventional FCM technique is eliminated by modifying the standard technique. The Modified FCM algorithm is defined by modifying the distance measurement of the standard FCM algorithm to permit the labeling of a pixel to be influenced by other pixels and to restrain the noise effect during segmentation. Rather than having one term in the objective function, a second term is included, forcing the membership to be as high as possible without a maximum frontier restraint of one. Experiments are carried out on real images to observe the performance of the proposed modified Fuzzy Possibility FCM technique in segmenting the medical.

Table-4: Summary of techniques used

Reference/year	Author name	Techniques used	Accuracy
[2017]	Alakwaa, W., Nassef, M. and Badr, A	-CAD -Kaggle’s CT images -Thresholding	86.6%
[2016]	Bhagyarekha U.Dhaware, Anjali C.Pise	-Bayasein classifier - Fuzzy c-means - CT scan	90%
		-neural network	

[2005]	Daw, Chung, Wen	-Fuzzy model	89.3%
[2016]	Gruetzemacher. R and Gupta,	-image recognition algorithm -neural network	82%
[2010]	M. Gomati	-CT image -Fuzzy c-means	89.3%
[2013]	Hashemi et.al	-CT -linear filtering -contrast enhancement -Fuzzy logic	Compared with Artificial neural network for better performance
[2018]	Jayshree Talukdar , Dr. Parismita Sarma	-Artificial neural network -Gaussian noise -Otsu's Thresholding	<80%
[2016]	Manasee KurKure, Anuradha Thakare	-CT -X-ray -PET -Naïve Bayes -Genetic algorithm	80%
[2012]	Mokhled S. AL-TARAWNEH	-Gabor filter -Watershed algorithm -masking approach	80.95 85.375 85.75
[2016]	Nastaran Emaminejad, Wei Qian, Yubao Guan, Maxine Tan, Yuchen Qiu, Hong Liu, and Bin Zheng	-Naïve Bayesian network based classifier - Computer-aided detection -Multilayer perceptron classifier	>90%
[2016]	Rodiah, Emy Haryatmi, Fitrianinghsih, Herio Susanto	-Dataset -Fuzzy logic -PHP -MYSQL	Best support medical tool
[2015]	Suman Shrestha, Jeff Petermann, Tannaz Farrahi, Aditi Deshpande, and George C. Giakos	-Liquid crystal optical polarimeter instrument	Best for medical lab
[2014]	Mr. Vijay A. Gajdhane, Prof. Deshpande L.M	-Gabor Filter -Watershed segmentation algorithm -CT scan -support vector machines	>90%
[2018]	Xuechen Li, Linlin Shen, and Suhuai Luo	-Chest X-ray radiographs -Stationary wavelet transform -convergence index filter -Ada boost classifier	80%

CONCLUSION

This paper presents in detail about the related works to find the earlier stage of lung cancer by using various techniques that have been used in preprocessing, feature extraction, segmentation and classification. Some works shows accurate results, moderate results. The future works may carry with more features could be employed to identify the real lung nodule. In this review result and discussion in various classification techniques. Thus the classification will acts as main role in image processing system it will be used to classify cancer and non-cancer images. So, the earlier stage of lung cancer detection will prevent a person from the death.

REFERENCES

- Alakwaa, W., Nassef, M. and Badr, A., (2017). "Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN)".
- Bhagyarekha U. Dhaware, Anjali C. Pise (2016) "Lung cancer Detection using Bayasein classifier and FCM segmentation" International conference on Automatic control and Dynamic Optimization technique,

-
-
- Daw-Tung Lin, Chung-Ren Yan, Wen Tai Chen , "Autonomous detection of pulmonary nodules on CT images with a neural network-based Fuzzy system," *Computerized medical imaging and graphics* 29(2005) 447-458.
 - Gruetzemacher, R and Gupta, A., 2016. Using deep learning for pulmonary nodule detection & diagnosis
 - M. Gomati, "A Parameter Based Modified Fuzzy Possibility C-Means Clustering Algorithm for Lung Image Segmentation", *Global Journal of Computer Science and Technology* Vol. 10 Issue 4 Ver. 1.0 June 2010.
 - Hashemi, A., Pilevar, A. H., & Rafeh, R. (2013). Mass Detection in Lung CT Images Using Region Growing Segmentation and Decision Making Based on Fuzzy Inference System and Artificial Neural Network. "International Journal of Image, Graphics and Signal Processing (IJIGSP), 5(6), 16.
 - Jayshree Talukdar, Dr. Parismita Sarma [2018], " A survey on lung cancer Detection in CT scan using Image Processing Techniques" *International journal on current trends and technology*.
 - Manasee KurKure, Anuradha Thakare [2016] "Lung Cancer Detection using Genetic Approach" DOI: 10.1109/ICCUBEA.2016.7860007, NSPEC Accession Number: 16693054
 - Mokhled S. AL-TARAWNEH "Lung Cancer Detection Using Image Processing Techniques" *Leonardo Electronic Journal of Practices and Technologies* , ISSN 1583-1078 , Issue 20, January-June 2012.
 - Nastaran Emaminejad, Wei Qian, Yubao Guan, Maxine Tan, Yuchen Qiu, Hong Liu, and Bin Zheng "Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients", *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING*, VOL. 63, NO. 5, MAY 2016.
 - Rodiah, Emy Haryatmi, Fitrianinghsih, Herio Susanto (2016) "Web Based Fuzzy Expert System for Lung Cancer Diagnosis", 2016, 2nd International conference on Science in Information Technology, 978-1-5090@2016 IEEE.
 - S. Swensen et al., (2005) "CT screening for lung cancer: Five-year prospective experience" *Radiology*, vol. 235, no. 235, pp. 259–265, Apr.
 - Suman Shrestha, Jeff Petermann, Tannaz Farrahi, Aditi Deshpande, and George C. Giakos, (2015) "Design, Calibration, and Testing of an Automated Near-Infrared Liquid-Crystal Polarimetric Imaging System for Discrimination of Lung Cancer Cells" *Fellow, IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT*, VOL. 64, NO. 9, SEPTEMBER 2015
 - Mr. Vijay A. Gajdhane, Prof. Deshpande L.M "Detection of Lung cancer on CT scan image by using various Image processing Techniques, *IOSR journal of Computer Engineering* [2014].
 - Xuechen Li, Linlin Shen, and Suhuai Luo (2018). "A Solitary Feature – Based Lung Nodule Detection Approach for Chest X-Ray Radiographs" *IEEE TRANSACTIONS ON BIOMEDICAL HEALTH INFORMATICS* VOL.22, NO 2, March 2018.
-
-

INTELLIGENT ANALYSIS ON BREAST CANCER THROUGH KNOWLEDGE REPRESENTATION USING AND – OR GRAPH

Swetha. S, Saraswathi. S and Ayesha Begam. S

Department of Computer Science & Applications, Sri Krishna Arts and Science College, Coimbatore

ABSTRACT

Data Mining is the progressive area of categorizing over enormous collection of attributes that exist in datasets to classify meaningful arrangements create link to clarify problems over data analysis. Attributes that reflects object behavior becomes huge collection. It is tedious to determine the optimal solution among the collection in predicting the occurrence of Breast Cancer, the effort to predict the objective is done through preprocessing, filtering and data analysis so, the Data mining tools grants to conclude forthcoming progression.

Knowledge engineering is a pasture of Intelligent System. This technique is applied to analyze the data of breast cancer by using representation logics AND-OR Graph. The dataset consist of 286 instances and 10 different attribute are considered for study new assertion are formulated to enhance the prediction.

Keywords: Data Mining, Breast cancer dataset.

1. INTRODUCTION

The standard shield Significantly boost precision and aspect of the database. The outline of the results is authentic and advantageous in case extra genes or cancer types are examined at same time. [10]

Breast cancer is the specific dominant cancer for women in advance countries as well as India. It is second biggest familiar cause of cancer death in women. The huge extent of breast cancer in women has extended naturally in the previous year's [2].

In this paper the classification of data will be based on breast cancer patient data set developed by collection of data from hospital repository which consists of 286 instances with 10 different attributes. Facts are formed as data for organizing and analyzing to extract information. The information's are evaluated to gain knowledge the knowledge engineering process is applied to reformulated the initial knowledge.

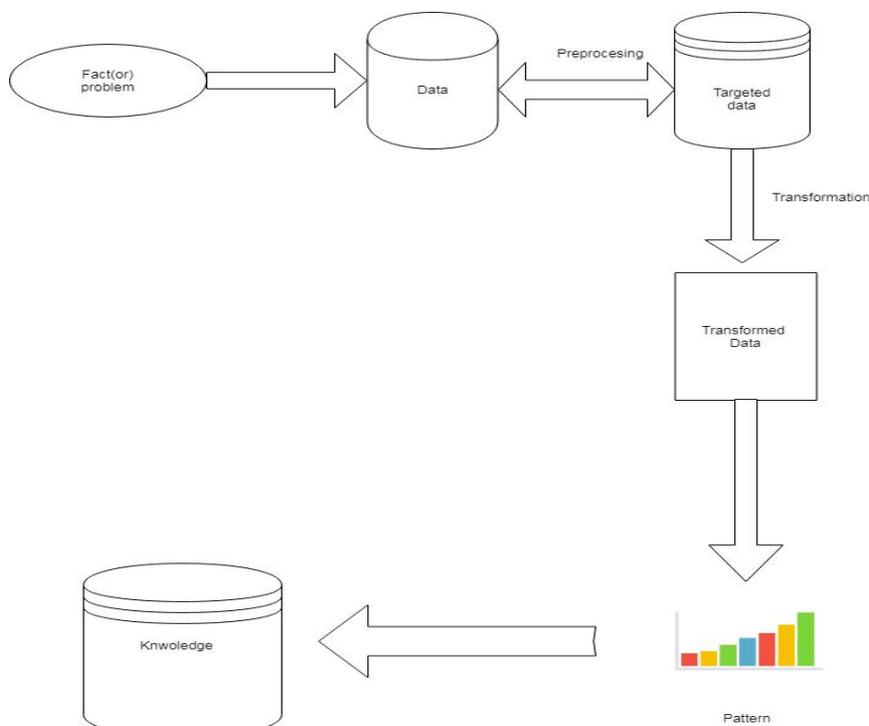


Fig-1: Knowledge Engineering Diagram

In the above representation are many facts that exists in real applications are formed into data. The data was organized, analyzed for further computation it was splited as information it's also processed as interpretation and evaluation was formed as knowledge is further processed into knowledge base acquire knowledge through knowledge engineering it base for building intelligence system. Here fact can be represented in the form of 'f' it is considered as fact and data can be represented in the form off 'I'.

$f(i) \rightarrow d$

Datasets (d) are formed by collection of data with real attributes, reflecting the object behavior and denotes the nature of the domain, by using this attributes values data analysis are carried out to extract and meaningful patterns. After the completion of this analysis process, the simplified clustered or classified dataset is considered for organizing the sequence. The data organization provide unique information.

$p(d) \rightarrow td$

Processed data is represented as p(d), shows filtered essential attribute from the dataset. The distinct data for the problem will be provided by (td), targeted data.

The newly generated dataset will proceed to next step of analyzing to result in accurate extraction of patterns.

In this process the entire dataset will get analyzed and it will organized relevant data

$td \rightarrow k(i)$

The knowledge is gained from the targeted data (td) go through process like discharge of the data using analytical tools.

1.1 To Represent knowledge

- Propositional calculus.
- 1st Order Predicate logic.
- Probabilistic logic, analytical techniques.
- Fuzzy Logic, Evolutionary computing.
- Decision trees, Soft computing
- Data Mining Techniques.
- These are some types of techniques to represent knowledge.

Since Natural language is existing to have machine interpretation, for performing required computation. So the knowledge is represented using “**first order predicate logic technique**”.

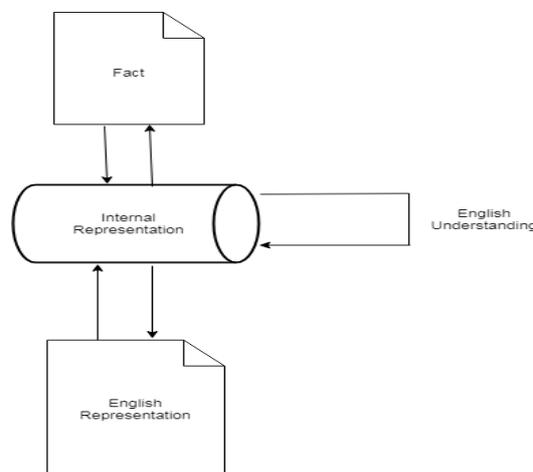


Fig-2: English Understanding

Facts which are in the English representation cannot understand by the machine, but give more clarity in understanding the problem. So it's converted into internal (multiple) representation, it is nothing but a machine language (0's & 1's) binary code by doing this machine can easily understand the facts and it can interpret to carry out the process, it will reduce the difficulties exists in machine computation.

1.2 Forward Mapping and Backward Mapping

The Mapping is the process of converting natural language representation into formal language representation, machine to interpret the denoted facts by this conversion. It is used to map the entities (or) conversion by using this method. The mapping is divided into two stages, such as forward and backward mapping and used based on the user requirements. The user will load (or) give the input dataset in the form of English representation and it will be converted into formal language this is used for read the data, interpret and mapped. Another method is used to display the output in the form of how user can understand, it will be converted from formal representation to English language representation.

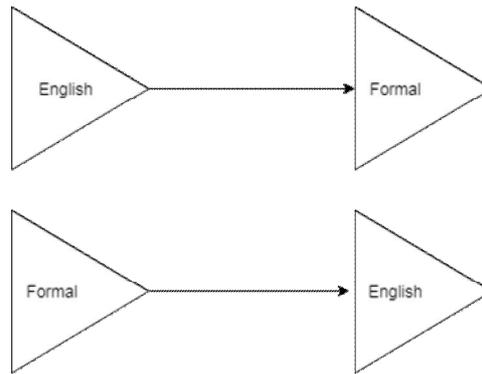


Fig-3: Mapping

Fact	Representation
X women's has affected by breast cancer	English Representation
Women's (x)	Logical Representation
X:Women's(x) affected by breast cancer(x)	Formal Logical Representation
Has affected(women)	Deductive Mechanism
Women's are affected by breast cancer	English Representation (background representation)

Table-1: Fact 1

Fact	Representation
Y Men's has affected by breast cancer	English Representation
Men's (Y)	Logical Representation
Y:Men's(y) affected by breast cancer(y)	Formal Logical Representation
Has affected(Men)	Deductive Mechanism
Men's are affected by breast cancer	English Representation (background representation)

Table-2: Fact 2

Fact	Representation
Women(X) age above(40+) has affected by breast cancer	English Representation
Women's (x)	Logical Representation
X:Women's(x) above (40+) affected by breast cancer(x)	Formal Logical Representation
Above (40+) has affected(women)	Deductive Mechanism
Women's are above (40+)affected by breast cancer	English Representation (background representation)

Table-3: Fact 3

Fact	Representation
1 Million(X) women's has affected by breast cancer in India	English Representation
Women's (x)	Logical Representation
X:Women's 1 million (x) affected by breast cancer(x)	Formal Logical Representation
1 million has affected(women)	Deductive Mechanism
1 million women's are affected by breast cancer	English Representation (background)representation)

Table-4: Fact 4

Fact	Representation
Less than 1 million Men's(y) has affected by breast cancer in India	English Representation
Men's (x)	Logical Representation
Less than 1 million Y:Men's(y) affected by breast cancer(x)	Formal Logical Representation
Less than 1 million has affected(Men)	Deductive Mechanism
Less than 1 million men's(y) are affected by breast cancer	English Representation (background representation)

Table-5: Fact 5

2. PROPERTIES OF GOOD KNOWLEDGE BASE SYSTEM

It is essential to represent the derived knowledge in different perspective. The method which is used to encode knowledge in intelligent system knowledge base is knowledge representation. It shows the dependency of one from other variable. It also analyze attributes and finally represents the finding of major controls.

2.1 Representational Adequacy

The skill to perform all the distinct kinds of knowledge that might be desired in that domain. The Representation is authentic, once sufficient domain knowledge is accessible. It focuses of completeness of information.

➤ **Inferential Adequacy**

The skill to manipulate the representational structures to acquire contemporary structures (comparable to new knowledge) from extant structures. The new knowledge structure is derived from existing one to have meaning full analysis.

➤ **Inferential Efficiency**

The skill to aerial new information into the knowledge structure which can be used to aim the consideration of the inference mechanisms in the most auspicious direction.

➤ **Acquisitional Efficiency**

The skill to knowledge new information easily. Admirably the agent should not able to control its own knowledge acquisition, but direct assertion of information by a ‘Knowledge Engineer’ would be adequate.

➤ **Knowledge Acquisition**

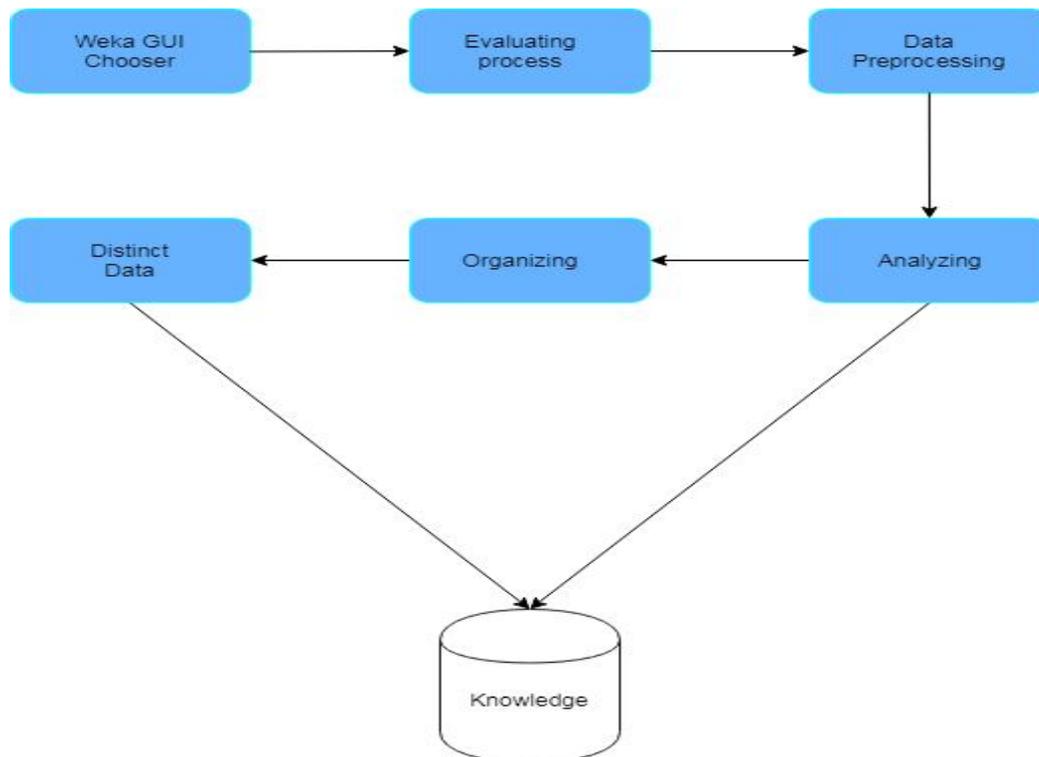


Fig-4: Knowledge Acquisition

The above flowchart domain is said as the sequence of process which converts the data from one form to another. Each form is having some incorporated technique that are essential to improve the performance of the process. Associate methods at the interchange of machine learning database and statistics system the step analysis is called knowledge discovery(or) identify domain after the completion of this process. The next is domain boundaries, here there will be some limitations for dataset so finding that limitation is called domain boundaries.

The final part of the process is to select and applying elicitation technique. An elicitation technique is several statistics of data compilation technique used in knowledge engineering. Final step of the process is review and acquire knowledge, it checks all the process once again and it moves to the end domain knowledge base it is similar to the locate knowledge source. So this process came to an end then it moves for data manipulation part here the data will be manipulated.

3. KNOWLEDGE MANIPULATION

Knowledge manipulation are if independent then corollary, has two primary equitable of prediction and description.

Ex: If age $> 40 \leq 60$ there is chances for occurrence of breast cancer for women.

If temp > 100 then it's sign of high fever.

AND – OR Graph

AND – OR Graph shows structured representation and knowledge, this graph supports problem disintegration like divide and over throw, finding solution to small parts and ploy to upper node.

AND can be represented as (\wedge)

OR can be represented as (\vee)

4. COMBINATION OF AND NODE & OR NODE

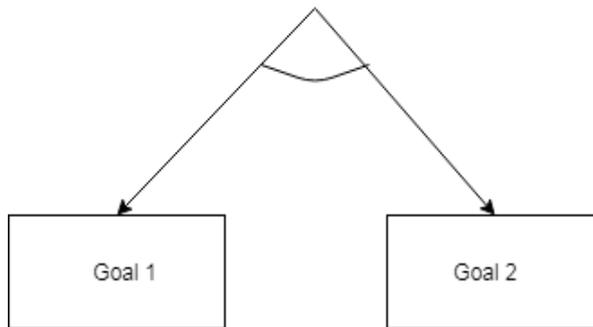


Fig-5: OR Node

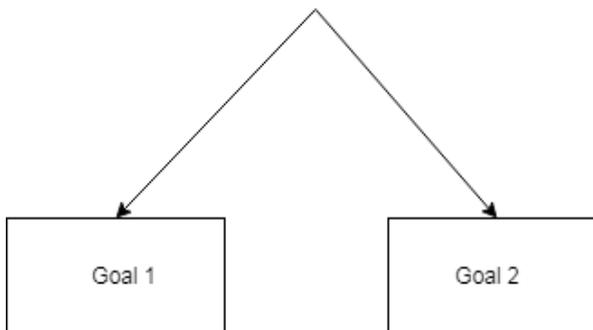


Fig-6: AND Node

Problem: The problems are described with age and advice given by the doctor. One is to control weight and other one is to do physical work and finally doing both. These two are small prevention for breast cancer.

Women's with the age of $> 40 \leq 60$ can affected by breast cancer to prevent this doctor advised them to control weight, do physical activity or both.

Assertion: Women's age and control weight, do physical activity or both.

Step-1: To create the top node and name it as doctor advice

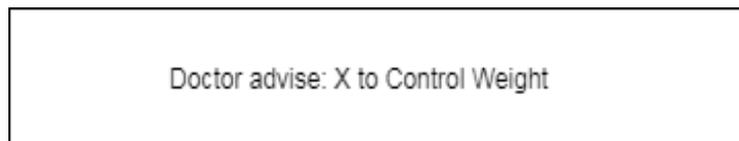


Fig-7: Top node

Step-2: Top node is among sub goal as other 3 nodes to control weight, do physical activity and both.

Predicate Logic	Representation
Breast cancer is a name of cancer	Cancer(Breast Cancer)
All Cancer belong to the class of diseases	$\forall : C(\forall) \rightarrow \text{diseases}(\forall)$
All diseases either cancer or breast cancer	$\forall E: \text{diseases}(E) \rightarrow \text{Cancer}(E, \text{Breast Cancer})$ $\vee \text{Cancer}(X, \text{Cancer})$

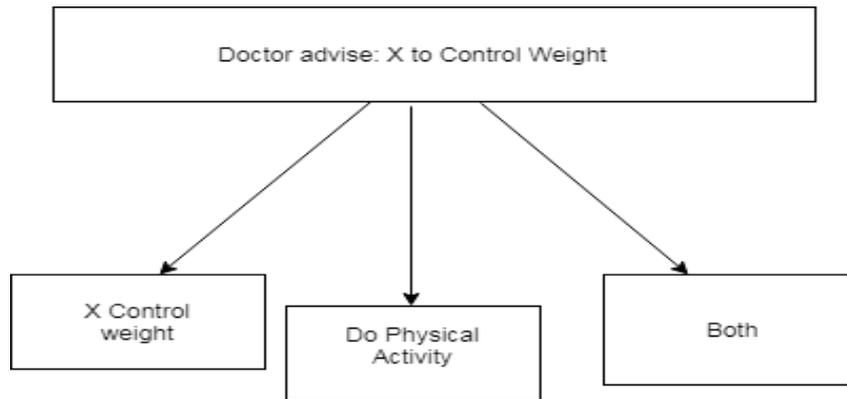


Fig-8: Doctor’s advice

Step-3: Again nodes are split up, like sufficient and in sufficient of problem (or) cases

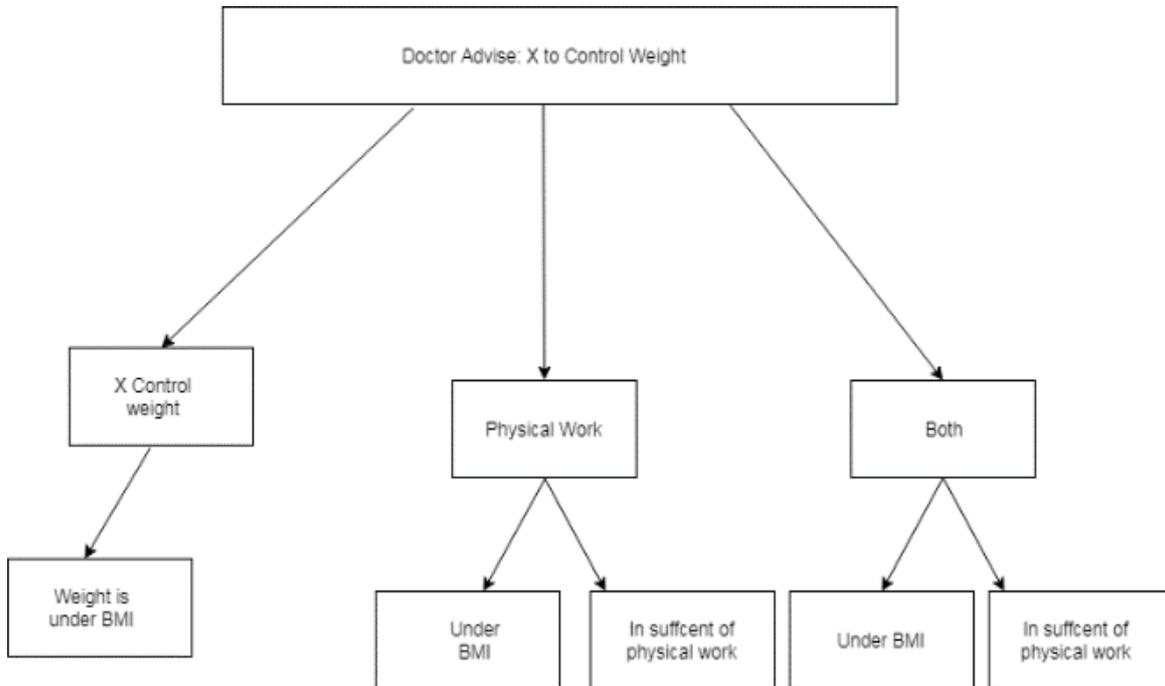


Fig-9: Problem Solving using graph

Case-1: It is only “true” that X to control weight “if weight is more than BMI that will be splitted up under “ X is over weight”.

Case-2: If weight is under BMI and “Physical work is sufficient” then is only true “X is both”.

Case-3: “X is mixture”.

5. TYPES OF KNOWLEDGE

Study of knowledge is called as **epistemology**.

➤ **Relational Knowledge**

Relational Knowledge compares two object based on attributes

Age	Tumor- Size	Deg- Malig
54	0.4	10
45	0.2	12
60	0.7	11

➤ **Inheritable Knowledge**

Inheritable Knowledge consists of object and attributes and inherits new object from existing object.

➤ **Inferential Knowledge**

Inferential Knowledge consist of AND – OR Graphs

6. CONCLUSION

By using Intelligent Analysis on Breast Cancer through Knowledge Representation using AND – OR Graph shows that the occurrence of Cancer either by means of food habits, lack of physical activity or incompleteness in day to day task.

The Outcome represents the knowledge gained by using AND-OR graph and the successful formulation of facts are tabulated. Different possibilities are suggested based on the formulated assertions.

REFERENCE

1. KB Cohen, L Hunter, Getting started in text mining - PLoS computational biology, 2008 - journals.plos.org
2. S Kharya Using data mining techniques for diagnosis and prognosis of cancer disease - arXiv preprint arXiv:1205.1923, 2012 - arxiv.org
3. D Delen, G Walker, A Kadam - Predicting breast cancer survivability: a comparison of three data mining methods Artificial intelligence in medicine, 2005 - Elsevier
4. M Elter, R Schulz-Wendtland, T Wittenberg , The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process- Medical physics, 2007 - Wiley Online Library
5. A Özgür, T Vu, G Erkan, DR Radev Identifying gene-disease associations using centrality on a literature mined gene-interaction network- Bioinformatics, 2008 -academic.oup.com
6. JD Wulfkuhle, KC McLean, CP Paweletz New approaches to proteomic analysis of breast cancer PROTEOMICS, 2001 - Wiley Online Library
7. HY Chuang, E Lee, YT Liu, D Lee... - Network-based classification of breast cancer metastasis Molecular systems, 2007 - msb.embopress.org
8. GI Salama, M Abdelhalim, MA Zeid Breast cancer diagnosis on three different datasets using multi-classifiers- Breast Cancer (WDBC), 2012 - Citeseer
9. X Liu, A Krishnan, A Mondry An entropy-based gene selection method for cancer classification using microarray data- BMC, 2005, bmcbioinformatics.biomedcentral .
10. M Ongenaert, L Van Neste, T De Meyer PubMeth Nucleic acids: a cancer methylation database combining text-mining and expert annotation 2007 - academic.oup.com

USES OF ICT LEARNING TECHNIQUES TO IMPROVE STUDENTS LEARNING SATISFACTION**Nirmala Grace Rani S.¹ and Dr. Joseph Catherine²**Research Scholar¹ and Principal², Stella Matutina College of Education, Chennai**ABSTRACT**

Learning is a never-ending process one starts right after he / she enter this world. Consciously or unconsciously every human being take part is in this tedious process. The educational institution play a pivotal role in teaching moral and social behaviour to students along with vocational and language education. As the students of present generation are identified as tech savvy kids, the learning fraternity should have above the talents of the students. Hence, the curriculum demands novel teaching techniques to teach the students. Therefore different types of teaching – Learning techniques are discussed in this paper.

Keywords: Bookends, Diagnostic, Learning, Muddiest Point, Traveling Files.

INTRODUCTION

Learning is the process of acquiring new or modifying existing, knowledge, behaviors, skills, values or preferences. The ability to learn is possessed by humans, animals and some machines; there is also evidence for some kind of learning in some plants. Some learning is immediate, induced by a single event (e.g. being burned by a hot stove), but much skill and knowledge accumulates from repeated experiences [1]. The changes induced by learning often last a lifetime and it is hard to distinguish learned material that seems to be “lost” from that which cannot be retrieved.

Many students are being left behind by an educational system that some people believe is in crisis. The techniques include Learning Jigsaws, Think-Pair-Share, Diagnostics, Bookends, Traveling Files, Minute Paper, Team Paper Activity, Numbered Heads Together, Prairie Fire, Take a Stand, Cocktail Party, Meet Your Match, Muddiest Point technique, One Sentence Summary, Concept Maps, Elaborative interrogation, Self-explanation, Summarization, highlighting (or underlining), the keyword mnemonic, imagery use for text learning, rereading, practice testing, distributed practice and interleaved practice.

To offer recommendation about the relative utility of these techniques, we evaluated whether their benefits generalize across four categories of variables: learning conditions, student characteristics, materials and criterion tasks [2]. Learning conditions include aspects of the learning environment in which the technique is implemented, such as whether a student studies alone or with a group. Student characteristics include variables such as age, ability and level of prior knowledge.

LEARNING

In the present competitive world, the necessity to develop strategies to compensate for an individual’s weakness and capitalize on his / her strengths has increased to a great extent. It is important here that the implications of Learning Styles as a means to learn effectively are evident [3]. Learning occurs in stage, and at each stage students learn in different ways. Difficulties that arise at home and in school are often due to differences in Learning Styles. Children’s academic performance and success in life depend on the thinking and problem solving skills they develop in early childhood. Psychologists define learning as “a relatively permanent change in behaviour, which occurs as a result of activity, training, practice or experience”.

This definition of learning has three important elements.

- Learning results in change of behaviour.
- It is a change that takes place through practice or experience.
- Before it can be called learning, the change must be relatively permanent. It must last a fairly long time.

CHARACTERISTICS OF LEARNING

- Learning is universal. All living beings learn.
- Learning is continuous; it is a perpetual activity that takes place from ‘womb to tomb’.
- Learning results in improved performance.
- Learning is purposive: A child’s learning in and out of school is closely linked up with its goals, purposes and satisfactions. Nobody learns anything without a purpose.
- Learning is contingent upon experience.

- ‘Learning’ is not something to be given; it is to be gained by self-experience. A person’s knowledge or learning is the result of that person’s experience.

SATISFACTION

Satisfaction - a fulfillment of need or desire, the pleasure obtained by such fulfillment. “Satisfaction is the feeling of pleasure or disappointment attained from comparing a product’s perceived performance (outcome) in relation to his or her expectations. If the performance falls short of expectations, the customer is dissatisfied[4].

According to Merriam - Webster’s dictionary and Thesaurus (2012), satisfaction means

- (a) Fulfillment of a need or want
- (b) The quality or state of being satisfied (contentment) and

USES OF ICT FOR STUDENTS SATISFACTION

In the era of technology, ICT aids plenty of resources to enhance the teaching skills and learning ability. With the help of ICT audio visual education has become easier. The learning resources are widening. Now with this vivid and vast technique as part of the ICT curriculum, learners are encouraged to regard computers as tools to be used in all aspects of their studies. In particular, they need to make use of the new multimedia technologies to communicate ideas, describe projects, and order information in their work. ICT has provided immediacy to education.

Audio-Visual Education, planning, preparation, and use of devices and materials that involve sight, sound, or both, for educational purposes. Among the devices used are still and motion pictures, filmstrips, television, transparencies, audiotapes, records, teaching machines, computers, and videodiscs.

Higher education systems have grown exponentially to fulfill the demands of quality education for all due to advancements in Information and Communication Technology (ICT). Demand for skilled and competent labour is ever increasing in the contemporary globalized society. Recently the National Knowledge Network (NKN) has been setup with the objective to interconnect all institutions with a high speed data communication network to facilitate knowledge sharing and collaborative research.

The main applications NKN is focusing to cover in coming time are

- Virtual Classroom
- Immersive Classroom
- Remote Synchrotron Access
- Country Wide Classroom
- Collab CAD Design
- Webcasting Services (Live & VOD)
- CDN (Content Delivery Network)
- URL Filtering Services (out of band)
- DDOS (Distributed denial of service)
- Protection Services

The Mission has two major components viz., (a) content generation and (b) connectivity along with provision for access devices for institutions and learners. It plans to focus on appropriate pedagogy for e-learning, providing facility of performing experiments through virtual laboratories, on-line testing and certification, on-line availability of teachers to guide and mentor learners, utilization of available Education Satellite (EDUSAT) and Direct to home (DTH) platforms, training and empowerment of teachers to effectively use the new method of teaching learning etc. in access and equity in Higher Education[5].

Classroom Climate is very important to promote positive learning and stimulate students who want to learn. The Classroom Climate needs to address all types of learning style.

Positive feeling of students enhances confidence and encouragement in learning. Effective educational contacts demands higher satisfaction level of students, teachers and parents. Satisfaction is linked with the achievements. Students’ satisfaction is their perception of how well a school atmosphere supports educational success.

Strong satisfaction activates students’ analytical reasoning and learning. Students’ satisfaction influences their motivation which is an important psychological factor in their success.

LEARNING TECHNIQUES**IMPLEMENTATION OF LEARNING JIGSAW**

Learning Jigsaws are used to help students become actively engaged in their learning. Students learn a piece of content and then teach it to their fellow students. This activity takes about an hour of class time. Learning jigsaws work best when the content can be chunked into smaller pieces.

1. Divide the class into home groups
2. Have students number off in their home groups
3. Have groups of liked numbered students gather in a specific location called corner groups
 - Have handouts (enough for every student in class) at each corner group as well as instruction on what the students are to complete at the location
4. Students gather enough handouts and return to their home groups
5. Students then teach their newly learned material to their home groups
6. Conduct a class debrief at the end of class to ensure students have no questions regarding the content from the jigsaw.

IMPLEMENTATION OF THINK-PAIR-SHARE

Think-Pair-Share is best used for questioning throughout class. Present the question. Invite students to silently prepare an answer for 1-2 minutes. Then, seek answers from various groupings; invite groups to vote on the answer etc.

IMPLEMENTATION OF DIAGNOSTICS

Diagnostics are low-risk activities that are great for introducing a new topic. Prior to beginning a new topic present a case to the students that can only be solved with information from the new topic.

IMPLEMENTATION OF BOOKENDS

Bookends learning technique is best used for organizing activity; helping students connect to a topic; assessing pre-existing knowledge.

Prior to beginning a new topic, have each student or group of students call out questions they would like answered about this topic; record on chart paper in full view.

IMPLEMENTATION OF THE TRAVELING FILES LEARNING TECHNIQUE

Traveling Files learning technique is best used for skill application practices, building analytical skills and building critique and feedback skills. Put the students into small groups. Each group receives a different case study or a problem requiring a solution, which they complete using guided questions.

IMPLEMENTATION OF MINUTE PAPER

Minute Paper learning technique is best used as a summary activity. Ask students to spend one minute at the end of class summarizing a key idea from the lesson. It could be answering a key question, defining a key term, or expressing what they found most surprising, most significant about the lesson, or how they will apply the lesson in their lives whatever best suits the content.

IMPLEMENTATION OF TEAM PAPER ACTIVITY

Team Paper learning technique is best used to support small group discussion; scaffolding to large group discussion. Provide chart paper and markers. Students sit in small groups surrounding their paper. Pose the issue to be discussed. In silence, students write point form responses all over their paper for a set time period (5minutes). Once time is up, small groups discuss their results and add more as they talk. After the discussion is over, post the papers for a "poster walk" for groups to see others' results.

IMPLEMENTATION OF NUMBERED HEADS TOGETHER

Numbered Heads Together learning techniques is best used for content review prior to an assessment or to liven up a lecture session. Put the students into small groups (4-5) and number off within each group. Post a multiple choice (review or application) question on PowerPoint or overhead. Give 1 minute for groups to choose an answer and then call out one of the numbers (1-5). All students with that number in the group stand and call out the answer.

IMPLEMENTATION OF PRAIRIE FIRE LEARNING TECHNIQUE

Prairie Fire learning technique is a great way to review content or application exercises. This technique will help students develop a study sheet.

Develop a number of review or application questions and prepare handouts with all questions and room to write answers. Divide the class into small group; assign one or more questions per group.

IMPLEMENTATION OF TAKE A STAND LEARNING TECHNIQUE

Take a Stand learning technique encourages students to express an opinion and provide their rationale[6]. Students stand and form a line across the classroom, with one extreme opinion at one end, and the opposite opinion at the opposite end (neutral or undecided students stand near the middle). Walk along the line with a “microphone” (a chalk brush will do), interviewing students on their opinions and querying their reasons.

IMPLEMENTATION OF COCKTAIL PARTY

Cocktail Party learning techniques is best used for helping students deduce underlying principles from multiple examples; topics that students are likely to have personal experience with. Students stand, and form into pairs. When directed, they exchange stories or experiences in relation to the given topic. On the signal, students mill around the room, re-forming into new pairs and exchanging stories again. Repeat 5-6 times.

IMPLEMENTATION OF MEET YOUR MATCH

Meet Your Match learning technique is best used for reviewing terms and definitions prior to assessment. Prepare file cards, half with terms, half with matching definitions (only one card pair per term, do no repeat terms). Divide the class in half and randomly distribute terms to one half, definitions to the other. Students must move about the room, matching terms and definitions. As students pair up, they stand close together. When everyone is paired, terms and definitions are read aloud.

IMPLEMENTATION OF MUDDIEST POINT TECHNIQUE

Muddiest Point learning technique is a simple yet effective technique you can use. This technique provides a high information return for a very low investment of time and energy.

The technique consists of asking students to jot down a quick response to one question: “What was the muddiest point in?” The focus of the Muddiest Point assessment might be a lecture, a discussion, a homework assignment, a play or a film.

- ❖ Determine what you want feedback on: the entire class session or one self-contained segment? A lecture, a discussion, a presentation?
- ❖ If you are using the technique in class, reserve a few minutes at the end of the class session. Leave enough time to ask the question, to allow students to respond, and to collect their responses by the usual ending time
- ❖ Let students know beforehand how much time they will have to respond and what use you will make of their responses.
- ❖ Pass out slips of paper or index cards for students to write on

IMPLEMENTATION OF STUDENT GENERATED TEST QUESTIONS

Student Generated Test Questions allow teachers to see what their students consider important or memorable, what they consider to be fair and useful test questions and how well they can answer the questions other students have posed[7]. Learner’s responding to these questions helps them to assess how well they know the material.

Timing

This particular technique should be used at the end of any of the following

- ❖ Presentation or lecture
- ❖ Discussion
- ❖ Reading assignment
- ❖ Unit

The Student Generated Test Questions learning techniques is best administered two or three weeks before a major test, this will allow students adequate time for feedback and studying adjustments.

Feedback

It is extremely important to give students feedback on how their questions compare to the actual test questions.

- ❖ Provide e-mail feedback to individual students
- ❖ Provide additional resources / review materials if needed
- ❖ Students answer each other’s questions

IMPLEMENTATION OF ONE SENTENCE SUMMARY

One Sentence Summary techniques involves having students answer the questions “Who does what to whom, when, where, how and why?” (WDWWWWHW) about a given topic. The students must do this in one informative and grammatical sentence (usually a long one). This technique will help faculty find out how well students can concisely and appropriately summarize information on a selected topic.

- ❖ Select an important topic that you expect your students to be able to summarize
- ❖ Try to answer the WDWWWWHW question yourself, as quickly as you can
- ❖ Give students about twice as much time as it took you to come up with the sentence

IMPLEMENTATION OF CONCEPT MAPS

Concept Maps are drawings or diagrams showing the mental connections that students make between a major concept stressed in class and other concepts they have learned. This technique provides an observable and assessable record of the students’ conceptual schemata (the patterns of associations they make in relation to a given focal concept)[8]. Select a concept that is both important to understanding the course and relatively rich in conceptual connections to use as the stimulus or starting point for the Concept Map.

1. Before class, create your own concept map to determine if the topic lends itself to the mapping process.
2. Proceed to have your students draw their own maps, either individually or in groups. Give them the directions and show a simple example of a concept map
3. Begin the process by brainstorming for a few minutes, writing down terms and short phrases closely related to the stimulus.
4. Draw a concept map based on your brainstorming, placing the stimulus in the center and drawing lines to other concepts. It can look roughly like a wheel with spokes, or it might take other forms such as a geographical map, a hierarchical chart, a flowchart etc...

CONCLUSION

So, these techniques play an essential role for students to improve their knowledge and satisfaction. Without these techniques they feel bored to read and understand their lessons.

REFERENCES

1. Angadi G.R. (2016) Professional Development and ICT in Education. A.P.H. Publishing Corporation, New Delhi.
2. Arulsamy S. and Sivakumar P. (2009) Application of ICT in Education. Neelkamal publications Pvt. Ltd., New Delhi, Hyderabad.
3. Binu. B.L. and Sulabha (2013). “Information Technology in Educational Management.” *Edutracks*, Vol. 12(9), pp 15-17.
4. Ming-Lung Wu, Pai-Lu Wu, and Pei-Ving Tasi (2014). “The Relationship between Classroom and Learning Satisfaction on Senior High School Students.” *Asian Journal of Management Sciences & Education*. ISSN: 2186-845X ISSN: 2186-8441 Print Vol. 3, No.1, 158-168.
5. Muthuchamy, I., & Thiyagu, K. (2010). “Higher Secondary Students’ Perception towards Information and Communication Technology at the school level.” *Journal of Educational Research and Extension*, Vol. 47, Iss:1, pp 46 – 57.
6. Nachimuthu, K. and Vijayakumari, G. (2007). “Modern ICT trends in teaching technology” *Edutrack* Feb 6(6).
7. Prashanthakumara. T.M. (2015). “Enhancing Process of Teaching – Learning through Technology in Classroom.” *Edutracks*, Vol. 15(4), pp 21- 23.
8. Surabhi Aggarwal (2015). “Impact of Modern Technology in Improving the Quality of Education.” *Journal of Educational & Psychological Research*, Vol. 5(2), ISSN: 2230-9586, pp 222- 224.

ANATOMIZATION OF MACHINE LEARNING TECHNIQUES AND ITS APPLICATIONS

R. Shinoj Robert¹ and M. Maria Dominic²

Research Scholar¹ and Assistant Professor², Department of Computer Science, Sacred Heart College, Tirupattur

ABSTRACT

Machine learning is a widely accepted and in practice in Artificial Intelligence. It teaches the machine to adopt different concepts and patterns to adapt the finest situation and to hold it. Machine Learning improves performance by acquiring the knowledge from the experience. This paper provides a broad area of studies and the evolution of Machine Learning Techniques (MLT) and also confronts the most commonly used Machine Learning (ML) Algorithms with real time demonstrations. It reviews the different studies and the applications used in Machine learning and the best suited for it.

Keywords: Artificial Intelligence (AI.), Machine Learning (ML), Machine Learning Techniques (MLT)

I. INTRODUCTION

Machine Learning is being an essential and backbone of Information Technology. Machine Learning is used to make exact decisions based on observation and predictions. At present the concept of machine learning is used in many applications of intelligent systems. Machine learning systems are used to identify objects, images, record speech into text, game and news [2]. They are also practiced in pattern recognition techniques such as fingerprint image, handwriting word, human face, speech signal and in DNA sequencing [3]. Machine Learning examines the areas of algorithms that can make high end predictions on data [4]. The idea behind the Machine Learning is if the machine could learn from its experiences its value gets increased. This lead to automatic predictions of values based on the experiences handled by the machines. It leads to accurate predictions of data [5]. The main we focus on this paper is machine learning in which machines are used and trained for correct predictions and to analyze about the different techniques and different algorithms. It is efficient to implement Machine Learning for a researcher to seek an approach in which the algorithm comes with its own set of solutions based on the example or trained set of data, rather than designing an algorithm to the address the specific problem. The machine learning paradigms are classified under supervised learning focuses to learn the input and output that results to specified type of solutions, unsupervised on discovering the patterns which includes clustering and association techniques working on frequently learning techniques and the reinforcement on the learning controls.

II. MACHINE LEARNING MODEL

Learning Process in Machine Learning is classified into Training and Testing [7]. The Training data or set is used to build the model and the testing set is to validate the model. In training process, the input is taken as training data and it is trained by learner (learning algorithm) from which the model is arrived or built. In testing the execution engine is used to make the prediction from learning model. The final predicted data or production data is classified a used to build the Tagged data.

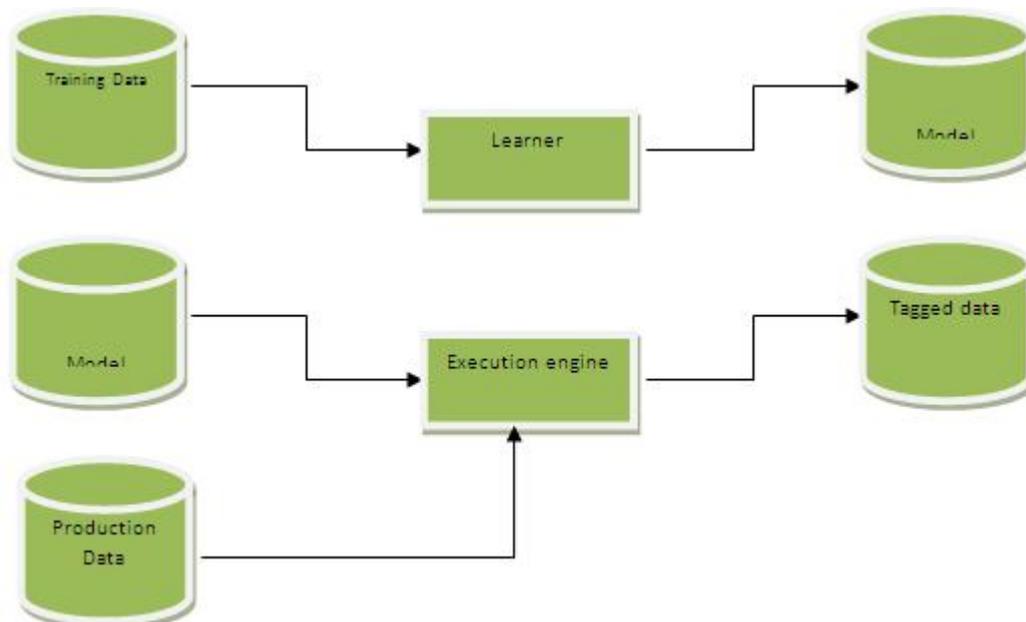


Fig-1: Machine Learning Model [7]

III. MACHINE LEARNING STYLE

There are number of ways that an algorithm can model a problem which is based on the interaction with an experience or environment or input data [4]. First we need to adopt a learning style that an algorithm can adopt. An Algorithm can adopt only few learning models. The steps involved in Machine Learning is much useful which makes us to think on the different opinions about the type of input data that fed inside and selecting the most correct data to achieve the target result. Machine Language is to overcome heterogeneous types of problems and to give the solutions as a clean prototype [6]. Let see the different learning style and in machine learning algorithms and different components.

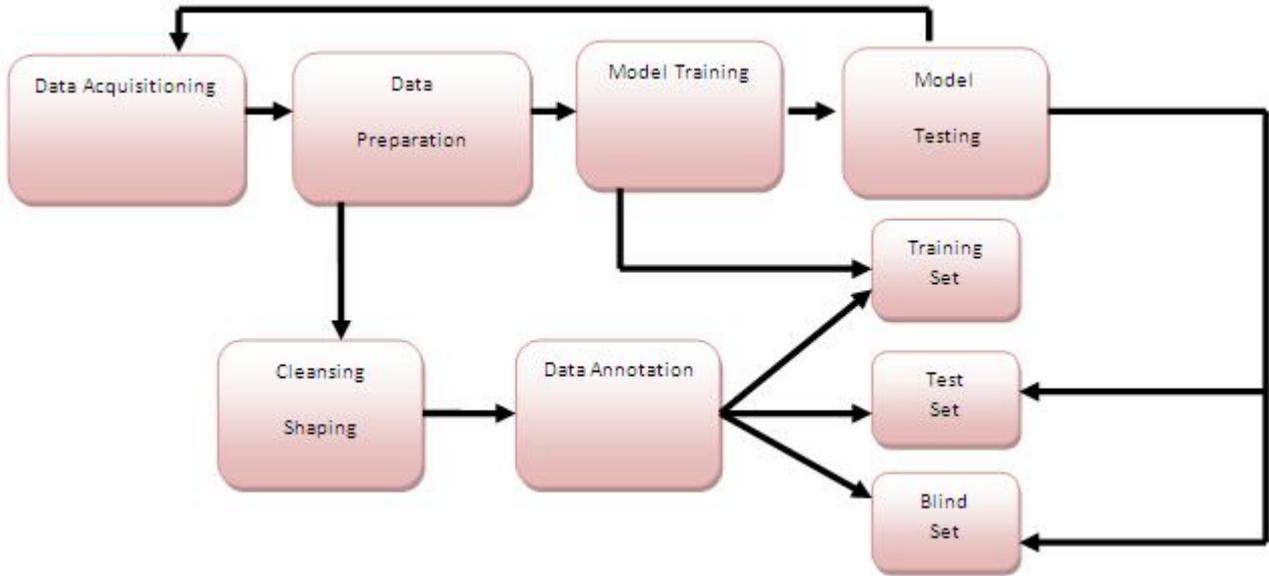


Fig-2: Machine Learning Flow Diagram

IV. MACHINE LEARNING TECHNIQUES

The Machine Learning Techniques are classified under four major categories depending upon the nature of the learning the system differs.

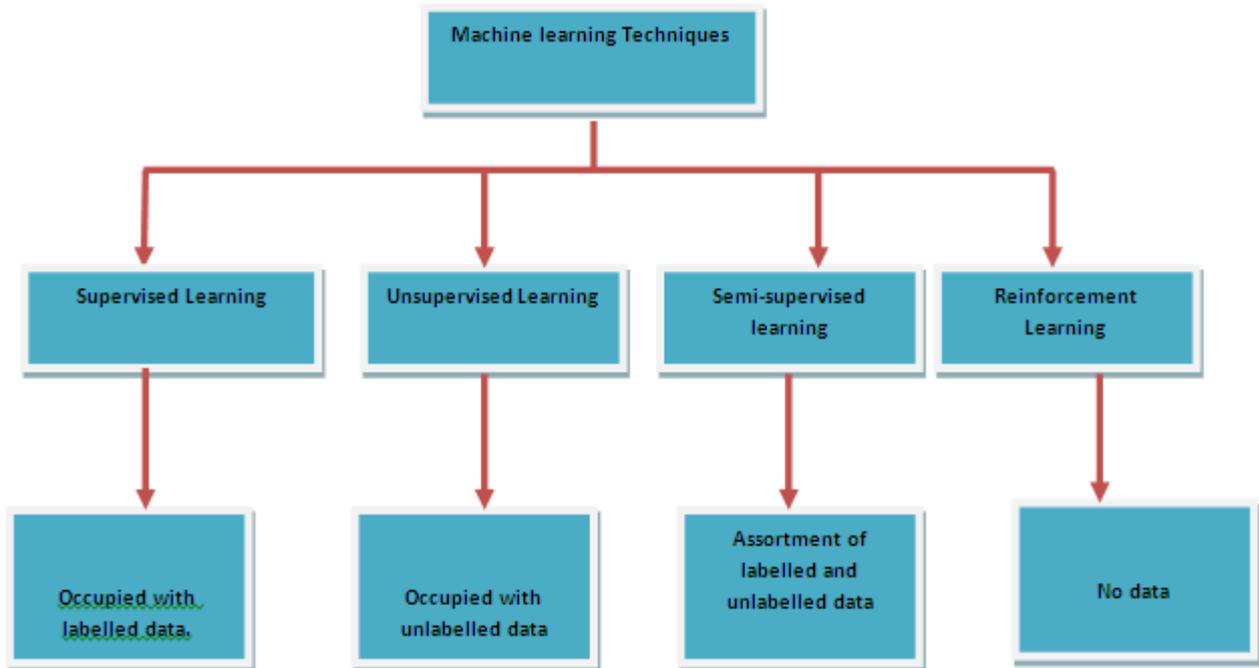


Fig-3: Machine Learning Techniques and the required data

SUPERVISED LEARNING

Supervised learning focused on monitoring. Supervised Learning (SL) is to train the particular set and get the corresponding answer to it. It is used to predict the values for the target function and the desired output is known. The concept of approximation is fixed in supervised learning. The Input data or training data has a pre-determined labelled such as True or False, positive or negative, spam or not spam etc [7]. The target results are

known and given as input to the model during the learning process. In these supervised learning the raw data that may be in text, document or images supervised learning (SL) allows dataset consisting of both features and labels. The purport of this method is to find by comparing the actual output with the “learned” output to find out the errors, and to do the modifications accordingly by the learning algorithm. It used patterns to predict label values on an unlabelled data. Eg. In supervised learning, an algorithm is fed with the images of leopard labelled as cat and images of forest labelled as trees. By trained on this data, the supervised learning algorithm should be able to later identify unlabelled leopard images as cat and unlabelled forest images as trees. A real time example for SL is prediction for rain. To see the predicted results it need to be filled with parameters inside the label such as the level of the humidity, temperature and the possible directions of the wind. If all these scenarios match there is a possibility of raining. Inputs are trained to the computer with the data. Each data is assigned with some representations as 0 if it does not rains today and 1 for the possibility of raining. If the humidity is high and the temperature is high it rains. Basically supervised learning is provided with answers and the input. SL is used to predict historical data predict statistically future events. Sentiment analysis, image recognition and speech technologies are the areas that have made improvement by applying supervised techniques by equating human performance. In the case medical field for the event of tumour detection and for the expensive expert’s services where human label data are not sufficient, the techniques of SL are applied.

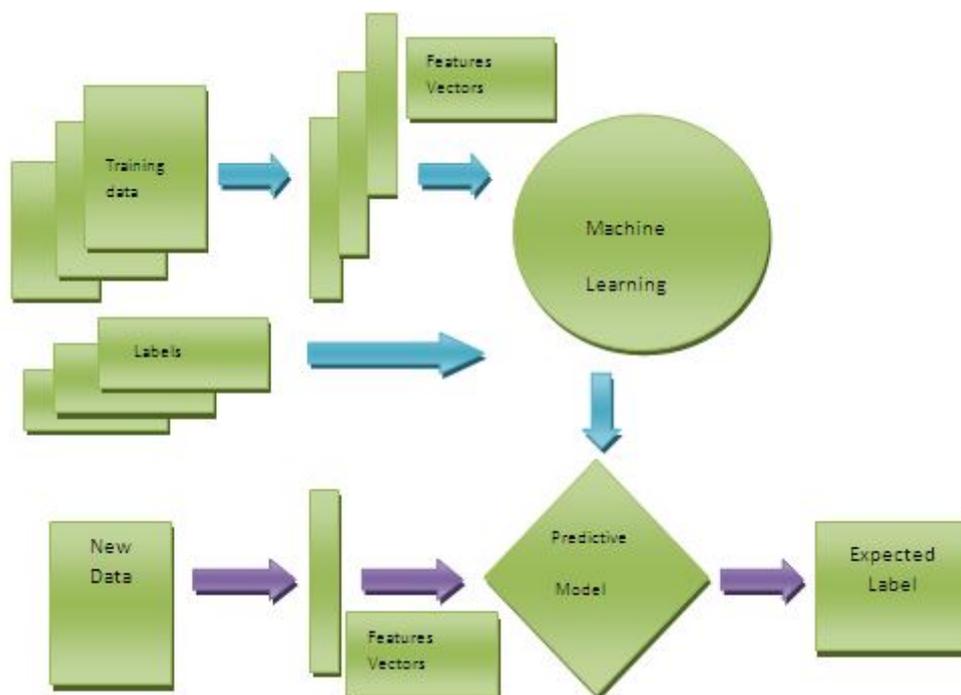


Fig-4: Supervised Learning framework

UNSUPERVISED LEARNING

In Unsupervised Learning there is a deficiency of training data. The data is unlabelled in unsupervised learning. The approximation is to find the hidden structure in this data [14]. The learning algorithm is used to find out the similarities between the objects from the training (input) data. Here the labels of the data are unknown and discovered from the data itself. In clear, the learning algorithm uses a single n training data sets and n features without fixing any labels. In Learning algorithm of the training process the unsupervised learning builds the predictive model which will try to find its predictive model situated to it. The typical unsupervised learning is the form of clustering where the set of inputs is divided into groups [8]. The unsupervised techniques are included in self-organizing maps, nearest neighbour mapping, k-means clustering and singular value decomposition. Unsupervised learning provides inputs based on patterns. When structure is needed in classification of data unsupervised learning is recommended. Here the data are not labelled. An example for unsupervised learning is some inputs based on fruits. Here the inputs are not given in explicit, instead the parameters based on the size, colour and taste of the fruit. When the set of conditions or parameters is given to computers, it groups the fruit based on size, colour and taste. The computer can identify through some patterns whereas in supervised the actual input is given. The problems in unsupervised learning are it could not correlate the data or structure of the data. When there is a need of finding a structure unsupervised learning is preferred. It's the job of the algorithm to give pattern and to move ahead. Unsupervised is used efficiently once the structure is unknown.

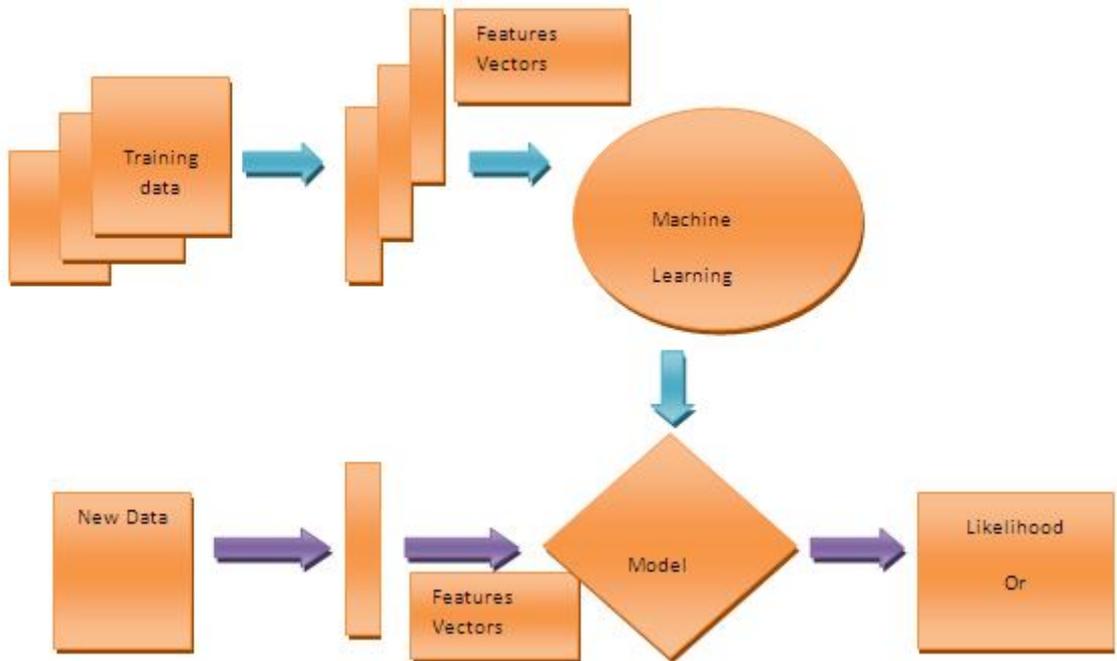


Fig-5: Unsupervised Learning framework

SEMI-SUPERVISED LEARNING

Semi-Supervised Learning (SSL) algorithm is a combination of both labelled and unlabelled to generate appropriate functions or classifier [9]. In this type of Learning the given data is a mixture of both classified and unclassified data. The combination of these two generates an appropriate model for the better classification of the data. In supervised it add the labelled data and in unsupervised it concentrated more on unlabelled data, combining both in semi-supervised it uses both labelled and unlabelled data for training. Semi-supervised learning endeavours the predictive model using both labelled and unlabelled data. The expected result that derived from semi-supervised learning is better than using any one labelled data set. The SSL using trained algorithm tries to predict the value of unlabelled algorithm. It could be explained with real time example. To identify the result analysis of the student with the current semester and prediction for the next semester can be made by using SSL.

Table-1: Marks of the student in a semester

Student (Gender)	Mark in Particular Subject	Semester
B	55	3
G	67	3
G	77	3
B	43	3
B	59	3
?	62	4
?	67	4
?	40	4
?	55	4

The student gender is classified as Boy and Girl. The marks scored by the particular Boy or Girl are mentioned and it is called as labelled data. The unlabelled data are classified without the reference of the gender. For classifying the student mark list, use the supervised algorithm which includes labelled data and train them to get a model data. With the estimated model in unlabelled data find the class probability and try to find the class probability that devolves as expected category. Later train the supervised algorithm with the expected value from the unlabelled data. Repeat the step until convergence. Another real time scenario for semi supervised learning is document classification. It works with the keywords, attributes and assigns the page classification; with the page classification the labelling of the pages could be found. The student could predict the subject book with the indications of word classifications “Constraints”, “Primary key”, “Auto Increment” that the page speaks about specifically database. Semi-supervised learning can be used with methods such as classification, regression and prediction. Google Expander is the progression from semi-supervised learning [10]. Speech analysis is another application of from semi-supervised learning here audio file labelling is done through SSL

(Semi-Supervised Learning) instead by human resources that can help to improve traditional speech analytic models. Protein sequence classification is a learning technique of SSL and its help to deduce the functions of proteins in human body which involves human interferences. Identifying persons face through the web cam is a current technique applied in SSL techniques. The disadvantage prevails in semi supervised learning is the unlabelled data is not equal to the labelled data.



Fig-6: Semi-supervised Learning Framework

REINFORCEMENT LEARNING

It is a type of Machine Learning, where the computer trying to take decision. It acts as a subdivision of Artificial Intelligence. Reinforcement Learning (RL) is sometimes called as unsupervised learning. Reinforcement is a learning situation where a machine can learn to behave it situation where feedback is not immediate [11]. The goal of RL algorithm is to learn the policy which maximizes the long term performances. RL provides the trial and error component and does not totally rely on unsupervised learning. Some amount of supervised learning input is given as feedback from the environment. The ability of learning to control the system through trial and error and minimal feedback is the actual task performed in RL. The RL Framework consist of learning agent which could be performed as supervised or unsupervised learner, the environment which works along with the agent as the close interaction. The agent senses the state of environment and performs action. The action in turns changes the state of the environment. The learning process of agent starts from the close interaction with the environment. The result of the action is not constant and the subject to the nature of changes in the environment. The performance of environment is measurable in each task in the form of scalar evaluation through rewards. RL has a reward function that act as an agent where an agent learns how to behave in the environment by performing actions and seeing the result. The algorithm is trained to map the actions according to the situations in order to maximise the performance. It learns from the closer interactions with the environment and results in rewards. The reinforcement learning (RL) has three major components: the learner, environment and action. Here the learning and action are a simultaneous.

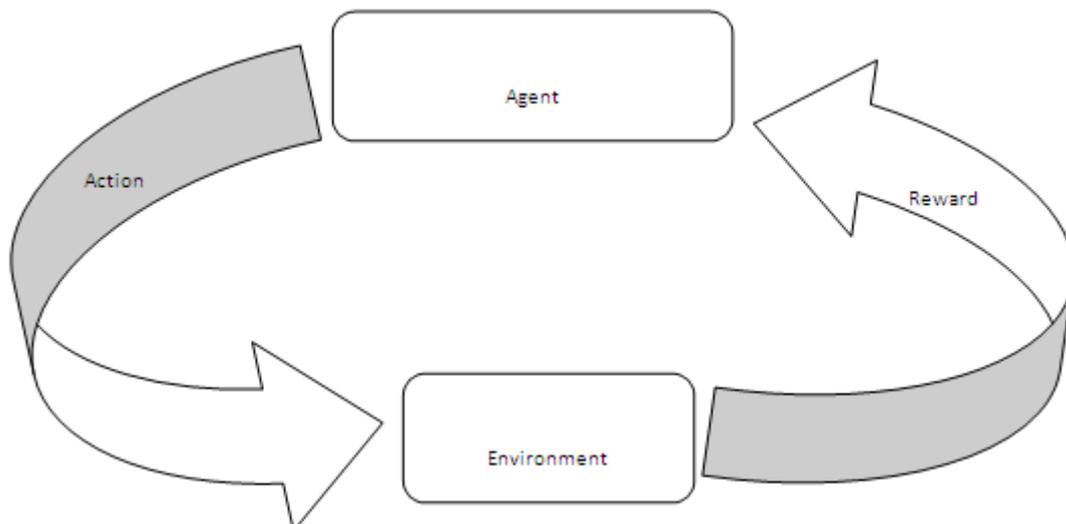


Fig-7: Reinforcement learning framework

The real time example of RL is applied in the personalization of adaptive skills of the student behaviour in tutoring systems. The students are classified according to their IQ level. Different set of pictures are given to the students and make them to identify using their skills, when some level of student able to perform better are given positive reinforcement by rewarding them through encouragement and for the other set of student reinforcement is provided through breaks and hints which also helps for the completion of actions. Here the environment used inside the class room and the materials. The algorithm is trained to react to an environment. The applications of RL include games with the features of artificial intelligence like transportation car games

where the input is taken from the screen as pixels and the output is through the joystick controls. Robotics as the agent that performs autonomous action and navigation, adaptive control in the field of aviation where the helicopter is trained to fly at a human level competence RL is applied. The RL algorithms are applied in the applications of optimization problems and personalization like adaptive in students behaviour and intelligent tutoring systems. In specific example for RL is classified with playing chess, the computer cannot be trained to move all the stages. The only possible explorer is right move or wrong. A classic example is training a dog where it does not understand instead you reward when it does well. The computer is used to maximize it works. Some of the applications are Temperature Controller system

IV. MACHINE LEARNING ALGORITHM (grouped by similarity)

Table-2: Model of Machine Learning and its techniques

Machine Learning (Types)	Methods	Algorithm	Applications
Supervised Learning	Regression Support Vector Machine (SVM)	Linear Regression Logistic Regression Stepwise regression Multivariate Adaptive Regression Splines (MARS) Locally Estimated Scatterplot Smoothing (LOESS) Gaussian process regression Multinomial regression.	Weather Report Analysis Detecting Credit card fraud Stock market analysis Bioinformatics
	Bayesian statistics ANOVA(Analysis Of Variance)	Naive Bayes Bayesian Belief Network Gaussian Naive Bayes Multinomial Naive Bayes Bayesian Networks Average-one Dependence estimators (AODE)	Predicting the weather feature.
	Decision Tree Artificial Neural Networks (ANN)	Classification and Regression Tree (CART) C4.5 C5.0 Chi-squared Automatic Detection Decision Stump Conditional Decision Tree M5	Risk Assessment Threat Management System Predicting Financial Results
Unsupervised Learning	Clustering	1.1 K-Means 1.2 K-Medians 1.3 Mixture Models 1.4 hierarchical clustering Expectation Maximization(EM)	Image recognition Natural language processing Medical Imaging, Healthcare monitoring
	Neural Networks Method Moments Blind Signal Separation Techniques	Dimensionality Reduction Hebbian Learning Principal Component analysis, independent component analysis, Non-negative matrix factorization, singular value decomposition.	
Semi-Supervised Learning	Generative Methods Low-Density Separation Graph Based methods Heuristic approaches	Supervised Classification Unsupervised Clustering Semi-supervised classification.	Adaptive website Webpage classification Search engines
Reinforcement Learning	Brute Force Temporal difference (TD) methods Monte Carlo methods SARSA Direct Policy Search	Q-Learning - Policy Iteration - Value Iteration State-Action-Reward-State-Action (SARSA) Deep Q-Network. Deep Deterministic Policy Gradient. Pendulum problem	Robotics and industrial automation Inventory management Delivery Management Power systems Finance Sector

Deep Learning	Artificial Neural Networks	Gradient Based Learning Deep Boltzmann Machines Back-Propagation Stochastic Gradient Descent Learning Rate Decay Dropout Max Pooling Batch Normalization Long-short-term Memory Ski gram Continuous bag of words.	Computer Vision Speech recognition Image Caption Generation Character Text Generation Automatic Machine Translation Colorization of black and white images.
----------------------	----------------------------	---	--

There are many algorithms that the machine can learn and here are the some machine learning techniques. The algorithms are classified according to the learning styles. The learning algorithms are grouped by similarity.

I. Supervised Learning

A) Regression Algorithms

Regression is pertained with modelling the relationship between variables. Regression is considering about predicting the value that continuously varies within the relationship. The dependent variable is a target variable and the independent variable is a predictor. A real time example relationship between students learning and number of test conducted by a teacher is analyzed through regression.

1.1 Linear Regression

It is one of the most indeed and inferred algorithm in machine learning. Linear Regression is one of the most former methods of machine learning. It is a statistical approach of fitting a line to some data. Regression comes under the supervised learning category. Linear regression attains to model the relationship between two variables by fitting the linear regression to the observed data. One variable is considered to be explanatory variable and the other variable is considered to be depended variable. Here the dependent variable are need to be explained or focused. The explanatory or independent variable which explains the other variable and the values are independent. The dependent variables can be denoted as (y) and the independent variables are denotes as (x). Here both the input and output values are numeric. Linear regression can be applied as two ways. Two find out whether there is any relationship between two variables and two find out the whether they have statistical significant relationship between two variables. The real time examples for linear regression are to find out the co-relationship between the employee experience and his/her salary. The Salary is dependent upon the experience of the employee. Linear Regression can be used for predictions. The best name to put forward to linear regression as line of best fit. The line is not at straight hence the best fit line between the patterns. The minimal point in the data series is named as residual. The linear regression applications are used effectively in analyzing, market strategy, stock market pricing, promotion of sales of a product, forecasting the business strengths, weather data. Linear Regressions are used to predict the probability of an event, financial predictions, finding the cost estimation for software products, helps in organizing the structure of an organization in financial needs.

1.2 Logistic Regression

Logistic regression is a classification technique from Machine Learning. The logistic regression is used for binary classifications. It carries the similar features of Linear Regression. The algorithm carries with a base line. The logistic function has a value between 0 and 1(True and False). Unlike Linear regression, the prediction of the output is transformed using a non-linear function called logistic function. Logistic Regression is used for categorized target whereas linear regression is in use of achieving continuous target. Some variations in the names are called similar functions or logit. The logistic function also called similar function developed by a statistician David Cox to describe the properties of the population growth in ecology rising quickly. Logistic regression is applied in healthcare, Geographic image processing, image segmentation and categorization. There are less natural manner regression types like multiple regressions where only one outcome variable and many covariates, these applications are applied in used to find out the effect of many covariates in hospitalizations [15]. The applications of Logistic are forecasting market trends, to find the success and failure rates in results, true or no category in recruiting employees based on their performance in need of employment in a company, image categorization, and healthcare to analyze a group of people affected by Myocardial Infarction, used as prediction to find out the student prediction about the depression towards result.

1.3 Stepwise Regression

The stepwise regression is used to deal with multiple independent variables which are the most essential variable to be considered. The stepwise technique is an automatic process without a human intervention. The

goal of a stepwise regression is to come with a model that is able to predict the variables which were considered to be the fewest variables. Is simple the ability to derive the most outcome with the fewest parameters. The basis for selection of a variable is classified as choosing the preferred variable that satisfies the criteria or ignoring the variable that considered being least satisfactory. The variable that is held back in the model is grounded on the absolute value of t-statistic. This value is considered to be the highest absolute t value. The highest absolute value is seems to be used as most predicted model called as forwarded-selection process. An example for stepwise regression is with y, the number of hours students attend the class per day with the variables. The below variables X can help us to explain the variables to explain the variation in number of hours.

Table-3: Stepwise Regression model of students attended various activities

Y	No. of Hours Attended Student Per day (Independent variable)
X1	No. of Test Taken
X2	No. of class work processed
X3	No. of Written works covered
X4	No. of practical classes attended
X5	No. of hours notes taken
X6	No. of hours listened to the classes
X7	No. of hours spent in the canteen

The variables listed (X) are too many to predict the nearest variable that models the Y variable. It needs to be classified with the datasets where each explanatory variable compared with the independent variable (Y) individually. The comparison can be made individually to find the nearest predicted model. Repeat the same for(X2, X2...X7). The X5 is the first variable as the largest absolute t-value that predicts the model in the first run and added to the model. The next predicted model and the absolute value is continued and added to the model. The disadvantage of the stepwise regression is there is a chance of selecting unimportant variables which has only absolute t-values but not the real predictive model and excluding important ones. The values of t could lead to mislead.

B) Support Vector Machine

Support Vector Machine solely classified under supervised learning model. It is a classification technique. SVM (Support Vector Machine) is a popular machine learning technique that can be a group of its own [12]. The SVM algorithm is a best frontier which best segregates two classes (Hyper-Plane and line). SVM can be used in large datasets. It is also known as LSVM (Linear Support Vector Machine). SVM can be better explained with the decision boundary. Since SVM in classification technique, decision boundary need to be drawn. The question is to make a decision boundary as best.

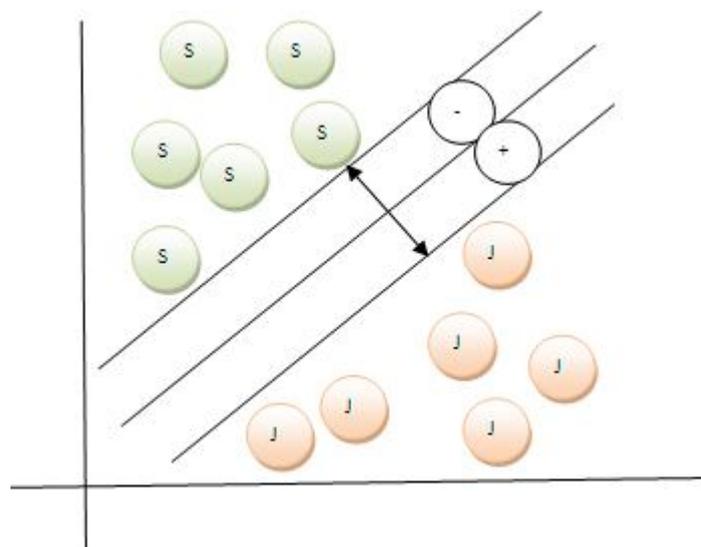


Fig-8: Support Vector Machine classification of students

With an example student in comparison of two classes as a real time example, where the S indicates the Senior class student and J indicates junior set of students. The line segregating the two different classess matching to the height and weight need to be matched. While classifying sometimes optimized decision boundary could result in greater misclassification. The line on decision boundaries that segregates the two boundaries is commonly referred as a hyper-plane which act as an extreme point in dataset. The distance between the hyper-

plane and the closest data set is denoted as margin. SVM define a state of point margin position. The algorithm that basically support vectors are important whereas other training examples are ignorable. Here D^+ denotes the closest positive points. D^- The shortest distance to the closest negative. The advantages of support vector machine are effective in high dimensional spaces and different kernel functions for various decision functions. The disadvantages are poor performance when features are more than samples SVM (Support Vector Machine) do not provide probability estimates. The applications where SVM is applied are Medical Emerging, Medical classification doubts, Time series prediction, Financial Analysis, pattern recognition for Machine diagnosis, Page ranking in Google and object recognition.

C) Bayesian Algorithms

Machine Learning is to be a combination of statistics and computer science applied algorithms. The cases of uncertainty are represented by statistically in order to find the quantification of results, the Bayesian algorithms are used to represent the uncertainty grounded on probability theory [7].The Bayesian algorithm solves classification like text classification and regression type problems. Some of the algorithms supported by Bayesian types are Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AOE), Bayesian Belief Network (BBN), Bayesian Network (BN) etc. The Naive Bayes classification are applied in spam classification to predict an email whether the received mail is spam or not, Medical diagnosis to help the patient about diagnosing the type of disease affected the patient. To predict weather based on temperature, humidity, and the rainfall possibilities. The regularly used Bayesian algorithms are listed below.

1.1 Naive Bayes

The naive Bayes algorithm is based on bayes theorem, which is also called as conditional or trust theorem. Naive Bayes derive from the classification techniques based on Bayes’ theorem. Bayes theorem is a mathematical formula to evaluating the level of trust evidence. Naive Bayes is a simple but powerful algorithm for predictive modelling. In Naive Bayes each input variable is considered as independent. Decisions are taken from the predictive models which affects both the scenarios.

Table-4: Probability type in Naive Bayes

S. no	Probabilities
1	The probability of each class
2	The conditional of each probability

With real time example Education institution are more likely to select student for the opted discipline based on the cut off marks. If the student scored a good mark only in language papers (English & Tamil), it is not worth in selecting him or she for science and with the marks of science, maths to choose them for literatures is not worth. The circumstances for selecting the students requires datasets, from which the objective of estimating the likelihood of student selected or not through the yes or no condition. These predictive models of each class independent variable will decide the dependent variable whether or not to select the students. Based on these data the problem of selecting the students is approached through Naive Bayes algorithm. Naive Bayes consist of two major probabilities that can be calculated directly from the given training data

Table-5: Independent data sets model of grades of the students

Sno	Maths	English	Science	Social	Tamil	Selected
1	A	B.A	A	E	G	No
2	G	E	A	G	A	Yes
3	A	A	G	A	G	Yes
4	B.A	A	E	G	G	No
5	A	G	A	E	E	Yes
6	E	E	G	A	G	Yes
7	G	G	B.A	E	G	No
8	A	A	A	G	A	Yes
9	E	E	G	A	B.A	No

(A) = Average, (B.A) = Below Average, (G) = Good, (E) =Excellent.

First the determination of yes and no possibilities need to identified. The probability of class selected is given as $P(C)$. The probability of Yes is given as $P(Y)$ and no as $P(N)$. Count the probability of yes separately $P(Y) = 5/9$ which is five total yes in nine possibilities. In the possibilities of $p(N)=4/9$, which is four total no in nine possibilities. The individual probabilities with respect to each features of selecting the students in the datasets

are calculated for the other possibilities. It makes easy to predict the class of test data. Comparing to other predictive models Naive Bayes performs better comparison results. The cons in Naive Bayes is seems to be bad estimator, the outputs of the probability are not taken serious. The applications of Naive Bayes are applied in decision making factors especially on games, institutions on the possibilities of conducting an event, predicting the weather temperature, the algorithm can be used in multi class prediction of target variable. It is mostly used in text classification to give better results in multiclass problems, equating to other algorithm it has a higher success rates in finding out the independence variable. Naive Bayes algorithm used in recommender systems, credit scoring and medical date classifications. It is also used spam filtering to identify the spam emails and sentiment analysis in social media analysis to identify positive and negative customer sentiments.

1.2 Gaussian Naive Bayes

A Gaussian Naive Bayes algorithm is known and a special type of Naive Bayes classifier type. Gaussian Naive Bayes is applied when the features have continuous values. The class density follows normal distribution. The Gaussian Naive bayes can be explained with an example of the height of the student that can be compared with the gender. For the height of given student need to find out the gender. The two classes are height and gender. In Gaussian Naive Bayes algorithm the continuous variable is need to be found out. When attributes are continuous, based on the assumption the values are associated with each class allowing them by normal distribution [16].

Table-6: Dataset Model of Student Heights and gender

Height of student	gender
5.5	M
5.3	F
5.7	M
5.5	F
5.3	F
5.6	M
5.6	M

The data is given in the class of height and gender of the student. For the height of 5.8 the gender of the student is to be identified. The probability of finding the gender can be done through $P(\text{class}/\text{height})$. The Gaussian classifier can be applied by combining these two classes $P(G/\text{height})$ and $P(B/\text{height})$. Computing these two probabilities the gender could be identified. To find the height it follows Gaussian distribution. After computing the distribution with the average height of the girl and the boy, it could be concluded that the given height belongs to the gender male. The applications are to find the continuous set of problems, Recommendation systems, predicting the weather conditions. The advantages of Gaussian Naive Bayes algorithm is can make the predictions fast and continuous. The cons seen in the Gaussian Algorithm is cannot perform regression and more time spent on unknown independent values.

1.3 Multinomial Naive Bayes

Multinomial Naive Bayes in one of the standard classic algorithm preferred to use multinomial distributed data. The applications of Multinomial Naive Bayes are used in text categorization or text classification which is used to constitute the occurrence of the word in a document. It explicitly calculates the word counts. Multinomial naive bayes are applied to find the pattern recognition, Human Language processing, word disambiguation it deals with the finding of correct translation of a word in a sentence from the given language, Text classification [17]. The multinomial model captures the word frequency information in documents; the document is ordered sequence of events. The length of the document is independent class [18]. In text classification the target can be either positively classified or negative classified. The documents can be represented as V , an array of words or Vector of word. The training examples are used for classifying text documents. The training examples are predicting the possible outcomes, the probability of each word with a document either positive or negative. The probability of the document belonging to the certain classes is denoted as $P(\text{doc}|V) = \text{length}(\text{doc})$. The probability of a word in a position is referred as W . The text classification could be explained with an example with the set of students review upon the subject as a real time one.

Table-7: Text classification model using multinomial naive bayes

Doc	Text	class
1	I love the subject maths	+
2	I hate the subject maths	-
3	Good subject. Interesting subject	+

4	Poor subject maths	-
5	Poor subject.	-

The set of student review is called set of documents. The list of document tells about the student review about math subject. When the review is positive it denoted with + and for negative review -.

Table-8: Text classification model using multinomial naive bayes

Unique words
I, love, the, subject, maths, hate, good, interesting, poor

Count the unique words and separate for converting the documents into feature vectors. Here the attributes are mentioned as possible words and the values are occurrence of the word in a document.

Table-9: Text classification model using multinomial naive bayes

Doc	I	Love	The	Subject	Math	Hate	Good	Interesting	Poor	Class
1	1	1	1	1	1					+
2	1		1	1	1	1				-
3				2			1	1		+
4				1	1				1	-
5				1					1	+

The first document contains the review about the student which was made. It occupies the four classifications. The following condition is applied for the rest of the document. Some documents walk out with minimum classification. The Vectors 1,3and 5 gives the positive outcome. N be the number of words in the cases and K the number of times the word occurs in the cases. The Multinomial Naive Bayes are applied in text classification, information retrieval, it is used in medicals to find out the blood types and bone marrow types for finding the match types and calculating probabilities.

D) Decision Tree

Decision tree are from the type of supervised Machine learning used in classification problems. Decision trees applied to the situations to find the possible solutions to a problem based on certain constraints. Decision tree algorithms results in the quick and precise data formats. The entities in the trees are decision nodes and leaves, the leaves represent the final outcomes and the nodes represent the decision represents the data [19]. Decision tree classifications are easy to read and understand, and the models used in the decision tree are explainable. It classifies the result based on the model that is easy to understand. Some of the popular decision tree algorithms are classified below. The Decision Tree in Machine Learning are classified by two main types Classification and Regression.

1.1 Classification and Regression Tree (CART)

A Decision tree is to represent visually and explicitly represent decisions and decision making.

Decision Tree is a flow chart like structure where each internal node denotes the attributes. The attributes represents the branch, the branch represents the outcome of the test. The leaf is considered to the terminal or the label with the dataset. The top most nodes in a tree are root node.

Table-10: Model of Decision Tree Types Table

S. no	Decision Tree	Types
1	Classification	Yes/No
2	Regression	Continuous

Similarities

The similarity between classification and regression trees is Regression type decision trees are left with continuous data types whereas classification trees are used when the variables are independent and categorical. Regression tree values are obtained by terminal nodes using the mean or average response of the observation the region. In the case of classification trees the values of class receive by terminal mode use mode value. The region of splitting process reaches the full height in both the trees. The advantages of CART are simple to understandable, explainable and to visualize. It can handle numerical and categorical data. The non linear relationships data will not affect or disturb the performance of the result. The cons seen in the CART are it leads to over complex where too many results were brought more than the required one, may also known as over fitting. The data collected using decision tree sometimes result in variance, because little variance in data may result in complete different tree structure. For the serious set of problems or solutions better to avoid Decision tree structures. There is a chance of domination in decision trees if there is no balance between the datasets.

The applications used are in Insurance, whether the customer will renew the insurance or not, credit card promotions. In calculating population census, predicting the percentages of gender. Price prediction of housing loans and height predictions of heights gender classification. Decision tree are utilized effectively when it is used with advanced machine learning algorithm like Random Forest and Boosting. The commonly used algorithm in decision tree is listed below.

Table-11: Decision Tree model Classification Table

Decision Tree Algorithms
Chi-squared Automatic Detection
Reduction in variance
Conditional Decision Tree
Gini Index
Information Gain
C4.5 and C5.0
Decision Stump
M5

II. Unsupervised Learning

A. Clustering

Clustering is the type of unsupervised learning in a machine learning technique. It is a form of grouping objects into similar groups. Clustering performs the identification of smaller groups of data in a data set. Clustering is used for analysing statistical based data. The methods of clustering are classified as hierarchical and partition clustering. The centroid- based approaches are applied in the clustering algorithm. Clustering could be explained with real time example, if the class teacher wants to familiar student knowledge in the specific subject, in order to improve the teaching or coaching. It is difficult to look at each student level personally and work out a methodology for every individual. Instead, the student could be classified as a group depending upon the size and analyse their knowledge level, writing skill, listening and understanding and can apply a methodology for each group separately. The each student activities are more similar to other students within the group. The basic technique behind clustering is to group the students and breakdown the larger or heterogeneous group of student into more specific or homogenous type.

Table-12: Model of Clustering Algorithm Classification

Clustering	Algorithms
Hierarchical	Agglomerative Divisive
Partitive	K-means Self Organizing Map

The applications of clustering in the fields of Marketing used to bring out characterization of customers for marketing purposes. Medical imaging for the classification of blood cell relationship, on educational institution in clustering different details and books according to topics, Streaming analysis in IoT, Image segmentations, Financial Transactions, Anomaly detection of fraudulent activities in Insurance. Analyzing the earthquake affected areas.

Clustering Algorithms

1. Hierarchical Clustering

Hierarchical Clustering algorithm implies in constructing from top to bottom or bottom up approach. This algorithm that build each data act as an own cluster. It builds the hierarchical clustering and blends the nearest data point by merging together as one cluster. The representation of each relationship in hierarchical clustering is created by dendograms or binary tree. The dendograms is a multilevel hierarchical view where the clusters at different next to the data point are joined together and form as a cluster to next level. Each level generates new heights of clustering and leads to take the decision for the final clustering. Hierarchical clustering could be explained with a real time example to find out the uniqueness of each student relationship in the class. The features to find the dissimilarities amongst the students could be measured with the closest distance to each cluster and into complex genetic code. With the data sets of cluster it is easy to represent in dendograms. Looking at the students, find out the possibilities of sibling, twins and with the external look that matches each other. The two nearest clusters are matched together to under a single cluster. A vertical line is represented to show the dissimilarities between the relationships.

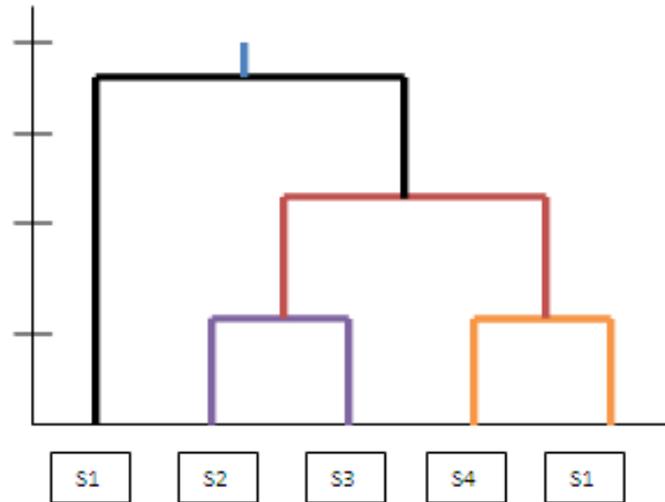


Fig-9: Hierarchical Clustering dendrograms of student relationship

The grouping of student teacher relationship compared with family ways relationship such as father son mother daughter, siblings, twins and in any other existent. The relationships between the two students in the same class consider brother and sister will be clustered under the siblings, the student close to each other could be clustered under the one clustering. Each item with the own cluster finds the best way to merge with new cluster in hierarchical clustering. The applications of hierarchical clustering are applied in recommendation engines, market segmentations to group them as behavioural, geographic, demographic, psychographic, social media analysis to find the positive and negative analysis of statements, search result grouping, to evaluate the student performance at different levels. Hierarchical clustering has two algorithm classifications.

1.1 Agglomerative Hierarchical Clustering

It is a type of hierarchical clustering with a bottom up approach. The Agglomerative Clustering starts with very small object or component and each component to its own cluster. Clustering is performed by computing the distance between each of the cluster. The similarities between each component will merge the process to the next level until it reaches to big components. At last single component is left. The distance between each clusters in evaluated by the proximity matrix.

Table-13: Methods to evaluate the distance between the clusters

Clustering in Agglomerative	
Proximity matrix	Single Linkage
	Complete Linkage
	Average Linkage
	Minimum Variance
	Centroid Method

The dendrograms, showing the tree diagram structure of relationships are used for pictorial representation of Hierarchical Agglomerative Clustering. The procedure to create the list of n students one for each data point assigned to be a cluster, with the idea of leaving only one cluster left. Merge the nearby clusters and the major clustering is to make the computation between two clusters. The proximity matrix contains the distance between each point. The distances between two approaches between the algorithms leave the differences. The classification of student relation can be too taken as real time example. The applications of Agglomerative are used in medical imaging, market segmentation, and anomaly detection.

1.2 Divisive Hierarchical Clustering

The Divisive clustering starts from top to bottom or top-down from the family of hierarchical clustering. Comparing to agglomerative approach the divisive algorithm it starts from big process and break down into smaller process. In a divisive clustering algorithm the clusters performs as a single and then break down into two least similar clusters. This partition continues recursively until there remains one cluster for each observation. Each object forms a cluster of its own. It is less popular than the agglomerative approach. Here the single cluster is partitioned into two least similar clusters. At last it remains one cluster for each observation. The advantages of using divisive clustering algorithm more precise results could be derived than the bottom –up approach of agglomerative algorithm. Since the algorithm of top-down approach is linear in the flow of cluster and components, it produces faster results that are least quadratic [20].

1.3 K-Means Clustering

K-means clustering is one the unsupervised learning type from the Partitive clustering algorithm. It requires prior knowledge. The data chosen in this type of algorithm is not grouped or categorized. The K-is the number of clusters in the data. The k-means algorithm addresses the large sets of data points. It acquires the Euclidean distance between the two clusters. K is the given cluster and to initialize the K clusters by picking one point near to the cluster. The random selection could be done for picking the K point. Place the nearest cluster whose centroid in closer. Find the available data point in the cluster. Each point to place it in the cluster is to choose the centroid to its closest. Find the cluster whose centroids are closest to the data point and the data point are assigned to the cluster. Repeat the process for each data point. When all the data points are assigned to the K-cluster, the possible of changing the centroid to another cluster could arise due to arrival of new data points in the cluster. Update the each location of the centroid by taking into the count whenever new data point is added to the K-cluster. Reassign all the points to the closest centroid. The processes are repeated until the convergence took place. At the stage of convergence the point does not make a move between the centroid and makes to stabilize [22]

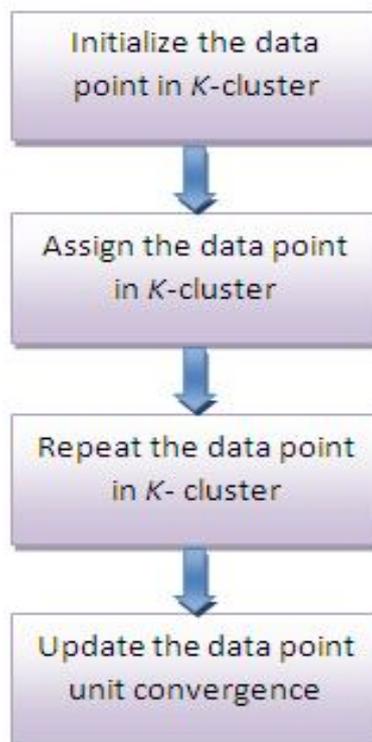


Fig-10: Flow chart representation of K-Means clustering

The K-means algorithm can be applied in the field of behavioural segmentation of the customer like the customer purchase history, to classify the customer or use website search or the application which user preferred most. In the field of inventory categorization like metrics, it is applied to find the manufacturing details, sales and delivery activities. To predict the sensor measurement, to use in audio separation, healthcare monitoring.

1.4 Self Organizing Map

The self Organizing Map (SOM) is a classification of unsupervised learning from the type of Partitive clustering. It is used for visualization and high dimensional datasets. The SOM does not require any target vector and it could classify without any update or monitoring. It is used for high dimensional data includes large number of features or variables. An SOM representation is formed from a grid of nodes in which the input data are represented. SOM nodes are not connected to each other and every node is connected to the input. The weight of each node has inputs. Best Matching Unit (BMU) is implemented to examine which node has the equal weight to the input vector. The distance between the input vector and the weights of the nodes is used to find the BMU. Until the BMU is attained it shrinks the each iteration. The advantages of using SOM algorithm are to represent the invisible patterns in the data. The applications of self Organizing maps are applied in atmospheric sciences [22].

SEMI-SUPERVISED ALGORITHM

1.1 Generative Methods

The generative methods are best in for classification which falls into two main categories.

Table-14: Generative method model types

Methods
Generative Model
Discriminative Model

1.1.1 Generative Model

Generative model a learning algorithm used in classification. Generative model are basically try to build a model of the positive and a model of the negative data classifications. The model here focuses on the particular characterization. The individual classes in the model are distributed as a separate model. The classification of generative model looks for a boundary between the spaces, where one model becomes more likely than another model. In Generative method more probabilistic model for each class is classified. Here both labelled and unlabelled data for estimating the distribution of data. In real time example, the generative model is applied to model the distribution of individual classes for student hand writing recognition. For that the data model is created is every single alphabet and number and it would store that model in every single time when the new input is received the model is represented and cross reference is made with the historical model. The applications of generative models are applied in Text classifications for representation of bag of words [24]. The cons of generative methods are very easy to use since it is handled as unlabelled data as the maximum, useful as model representations. The boundaries are determined in a frequent time. Discriminative models are good at predictive analysis. The disadvantages of using generative models are sometimes the model could lead wrong interpretation.

1.1.2 Discriminative Model

Discriminative model attempts in creating a boundary between the data points instead of creating a class model. It focuses on creating a boundary between positive and negative model. The discriminative model tries to find out the key differences between each model. It focuses about the main task is separation of positive and negative data by acting as a supervised task. Discriminative approach ignores the labelled data and focus on important or needed data. It mightily supported in training algorithm. Discriminative model do not focus on unsupervised learning. The real time example taken for generative model of classification of handwriting techniques could be applied here, instead of creating a data model between the classes, the data are tries to separate boundaries between each alphabet and numbers by finding the key differences. The applications of discriminative model are applies in Object tracking by the models to track the objects in a dynamical situation [25]. The example for discriminative algorithm is Support Vector Machine.

1.2 Graph Based methods

Graph based learning are widely used in machine learning for the classification techniques such as clustering in semi supervised learning. In familiar every data point in the data set is represented as node in the graph. The face recognition of the student could be taken as a running example. The face of each student is assigned as node in the graph; the pixels in the graph act as an edges representing the similarities between the faces and the vertices representing the data points. Here the SSL technique can match with the unlabelled data with the labelled data. The datasets are normally represented by a graph [26]. For the consistent in dataset the graph based methods in SSL are used, the results are shown as heterogeneous. The data produced through graph based methods are scalable in manner of large data. The graph matching is used to compare the graph and find the similarities between them. Comparing with SVM learning techniques, the graph based semi supervised learning with the unlabelled data able to produce better results [27]. The real time example could be match with student question and answering system in the exam. The process of question and answers are extracted and matched according to the student answering skill and questioning skill in the class room. The applications of graph based methods are applied in Web, citation networks, semantic parsing, Multilingual POS Tagging, Sentiment Analysis in social networks, text categorization. Some of the well known algorithms practiced in machine learning process for graph-based are given under the table representations.

Table-15: Tabular representation model of Graph-based algorithms

Graph Based Machine Learning Algorithms	Page Rank
	Graph Matching
	Shortest Path
	Depth-First Graph Traversal
	Breadth – First Graph Traversal
	Minimum Spanning Tree
	Node2vec

1.3 Heuristic approaches

Heuristic approaches are used in machine learning to find the possible solution among all the available and with the potential one. The heuristic approaches hint some estimated results to the larger set of problems. In this scenario the solution is to find the possibilities for the serial set of problems rather than the objective of the accurate result. The maximum possibilities will be the appropriate to the solutions rest of problems. The objective behind the heuristic approach is a quality which lays as adequate solution to the general set of problems. The goal of the heuristic approach is to find a solution that best fits for all the possible problems [28]. The heuristic approaches are applied with a real time example with a student dress code. Every time the student used to make judgement about getting dressed well upon, sometimes they are able to think about their dressings carefully, in rest of the times the decisions are carried out using little information. Instead of thinking the best outfit to wear a day, the students would go with hundreds of choices around their mind but the default outfit is a uniform for the students. Considering the example of list of menu for the food items in the canteen, the student may have list of choices to decide. Instead of time efficiency and from the past history of tasting the food, the students would normally prefer to have the food which they enjoy in the regular intervals of time. Heuristic approaches help to take decisions based on the history and act like a cognitive tool from which the decisions and judgement could be made quickly.

Let's take a another example with real world facts applied on students a positive for scoring centum in all the subject that only one in a thousand student can do that. The llikelihood of the possibilities for the student could be guessed as 10% based on the test results.

Table-16: Representativeness Heuristic model of Student scoring percentage

Possibilities	Percentage (%)
Subject knowledge	5
Regular exercise	2
Writing skill	1
Creative skill	2
Total	10

To find the likelihood of scoring the possibilities of centum in all the subjects we can use the representatives heuristics with the result inaccurate to be guessed as 90 percent and the positive to be 10 percent, if the false positive is 90 percent a common number in student test, then it is highly unlikely for the student to score centum. But the accurate result of 10 percent of all the time with the test conducted previously then it is highly unlikely to score a centum in all the subjects. This is because the odds of the getting a positive result one in 10 are much higher than the odds of actually who has not scored a centum in all one in thousand. The algorithms used for heuristic approaches are not called to furnish a right solution or to give the conclusion instead the results are based are traditional techniques. Some of the algorithms developed to solve the heuristic techniques are tabulated below.

Table-17: Representation model of heuristic algorithm and techniques

Type	Algorithms
Heuristic Algorithms	Exhaustive search
	Local Search
	Divide and Conquer
	Branch and Bound
	Dynamic programming
Heuristic Techniques	Simulated annealing
	Tabu search
	Swarm intelligence
	Evolutionary Algorithms

Applications of heuristic are applied using algorithms real time on the identification of various resources in land specifically water resources and the procedure of extracting it [29]. The trial and error methods are applied logically for setting the possibilities in different fields or areas. It is used in business enterprises by the entrepreneurs for the development of the business by promoting the risk factors where customers are not shown any interest of profit. It is used for comparing the web page classifications in order for the developments of the marketing heuristic approaches are used. Heuristic methods are used to test the product development through the SDLC approaches.

REINFORCEMENT LEARNING

1.1 Temporal Difference method

The Temporal Difference (TD) is a classification under Reinforcement Learning in the Machine Learning Technique. TD is multi step prediction process considering the value of each function. It is method learning to predict and it is an iterative way of learning until the accurate value of estimated returns. The value of function could be changed in every step. The value of each function is derived from the state and it is rewarded in return of the state. TD is said to be an incremental approach of learning a guess from the guess. The TD learning makes a different between the prediction and moves to the next state of prediction like learning from the prediction. TD is a long term prediction consists of multistep prediction where Motivation factor is an important for the next step of prediction.



Fig-11: State of a Temporal Difference method [29]

In the above the states of Temporal Difference is pictured. TD applies the current count of the next state. The state of an actions noted as C1 & C2, the agent takes an action and transit to the next state. The value of the current state C2 is used to update the value of the previous state C1. Repeat the state of the exercise for each episode by initializing the state current state as C, select the action using agent and in order to reward use the current state of the next state. The advantage in TD is the expected values are derived since it uses the current estimation for the next state of prediction. The result of the model is being updated on each stage of action; the possibility of waiting for an long hour is reduced in TD approach. The disadvantages using TD is the target is never known, since the prediction of the state becomes the input to the next state of process like winning a move in a chess game as we never think of winning at a state instead it keeps on moving to the next stage. Perform the update and make the new state as the present state update and repeat the state of convergence until it reach the target result. Let's take a simplest real time example of the games conducted for students at university level. The goal of the team is to win the opposite rather than estimating their opponent to face off next level. The winning team decides the level to the tournament. So the current situation is the estimator for the next level. The current goal of the participated team is to win not worried about the next level. In each situation the level of iteration is maintained.

Table-18: Representation model of algorithms in TD [30]

Temporal Difference Methods (left) and Algorithms (right)	
Markov Decision Process (MDP)	State-action-reward-state-action (sarsa)
Linear Function Approximation	
Temporal difference in prediction learning Least Square Temporal Difference Learning	Temporal Difference Learning
Q learning	Deep Reinforcement learning

The popular learning algorithms based on the reinforcement technique in TD is mentioned above with their classifications. The above classified algorithms are fit into the requirements of reinforcement techniques.

1.2 Monte carlo methods

Monte Carlo in Reinforcement Learning type algorithm from the Machine Learning techniques follows the method of random sampling. Monte Carlo takes more random points inherently. The approximation of the values lies in choosing the most possible random points. The Monte Carlo methods are helps to find the better approximation with more possible sampling methods. Monte Carlo methods are broad classified computational algorithms that rely on repeated random sampling to obtain numerical results. The more random points come in, the better approximation of results comes to known in using Monte Carlo algorithm. Monte Carlo idea of necessity is using random list to puzzle out the unsolved problems. The two main classifications of Monte Carlo methods are listed below.

Table-19: Model of Monte Carlo representation

Monte Carlo Methods	
1.	Monte Carlo prediction
2.	Monte Carlo Control

Monte Carlo methods are mainly used in the classification of three clear-cut problem classes Optimization, Numerical Integration and Probability Distribution. Monte Carlo methods mostly used in most impossible state of other approaches as a mathematical representation. The Monte Carlo predictions are applied in the field of predicting the future value of stocks in the market with the probabilistic values, in the field of space networks for the image corrections in satellite, in partial different equations. Monte Carlo Control methods applied in the physics related problems solving the simulation techniques. The other approaches are generating draws from the sequence of probability distributions satisfying non linear equation. Monte Carlo methods used as prediction as the methodologies relies on the sequential interacting samples. The terminologies reflecting the samples also known as a particle, individuals, walkers, agents, creatures interacts with the empirical measures of the process. When the size of the system turns to infinity these random empirical measures converge to the deterministic distribution of the random state of the non linear actions.

CONCLUSION

This paper provides a brief introduction about the machine learning techniques and the popular algorithms used to handle the methods in ML for the novice. From above applications we are able to give complete hope about the machine learning and its current trend applications by the explanation of the concepts. Demonstration of diagrams with real time examples will make the reader to understand each algorithm.

REFERENCES

1. Rob Schapire. (2010), "Machine learning algorithm for classifications" Princeton University, 1023-00-1907.
2. Yi-Fan Chang Adviser, and Prof. Ding J. J. (2011), "An Overview of Machine Learning".
3. Prof. Richard Zanibbi. (2004) "Pattern Recognition an overview" by L. Kuncheva, Wiley.
4. Rehana Rajam and K.G.sathish kumar. (2015), "The best open source Machine Learning Framework" Volume 3, Issue 11.
5. Sunpreet Kaur and Sonika Jindal. (2016), "A survey on Machine Learning Algorithms". A. Smola and S. Vishwanathan. (2010), "Introduction to Machine Learning". United Kingdom at the University Press, Cambridge.
6. Sandhya N. Dhage and Charanjeet Kaur Raina. (2016), " A review on Machine Learning Techniques" Volume: 4 Issue: 3.
7. Yogesh Singh, Pradeep Kumar Bhatia and Omprakash Sangwan. (2007), "A review of studies in machine learning technique". International Journal of Computer Science and Security, Volume (1): Issue (1) 70 – 84, June 2007.
8. Taiwo Oladipupo Ayodele. (2010) ,"Types of Machine Learning Algorithms" University of Portsmouth United Kingdom, February. DOI: 10.5772/9385.
9. Dated on: 07/06/18 -<https://medium.com/@jrodthoughts/understanding-semi-supervised-learning-a6437c070c87> (accessed 7th July 2018).
10. Carl Burch. (2001), "A survey on machine learning", Pennsylvania Governor's School for the Sciences 2001.
11. Kajaree Das and Rabi Narayanan Behera. (2017), "A survey of Machine Learning: Concept, Algorithms and Applications", Vol.5, Issue 2, ISSN: 2320-9801, February 2017.

12. Ayodele, T.O. (2010), “Types of Machine Learning Algorithms”,(Feb 1, 2010).
13. Mohssen Mohammed, Muhammad Badruddin Khan and Eihab Bashier Mohammed Bashier. (2017), “Machine Learning algorithm and its applications”, ISSN: 13: 978-1-4987-0538-7, 2017 by Taylor & Francis Group, LLC.
14. Jonathan Walker and Hughes Training. (1996), “Methodology Application: Logistic Regression Using he codes data,” 1996.
15. Rahul Saxena. (2017), “How the Naive Bayes Classifier works in Machine Learning”, Data Science, Machine Learning, February 2017.
16. Vincent Alabau and Jesu’s Andre’s. (2006), “The Naive Bayes model, generalisation and applications”, Universitate politecnica de Valencia, Spain.
17. Andrew McCallum and Kamal Nigam. (1998), “A comparison of Event Models for Naive Bayes Text Classification”.
18. Mayur Kulkarni. (2018), “Decision Trees for classification: A Machine Learning Algorithm”, September.
19. Christopher D Manning and Prabhakar Raghavan. (2008), “Introduction to Information Retrieval”, Cambridge university Press, ISBN:0521865719.
20. Tapas Kanungo and David M Mount. (2002) “An efficient K-Means Clustering Algorithm”, IEEE transactions on pattern analysis and machine intelligence, Vol 24, No.7.
21. Magnus Johnson. (2012), ” Application of self organizing Maps” Lund University, Sweden, Nov 2012, ISBN: 9789-953-51-0862-7.
22. Olivier Chapelle. (2005), “A Taxonomy of Semi-Supervised Learning Algorithms”, Max Planck Institute for Biological Cybernetics.
23. Nigam.K et al. (2000), “Text Classification from Labelled and unlabelled Documents Using EM, Machine Learning”, May-June, Volume 39 Issue 2-3: 103-134.
24. Ruei-Sung Lin et al. (2004), “Adaptive Discriminative Generative Model and its Applications”, International Conference on Neural Information Processing System, MIT Press Cambridge, MA, USA: 801-808
25. Amarnag subramanya and Partha Pratim Taludkar. (2012), “A tutorial on Graph-Based semi-supervised Learning Algorithms for NLP”, south Korea August 2012.
26. Jingrui Hi and Jaime Carbonell et all. (2007), “Graph Based Semi-Supervised Learning as a Generative Model”, International joint conference on Artificial intelligence, Carnegie Mellon university :2492-2497.
27. Natalliah Kokash. (2005), “An introduction to heuristic algorithms”, University of Trento, Italy.
28. Richard Sutton and Anfre Barto. (1998), “Reinforcement Learning: An Introduction”, MIT Press, Cambridge, 1998.
29. Florian Kunz. (2000). ” An Introduction to Temporal Difference Learning”, Department of Computer Science, TU Darmstadt.

COMPARATIVE STUDY ON CLUSTERING TECHNIQUES

Sathiyapriya. S¹ and Dr. Kanagaraj. A²Research Scholar¹ and Assistant Professor², NGM College, Pollachi, Coimbatore**ABSTRACT**

The overall goal of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering is the method of grouping a set of similar data objects within the same group based on similarity criteria. This paper reviews five types of clustering techniques in Data Mining - Partitioning Clustering, Hierarchical Clustering, Grid based clustering, Model based clustering, and Density based clustering.

Keywords: Clustering, Data Mining, Hierarchical, Partitioning, Unsupervised Learning

I. INTRODUCTION

Data mining is a technique to analyze and retrieve knowledge from large amount of database and transform it into useful information for future use [1]. Data mining is used in classification, clustering, regression, association rule discovery, sequential pattern discovery, outlier detection, etc.

In data mining, mining of data can be done using two learning approaches- Supervised and Unsupervised learning. Clustering is an unsupervised learning in data mining applications. Clustering is the task of grouping a set of objects in such a way that objects in the cluster are more similar to each other than to those in other clusters[4]. Clustering techniques have numerous applications in various fields including, artificial intelligence, pattern recognition, bioinformatics, segmentation and machine learning.

II. LITERATURE REVIEW

T.Sajana, C.M.Sheela Rani and K.V.Narayana, [10] analyzed different clustering algorithms required for processing Big Data. The study revealed that to identify the outliers in large data sets, the algorithms that should be used are BIRCH, CLIQUE, and ORCLUS. To perform clustering, various algorithms can be used but to get appropriate results the present study suggests that – by using CURE and ROCK algorithms on categorical data, arbitrary shaped clusters will be created. By using COBWEB and CLASSIT algorithms on numerical data with model based, non-convex shape clusters can be formed. For spatial data STING, OPTIGRID, PROCLUS and ORCLUS algorithms when applied yield arbitrary shaped clusters.

Tajunisha and Saravanan, [11] proposed a method for finding the initial centroid and cluster the data in low dimensional space. Analyzed the performance of proposed method with the existing works. By comparing the results on iris data set, it was found that the results obtained by the proposed method are more accurate and efficient compared to the existing method. They used Principal Component Analysis (PCA) for dimension reduction and to find the initial centroid for k- means.

Singh Vijendra, Sahoo Laxman and Kelkar Ashwini, [12] proposed a genetic clustering technique (GACR). Experimental results on different data sets demonstrate the comparison of GACR algorithm, GCA algorithm and Kmeans algorithm. GACR is found to provide satisfactory performance where GCA and K-means fails. Results on various artificial and real-life data sets demonstrate that GACR has better performance than GCA and K-means clustering algorithm.

T. Soni Madhulatha, [13] presented that when applying a cluster analysis we are hypothesizing that the groups exist. But this assumption may be false or weak. Clustering results “should not be generalized. Cases in the same cluster are similar only with respect to the information cluster analysis was based on i.e., dimensions / variables inducing the dissimilarities.

P. IndiraPriya, Dr. D.K.Ghosh, [14] proposed a different Clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. There is also some scope for applying the clustering procedure to large datasets. In large datasets, the clustering efficiency is degraded and also need to improve time and scalability values.

Rahumath Beevi A, Remya R, [15] presented that, among the different fuzzy clustering techniques FRECCA algorithm is superior to others. When time is critical factor then we cannot adopt fuzzy based approaches fuzzy based clustering approaches provide significant performance. But, fuzzy approaches do have certain drawbacks which have to be eliminated.

Saroj, Tripti Chaudhary , [16] presented that, Clustering is the technique of data mining which is used to extract the useful information from raw data. We can say that raw data is useless without the different clustering techniques. They also discussed in detail about the types of clusters.

Sukhdev Singh Ghuman, [17] presented the review of k-means algorithm, BIRCH algorithm, CLIQUE algorithm, DBSCAN algorithm. Clustering techniques are compared based on shape of cluster, clustering algorithm used and outlier handling.

Osama Abu Abbas, [18] intended to compare different data clustering algorithms based on size of dataset, number of clusters, type of dataset and type of software used. Some conclusions that belong to the performance, quality and accuracy of the clustering algorithms.

Shivangi Bhardwaj, [19] presented that different data mining techniques can be used or applied not only on small data but on large and large amount of data sets. Automatic work is also discussed. So that less time is consumed to collect the data to form clusters.

III. GENERAL TYPES OF CLUSTERS

1. Well-separated clusters

A cluster is a set object in which each object is closer (or more similar) to every other object in the cluster than to any object not in the cluster. The distance between any two points in different groups is larger than the distance between any two points within a group. Well-separated clusters do not need to be globular, but can have any shape.

2. Center-based clusters

A cluster is a set of objects such that an object in a cluster is nearest (more similar) to the “center” of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid.

3. Contiguous clusters

A cluster is a set of points so that a point in a cluster is nearest (or more similar) to one or more other points in the cluster as compared to any point that is not in the cluster.

4. Density-based clusters

A cluster is a dense region of points, which is separated by according to the low-density regions, from other regions that is of high density. This definition is more often used when the clusters are irregular or intertwined, and when noise and outliers are present.

5. Shared Property or Conceptual Clusters

Finds clusters that share some common property or represent a particular concept. Objects in a center-based cluster share the property that they are all closet to the same centroid or medoid.

IV. CLUSTERING TECHNIQUES

Clustering is considered as an unsupervised classification process [4]. Clustering is that the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. A large number of clustering algorithms have been developed for different purposes [3][4].

Based on the strategy of how data objects are distinguished, clustering techniques can be broadly divided in two classes: hierarchical clustering techniques and partitioning clustering techniques [2]. However there is no clear boundary between these two classes. Some efforts are done on the mixture of various clustering methods for addressing specific applications. Beyond the two traditional hierarchical and partitioning classes, there are several clustering techniques that are categorized into independent classes, for example, density-based methods, Grid-based methods and Model based clustering methods [1][2]. A short review of these methods is described below.

Now a day, people come across a large quantity of information. Then store or represent it as data [3]. One of the imperative suggests that in addressing these information is to group them into a set of clusters. Clustering involves making groups of objects that are same, and those that are not same. The clustering problem lies to find groups of same objects within the information.

1. Partitioning Method

Partitioning algorithms divide data into several subsets. The reason of dividing the data into many sets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes

are used in the form of iterative optimization. Specifically, this implies totally different relocation schemes that iteratively reassign points between the k clusters. Relocation algorithms gradually improve clusters.

Relocation Algorithms, Probabilistic Clustering, K-Medoids Methods, K-Means Methods are the algorithms used in the partitioning method.

2. Hierarchical Method

Hierarchical Clustering method merged or splits the similar data objects by constructing hierarchy of clusters also known as dendrogram [4]. Hierarchical Clustering method forms clusters progressively.

Agglomerative hierarchical clustering is a bottom up method which starts with every single object in a single cluster. Then, in each successive iteration, it combines the closest pair of clusters by satisfying some similarity criteria, till all of the data is in one cluster or specify by the user [4].

Divisive hierarchical clustering is a top down approach. Divisive hierarchical clustering starts with one cluster that contain all data objects. Then in each successive iteration, it divide into the clusters by satisfying some similarity criteria until each data objects forms clusters its own or satisfies stopping criteria.

3. Density Based Method

The density-based methods follow the expanding the cluster until a density threshold is reached. [5] For these methods a “neighborhood” has to be defined and the density must be calculated according to number of substance in the neighborhood.

A cluster is a dense region of points that is separated by low density regions from the tightly dense regions. This clustering algorithm can be used when the clusters are irregular It finds core objects i.e. objects that have dense neighborhoods. It connects core objects and their neighborhoods to form dense regions as clusters. Clusters are formed as maximum sets of density connected points and can detect noise and used when outliers are encountered.

4. Grid Based Method

In this technique measures the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed [9]. It is based on clustering oriented query answering in multilevel grid structures. In upper level stores analysis of the information of it's after that level, therefore the grids make cells between the connected levels.

Grid based clustering is the fastest processing time that typically depends on the size of the grid instead of the data. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes from the objects.

5. Model Based Method

Model-based clustering strategies are based on the assumption that data are generated by a combination of underlying probability distributions, and they optimize the fit between the data and some mathematical model. Examples are statistical approach, neural network approach and other AI approaches. When facing an unknown data distribution, choosing a suitable one from the model based candidates is still a major challenge. On the other side, clustering based on probability suffers from high computational cost, especially when the scale of data is very large.

Set of data points are connected along supported various methods like statistical methods, conceptual methods, and robust clustering methods. There are two approaches for model based algorithms one is neural network approach and another one is statistical approach.

IV. CONCLUSION

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Clustering is a significant task in data analysis and data mining applications. In this paper, various clustering approaches for clustering research papers were compared in detail.

V. REFERENCES

1. J. Han and M. Kamber, “Data Mining: Concepts and Techniques,”Morgan Kaufmann Publishers, 2001.
2. P. Berkhin, “A Survey of Clustering Data Mining Techniques” Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press (2006) 25-72.
3. P. IndiraPriya, Dr. D.K.Ghosh ” A Survey on Different Clustering Algorithms in Data Mining Technique” IJMER- International Journal of Modern Engineering Research, Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274.

4. Dr.N.Rajalingam K.Ranjini, "Hierarchical Clustering Algorithm - A Comparative Study", International Journal of Computer Applications, Volume 19– No.3, April 2011, pp.42-46.
5. Gonzalo E. Paredes,Luis S. Vargas "Circle-Clustering: A New Heuristic Partitioning Method for the Clustering Problem" of WCCI 2012 IEEE World Congress on Computational Intelligence
6. A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall,1988.
7. A. K. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Volume 31(3), 1999, pp. 264-323.
8. P. Berkhin, "A Survey of Clustering Data Mining Techniques" Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds) Grouping Multidimensional Data, Springer Press (2006) 25-72
9. Suman and Mrs.Pooja Mittal "Comparison and Analysis of Various Clustering Methods in Data mining On Education data set Using the weak tool" IJETTCS - International Journal of Emerging Trends & Technology in Computer Science, Volume 3, Issue 2, March – April 2014.
10. T.Sajana, C.M.Sheela Rani and K.V.Narayana, "A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology, Vol.9(3), Jan-2016, pp.1-12.
11. Tajunisha and Saravanan," Performance analysis of k - means with different initialization methods or high dimensional datasets, " IJAIA- International Journal of Artificial Intelligence & Applications, vol. 1, no.4, Page(s):44 - 52,Oct. 2010.
12. Singh Vijendra, Sahoo Laxman and Kelkar Ashwini, "Mining Clusters in Data Sets of Data Mining: An Effective Algorithm", International Journal of Computer Theory and Engineering", Vol.3, No.1, February-2011, pp.171-177.
13. T. Soni Madhulatha " An overview of clustering methods" IOSR Journal of Engineering, Apr-2012, Vol. 2(4) pp. 719-725
14. P. IndiraPriya,Dr. D.K.Ghosh " A Survey on Different Clustering Algorithms in Data Mining Technique" IJMERE-International Journal of Modern Engineering Research,Vol.3, Issue.1, Jan-Feb-2013 pp.267-274.
15. Rahumath Beevi A, Remya R. "A Comparison of Clustering Techniques in Data Mining", IJCAT-International Journal of Computing and Technology, Volume 1, Issue 4, May 2014.
16. Saroj, Tripti Chaudhary, "Study on Various Clustering Techniques", IJCSIT –International Journal of Computer Science and Information Technologies, Vol.6(3), 2015, pp.3031-3033.
17. Sukhdev Singh Ghuman, "Clustering Techniques – A Review", IJCSMC – International Journal of Computer Science and Mobile Computing, Vol.5, May-2016, pp.524-530.
18. Osama Abu Abbas, " Comparisons Between Data Clustering Algorithms ,"The International Arab Journal of Information Technology, vol. 5, no. 3, Page(s):320 - 326,Jul. 2008.
19. Shivangi Bhardwaj, "Data Mining Clustering Techniques – A Review", IJCSMC – International Journal of Computer Science and Mobile Computing, Vol.6, Issue.5, May-2017, pp.183-186.

INTERNET OF THINGS: SECURITY ISSUES, CHALLENGES AND COUNTER MEASURES

D. Sakthivel¹ and R. Kumaramoorthy²Assistant Professor¹ and Student², Department of Computer Science, Sree Saraswathi Thygaraja College, Pollachi

ABSTRACT

The paper presents a survey and analysis on the current status and concerns of Internet of things (IoT) security and challenges. There are many solutions and products marketed today under the label IoT that lack basic security architectures. It is very easy for a knowledgeable person to take control of devices for malicious purposes. Not only devices at home are at risk, but cars, trains, airports, stores, ships, logistics applications, building automation, utility metering applications, industrial automation applications, health services, and so on, are also at risk because of the lack of security measures in their underlying architecture. It has gone so far that many western countries have identified the lack of security measures in automation applications as a risk to national security, and rightly so. It is just a matter of time before somebody is literally killed as a result of an attack by a hacker on some vulnerable equipment connected to the Internet. This paper presents an overview of security principles, technological and security challenges, proposed countermeasures, and the future directions for securing the IoT.

Keywords: Attacks, Challenges, IoT, Privacy, Tools

I. INTRODUCTION

A global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving, interoperable information and communication technologies [1].

An IoT is a network that connects uniquely identifiable "things" to the Internet. The "things" have sensing/actuation and potential programmability capabilities. Through the exploitation of the unique identification and sensing, information about the "thing" can be collected and the state of the "thing" can be changed from anywhere, anytime, by anything.[2]

Every connected device creates opportunities for attackers. These vulnerabilities are broad, even for a single small device. The risks posed include data transfer, device access, malfunctioning devices, and always-on/always-connected devices. The main challenges in security remain the security limitations associated with producing low cost devices, and the growing number of devices which creates more opportunities for attacks.[4]

a. Security Spectrum

The definition of a secured device spans from the simplest measures to sophisticated designs. Security should be thought of as a spectrum of vulnerability which changes over time as threats evolve. Security must be assessed based on user needs and implementation. Users must recognize the impact of security measures because poorly designed security creates more problems than it solves.

II. IOT SECURITY CHALLENGES**Beyond costs and the ubiquity of devices, other security issues plague IoT****a. Unpredictable Behavior**

The sheer volume of deployed devices and their long list of enabling technologies mean their behavior in the field can be unpredictable. [4] A specific system may be well designed and within administration control, but there are no guarantees about how it will interact with others.

b. Device Similarity

IoT devices are fairly uniform. They utilize the same connection technology and components. [6] If one system or device suffers from vulnerability, many more have the same issue.

c. Problematic Deployment

One of the main goals of IoT remains to place advanced networks and analytics where they previously could not go. [7] Unfortunately, this creates the problem of physically securing the devices in these strange or easily accessed places.

d. Long Device Life and Expired Support

One of the benefits of IoT devices is longevity; however, that long life also means they may outlive their device support. Compare this to traditional systems which typically have support and upgrades long after many have stopped using them. Orphaned devices and abandon ware lack the same security hardening of other systems due to the evolution of technology over time. [4]

e. No Upgrade Support

Many IoT devices, like many mobile and small devices, are not designed to allow upgrades or any modifications. Others offer inconvenient upgrades, which many owners ignore, or fail to notice.

f. Poor or No Transparency

Many IoT devices fail to provide transparency with regard to their functionality. Users cannot observe or access their processes, and are left to assume how devices behave. They have no control over unwanted functions or data collection; furthermore, when a manufacturer updates the device, it may bring more unwanted functions. [4]

g. No Alerts

Another goal of IoT remains to provide its incredible functionality without being obtrusive. This introduces the problem of user awareness. Users do not monitor the devices or know when something goes wrong. Security breaches can persist over long periods without detection. [6]

III. SECURITY RISK and MODES OF ATTACK

Engineers working on machine-to-machine (M2M) communication paradigms, such as industrial automation, might have considered the problem solved when they discovered that machines could talk to each other over the Internet, that is, when the message-exchanging problem was solved.[5] This is simply relabeling their previous M2M solutions as IoT solutions because the transport now occurs over the IP protocol. But, in the realm of the Internet, this is when the problems start. Transport is just one of the many problems that need to be solved[3].

A. Denial of Service

A Denial of Service (DoS) or Distributed Denial of Service (DDoS) attack is normally used to make a service on the Internet crash or become unresponsive, and in some cases, behave in a way that it can be exploited.[6] The attack consists in making repetitive requests to a server until its resources get exhausted. In a distributed version, the requests are made by many clients at the same time, which obviously increases the load on the target. It is often used for blackmailing or political purposes. The attack is distributed and the target centralized, the attack gets less effective if the solution itself is distributed. To guard against this form of attack, you need to build decentralized solutions where possible. In decentralized solutions, each target's worth is less, making it less interesting to attack.[3]

B. Guessing the credentials

One way to get access to a system is to impersonate a client in the system by trying to guess the client's credentials. To make this type of attack less effective, make sure each client and each device has a long and unique, perhaps randomly generated, set of credentials. Never use preset user credentials that are the same for many clients or devices or factory default credentials that are easy to reset. Furthermore, set a limit to the number of authentication attempts per time unit permitted by the system; also, log an event whenever this limit is reached, from where to which credentials were used.[8] This makes it possible for operators to detect systematic attempts to enter the system.

C. Getting access to stored credentials

One common way to illicitly enter a system is when user credentials are found somewhere else and reused. Often, people reuse credentials in different systems. There are various ways to avoid this risk from happening. One is to make sure that credentials are not reused in different devices or across different services and applications. Another is to randomize credentials, lessening the desire to reuse memorized credentials. A third way is to never store actual credentials centrally, even encrypted if possible, and instead store hashed values of these credentials. This is often possible since authentication methods use hash values of credentials in their computations. Furthermore, these hashes should be unique to the current installation. Even though some hashing functions are vulnerable in such a way that a new string can be found that generates the same hash value, the probability that this string is equal to the original credentials is miniscule. And if the hash is computed uniquely for each installation, the probability that this string can be reused somewhere else is even more remote.

D. Man in the middle

Another way to gain access to a system is to try and impersonate a server component in a system instead of a client. This is often referred to as a Man in the middle (MITM) attack. The reason for the middle part is that the attacker often does not know how to act in the server and simply forwards the messages between the real client and the server. In this process, the attacker gains access to confidential information within the messages, such as client credentials, even if the communication is encrypted. [3]The attacker might even try to modify messages for their own purposes. To avoid this type of attack, it's important for all clients (not just a few) to always validate the identity of the server it connects to. If it is a high-value entity, it is often identified using a certificate. This certificate can both be used to verify the domain of the server and encrypt the communication.

Make sure this validation is performed correctly, and do not accept a connection that is invalid or where the certificate has been revoked, is self-signed, or has expired. Another thing to remember is to never use an unsecure authentication method when the client authenticates itself with the server. If a server has been compromised, it might try to fool clients into using a less secure authentication method when they connect. By doing so, they can extract the client credentials and reuse them somewhere else. By using a secure authentication method, the server, even if compromised, will not be able to replay the authentication again or use it somewhere else. The communication is valid only once.[9]

E. Sniffing network communication

If communication is not encrypted, everybody with access to the communication stream can read the messages using simple sniffing applications, such as Wireshark. If the communication is point-to-point, this means the communication can be heard by any application on the sending machine, the receiving machine, or any of the bridges or routers in between. If a simple hub is used instead of a switch somewhere, everybody on that network will also be able to eavesdrop. If the communication is performed using multicast messaging service, as can be done in UPnP and CoAP, anybody within the range of the Time to live (TTL) parameter (maximum number of router hops) can eavesdrop. Remember to always use encryption if sensitive data is communicated. If data is private, encryption should still be used, even if the data might not be sensitive at first glance. A burglar can know if you're at home by simply monitoring temperature sensors, water flow meters, electricity meters, or light switches at your home. Small variations in temperature alert to the presence of human beings. Change in the consumption of electrical energy shows whether somebody is cooking food or watching television. The flow of water shows whether somebody is drinking water, flushing a toilet, or taking a shower. No flow of water or a relatively regular consumption of electrical energy tells the burglar that nobody is at home. Light switches can also be used to detect presence, even though there are applications today that simulate somebody being home by switching the lights on and off.

F. Port scanning and web crawling

Port scanning is a method where you systematically test a range of ports across a range of IP addresses to see which ports are open and serviced by applications. This method can be combined with different tests to see the applications that might be behind these ports. If HTTP servers are found, standard page names and web-crawling techniques can be used to try to figure out which web resources lie behind each HTTP server. CoAP is even simpler since devices often publish well-known resources. Using such simple brute-force methods, it is relatively easy to find (and later exploit) anything available on the Internet that is not secured. To avoid any private resources being published unknowingly, make sure to close all the incoming ports in any firewalls you use. Don't use protocols that require incoming connections. Instead, use protocols that create the connections from inside the firewall. Any resources published on the Internet should be authenticated so that any automatic attempt to get access to them fails. Always remember that information that might seem trivial to an individual might be very interesting if collected en masse. This information might be coveted not only by teenage pranksters but by public relations and marketing agencies, burglars, and government agencies (some would say this is a repetition).[3][7][8]

G. Search features and wildcards :

Don't make the mistake of thinking it's difficult to find the identities of devices published on the Internet. Often, it's the reverse. For devices that use multicast communication, such as those using UPnP and CoAP, anybody can listen in and see who sends the messages. For devices that use single-cast communication, such as those using HTTP or CoAP, port-scanning techniques can be used. For devices that are protected by firewalls and use message brokers to protect against incoming attacks, such as those that use XMPP and MQTT, search features or wildcards can be used to find the identities of devices managed by the broker, and in the case of MQTT, even what they communicate. You should always assume that the identity of all devices can be found, and that there's an interest in exploiting the device. For this reason, it's very important that each device authenticates any requests made to it if possible. Some protocols help you more with this than others, while others make such authentication impossible. XMPP only permits messages from accepted friends. The only thing the device needs to worry about is which friend requests to accept. This can be either configured by somebody else with access to the account or by using a provisioning server if the device cannot make such decisions by itself. The device does not need to worry about client authentication, as this is done by the brokers themselves, and the XMPP brokers always propagate the authenticated identities of everybody who send them messages. MQTT, on the other hand, resides in the other side of the spectrum. Here, devices cannot make any decision about who sees the published data or who makes a request since identities are stripped away by the protocol. The only way to control that gets access to the data is by building a proprietary end-to-end encryption layer on top of the MQTT protocol, thereby limiting interoperability. In between the two resides protocols such as HTTP and CoAP that

support some level of local client authentication but lacks a good distributed identity and authentication mechanism. This is vital for IoT even though this problem can be partially solved in local intranets.

H. Breaking ciphers

Many believe that by using encryption, data is secure. This is not the case, as discussed previously, since the encryption is often only done between connected parties and not between end users of data (the so-called end-to-end encryption). At most, such encryption safeguards from eavesdropping to some extent. But even such encryption can be broken, partially or wholly, with some effort.[9]

Ciphers can be broken using known vulnerabilities in code where attackers exploit program implementations rather than the underlying algorithm of the cipher. This has been the method used in the latest spectacular breaches in code based on the OpenSSL library. To protect yourselves from such attacks, you need to be able to update code in devices remotely, which is not always possible. Other methods use irregularities in how the cipher works to figure out, partly or wholly, what is being communicated over the encrypted channel. This sometimes requires a considerable amount of effort. To safeguard against such attacks, it's important to realize that an attacker does not spend more effort into an attack than what is expected to be gained by the attack. By storing massive amounts of sensitive data centrally or controlling massive amounts of devices from one point, you increase the value of the target, increasing the interest of attacking it. On the other hand, by decentralizing storage and control logic, the interest in attacking a single target decreases since the value of each entity is comparatively lower. Decentralized architecture is an important tool to both mitigate the effects of attacks and decrease the interest in attacking a target. However, by increasing the number of participants, the number of actual attacks can increase, but the effort that can be invested behind each attack when there are many targets also decreases, making it easier to defend each one of the attacks using standard techniques.

IV. TOOLS FOR ACHIEVING SECURITY

There are a number of tools that architects and developers can use to protect against malicious use of the system.

A. Virtual Private Networks

A method that is often used to protect unsecured solutions on the Internet is to protect those using Virtual Private Networks (VPNs). Often, traditional M2M solutions working well in local intranets need to expand across the Internet. One way to achieve this is to create such VPNs that allow the devices to believe they are in a local intranet, even though communication is transported across the Internet. Even though transport is done over the Internet, it's difficult to see this as a true IoT application. It's rather a M2M solution using the Internet as the mode of transport. Because telephone operators use the Internet to transport long distance calls, it doesn't make it Voice over IP (VoIP). Using VPNs might protect the solution, but it completely eliminates the possibility to interoperate with others on the Internet, something that is seen as the biggest advantage of using the IoT technology.[10]

B. X.509 certificates and encryption

The use of certificates to validate the identity of high-value entities on the Internet. Certificates allow you to validate not only the identity, but also to check whether the certificate has been revoked or any of the issuers of the certificate have had their certificates revoked, which might be the case if a certificate has been compromised.[10] Certificates also provide a Public Key Infrastructure (PKI) architecture that handles encryption. Each certificate has a public and private part. The public part of the certificate can be freely distributed and is used to encrypt data, whereas only the holder of the private part of the certificate can decrypt the data. Using certificates incurs a cost in the production or installation of a device or item. They also have a limited life span, so they need to be given either a long lifespan or updated remotely during the life span of the device. Certificates also require a scalable infrastructure for validating them. For these reasons, it's difficult to see that certificates will be used by other than high-value entities that are easy to administer in a network.

C. Authentication of identities

Authentication is the process of validating whether the identity provided is actually correct or not. Authenticating a server might be as simple as validating a domain certificate provided by the server, making sure it has not been revoked and that it corresponds to the domain name used to connect to the server. Authenticating a client might be more involved, as it has to authenticate the credentials provided by the client. Normally, this can be done in many different ways. It is vital for developers and architects to understand the available authentication methods and how they work to be able to assess the level of security used by the systems they develop. Some protocols, such as HTTP and XMPP, use the standardized Simple Authentication and Security Layer (SASL) to publish an extensible set of authentication methods that the client can choose

from. This is good since it allows for new authentication methods to be added. But it also provides a weakness: clients can be tricked into choosing an unsecure authentication mechanism, thus unwittingly revealing their user credentials to an impostor. Make sure clients do not use unsecured or obsolete methods, such as PLAIN, BASIC, MD5-CRAM, MD5-DIGEST, and so on, even if they are the only options available. Instead, use secure methods such as SCRAM-SHA-1 or SCRAM-SHA-1-PLUS, or if client certificates are used, EXTERNAL or no method at all. If you're using an unsecured method anyway, make sure to log it to the event log as a warning, making it possible to detect impostors or at least warn operators that unsecure methods are being used.

Other protocols do not use secure authentication at all. MQTT, for instance, sends user credentials in clear text (corresponding to PLAIN), making it a requirement to use encryption to hide user credentials from eavesdroppers or client-side certificates or pre-shared keys for authentication. Other protocols do not have a standardized way of performing authentication. In CoAP, for instance, such authentication is built on top of the protocol as security options.

D. Usernames and passwords

A common method to provide user credentials during authentication is by providing a simple username and password to the server. This is a very human concept. Some solutions use the concept of a pre-shared key (PSK) instead, as it is more applicable to machines, conceptually at least. If you're using usernames and passwords, do not reuse them between devices, just because it is simple. One way to generate secure, difficult-to-guess usernames and passwords is to randomly create them. In this way, they correspond more to pre-shared keys. One problem in using randomly created user credentials is how to administer them. Both the server and the client need to be aware of this information. The identity must also be distributed among the entities that are to communicate with the device.

The XMPP Protocol the device creates its own random identity and creates the corresponding account in the XMPP server in a secure manner. There is no need for a common factory default setting. It then reports its identity to a Thing Registry or provisioning server where the owner can claim it and learn the newly created identity. This method never compromises the credentials and does not affect the cost of production negatively. Furthermore, passwords should never be stored in clear text if it can be avoided. This is especially important on servers where many passwords are stored. Instead, hashes of the passwords should be stored. Most modern authentication algorithms support the use of password hashes. Storing hashes minimizes the risk of unwanted generation of original passwords for attempted reuse in other systems.[8]

E. Using message brokers and provisioning servers

Using message brokers can greatly enhance security in an IoT application and lower the complexity of implementation when it comes to authentication, as long as message brokers provide authenticated identity information in messages it forwards. In XMPP, all the federated XMPP servers authenticate clients connected to them as well as the federated servers themselves when they intercommunicate to transport messages between domains. This relieves clients from the burden of having to authenticate each entity in trying to communicate with it since they all have been securely authenticated. It's sufficient to manage security on an identity level[8].

F. Centralization versus decentralization

Comparing centralized and decentralized architectures is like comparing the process of putting all the eggs in the same basket and distributing them in many much smaller baskets. The effect of a breach of security is much smaller in the decentralized case; fewer eggs get smashed when you trip over. Even though there are more baskets, which might increase the risk of an attack, the expected gain of an attack is much smaller. This limits the motivation of performing a costly attack, which in turn makes it simpler to protect it against. [3][9]

When designing IoT architecture, try to consider the following points

- Avoid storing data in a central position if possible. Only store the data centrally that is actually needed to bind things together.
- Distribute logic, data, and workload. Perform work as far out in the network as possible. This makes the solution more scalable, and it utilizes existing resources better.
- Use linked data to spread data across the Internet, and use standardized grid computation technologies to assemble distributed data (for example, SPARQL) to avoid the need to store and replicate data centrally.
- Use a federated set of small local brokers instead of trying to get all the devices on the same broker. Not all brokered protocols support federation, for example, XMPP supports it but MQTT does not.

CONCLUSION

The IoT is susceptible to attacks hence there are many security challenges and requirements that need to be addressed. Current state of research in IoT is mainly focused on types of attack and tools to counter measure, but with the rapid advancement of technology it is essential to incorporate new networking protocols like IPv6 and 5G to achieve the dynamic mash up of IoT topology. The major developments witnessed in IoT are mainly on small scale i.e. within companies, some industries etc. To scale the IoT framework from one company to a group of different companies and systems, various security concerns need to be overcome. The IoT has great potential to transform the way we live today. But, the foremost concern in realization of completely smart frameworks is security. If security concerns like privacy, confidentiality, authentication, access control, end-to-end security, trust management, global policies and standards are addressed completely, we can witness the transformation of everything by IoT in the near future.

REFERENCES

1. <http://www.itu.int/ITU-T/recommendations/rec.aspx?rec=y.2060>
2. http://iot.ieee.org/images/files/pdf/IEEE_IoT_Towards_definition_Internet_of_Things_Revision1_27MAY15.pdf
3. Peter Waher (2015), "Learning Internet of Things", 2015.
4. <http://www.tutorialspoint.com>
5. Rwan Mahmoud, Tasneem Yousuf, Fadi Aloul, Imran Zualkernan(2015)," Internet of Things (IoT) Security: Current Status, Challenges and Prospective Measures", The 10th International Conference for Internet Technology and Secured Transactions (ICITST-2015).
6. Ashvini Kamble, Sonali Bhutad (2013)," Survey on Internet of Things (IoT) Security Issues & Solutions", Proceedings of the Second International Conference on Inventive Systems and Control.
7. M. Abomhara and G. M. Koien (2014), "Security and privacy in the Internet of Things: Current status and open issues," in Int'l Conference on Privacy and Security in Mobile Systems (PRISMS), 1-8, 2014.
8. K. Zhao and L. Ge (2013), "A survey on the internet of things security," in International Conference on Computational Intelligence and Security , 663-667, 2013.
9. W. H. Chin, Z. Fan, and R. Haines (2014), "Emerging technologies and research challenges for 5G wireless networks," Wireless Communications, vol. 21,106-112.
10. Yiyang Zhang, Weifu Zou, Xi Chen Chengyue Yang, Jinping Cao (2014)," The Security for Power Internet of Things: Framework, Policies and Countermeasures", International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.

A SURVEY ON SOFTWARES FAILURE IN 2018**Ashin Thejas. J, Dinesh kumar. R, Sathyabama B and Dr. V. Kavitha**Student^{1,2}, Assistant Professor³ and Associate Professor⁴Department of MCA, Hindusthan College of Arts and Science (Autonomous), Coimbatore

ABSTRACT

Over the past few years, technology has made many great advances but has also suffered some spectacular failures. Ransomware attacks, IT outages and data leakages have wreaked havoc at banks, British Airways, bitcoin and the NHS, affecting millions of their users. Most software projects can be considered at least partial failures because few projects meet all their cost, schedule, quality, or requirements objectives. Failures are rarely caused by mysterious causes, but these causes are usually discovered post-mortem, or only after it are too late to change direction. This paper is based on the software's which are failed due to improper maintainance, poor quality, Use of immature technology, Badly defined system requirements, ect.,

Keywords: Software Failure, Software Issue.

INTRODUCTION

The Software Failure is occurred when the customer feels uncomfortable with the product. But, for last few years many software's getting fail due to loads of reasons. The Software failure escort their companies to financial lost. By these financial losses, many employers are facing jobless situations.

O2

More than 30 million O2 users in the UK lost access to data services after a software issue left them unable to use 3G and 4G services. Customers of O2 and the other companies using its network including Tesco Mobile, Sky Mobile, Lycamobile, O2 subsidiary Giffgaff were affected after a software glitch in equipment supplied by Ericsson brought the network to its knees. The nationwide outage also affected Transport for London's live electronic timetables at bus stops.

TSB Bank

Millions of TSB customers were locked out of their accounts after an IT upgrade led to an online banking outage. A planned system upgrade was expected to shut internet and mobile banking services down for one weekend in April 2018, but ended up causing months of disruption. The problems arose from TSB's move to a new banking platform following its split from Lloyds Banking Group.

Immediately after the new system was switched on, many customers experienced problems logging in, while others were shown details from other people's accounts or inaccurate credits and debits on their own. Customers remained locked out of their accounts two weeks after the initial outage. In July, TSB was still working its way through the backlog of complaints, when another outage struck, locking customers out of their online accounts once again.

TSB claimed that the problem was resolved later that day, but the debacle will further rupture the bank's relationship with parent company Sabadell. The Financial Times reports that the Spanish banking group is now considering selling TSB.

Welsh NHS IT failure

In 2018, doctors and hospital staff of the Wales NHS experienced a widespread computer failure that led to them being unable to access patient files. According to the National Cyber Security Centre, the failure was due to technical issues as opposed to a cyber attack yet it still caused wide disruption as GPs were unable to access blood and X-Ray results. It also caused a backlog as patients could not be contacted to cancel appointments, and notes could not be typed up and saved on NHS systems.

Meltdown & Spectre

At the start of 2018, Google researchers revealed CPU hardware vulnerabilities called Meltdown and Spectre had affected almost all computers on the market. Meltdown primarily affects Intel processors, while Spectre affects Intel, AMD and ARM processors. Daniel Gruss, one of the researchers that discovered the flaw at Graz University of Technology described Meltdown as "one of the worst CPU bugs ever found".

Although these are both primarily hardware vulnerabilities, they communicate with the operating system to access locations in its memory space. Meltdown, Google says, "breaks the most fundamental isolation between user applications and the operating system. This allows a program to access the memory, and also the secrets, of

other programs and the operating system". Spectre meanwhile "breaks the isolation between different applications" – "it allows an attacker to trick error-free programs, which follow best practices, into leaking their secrets.

WannaCry

In May 2017, a large ransomware attack called WannaCry (also known as WannaCrypt0r and WCry) hit NHS England and various organisations in the UK and around the world. The attack was due to vulnerabilities found in Microsoft operating systems installed in millions of computers around the world. According to Microsoft, the Windows versions that were vulnerable to the attack were versions which were no longer supported by Microsoft such as Windows 8 and Windows XP, which the NHS trusts and affected companies seemed to be running.

Cloudbleed

In February 2017, Cloudflare faced a major software bug that led to sensitive customer data like passwords, cookies and authentication tokens to get leaked from customer websites. Cloudflare is known to provide performance and security services to millions of customer websites and although the bug was patched within hours, it is expected that the data leakage could have started as early as September 2016.

Bitcoin Unlimited

Bitcoin Unlimited suffered a serious memory leak which caused several nodes to fall from 800 to about 300. This is almost 70 percent of the nodes run by Bitcoin Unlimited at the time. Although the memory leak was patched fairly quickly, this appeared to be the third memory leakage to crash the preferred method for Bitcoin Unlimited.

British Airways

For the sixth time in a year, British Airways faced a massive global IT failure which led to the airline cancelling all flights from Heathrow and Gatwick in May 2017. The IT failure affected over 1,000 flights, British Airways call centres, the website and mobile app. According to the GMB Union, the failure could have been avoided if the airline had not made hundreds of its IT staff redundant in 2016.

Nest thermostat leaves users in the cold

In mid-January 2016 the Nest 'smart' thermostat (owned by Google) was hit with a software glitch which left users, literally, in the cold. A software update went wrong, forcing the device's batteries to drain and leaving it unable to control temperature – so customers were unable to heat their homes or get hot water on one of the coldest weekends of the year so far. Nest said the fault had been caused by a December 4.0 firmware update, plus issues such as old air filters or incompatible boilers. It has since rolled out a 4.0.1 software update which it says has solved the issue for 99.5 percent of affected customers.

HSBC suffers major outage

Less than a week into 2016, HSBC became the first bank to suffer a major IT outage. Millions of the bank's customers were unable to access online accounts. Services only returned to normal after a two-day outage. The bank's chief operating officer Jack Hackett blamed a "complex technical issue" with its internal systems.

HSBC business banking glitch

HSBC again! In August 2015 a reported 275,000 individual payments failed to be processed by HSBC, which left many potentially without pay before the Bank Holiday weekend. The cause of this major failure was a problem with its electronic payment system for its business banking users which affected salary payments. Bacs is the payment system that is used for payment processes across the UK is reported to have picked up on the issue but noted that it was an 'isolated issue'.

Government divorce software failure

The government's online calculator for calculating spouses' financial worth was hit with a Form E fault meaning that those calculations were wrong for thousands of couples that have been divorced in the past 20 months. According to reports this error has been inflating spouses' finances since April 2014 but was only brought to light in December 2015.

According to the Office of National Statistics there were 114,720 divorces in England and Wales in 2013 so the damage that this fault has caused is yet to be fully revealed.

Bloomberg cancels debt issue

In April this 2015, Bloomberg's London office suffered a software glitch resulting in their trading terminals going down for two hours. This came at a very bad time as the UK's Debt Management Office (DMO) was set

to auction a series of short-term Treasury bills (these bills are short-term Government borrowing). In a statement Bloomberg said: "Service has been fully restored. We experienced a combination of hardware and software failures in the network, which caused an excessive volume of network traffic."

Starbuck's software bug

Back in April 2015, Starbucks witnessed a register malfunction which according to Starbucks, was caused by an 'internal failure' during its routine refresh. This resulted in 60 percent of stores in the US and Canada being forced to close. The affected stores were unable to process payment transactions and at one point many stores were giving the coffee away for free.

Heartbleed security flaw uncovered

In April 2014, a member of Google's security team found a flaw in the encryption library Open SSL, which hosts 66 percent of all websites. Although it was quickly patched by most IT firms, the sheer scale of the services affected means it is likely that there are still servers out there which remain vulnerable to attack.

Apple forced to pull iOS 8 update

Despite being the world's most profitable company, Apple had a major embarrassment in September 2014 when it had to pull the update for its new iOS 8 operating system a mere hour after release. Users complained of lost phone signal, frozen updates and unlocking problems. A study by Bloomberg claimed the latest operating system crashes 67 percent more often than its predecessor.

Emergency numbers go offline for six hours

Emergency services were unavailable for six hours across seven US states last April. The incident affected 81 call centres, meaning about 6,000 people made 911 calls that were unable to connect in Washington and parts of six other states. A study from Federal Communications Commission found an entirely preventable software error was responsible for causing the service to drop.

iCloud hacked

Compromising photos of A-list stars like Jennifer Lawrence and Kirsten Dunst appeared on 4chan and other internet forums in September 2014. Hackers had gained access to the stars' iCloud accounts using phishing schemes and brute-force guessing. Apple CEO Tim Cook promised to beef up iCloud's security features to restore user confidence after the hack.

CONCLUSION

Testing is a critically important verification method that takes up a very large portion of a project's resources, including schedule, budget, staffing, and facilities. Unlike the many constructive activities of systems engineering, testing is relatively unique because it is inherently destructive. Its primary purpose is to force the system or its components to fail so that the defects that caused the failure can be uncovered and then fixed. In addition to defect detection, testing is also performed to provide sufficient objective evidence to justify confidence in the system's quality, fitness for purpose, and readiness for being accepted and placed into operation.

REFERENCE

1. "Software Development Failures", Kweku Ewusi-Mensah MIT Press, 01-Aug-2003.
2. "Faulty Software May Have Doomed Mars Orbiter". Space.com. Archived from the original on July 24, 2008. Retrieved January 11, 2007.
3. "Software Bug Contributed to Blackout". Retrieved 2008-01-07.
4. "Software Failure Risk: Measurement and Management", Susan A. Sherer.
5. "Multimedia Networking: Technology, Management and Applications: Technology, Management and Applications" Syed, Mahbubur Rahman (July 1, 2001) Idea Group Inc (IGI).

CREDIT CARD FRAUD DETECTION TECHNIQUES AND EXISTING METHODS TO FRAUD DETECTION

S. Senthil Kumar¹, P. Balakrishnan² and Sathyabama B³

Research Scholar¹, Kongunadu Arts and Science College (Autonomous), Coimbatore
Assistant Professor², Dr. SNS Rajalakshmi College of Arts and Science (Autonomous), Coimbatore
Assistant Professor³, Hindusthan College of Arts and Science (Autonomous), Coimbatore

ABSTRACT

Fraud is one of the major ethical issues in the credit card industry. Firstly, to identify the different types of credit card fraud, and, secondly, to review alternative techniques that have been used in fraud detection. Credit card fraud may happen in various ways, which depend on the type of fraud concerned; it encapsulates bankruptcy fraud, theft fraud / counterfeit fraud, application fraud and behavioral fraud. Each of these sub-fraud categories has its own definition and specificity. With this extensive use of credit card, fraud appears as a major issue in the credit card business. Information in the credit bureau data is gathered from many different sources. Banks, consumer finance companies, credit unions, and collection agencies are some of the entities that periodically report to the credit bureaux. The main contribution of the proposed research is to introduce the framework which can accurately identify the credit card fraudulent activities.[8]

Keywords: Credit card, Credit card Bureau

TYPES OF FRAUD**Bankruptcy Fraud**

Bankruptcy fraud is one of the most difficult types of fraud to predict. However, some methods or techniques may help in its prevention. Bankruptcy fraud means using a credit card while being insolvent. In other words, purchasers use credit cards knowing that they are not able to pay for their purchases. The bank will send them an order to pay. The only way to prevent this bankruptcy fraud is by doing a pre-check with credit bureaux in order to be informed about the banking history of the customers.

Theft Fraud/ Counterfeit Fraud

Theft fraud means using a card that is not yours. The perpetrator will steal the card of someone else and use it as many times as possible before the card is blocked. The sooner the owner will react and contact the bank, the faster the bank will take measures to stop the thief. Counterfeit fraud occurs when the credit card is used remotely; only the credit card details are needed. At one point, one will copy your card number and codes and use it via certain web-sites, where no signature or physical cards are required.

Application Fraud

Application fraud is when someone applies for a credit card with false information. To detect application fraud, the solution is to implement a fraud system that allows identifying suspicious applications. To detect application fraud, two different situations have to be distinguished: when applications come from a same individual with the same details, the so-called duplicates, and when applications come from different individuals with similar details, the so-called identity fraudsters.

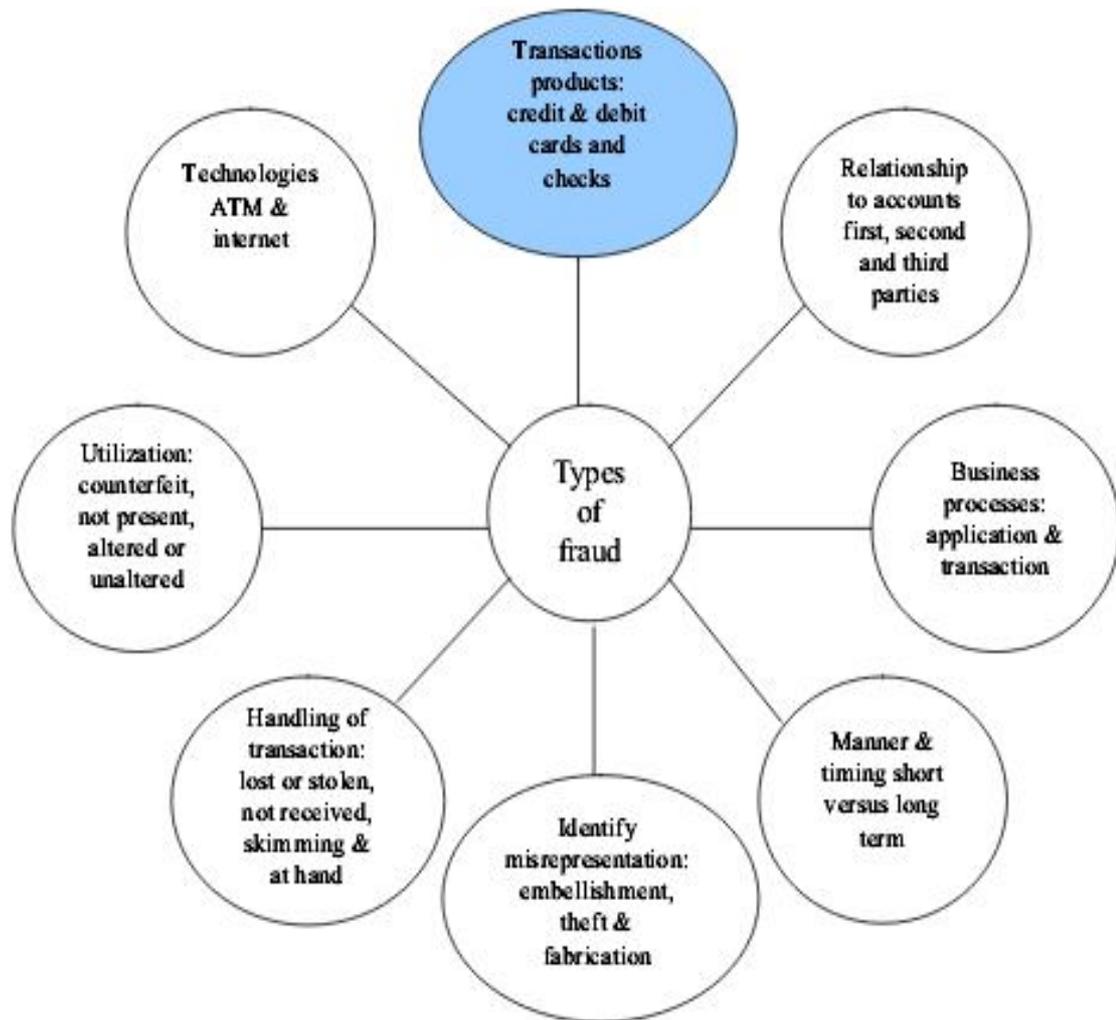


Fig: Types of Fraud

Behavioral Fraud

Behavioral fraud occurs when details of legitimate cards have been obtained fraudulently and sales are made on a ‘cardholder present’ basis. Behavioral fraud can be detected by implementing a fraud scorecard predicting which customers are likely to default. To build a scorecard, it is important to define what the profile of a fraudulent customer is, and especially the

card holder level profiles encapsulating normal transaction patterns, such as frequency of use, typical value range, types of goods purchased, To build a scorecard, it is important to define what the profile of a fraudulent customer is, and especially the cardholder level profiles encapsulating normal transaction patterns, such as frequency of use, typical value range, types of goods purchased, transaction types, cash usage, balance and payment histories, overseas spending patterns and daily, weekly, monthly and seasonal patterns.

EXISTING TECHNIQUES FOR CREDIT CARD FRAUD DETECTION

Decision Tree

The idea of a similarity tree using decision tree logic has been developed. A similarity tree is defined recursively: nodes are labeled with attribute names, edges are labeled with values of attributes that satisfy some condition and ‘leaves’ that contain an intensity factor which is defined as the ratio of the number of transactions that satisfy these condition(s) over the total number of legitimate transaction in the behavior (Kokkinaki, 1997). The advantage of the method that is suggested is that it is easy to implement, to understand and to display.

Genetic Algorithms

Algorithms are often recommended as predictive methods as a means of detecting fraud. One algorithm that has been suggested by Bentley et al. (2000) is based on genetic programming in order to establish logic rules capable of classifying credit card transactions into suspicious and non-suspicious classes. Basically, this method follows the scoring process. In the experiment described in their study, the database was made of 4,000 transactions with 62 fields. As for the similarity tree, training and testing samples were employed. Different types of rules were tested with the different fields. The best rule is the one with the highest predictability.

Other Algorithms

Chan et al. (1999) also developed an algorithm to predict suspect behavior. The originality of their research is that the model is evaluated and rated by a cost model, whereas other studies use evaluation based on their prediction rate/the true positive rate and the error rate/the false negative rate. They present different algorithms: diagnostic algorithms, diagnostic resolution strategies, probabilistic curve algorithms, best match algorithms, negative selection algorithms and density selection algorithms. They conclude from their investigation that neighborhood-based and probabilistic algorithms have been shown to be appropriate techniques for classification, and may be further enhanced using additional diagnostic algorithms for decision-making in borderlines cases, and for calculating confidence and relative risk measures.

Clustering Techniques

Bolton & Hand (2002) suggest two clustering techniques for behavioral fraud. The peer group analysis is a system that allows identifying accounts that are behaving differently from others at one moment in time whereas they were behaving the same previously. Those accounts are then flagged as suspicious. Fraud analysts have then to investigate those cases. The hypothesis of the peer group analysis is that if accounts behave the same for a certain period of time and then one account is behaving significantly differently, this account has to be notified. Break-point analysis uses a different approach. The hypothesis is that if a change of card usage is notified on an individual basis, the account has to be investigated. In other words, based on the transactions of a single card, the break-point analysis can identify suspicious behavior. Signals of suspicious behavior are a sudden transaction for a high amount, and a high frequency of usage.

Table-1: Different statistical techniques in credit card fraud

Study	Country	Method	Details
Aleskerov et al. (1997)	Germany	Neural networks	Card-watch
Bently et al. (2000)	UK	Genetic programming	Logic rules and scoring process
Bolton & Hand (2002)	UK	Clustering techniques	Peer group analysis and break point analysis
Brause et al. (1999a)	Germany	Data mining techniques & neural networks	Data mining application combined probabilistic and neuro-adaptive approach
Chan et al. (1999)	USA	Algorithms	Suspect behavioral prediction
Dorransoro et al. (1997)	Spain	Neural networks	Neural classifier
Ezawa & Norton (1996)	USA	Bayesian networks	Telecommunication industry
Fan et al. (2001)	USA	Decision tree	Inductive decision tree
Ghosh & Reilly (1994)	USA	Neural networks	FDS (fraud detection system)
Kim & Kim (2002)	Korea	Neural classifier	Improving detection efficiency and focusing on bias of training sample as in skewed distribution. To reduce "mis-detections".
Kokkinaki (1997)	Cyprus	Decision tree	Similarity tree based on decision tree logic
Leonard (1995)	Canada	Expert system	Rule-based Expert system for fraud detection (fraud modelling)
Maes et al. (2002)	USA	Bayesian networks & neural networks	Credit card industry, back-propagation of error signals
Quah & Sriganesh (2007)	Singapore	Neural networks	Self-Organizing Map (SOM) through real-time fraud detection system
Wheeler & Aitken (2000)	UK	Combining algorithms	Diagnostic algorithms; diagnostic resolution strategies; probabilistic curve algorithm; best match algorithm; negative selection algorithms; density selection algorithms and approaches
Zaslavsky & Strizhak (2006)	Ukraine	Neural networks	SOM, algorithm for detection of fraudulent operations in payment system

NEURAL NETWORKS

Neural networks are also often recommended for fraud detection. Dorransoro et al.(1997) developed a technically accessible online fraud detection system, based on a neural classifier. However, the main constraint is that data need to be clustered by type of account. Similar concepts are: Card watch (Aleskerov et al., 1997); Back-propagation of error signals (Maes et al., 2002); FDS (Ghosh & Reilly, 1994); SOM (Quah & Sriganesh, 2008; Zaslavsky & Strizkak, 2006); improving detection efficiency “mis-detections” (Kim & Kim, 2002). Data mining tools, such as ‘Clementine’ allow the use of neural network technologies, which have been used in credit card fraud (Brause et al., 1999a; Brause et al., 1999b).

Bayesian networks are also one technique to detect fraud, and have been applied to detect fraud in the telecommunications industry (Ezawa & Norton, 1996) and also in the credit card industry (Maes et al., 2002). Results from this technique are optimistic. However, the time constraint is one main disadvantage of such a technique, especially compared with neural networks (Maes et al., 2002). Furthermore, expert systems have also been used in credit card fraud using a rule-based expert system (Leonard, 1995).

CONCLUSION

Credit card fraud is an act of criminal dishonesty. This article has reviewed recent findings in the credit card field. This paper has identified the different types of fraud, such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, and discussed measures to detect them. Such measures have included pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms.

REFERENCES

1. Bolton, R. & Hand, D. 2002. ‘Statistical Fraud Detection: A Review’. *Statistical Science*, 17; 235-249.
2. Maes, S., Tuyls, K., Vanschoenwinkel, B. & B Manderick. 2002. Credit Card Fraud Detection using Bayesian and Neural Networks, Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies.
3. Quah T. S, & Sriganesh M. 2008. ‘Real-time credit card fraud using computational intelligence’. *Expert Systems with Application*, 35:4, 1721-1732.
4. Fan, W., Miller, M., Stolfo, S., Lee, W. & P Chan. 2001. Using Artificial Anomalies to Detect Unknown and Known Network Intrusions, Proc. of ICDM01; 123-248.
5. Bentley, P., Kim, J., Jung. G. & J Choi. 2000. Fuzzy Darwinian Detection of Credit Card Fraud, Proc. of 14th Annual Fall Symposium of the Korean Information Processing Society.
6. Leonard K. 1995. ‘The development of a rule based expert system model for fraud alert in consumer credit’. *European Journal of Operational Research*, 80; 350-356.
7. Ghosh, S. & Reilly, D. 1994. Credit Card Fraud Detection with a Neural-Network, Proc. of 27th Hawaii International Conference on Systems Science, 3; 621-630.
8. SenthilKumar, S, Mythili, S, (2017). Survey on Exception Rules and Anomaly Detection. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 2(6), 521-525

PREDICTION OF OVARIAN CANCER CLASSIFICATION USING BACTERIAL FORAGING ALGORITHM WITH MULTIPLE SWARM OPTIMIZATION TECHNIQUES

C. J. Pavithra¹, G. S. Narmada and P. Prema³ and T. Guhan⁴Students^{1,2,3} and Assistant Professor⁴, Department of CSE, Sri Ramakrishna Engineering College

ABSTRACT

Ovarian cancer is a cancer that forms in or around the ovary. It mainly depends upon the cancerous cells that found inside, near the ovary. It results in abnormal cells that have the ability to invade or spread to other parts of the body. The level of the ovarian cancer depends upon the stages of the cancer cells affect in the body. This cancer affects women who can be in smaller age or above the age of 40 and women suffer from this a lot. The detection of ovarian cancer at the early stage would be able to save the life of women. This paper mainly involves predicting the stages of the ovarian cancer based on the gene values associated with the intensity value of the different random persons. Based on the prediction of the stages, it results in three kinds, they are early, middle and severe stage. From which the starting stage cancer patients can be cured easily. Earlier diagnosis of Ovarian Cancer saves enormous lives. Then the middle stage of the cancer may be cured or may not. Then the final stage of the stages is very severe to cure the cancer. This work will cluster the gene value and its intensity of the ten person using k means and predict the optimal level with the help of bacterial foraging optimization. Finally the stages will be predicted with the help of the optimal level using particle swarm optimization. This can be predicted with the help of particle swarm optimization. .

Keywords: Ovarian Cancer, Optimal Level, bacterial foraging, particle swarm.

INTRODUCTION

Ovarian cancer is a Cancerous ovary growth and the growth of the part of the ovary may lead to spread of cancer to other parts of the body. Women who have a strong family history of breast or ovarian cancer are more likely to be affected by cancer of the ovaries, fallopian tube, or peritoneal cavity. It is estimated the 10-15% of ovarian, fallopian tube, or peritoneal cancers are associated with an inherited genetic mutation. The remaining majority of cases of cancer are linked to a genetic mutation that is acquired by the individual in their lifetime. DNA carries the hereditary information from parents to offspring. This is the reason why children look like their parents. Sometimes, this DNA carries defective genes. The defective genes may result in many diseases like cancer. This is thought to be due to a mutation in one of the genes that are involved in the regulation of cell growth and replication in these areas, which can be inherited from the parents.. Ovaries are responsible for the production of eggs, estrogen and progesterone. The cancer that attacks ovary is called ovarian cancer..Due to ovarian cancer many number of people are died due to predicting the cancer at the final stage. Diagnosis of cancer is mainly at the final stage of the cancer so that early detection and recovery of the cancer is not common among the people affected by cancer. The detection of cancer at the early stage may help people to treat the cancer with proper diagnosis. Several lives may be avoided from loss due to the early detection of cancer. Women who suffer from ovarian cancer die due to the final stage of predicting the cancer, out of 10% only 2-3% of women who are affected are escaped from death due to early detection of cancer.

Genes associated with ovarian cancer

The Breast Cancer 1 (BRCA1) and Breast Cancer 2 (BRCA2) genes have been identified as genes that are linked to an increased risk of the development of both breast cancer and ovarian cancer. Some of the genes that have been linked to an increased risk of ovarian cancer include:

CDH1: mutation is linked to a raised risk of ovarian and breast cancer.

MLH1 Gene: mutation is linked to a raised risk of both Lynch syndrome and ovarian cancer.

MLH2 Gene: mutation is linked to a raised risk of both Lynch syndrome and ovarian cancer.

PALB2 Gene: It encodes protein which may suppress tumors.

PTEN Gene: mutation is linked to a raised risk of Cowden syndrome and ovarian cancer.

STK11 Gene: mutation is linked to a raised risk of Peutz-Jeghers syndrome and ovarian cancer.

TP53 Gene: somatic mutation is present in almost half of all cases of ovarian cancer.

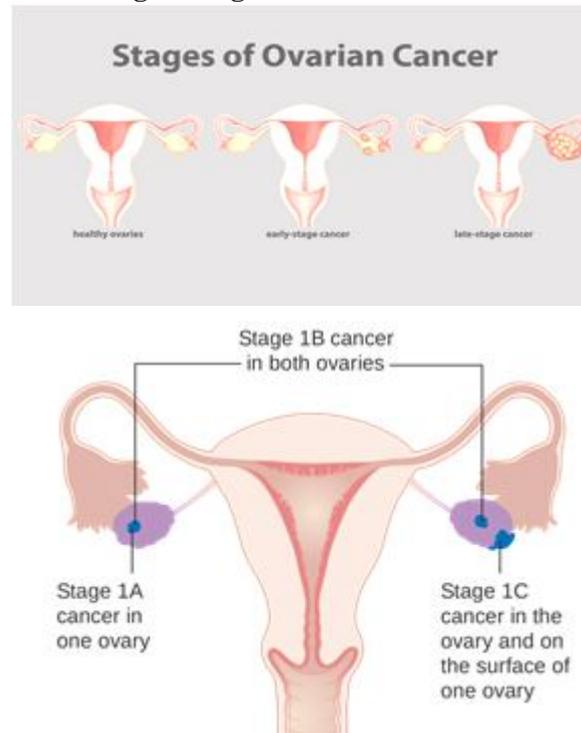
LITERATURE SURVEY

The cancer identification of genetic and environmental factors is very important in developing novel methods to detect and prevent cancer. Here, novel method layer combining clustering and decision tree technique to build

cancer risk prediction system which predicts lung, breast, oral, cervix, stomach and blood cancer. This research uses data mining technology such as classification, clustering and prediction to identify the potential cancer patients. Here, the data is clustered to separate the cancer patients and non cancer patients. Finally the prediction system is developed to analyze the risk levels which help in prognosis. This detection helps patients to know about their predisposition before going for clinical. The drawbacks of this paper are it doesn't predict the levels and stages of the cancer affected patients.

STAGES OF OVERIAN CANCER

Fig-1: Stages of ovarian cancer



Existing Idea

Ovarian cancer is a most lethal gynecological cancer which is increasing day by day in developing countries. Therefore, identification of genetic factors including mutations in BRCA1 and BRCA2 gene[3]. The algorithm used for the prediction of ovarian cancer are: K-means clustering algorithm for identifying the relevant data, Significant Frequent Pattern were discovered using Apriori tid and Decision tree.

The drawbacks of this system are:

- Predicting accuracy is to a small extent in the existing system.
- Stages of cancer was not predicted in this system.

Proposed System

Ovarian cancer may occur in all the ages of women. Cancer may occur in any part of the body and may spread to several other parts. This can be cured only when the cancer is identified in the earlier stage. This can prevent from spreading to other parts in malignant stage could save a person's life. By predicting the stages of the cancer, there may be a chance to prevent the middle stage to reach the severe stage. In this proposed system, the early, middle and the final stages of the ovarian cancer will be predicted.

In the proposed system the algorithms used are,

- Clustering can be done with the dataset by using k means algorithm
- Predicting the optimal feature selection for the levels of the ovarian cancer using bacterial foraging optimization.
- Predicting the early, middle and the final stage of the ovarian cancer using particle swarm optimization.

Dataset Description

- The dataset consists of gene values and intensity of the genes in different random persons
- Gene value: includes the different gene values which mainly causes ovarian cancer.

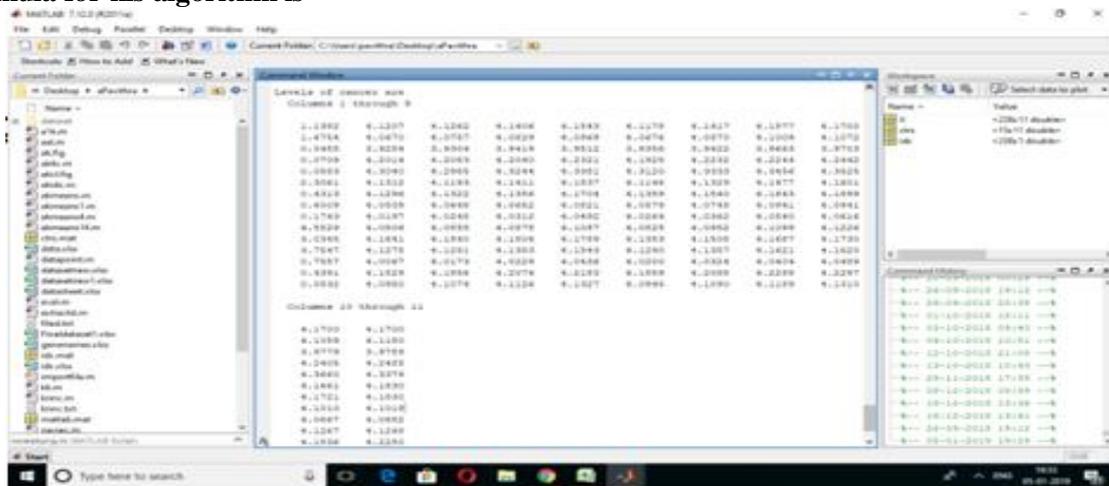
CLUSTERING USING KMEANS

k-means clustering is the process of combining related data set values into clusters without analysing the relationship between them. It is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition ‘n’ observations into ‘k’ clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into different cells. The algorithm works iteratively and assigns each data point to one of K groups based on similarity of features. The results of the K-means clustering algorithm are:

K-Means Clustering Algorithm

K-MEANS clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into different cells. In cluster analysis, the k-means algorithm can be used to partition the input data set into k partitions (clusters). However, the pure k-means algorithm is not very flexible, and as such is of limited use. In particular, the parameter k is known to be hard to choose when external constraints are not provided. Another limitation of the algorithm is that it cannot be used with arbitrary distance functions or non- numerical data.

The formula for his algorithm is



$$J(V) = \sum_{i=1}^c \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

Where,

$X_i - V_j \rightarrow$ The euclidian distance between X_i and V_j

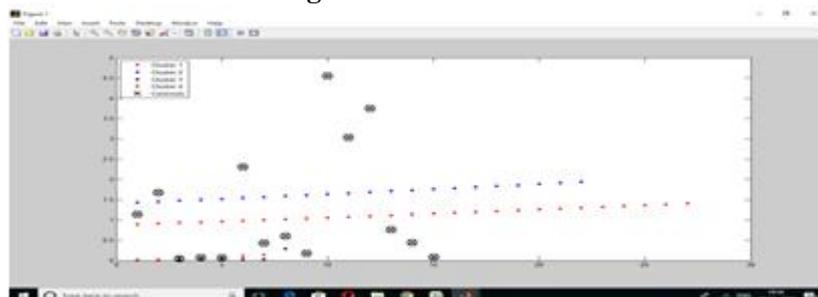
$C_i \rightarrow$ The number of points in i^{th} cluster

$c \rightarrow$ The number of cluster centers.

Fig 2 : ovarian cancer dataset

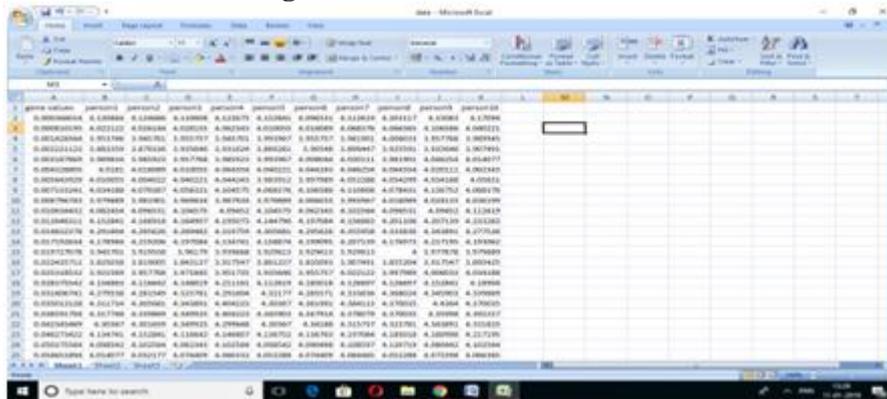
Dataset used for predicting the stages of ovarian cancer

Fig-3: k means cluster



Which forms the cluster using k means clustering algorithm using the gene values and the persons intensity level.

Fig-4: levels of ovarian cancer



Various levels of ovarian cancer for the given data set using k means clustering algorithm.

CONCLUSION

The levels of the ovarian cancer can be predicted using k means clustering algorithm with the help of the gene values and the patients intensity value. This is further enhanced to predict the optimal level of the ovarian cancer using bacterial foraging optimization and predicting the stages of the ovarian cancer using particle swarm optimization algorithm.

FURTHER ENHANCEMENT

Bacterial Foraging Optimization:

- Bacterial foraging optimization algorithm (BFOA) has been widely accepted as a global optimization algorithm of current interest for optimization and control.
- BFO possesses a poor convergence behavior over complex optimization problems as compared to other nature inspired optimization techniques.

Particle Swarm Optimization

- Particle Swarm Optimization algorithm (PSO) which is mainly used to reduce the medical errors which uses the computational method to optimize the solution based on the problem of the person.
- Particle swarm optimization normally works with the genetic algorithms and produce the optimal solutions by updating the generations.

REFERENCES

1. Arindam Roy, Bipu Shyam Purkayastha, Sunita Sarkar, 2013, "Application of particle swarm optimization in data mining clustering: A survey", International Journal of computer applications, volume 65-no.25.
2. T.Bhuvaneshwari, N.Girija, P.Ramachandran, 2014, "Early detection and prevention of cancer using data mining techniques", International Journal of Computer applications, Volume 97-no.13.
3. Devesh C.Jinwala, Pawan R.Bhaladhare, 2014, "A clustering approach for the diversity model in privacy preserving data mining using fractional calculus-Bacterial Foraging Optimization Algorithm", Hindawi Publishing Corporation, Article ID 396529.
4. Guhan Thangavelu, S Selvarajan, April 2017, An Accurate Prediction and Classification of Ovarian Cancer Using Microarray Data, Journal of Advanced Research in Dynamical & Control Systems, 01-Special Issue
5. Md Shariful Islam et al, 2016,"Early detection of ovarian cancer(braca 1 and braca 2mutation) risk prediction for low income country using data mining technology",6th International Conference on Bioinformatics and Systems Biology.
6. Narmada.G.S, Pavithra.C.J, Prema.P, Guhan.T, April 2018, "Ovarian Cancer Prediction Using Bacterial Foraging Optimization", International Journal of Computer Science and Information Technology Research Volume-6.

A SURVEY ON DATA MINING TECHNIQUES FOR HEART DISEASE PREDICTION

K. Dheenathayalan¹ and Dr. K. K. Savitha²Assistant Professor¹, Department of Computer Science, Kamban College of Arts and Science, Coimbatore
Assistant Professor², Department of Computer Application, Bharathiar University, Erode

ABSTRACT

Cardiovascular diseases (CVD) are the major cause of death globally, as well as in India. The reason for CVD is disorders of the heart and blood vessels, and includes coronary heart disease (heart attacks), cerebrovascular disease (stroke), raised blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. The healthcare industry collects large amounts of Healthcare data; they are information rich and knowledge poor. Unluckily not all the data are mined which is required for discovering hidden patterns and effective decision making. Data mining techniques are effectively used for knowledge discovery in healthcare dataset. This paper presents a survey about different data mining techniques used to predict the heart disease in human.

Keywords: ANN, CVD, Decision Tree, Feature selection, kNN.

1. INTRODUCTION**1.1 Cardiovascular diseases (CVD)**

CVD are a group of disorders of the heart or blood vessels, and include mainly hypertensive heart disease, Ischaemic heart disease, rheumatic heart disease and cerebrovascular disease or strokes. Tobacco use, an unhealthy diet, physical inactivity and harmful use of alcohol increase the risk of heart attacks and strokes. These inferior lifestyle choices can lead to metabolic risk factors such as overweight and obesity, high blood pressure, high blood sugar and high blood cholesterol. These become important, yet preventable risk factors for heart attacks and strokes. People may either experience symptoms of underlying cardiovascular disease years before a heart attack or stroke, or none at all. The most common symptoms and signs are:

- Chest pain or discomfort which may last for more than a few minutes
- Shortness of breath or difficulty in breathing
- Tiredness and weakness
- Pounding or racing heartbeat.
- Bulge of the ankles, feet, legs or abdomen.
- Excessive sweating.
- Nausea or lack of appetite
- Chronic coughing or wheezing
- Confusion or impaired thinking.

Heart failure is mounting at an alarming rate in India and has almost attained outbreak proportions. The estimated heart failure trouble in India ranges from 1.3 million to 4.6 million cases per annum. Various types of heart diseases are:

- Coronary heart disease
- Cardiomyopathy
- Cardiovascular disease
- Ischaemic heart disease
- Heart failure
- Hypertensive heart disease
- Inflammatory heart disease
- Valvular heart disease

1.2 Risk factors of CVD**a) Non-Modifiable Risk Factors**

- Age
- Sex
- Family History
- Genetic Factors

b) Modifiable Risk Factors

- Smoking
- High Blood Pressure
- Diabetes
- Obesity
- Sedentary lifestyle/Lack of physical exercise
- Unhealthy food habits like oil fried foods
- Stress

2. DATA MINING IN HEALTHCARE

Data mining is the analysis of large data sets to find out patterns and use those patterns to forecast or predict the likelihood of future events. Data mining holds great potential for the healthcare industry to facilitate health systems to systematically use data and analytics to recognize inefficiencies and best practices that improve care and lessen costs.

Electronic health records (EHR) are quickly becoming more common among healthcare facilities. With increased access to a large amount of patient data, healthcare providers can now optimize the efficiency and quality of their organizations using data mining. Healthcare providers employ data mining and data analysis to find best practices and the most effective treatments.

Data mining techniques compare symptoms, causes, treatments and negative effects and then proceed to analyze which action will prove most effective for a group of patient's illness or condition. This is also a technique for providers to develop the best standards of care and clinical best practices. Another application of data mining is identifying strange patterns of medical claims by clinics, physicians, labs, or others and also used to identify inappropriate referrals or prescriptions and insurance fraud and fraudulent medical claims. This manuscript gives survey on given below data mining techniques for heart disease prediction in human.

- Naive Bayesian
- Decision tree
- KNN – k-Nearest Neighbour algorithm.
- Feature Selection based Least Square Twin Support Vector Machine (LSTSVM)
- Associative classification algorithm using genetic approach
- ANN –Artificial Neural Network.
- GA
- SVM (Support Vector Machine)
- Multilayer Perceptron (MLP)
- Weighted Associative Classifier (WAC)

2.1 Naive Bayesian (NB)

The Naive Bayesian Classifier technique is particularly suited when the dimensionality of the inputs is high. Naive Bayesian model identifies the characteristics of patients with heart disease. Naive Bayesian is a statistical classifier which assumes no dependency between attributes. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes. The advantage of using Naive Bayesian is that one can work with the Naive Bayesian model without using any Bayesian methods.

2.2 Decision tree (DT)

Decision Tree is a popular classifier which is simple and easy to implement. There is no requirement of domain knowledge or parameter setting and can high dimensional data can be handled. It produces results which are

easier to read and interpret. The drill through feature to access detailed patients’ profiles is only available in Decision Trees. Decision Tree includes various types of algorithms such as ID3, C4.5, C5, J48 (Java Implementation of the C4.5 Algorithm) and CART (Classification and Regression Tree.). A decision tree can be used to classify a case by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance.

2.3 kNN – k-Nearest Neighbour algorithm and GA

This approach combines KNN and genetic algorithm to improve the classification accuracy of heart disease data set. Genetic search is used as a goodness measure to prune redundant and irrelevant attributes, and to rank the attributes which contribute more towards classification. Least ranked attributes are removed, and classification algorithm is built based on evaluated attributes. This classifier is trained to classify heart disease dataset as either healthy or sick.

2.4 ANN – Artificial Neural Network

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural network. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers input layer, hidden layer & output layer. In feed-forward neural networks the neurons of the first layer forward their output to the neurons of the second layer, in a unidirectional fashion, which explains that the neurons are not received from the reverse direction. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input x_i into neurons in hidden layer. Neuron of hidden layer adds input signal x_i with weights w_{ji} of respective connections from input layer. The output Y_j is function of $Y_j = f(\sum w_{ji} x_i)$ where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

2.5 Multilayer Perceptron (MLP)

A multilayer perceptron (MLP) is a class of feed forward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. Feed forward neural network where each node outputs an activation function applied over the weighted sum of its inputs

2.6 Weighted Associative Classifier (WAC)

Weighted Associative Classifier is a new concept that uses Weighted Association Rule for classification. Weighted ARM uses Weighted Support and Confidence Framework to extract Association rule from data repository. The WAC has been proposed as a new Technique to get the significant rule instead of flooded with insignificant relation. Weighted Associative Classifiers is another concept that assigns different weights to different features and can get more accuracy in predictive modeling system like medical field etc.

2.7 Feature Selection based Least Square Twin Support Vector Machine (LSTSVM)

A predictive model for heart disease diagnosis using Feature Selection based LSTSVM. This model uses F-score for selecting significant input features and Grid Search method to obtain the best parameters for predictive model. F-score is used to evaluate the importance of each feature of heart disease dataset. The selection of significant features improves the accuracy of a classifier, so instead of taking all the features only the relevant features are considered. This results in a classifier model which is more predictive and involves lesser computation.

2.8 Support Vector Machine (SVM)

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. A Support Vector Machine performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors that define the hyperplane are the support vectors. The beauty of SVM is that if the data is linearly separable, there is a unique global minimum value.

Table-1: Comparison of various Data mining Techniques for Heart Disease Prediction

Reference	DM techniques used / compared	Accuracy (%)	No. of Attributes used	Result (Best technique)
Abhishek Rairikar <i>et al</i> (2017)	kNN	-	13	kNN
	Decision Tree	-		
	Naive Bayes	-		

Theresa Princy. R <i>et al (2016)</i>	kNN-ID3	40.3	5	kNN-ID3
		80.6	7	
K.Gomathi et al (2016)	Naive Bayes	79	5	Naive Bayes
	Decision Tree(J48)	77.03		
Manjusha B. <i>et al (2015)</i>	MLP –NN	96.69	13	MLP-NN with 16 attributes
	MLP –NN	98.16	16	
G. Purusothaman <i>et al (2015)</i>	Decision Tree	76		Hybrid
	Association Rule	55		
	kNN	58		
	ANN	85		
	SVM	86		
	Naive Bayes	69		
	Hybrid Approach	96		
Hlaudi Daniel Masethe <i>et al(2014)</i>	J48	99.07	11	J48 REPTREE And Simple CART
	REPTREE	99.07		
	Naive Bayes	97.22		
	Bayes Net	98.14		
	Simple CART	99.07		
Aditya Methaila <i>et al(2014)</i>	Naive Bayes	86.53	13	DT
	Decision Tree	89		
	ANN	85.53		
Aditya Methaila <i>et al(2014)</i>	Naive Bayes	96.53	6	DT
	Decision Tree	99.20		
	Classification via clustering	88.30		
Aditya Methaila <i>et al(2014)</i>	Decision Tree – Genetic Algorithm	99.62	15	DT - GA
Divya Tomar <i>et al(2014)</i>	Feature Selection Based LSTSVM	85.59	11	Feature Selection Based LSTSVM
B.Venkatalakshmi <i>et al(2014)</i>	Naive Bayes	85.03	13	NB
	Decision Tree	84.01		
S.J. Gnanasoundhari <i>et al(2014)</i>	Naive Bayes	52.33	Mainly focused on Age, sex, blood pressure and blood sugar	WAC
	WAC	81.51		
	NN	78.43		
	SVM	60.78		

The above table, Table1 shows the comparison of various data mining classification techniques used to diagnose the heart disease in human.

3. CONCLUSION

The main motivation of this study is to provide the insight of data mining techniques used for prediction of risk factors of heart diseases commonly known as CVD (cardio vascular disease). Various data mining techniques and classifiers are discussed in many studies to help the researchers for their research career. This paper gives introduction to some of the familiar data mining techniques such as Naive Bayesian, Decision tree, kNN, SVM, WAC, Feature Selection based Least Square Twin Support Vector Machine (LSTSVM) and ANN –Artificial Neural Network. From Table1, it has been seen that hybridizing two or more techniques can result in more accuracy. This work can be extended in future; other combinations of data mining techniques can be tested on more or less attributes than previous works.

4. REFERENCES

1. Aditya Methaila et al, Early Heart Disease Prediction Using Data Mining Techniques; CCSEIT, DMDB, ICBB, MoWiN, AIAP pp. 53–59, 2014.
2. Abhishek Rairikar et al, “Heart Disease Prediction Using Data Mining Techniques”, International Conference on Intelligent Computing and Control (I2C2) 2017,

3. B.Venkatalakshmi and M.V Shivsankar, Heart Disease Diagnosis Using Predictive Data mining; International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 3, March 2014.
4. Divya Tomar and Sonali Agarwal, “ Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease”, International Journal of Bio-Science and Bio-Technology Vol.6, No.2 (2014), pp.69-82.
5. G. Purusothaman and P. Krishnakumari, “A Survey of Data Mining Techniques on Risk Prediction: Heart Disease”, Indian Journal of Science and Technology, vol. 8, no. 12, (2015).
6. Hlaudi Daniel Masethe and Mosima Anna Masethe, “prediction of Heart Disease using Classification Algorithms”; Proceedings of the World Congress on Engineering and Computer Science 2014.
7. K.Gomathi and Dr. D. Shanmuga Priyaa “Multi Disease Prediction using Data Mining Techniques”, International Journal of System and Software Engineering Volume 4 Issue 2, December 2016.
8. Miss. Manjusha et al, “A Data Mining Approach For Classification Of Heart Disease Dataset Using Neural Network”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 4, Issue 5, May 2015 ISSN 2319 – 4847.
9. S. J. Gnanasoundhari, G. Visalatchi and M. Balamurugan, “A Survey on Heart Disease Prediction System Using Data Mining Techniques”, International Journal of Computer Science and Mobile Application, vol. 2, no. 2 (2014), pp. 72-77.
10. Theresa Princy. R and J. Thomas “Human Heart Disease Prediction System using Data Mining Techniques”, International Conference on Circuit, Power and Computing Technologies [ICCPCT] 2016.

WIRELESS ATTACKS**K. Baby¹ and Dr. M. Shanthakumar²**Research Scholar¹ and Assistant Professor, Kamban College of Arts and Science, Coimbatore
Assistant Professor², Department of Computer Science, Kamban College of Arts and Science, Coimbatore

ABSTRACT

In this modern era, wireless access is very usual one to all the people. Wireless attacks have become a very common security issue in networks. There are different type of hackers and different type of attacks are held in everyday. It is necessary to take security measures are taken to prevent these attacks. The different type of wireless attacks are rogue access points, jamming, Evil twin, war driving, blue snarfing, packet sniffing, near field communication, Replay attacks, Wi-Fi Protected Access, Wi-Fi Protected Setup attacks.

Keywords: blue snarfing, evil twin, jamming, packet sniffing, war driving.

1. INTRODUCTION**Rogue Access Points**

An access point is created by the hacker without the knowledge of user. This access point is invisible. This creates a back door for hackers. This makes some security issues to the users.

Jamming

Jamming means disruption of one's network. The wireless signals will always get disrupted or signal may be lost. This may be done by Bluetooth headset or other signals. These signals made transmission and receiving of wireless signals very difficult.

Evil Twin

It is a fraudulent access point created by hackers. Fake access points are set up by configuring a wireless card. they act as an access host network and it is very difficult to trace.

War driving

The other name of war driving is access point location. One can extend wifi range by fitting antenna in an office. The outside user will be able to intrude into the network. The outsider gets a free Internet connection, and possibly gains access to all the records and other resources of the company **Wireless Attacks**

Blue snarfing

It is a malicious attack happened in Bluetooth enabled devices. Stealing information like contacts, images, videos over the blue tooth network.

Packet sniffing

Packet sniffing means capturing the packets while one send data across the network. What kind of data that are sending one to individual. Packet sniffing is possible without encrypting a information.

Near field communication

It is a wireless communication between devices like smart phones and the apps like share it, cshare. The people are transferring the information between the smart phones without the need to bring the devices in contact. This allows hacker to collect information from the device in close range.

Replay attacks

The hacker spies on information being sent between a sender and a receiver. He also spies on conversations between the two people. Once he started spied on the information he intercepts it or retransmit it again. This makes delay in data transmission and lack of security issues.

WEP/WPA attacks

The hackers monitoring the network for a long period of time. He can gather enough the information from clear text to determine the encryption key, which never changes a new twist to WEP, TKIP. Breaking the Encryption standard makes the hacker easily to read the information.

Wi-Fi Protected Setup attacks.

It is a very dangerous attacks done using password guessing tools. With the help of these tools one can retrieve the wireless network passwords. And also gain access to data and information that is on one's network Wireless Attacks.

2. CONCLUSION

The Wi-Fi networks normally have lot of loopholes. This makes the attacker make it very easy to carry out their attacks. Therefore Identifying the wireless attacks are very important and it makes preventing such attacks.

3. REFERENCES

- Andrea Saracino ; Francesco Restuccia ; Fabio Martinelli , Practical Location Validation in Participatory Sensing Through Mobile WiFi Hotspots, 2018, 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)
- Guolong Liu ; Jinliang Liu ; Yan Li ; Liang Xiao ; Yuliang Tang, Jamming Detection of Smartphones for WiFi Signals , 2015 IEEE 81st Vehicular Technology Conference (VTC Spring)
- Matthew Knysz , Open WiFi networks: Lethal weapons for botnets?
- Omar Nakhila ; Cliff Zou , User-side Wi-Fi evil twin attack detection using random wireless channel monitoring, MILCOM 2016 - 2016 IEEE Military Communications Conference
- Wenlong Shen ; Bo Yin ; Xianghui Cao ; Lin X. Cai ; Yu Cheng, Secure device-to-device communications over WiFi direct , IEEE Network ,2016.

A REVIEW ON SECURE AND DYNAMIC SEARCH METHODS OVER ENCRYPTED CLOUD DATA

P. Muthukumar¹, S. Sasikala and Renjith K V²Assistant Professor^{1,2} and HOD³, Department of Computer Science, Nilgiri College of Arts and Science, Nilgiris

ABSTRACT

Today, due to the enormous growth of data technology in cloud computing, the data owners are stimulated to outsource their data in data management to reduce cost and for the convenient. But the main problem in outsourcing is the lack of security and privacy for the sensitive data. So, to overcome this, the sensitive data should be encrypted before it should be outsourced for privacy requirement. This obsoletes data utilization in some cases like keyword – based documentation retrieval. Moreover, the traditional methods such as plain text keyword method are results in ineffective retrieval and become overhand on the encrypted data availability. Various methods were proposed to preserve the privacy and to provide security to the encrypted cloud data. Most of the searchable encryption focuses on single keyword search or Boolean key word search. In this review paper, various searching methods are analyzed to identify the problems persists in providing the security and privacy to the encrypted data.

Keywords: Boolean keyword, Encryption, Text keyword method, Single keyword, Searching methods.

I INTRODUCTION

Nowadays, cloud computing has been taken into consideration as a new edition of IT infrastructure, that may put together large beneficial useful resource of computing, storage and programs, and allow users to revel in ubiquitous, convenient and on-demand network get admission to a shared pool of configurable computing sources with splendid overall performance and minimal monetary overhead . Attracted by way of those appealing features, each individuals and firms are stimulated to outsource their facts to the cloud, rather of buying software program and hardware to manipulate the facts themselves. No matter the various blessings of cloud services, outsourcing touchy information (which includes e-mails, non-public health statistics, enterprise business enterprise finance records, authority's files, and plenty of others) to far flung servers brings privacy worries. The Cloud Carrier Vendors (CSVs) that preserve the facts for users may also additionally get admission to users' sensitive information without authorization. A fashionable technique to shield the information confidentiality is to encrypt the statistics earlier than outsourcing but, this will motive a big rate in phrases of records usability. For an example, the present techniques on keyword-based totally information retrieval, which are widely used on the plaintext records, cannot be immediately implemented at the encrypted records. Downloading all of the statistics from the cloud and decrypt regionally is impractical. On the opposite, greater sensible unique motive solutions, including searchable encryption (SE) schemes have made unique contributions in phrases of performance, capability and safety. Searchable encryptions schemes permit the purchaser to save the encrypted data to the cloud and execute key-word seek over cipher text area. To this point, abundant works were proposed beneath different hazard fashions to gain numerous search functionality, which include single key-word search, similarity seek, multi-key-word Boolean search, ranked search, multi-key-word ranked seek, and so forth. Amongst them, multi-keyword ranked seek achieves an increasing number of interest for its practical applicability.

In recent times, a few dynamic schemes were proposed to assist inserting and deleting operations on document collection. These are vast works as its miles notably possible that the facts proprietors want to update their records on the cloud server. However few of the dynamic schemes help green multi-key-word ranked search. This paper proposes a relaxed tree-based totally scheme over the encrypted cloud records, which facilitates multi-key-word ranked seek and dynamic operation at the report collection. Specifically, the vector area version and the appreciably-used —Time Frequency (TF) \times inverse record frequency (IRF) model are blended within the index manufacturing and query generation to offer multi-key-word ranked seek. Which will attain excessive seek performance, we assemble a tree-based index shape and recommend a greedy intensity-first search algorithm primarily based on this index tree. Due to the unique structure of our tree-based totally index, the proposed seek scheme can flexibly collect sub-linear seek time and cope with the deletion and insertion of documents.

The below Fig-1 shows the concept of cloud computing

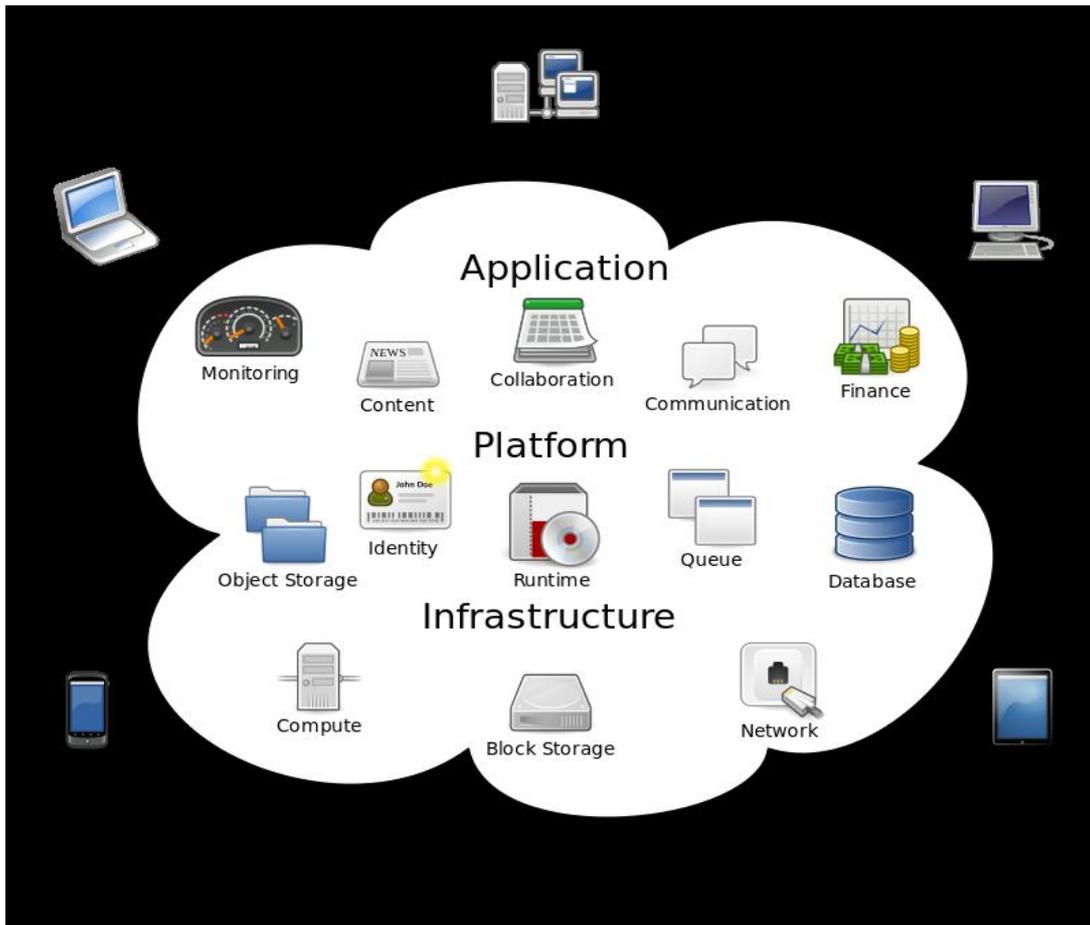


Fig-1: Concept of Cloud Computing

We assemble secure search schemes: the Basic Dynamic Multi-key-word Ranked Search (BDMRS) scheme inside the regarded cipher text model, and the improved dynamic multi-key-word ranked seek (EDMRS) scheme inside the diagnosed background model. Our contributions are summarized as follows:

- 1) We layout a searchable encryption scheme that enables each the accurate multi-keyword ranked search and bendy dynamic operation on file collection.
- 2) The proposed scheme can reap higher search performance thru executing our greedy intensity-first search set of rules. Moreover, parallel search may be flexibly carried out to in addition reduce the time cost of search technique.

This paper is organized as follows: In Section 2, the literature survey of the papers related to the searchable encryption techniques. Then the paper is concluded in Section 3.

II. LITERATURE SURVEY

This section of paper summarize the literature survey of various papers related to Multi-keyword ranked search over encrypted cloud data in cloud environment.

1. Secure and privacy preserving keyword search

Qin Liu et.al. [1] proposed a secure and privacy-preserving keyword search over the encrypted data for cloud storage applications using Elliptic Curve Cryptography (ECC) over FP. However, this scheme supports only Boolean key- word search i.e., either a keyword exists in a file or not, without considering the difference of relevance with the queried keyword of these files in the result.

Limitations: The correspondence and computational expense for encryption and decoding is more.

2. Secure and Efficient Ranked Keyword Search

Cong Wang [2] proposed a definition for ranked searchable symmetric encryption, and gives an efficient design by properly utilizing the existing cryptographic primitive, order-preserving symmetric encryption (OPSE)

Limitations: It doesn't play out numerous catch phrases seeks. Minimal overhead in record building

3. Single Keyword Search over Encrypted data on cloud

C Wang et al [3] proposed that searchable encryption plan agree to a client to solidly search for over scrambled information through watchwords without first applying decoding on it. The proposed systems reinforce just traditional Boolean catch phrase look, without catching any relevance of the records in the query item. Its next deficiency is at the point when straight forwardly connected in substantial joint information outsourcing cloud environment.

Limitations: Too slow in processing the request, not suitable for huge volume of data.

4. Privacy-preserving Multi-keyword Text Search

Wenhai Sun et.al.[4] proposed this inquiry that gives comparability based item positioning, catch phrase security, Index and Query classification and Query Unlink capacity. The scrambled record is worked by vector space model supporting solidified and particular document look. The searchable file is fabricated utilizing Multidimensional B tree. Proprietor makes scrambled question vector for document catch phrase set. Client gets the individual encoded inquiry vector from proprietor who is given to CS. Presently CS looks list by Merkle Damgård development calculation and thinks about cosine measure of document and question vector and returns top k scrambled records to client.

Limitations: The likeness rank score of the record vector completely relies on upon the kind of the report.

5. Secure Multi-catchphrase Top-k Retrieval Search

Jiadi et.al [5] proposed this pursuit utilizing Two Round Searchable Encryption (TRSE). In first round, clients presents various catchphrase "REQ" "Was encoded question for finishing information, watchword protection and make trapdoor (REQ, PK) as TW and sends to cloud server. At that point cloud server figure the score from encoded file for documents and returns the scrambled score result vector to client. In second round, client decode N with mystery key and ascertains the document positioning and afterward ask for records with Top k scores. The positioning of record is done on customer side and scoring is done on server side.

Limitations: The withdrawal and restricting is utilized to diminish figure content size, still the key size is too extensive. The correspondence elevated will be high, if the scrambled trapdoor's size is too vast. It doesn't make powerful searchable file redesign.

6. Privacy Preserving Multi-Keyword Ranked Search (MRSE)

Ning [6] proposed this quest for known figure content model and foundation model over encoded information giving low calculation and correspondence overhead. The direction coordinating is decided for multi watchword seeks. They utilized internal item likeness to quantitatively assess similitude for positioning documents. The downside is that MRSE have little standard deviation σ which debilitates catchphrase protection.

Limitations: Multi-watchword positioned look (MRSE) for known figure content model may deliver two diverse trapdoor which dubious the protection spillage issue of trapdoor unlink capacity which may debilitate the catchphrase security. MRSE has little standard deviation σ which thusly debilitates the watchword protection. The honesty of the rank request is not checked in MRSE.

7. Attribute-based Keyword Search

Wenhai Sun [7] proposed Attribute-based Keyword Search that gives conjunctive catchphrase look; watchword semantic security and Trapdoor unlink capacity. The proprietors makes file with all catchphrases and access list with strategy characteristics which determines the clients list approved for seeking. Presently proprietors scramble the archive, record with access list utilizing cipher text strategy characteristic based encryption procedure. To have client enrolment administration, they utilized intermediary re-encryption and languid re-encryption strategies to share the workload to CS. The client asks for the TW to CS utilizing its private key. Presently CS recovers TW and ventures the encoded files and return records just if the user's properties in TW fulfil access approaches in files which make coarse-grained dataset look approval.

Limitations: Trap door era will require additional time with the expanded number of traits.

8. Efficient and Secure Multi-Keyword Search on Encrypted Cloud Data

C Orencik and E Savas[8] proposed this technique which characterized and tackled the issue of successful yet protected and sound rank watchword seek over Encrypted cloud information. Positioned look enormously upgrades framework ease of use by giving back the coordinating records in a positioned request with respect to certain vital criteria (e.g. watchword recurrence) accordingly making one stage nearer towards sensible utilization of secure information facilitating administrations in Cloud Computing. These papers has characterized and tackled the testing issue of security saving and productive multi watchword positioned look

over scrambled cloud information stockpiling (MRSE), and set up an arrangement of strict protection prerequisites for such an ensured cloud information use framework to wind up a reality. The proposed positioning strategy ends up being effective to do a reversal to a great degree applicable reports comparing to submit seek terms. Proposed positioning strategy is utilized as a part of our future framework keeping in mind the end goal to improve the security of data on Cloud Service Provider.

Limitations: Dynamic redesigning and erasure of the record from the cloud is unrealistic.

9. DMRS: an efficient dynamic multi-keyword ranked search over encrypted cloud data

Lanxiang Chen, ET. Al [9] proposed a novel and efficient scheme which is improved from traditional secure kNN computation. It supports dynamic operation on document collection, allowing insertion and deletion operations in document. The future work consists of fine-grained access authorization and the revocation of the user is big challenges.

Limitations: Lack of granting access authorization and revocation.

III CONCLUSION

In this paper, various key word searching techniques used in retrieving data in cloud computing environment are discussed. The advantages and the disadvantages are also well discussed. From the above discussion, the following observations are made.

- ❖ In the recent years, researchers have proposed many cipher text search schemes by incorporating the cryptography techniques. In addition, the relationship between documents is concealed in the above methods. The relationship between documents represents the properties of the documents and hence maintaining the relationship is vital to fully express a document. For example, the relationship can be used to express its category. If a document is independent of any other documents except those documents that are related to sports, then it is easy for us to assert this document belongs to the category of the sports.
- ❖ Due to the blind encryption, this important property has been concealed in the traditional methods. Therefore, proposing a method which can maintain and utilize this relationship to speed the search phase is desirable.
- ❖ The use of Merkle hash tree and cryptographic signature to create a verifiable MDB-tree. However, their work cannot be directly used in our architecture which is oriented for privacy-preserving multiple keyword search. Thus, a proper mechanism that can be used to verify the search results within big data scenario is essential to both the CSPs and end users.
- ❖ Existing methods have been proven with provable security, but their methods need massive operations and have high time complexity. Therefore, former methods are not suitable for the big data scenario where data volume is very big and applications require online data processing.
- ❖ Some method has a high searching cost due to the scanning of the whole data collection word by word.
- ❖ Due to the lack of rank mechanism, users have to take a long time to select what they want when massive documents contain the query keyword. Thus, the order-preserving techniques are utilized to realize the rank mechanism,
- ❖ New architecture is also proposed which achieves better search efficiency. However, at the stage of index building process, the relevance between documents is ignored.

REFERENCES:

- Q. Liu, G. Wang, and J. Wu, "Secure and privacy preserving keyword searching for cloud storage services," *Journal of Networks and Computer Applications*, vol. 35, no. 3, pp. 927–933, May 2012
- Cong Wang, N. Cao, K. Ren and W. Lou, —Enabling secure and efficient ranked keyword search over outsourced cloud data, *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, issue 8, (2012) August, pp. 1467–1479.
- C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encrypted cloud data," in *Proc. of ICDCS'10*, 2010.
- W. Sun, B. Wang, N. Cao, M. Li, W. Lou, YT, Hou and H. L, —Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking, In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, (2013), pp. 71-82.

-
- N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 1, pp. 222–233, 2014.
 - Yu, Jiadi & Lu, Peng & Zhu, Yanmin & Xue, Guangtao & Li, Minglu. (2013). Towards Secure Multi-Keyword Top-k Retrieval over Encrypted Cloud Data. *IEEE Transactions on Dependable and Secure Computing*. 10.1109/TDSC.2013.9.
 - W. Sun, B. Wang, N. Cao, M. Li, W. Lou, YT ,Hou and H. L, —Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking, In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, (2013), pp. 71-82.
 - Cengiz Örencik , Erkey Savaş, An efficient privacy-preserving multi-keyword search over encrypted cloud data with ranking, *Distributed and Parallel Databases*, v.32 n.1, p.119-160, March 2014
 - Chen, Lanxiang, et al. "DMRS: an efficient dynamic multi-keyword ranked search over encrypted cloud data." *Soft Computing* 21.16 (2017): 4829-4841
 - C Orencik, M Kantarcioglu, E Savas A practical and secure multi-keyword search method over encrypted cloud data 2013 IEEE Sixth International Conference on Cloud Computing, 390-397
 - Peng, Tao, Qin Liu, and Guojun Wang. "a multilevel access Control scheme for Data security in Transparent Computing." *Computing in Science & Engineering* (2016).
 - Shen, Jian, et al. "An authorized identity authentication-based data access control scheme in cloud." *Advanced Communication Technology (ICACT), 2016 18th International Conference on*. IEEE, 2016.
 - Shen, Zhirong, Jiwu Shu, and Wei Xue. "Keyword Search with Access Control over Encrypted Cloud Data." *IEEE Sensors Journal* 17.3 (2017): 858-868.

A STUDY AND ANALYSIS OF IMAGE MINING TECHNIQUES AND ITS APPLICATIONS

N. Sundaravalli¹ and Dr. R. Vidya Banu²Research Scholar¹ and Assistant Professor², L.R.G Govt. Arts College for Women, Tirupur**ABSTRACT**

Data mining is the process of sorting through large database to identify patterns and establish relationships to solve problems through data analysis. A considerable part of data mining is image mining. Image mining deals with the extraction of image patterns from images stored in the large collection of image data base. Image plays an important role in every aspect of medical diagnosis, satellite image, space research, industries, remote sensing, even in the agriculture field. Analysing these images we get useful information. This paper is mainly focus on image mining process, image mining techniques, various extraction mechanisms used in image mining and its applications.

Keywords: Image mining, Data Mining, Medical Diagnosis, Agriculture images, CBIR.

I INTRODUCTION

In the current digitalization world, tremendous collection of digital data are stored in each and every second. Digital data are in different forms like Video, Audio, Image and Text. Image mining deals with the extracting image data and entrenched knowledge, image data relationship, or other patterns which is not clearly found in the images [1]. Image mining is additional than just an extension of data mining to image domain.

In image mining, the elemental challenge is how to expose out low-level pixel representation enclosed in a raw image or image sequence can be processed to recognize high-level image objects and relationships.[2] In general, image mining deals with study and growth of new technologies. The aspire of image mining is

To find out the image pattern from a given group of images [3].

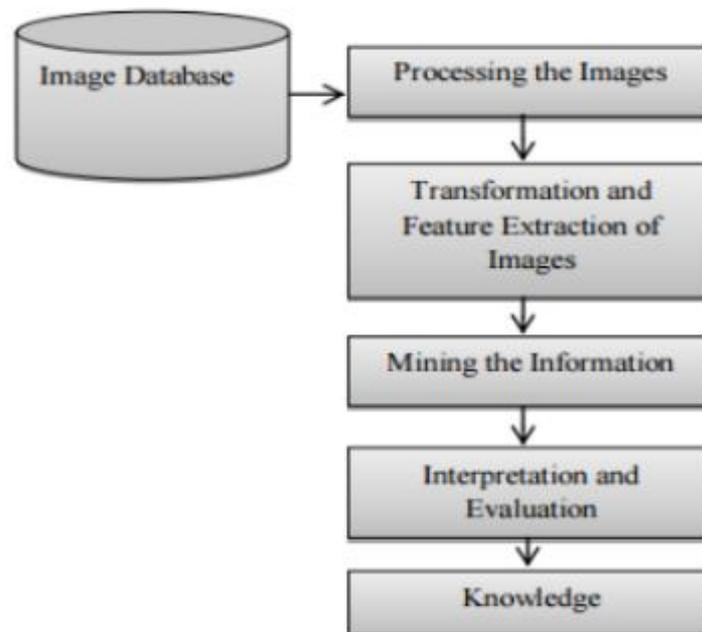


Figure-1: Process of Image mining

Figure 1 shows the process of image mining. To improve the quality of image, preprocess the image from massive image database is the first step. Then the images are go through no of transformations and feature extraction to create significant features. along with the developed features, mining process can be accomplish using data mining techniques to find out the considerable patterns.

The consequential pattern are then guesstimate and interpreted to find the final knowledge, which will be applied to applications [4][5].

II APPLICATIONS OF IMAGE MINING

In current world, image mining involves various fields like medical diagnosis, space research, DNA matching, Agriculture field, satellites image such image data are provides knowledge and information for decision making[6].

A. IMAGE MINING FOR

Agriculture

India is an agriculture country. around 80% of people depend on agriculture. Agriculture plays vital role in human life as well as Indian economy. Countries like india and china provide major priority to agriculture as many community depend more on agriculture. Now a days, agricultural organizations work with large no of data. Recently data processing techniques in agriculture essential evaluating. Data mining methods are aimed towards discover those schemes of work in data which are valuable and interesting for crop management[12].

Ramar .K., et al., [13]. provides image Data mining and classification techniques applied to soil image database can be winning in establishing significant relationships from the data.[14]

Dr K. Prakash Dr P. Saravanamoorthi Mr R. Sathishkumar, Dr M. Parimala et al., provides a method to control plant disease using image mining techniques.[15].and they provide a comprehensive survey on the subject, and gave the input of the various diseases affected leaves. so we can reduce the usage of weedicides, thus save the surroundings.

Image mining will be used for the following agriculture purposes

- i). To find diseased stem ,leaf,, fruit
- ii). To find and quantification of affected areaby disease.
- iii). Identification of intensity of diseases and their effect on productivity.

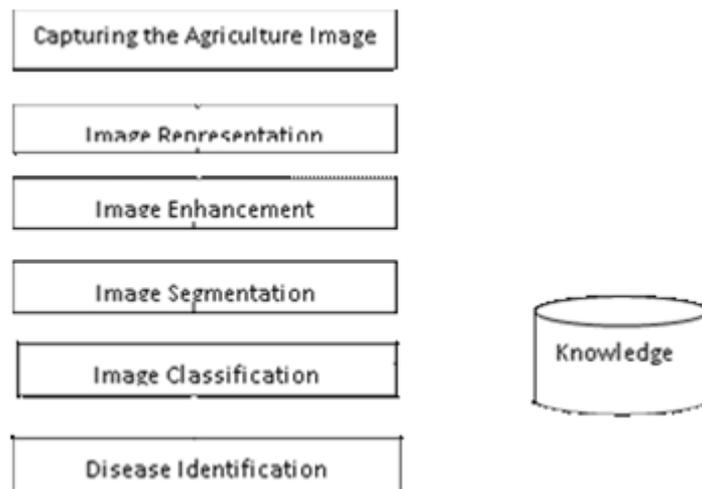


Figure-2: Crop Desease Identification



Figure-3: Symptoms of Crop Crops shell-eating disease

B. MEDICAL IMAGE MINING

Medical image refers to lot of techniques like X-ray, CT,MRI and altrosound these are creates visual representations of organs of the body for clinical analysis[7]. Image mining and data mining algorithms are already use for several applications. For example these applications are profitably in use of identification of defects and wrongly areas in medical image[8]. some of them are listed below

- i) Diagnosis
- ii) Locate tumors and pathologies
- iii) Finding cancer cell
- iv) Treatment planning
- v) X -Ray
- vi) Brain image and so on.

Image pre processing for medical images is normally done by change the converting the host images into binary images. and to identify the evasion region of host image. for example, tumor affected in brain image.

Nemat et al. (2010) offered an analytical study of 62 patient’s records with dissimilar age group. The age of the patients plays a key role for primary radiography image diagnosis.[9]

P. Rajendran, M.Madheswaran [10] projected a technique concerned with the classification of brain tumor in the CT scan brain images. The important steps concerned in this system are pre-processing,extraction, hybrid classifier andassociation rule mining.

Aswini Kumar Mohanty, Saroj Kumar Lenka[12] proposed image mining method such as breast mammograms to categorize and search the cancerous tissue. in this a cross approach of future selection using fast branch and bound algorithm and hybrid genetic algorithm are used and reduce 75% of the features and new decision tree is used for categorization and provide hopeful results

C. CONTENT BASED IMAGE RETRIVAL

Multimedia Data Mining can be defined as the process of finding interesting patterns from media data such as audio, video, image and text.

Image retrieval is the quick increasing and demanding research part with regard to both still and moving images[16]. content-based image retrieval (CBIR) was introduced in the early 1980s

In Content-based image retrieval (CBIR) Ying et al. (2007) is the accepted tools to search still or moving images from the image database based on the visual content of the query image. There are many applications and in different areas like searching cultural heritage, education, home entertainment, art collections,crime prevention, in medical imaging, etc. in this CBIR system.

Rao et al. explain In colour-based CBIR schemes require low processing cost, but these are not perfect since sometimes two similar images may not be same due to improper illumination.[17]

F. Long et al. provide a technique for mathematical expressions defined in subsequent sections and their descriptions and nomenclature are taken from To retrieve images, and users provide the retrieval system with example images or sketched figure[18].

In this, following are some feature of content based image retrival, they are

- a. Image Content Descriptors
- b. Color
- c. Color Histogram
- d. Color Moments
- e. Texture
- f. Shape
- g. User Interaction

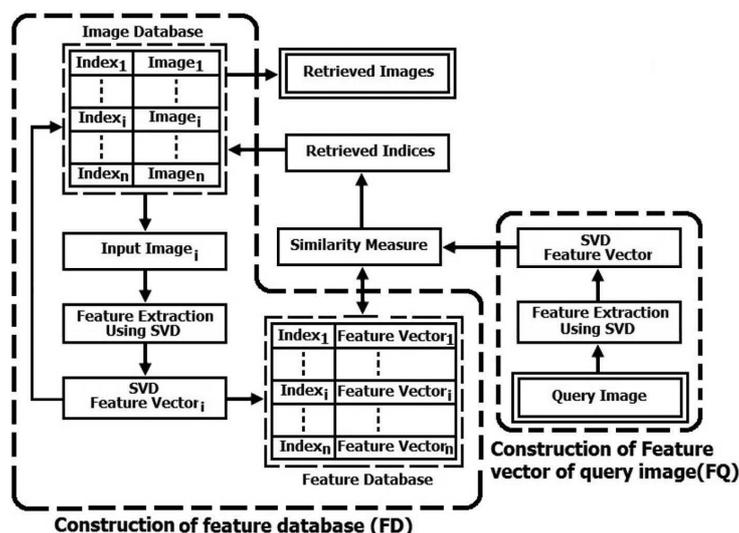


Figure-3: Block diagram of CBIR model

VI REFERENCES

1. Ji Zhang Wynne Hsu Mong Li Lee “ Image Mining: Issues, Frameworks and Techniques”
2. Ji Zhang, Wynne Hsu and Mong Li Lee “An Information-Driven Framework for Image Mining” Database and Expert Systems Applications in Computer Science, 2001, Volume 2113/2001, 232-242, DOI: 10.1007/3-540-44759-8_24
3. J. Fernandez, N. Miranda, R. Guerrero, F. Piccoli “Applying Parallelism in Image Mining” March 26, 2007
4. Peter Stanchev, “Image Mining for Image Retrieval”, In Proceedings of the IASTED Conference on Computer Science and Technology, pp 214-218, 2003.
5. Jiawei Han , Micheline Kamber , Jian Pei, "Data Mining; Concepts and Techniques", Reference text, Third edition.
6. J. Priya , Dr. R. Manicka Chezian , " A Survey on Image Mining Techniques for Image Retrieval ", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 7, July 2013
7. Carlton, R.R. and Adler, A.M. (2001) Principles of Radiographic Imaging: An Art and a Science, 3rd ed., Thomson Learning, Stanford, CT.
8. H. Seddik and E. Ben Braiek, Color Medical Images Watermarking, Based Neural Network Segmentation, GVIP Special Issue on Medical Image Processing, March, 2006.
9. Nemat, M.H., Hussein, A.M. and Juma, U.H. (2011) ‘Management of esophageal foreign bodies, retrospective study’, Facilities Medicine, Baghdad, Vol. 53, No. 1, pp.6–11.
10. P. Rajendran, M. Madheswaran “Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm” Journal Of Computing, Volume 2, Issue 1, January 2010, ISSN 2151-9617
11. Aswini Kumar Mohanty, Saroj Kumar Lenka, “Efficient Image Mining Technique for Classification of Mammograms to Detect Breast Cancer” Special Issue of IJCCT Vol. 2 Issue 2, 3, 4; 2010 for International Conference [ICCT-2010], 3rd-5th December 2010
12. Mucherino, A. and G. Ruß, 2011. Recent Developments in Data Mining and Agriculture. Retrieved November 30, 2012, from <http://www.antoniomucherino.it>: <http://www.antoniomucherino.it/download/papers/MucherinoRuss-dma11.pdf>
13. R. Vamanan and K. Ramar, “Classification of agricultural land soils a data mining approach,” International Journal of Computer Science and Engineering, vol. 3, no. 1, pp. 379-384, 2011.
14. I.V.Bragin , " Passive Microwave Complex Of Remote Sensing Ifor The Detection Of Objects In The Soil".
15. Dr K. Prakash Dr P. Saravanamoorthi Mr R. Sathishkumar, Dr M. Parimala “A Study of Image Processing in Agriculture” Int. J. Advanced Networking and Applications Volume: 09 Issue: 01 Pages: 3311-3315 (2017) ISSN: 0975-0290
16. S.Chidambaranathan “ Image Mining with CBIR” IJCSMC, Vol. 4, Issue. 3, March 2015, pg.12 – 15
17. Rao, M.B., Rao, B.P. and Govardhan, A. (2011) ‘CTDCIRS: content based image retrieval system based on dominant color and texture features’, International Journal of Computer Applications, Vol. 18, No. 6, pp.40–46.
18. Liu, G.H. and Yang, J.Y. (2013) ‘Content-based image retrieval using color’, Pattern Recognition, Vol. 46, No. 1, pp.188–198.

A COMPARATIVE STUDY ABOUT OUTLIER DETECTION TECHNIQUES**R. Sinivasan¹, S. Sivasankar², Dr. V. Kavitha³ and B. Sathyabama⁴**Student^{1,2}, Associate Professor³ and Assistant Professor⁴, Hindusthan College of Arts and Science, Coimbatore

ABSTRACT

Data mining is one of the significant technique for real time research which is used to extract fruitful knowledge from the wide range of database. Among the data mining techniques outlier detection is one of the major technique which is utilized to discover the deviated object than differ from the behavior of other data objects. Discovering an outlier objects from a collection of patterns is a major obstacle in the broad area of data mining. An essential challenge with outlier detection is that it is not well explained problem like other data mining techniques. Hence the technique of outlier detection requires more attention by the researchers. Outlier detection techniques can pinpoint errors and destroy their contaminating reaction on the data objects and as such to purify the information for processing. Discovering outliers and analyzing huge volume of data sets can lead to detect the hidden knowledge in the real time applications like telecommunications, web logs, web document and fraud detection. In this research paper focusing about various kinds of outlier techniques with various different approaches and also discuss about pros and cons of applications.

Keywords: Outlier Detection, Anomalies, High Dimensional

I. INTRODUCTION

Outlier detection technique is one of the data mining concept used is used to apply in numerous kinds of real world data sets like numerical dataset, graphical dataset, Audio and video datasets etc., Discovering outlier data points can lead to the identification of fruitful and significant knowledge. Outlier detection is the technique of discovering information with behaviors that are non identical from expectations. Like these data objects are referred anomalies or outlier. Identifying outliers from a collection of data objects pattern is a major obstacle in the broad area of data mining. An essential challenge with outlier detection is that it is not well explained problem like other data mining techniques. Hence the technique of outlier detection requires more attention by the researchers.

Outlier detection and analysis has numerous applications in various real world applications like criminal behaviors, fraud detection, discovering computer intrusion, credit card, medical applications, public health applications and industrial damage detection etc., Generally, the idea of these application is to discover data which deviates from usual standard behavior of data in the real dataset. For that reason it required a novel technique to discover outlier data object from the huge range of data sets. Outlier is defined as, it is a pattern which is not similar with respect to the remaining set of patterns in the whole data sets. It is referred as a data object that varies from other data objects in the data base. Outlier identification is a task of outlying observation. Outlier technique is one that appears to remarkably from the remaining data objects of the sample in which it prevail. Commonly the database contains data objects that do not comply with the common model or feature of the data object. These data objects are outliers. Outlier data objects are patterns in data base that is not conform to a well defined concept of usual behavior. The outlier data point is very divergent from the remaining of the data based on some other measuring factor. An outlier object can be elaborated as all information which may occurs to be non identical or distinct to the remaining data objects. An outlier is a monitoring process which appears to discovering outlier objects which is suitable on the data set. In data mining techniques, the outlier detection technique denotes to the identification of data object point which is not follow the assumption data behavior or pattern in a specific data set or is outstandingly different from the remaining data points in a data set.



Figure-1: Image of outlier

In figure 1 describes about the concept of outlier detection. In this figure there are collection of apples are there. Most of the apples are green in color. It denotes that all green apples have the features like color, shape, size and taste also. But the thing is only one apple is deviated from the rest of all the apples. Which is appeared in red color. Hence this apple is considered as an outlier. Because this must be different color then remaining of the apples. Each apple is considered as an individual data objects. So only one data object is deviated from the rest of the data object groups, which is specially called as outlier data point.

II. KINDS OF OUTLIERS

Outliers can be divided into various kinds. Namely erroneous, real, global, contextual and collective outliers respectively. Figure 2 depicts about the various kinds of outlier types.

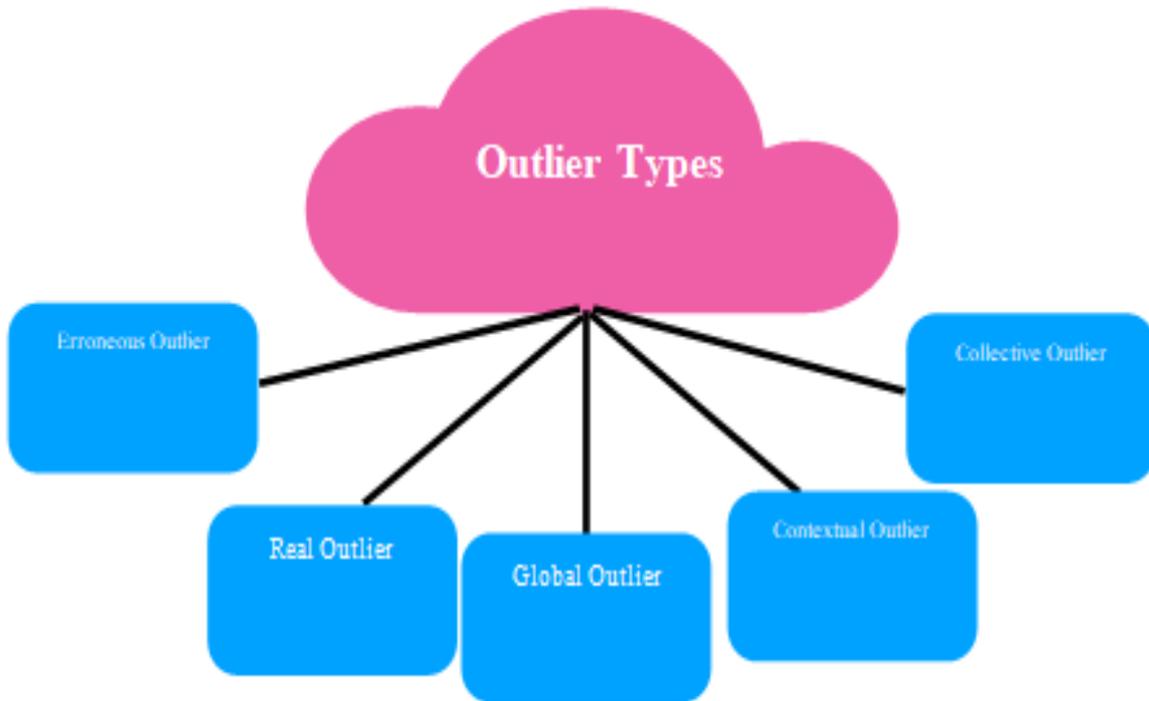


Figure-2: Kinds of Outlier Types

Erroneous Outliers

Some of the outlier observations are to be noticed that incorrect as an outlier, because of some catastrophic failure or any inherent obstacles, then these are considered as wrong outliers or illusive outliers. These kind of outlier are referred as erroneous outliers then it will remove as a noise.

Real Outlier

Monitoring data objects helps to discover the researcher something innovative and novel data object, that object must be removed anyhow, and must be completely left with the standard region. That data objects are declared as a real outliers, can not able to remove as a noise.

Global Outliers

Whether any individual data object instance are to be considered as anomalous with comparing the remaining data objects, then that particular data point is termed as outlier point. Global outliers are also referred as anomalies.

Contextual Outliers

Any data object is in rare appearance with respect to other specific context and it is a standard occurrence with remaining to other context, then such kinds of data objects are referred as contextual data objects.

Collective Outliers

Any data object is not anomalous but its collection with the hole dataset is anomalous, then that object is called as a collective outlier, the unique data instances which are referred as collective outlier could not be the outlier themselves but their appearance gathered as collective is anomalous, hence it is referred as collective outliers.

III. VARIOUS OUTLIER ANALYSIS APPROACHES

Outlier analysis approaches are in various types. Basically it is classified into classification based approach, cluster based approaches, statistical Approach, proximity based Approach, Frequency based approach. Figure 3 depicts about various kinds of outlier analysis approaches.

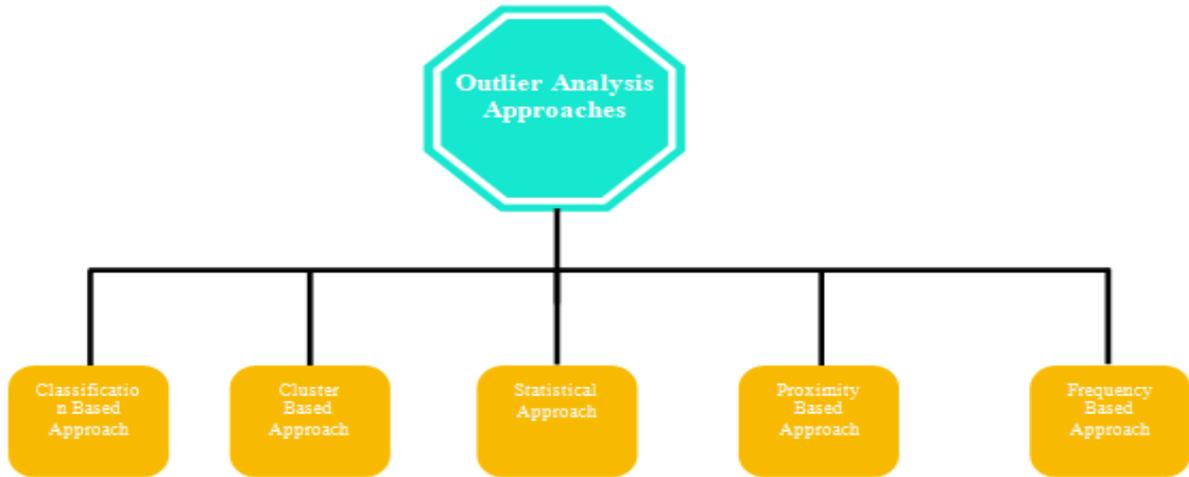


Figure-3: Outlier Analysis Approaches

Classification Based Approach

Classification is one of the significant technique in data mining, which is correlated with the concept of outlier detection. But for this classification process it needs the training data set to do the process. Hence, the idea behind the classification based outlier detection methods is to train up the classification model that can deviate the standard from outlier data instances. Due to the count of standard sample data likely far away of outlier sample data points, the training set is conclusively biased which is used to build an accurate classifier.

Cluster Based Approach

This kind of approach is used to detect outlier data object by testing the relationship among the data objects and clusters. An outlier data instance is an object which is based on a small and remote clusters, or it does not based to any specific cluster. Clustering based outlier detection using to find distance to the nearest cluster. Based on this approach, similarity among two data objects is calculated with the help of distance among the two data objects in data space, if this particular distance exceeds a specific fixed threshold, then the data instance will be referred as the outlier.

Statistical Approach

These kind of approaches were the historic algorithms utilized for outlier ascertainment. This approach guesses that data object follows any standard or predefined model or probability model, and it aims to discover outliers with respect to the specified model using a discordance experiment. This test is utilized to find whether a giver data object is an outlier point or not. The ideology of the statistical methods for outlier observations is to study a generative model fitting the specific data instances, and then discover those kind of objects in low likely regions of the data model as outliers. Statistical methods executes poorly on high dimensional data set. Statistical approach for outlier could be separated into two significant categories. They are named as parametric methods and nonparametric methods. Figure 4 explains about the categories of statistical approach.

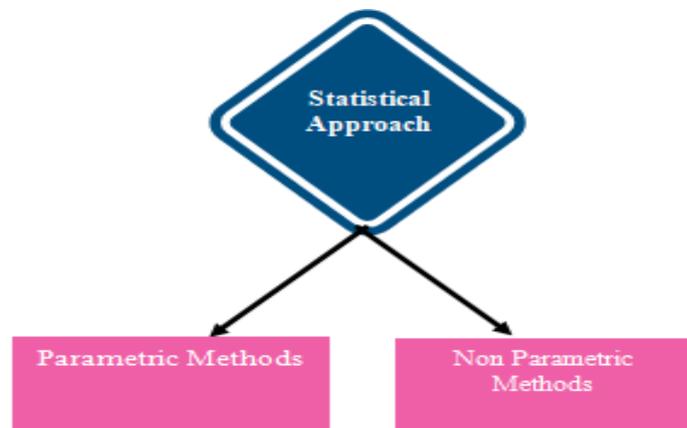


Figure-4: Categories of Statistical Approach

The former method of parametric method using parametric technique develop only with model difficulty not data size. Assumes that the standard data instances are produced by a parametric distribution. And the later one is nonparametric method is studied from the input data object. It could not make any expectation about statistical distribution of the data object.

PROXIMITY BASED APPROACHES

This approach expect that the proximity of an outlier data object to its close range neighbors outstandingly deviates from the proximity of the data object to the most of the other remaining data objects in the data base. Generally this approach is divided into three kinds namely density based approach, grid based approach and distance based approach.

The very first method of density based approach of outlier detection discovers the density allocation of the data and find out the outliers as those present less density region. Some of the clustering algorithms are structured to discover spherical shaped clusters but that is complicated to discover clusters of arbitrary shape. But the method of density based approach overcome this drawback. Moreover the density based outlier detection approach uses density distribution of data objects within data base.

The second method of distance based approach is another method is one of the most significant technique and widely accepted and used in the broad field of machine learning and data mining. It is mainly based on the concept of local neighborhood of data instances. Hence this is also referred as nearest neighbor analysis and it is Aldo used for various purposes like classification, decision making, clustering and association rules, outlier analysis.

Third method of grid based approach quantizes the location into a limited number of cells which form a grid structure on which all the manipulations for grouping are functioned. Most important advantage of utilizing this grid based method is speed processing time that based only on size of grids replaced of data size. The trouble with majority irregular data distribution, the declaration of grid must be too fine to attain better clustering quality. Table 1 interprets about the various techniques of outlier detection.

Table-1: Various Outlier Detection Techniques

Outlier Detection Techniques	Algorithms	Benefits	Drawbacks
Clustering Based Outlier Detection Techniques	<ul style="list-style-type: none"> * PAM Algorithm * CLARA Algorithm * CLARAN Algorithm * BIRCH Algorithm * Chameleon Algorithm 	<ul style="list-style-type: none"> * Minimal Complications. * Will support in high dimensional dataset. * Not necessary to assign the number of clusters. 	Whether the cluster shape is arbitrary then can not able to identify the number of clusters
Classification Based Outlier Detection Techniques	<ul style="list-style-type: none"> * Neural Networks * Support Vector Machine 	<ul style="list-style-type: none"> * Easily categories non linear and linear data objects * Able to handle noisy data effectively. * Classify the patterns even through the data is not trained. 	<ul style="list-style-type: none"> * Training times very poor. * Not good Interoperability.
Statistical Based Outlier Detection Techniques	<ul style="list-style-type: none"> * Regression * Scatter Point Method * Histogram * Kernel Density Function * Kernel Feature Space 	<ul style="list-style-type: none"> * Outlier detection techniques are better performed using statistical based approach. 	<ul style="list-style-type: none"> * Very complication in handling outlier technique. * Efficiency is very less. * It performs very poor in high dimensional data.
Proximity Based Outlier Detection Techniques	<ul style="list-style-type: none"> * DBSCAN * OPTICS * DENCLUE * K-Means * K-Nearest neighborhood * K-Mediod * STING * Wave Cluster * Clique 	<ul style="list-style-type: none"> * Interoperability level is too high. * Discover the cluster in arbitrary shape object also. * Technique is very easy and implementation is very simple. * Processing time is very fast. 	<ul style="list-style-type: none"> * Mechanisms very hard to understand. * Increase the curse of dimensionality. * Not obtain the well structured outcome when the data is irregular.

IV. OUTLIER DETECTION METHODS IN HIGH DIMENSIONAL DATABASE

In many real world applications are included with high dimensional domains in which the data set could contains hundreds or even dozens of dimensions. The techniques of outlier detection are reviewed in the preceding segments utilize concepts of proximity in order to discover the outliers depends on their relationship to the other data points in data set. In high dimensional space, the data are sparse and techniques utilizing the notion of failure proximity to attain most of their effectiveness. To handle the high dimensional data in research work is divided into two kinds. The first kind of outlier handling method is projecting the high dimensional data to lower dimensional data. These type of dimensionality detection techniques are PCA (Principal Component Analysis), ICA (Independent Component analysis), SVD (Singular Value Decomposition) could be applicable for high dimensional data bases before performing the outlier detection techniques. Feature selection is one of the preprocessing work for outlier detection which more essential for data mining concepts. The next category of method is more promising recent challenging.

Space Cube Method

Many researchers conducted some pioneering work against the high dimensional outlier detection technique. Hence the new technique is introduced which is used to find outliers by monitoring the density distributions of projections from the data. This new technique considers a data point to be an outlier point if in some lower dimensional projection it is placed in a local region of abnormal low density. Hence, the outlier points in these lower dimensional projections are discovered by simply searching for these projections featuring lower density.

Example Based Method

A recent approach of outlier examples provided by users are utilized to discover outliers in high dimensional data space. Purposely the man kind used or domain experts initially provide the systems with a few basic outlier samples. The technique discovers the subspaces attained in the initial steps. Then this technique divided the data subspace into equal depth cells and employs the sparsity coefficient proposed to calculate the non outlier samples in each subspace of the space. The fitness of the data subspace is the average sparsity coefficients of all data cubes in that data subspace to which it belonging to the outlier samples. All the data objects contained in the data cubes which are the outlier samples in the data subspace are discovered as outlier data points.

V. CONCLUSION

Discovering an outlier objects from a collection of patterns is a major obstacle in the broad area of data mining. An essential challenge with outlier detection is that it is not well explained problem like other data mining techniques. Hence the technique of outlier detection requires more attention by the researchers. Outlier detection techniques can pinpoint errors and destroy their contaminating reaction on the data objects and as such to purify the information for processing. Discovering outliers and analyzing huge volume of data sets can lead to detect the hidden knowledge in the real time applications like telecommunications, web logs, web document and fraud detection. Various kinds of outlier types and outlier analysis approaches are discussed. Moreover the specific approaches are suitable for the determined and distinct kinds of data sets. And also various outlier detection techniques are listed in the tables with specific advantages and also the disadvantages.

REFERENCE

1. Chiu, S. L. (1994). Fuzzy model identification based on cluster estimation. *Journal of Intelligent and Fuzzy Systems*, 2, 267–278.
2. Hodge, V. and J. Austin, A Survey of Outlier Detection Methodologies, *Artificial Intelligence Review*, Vol. 22, 2004, pp. 85–126.
3. Victoria J. Hodge and Jim Austin, A Survey of Outlier Detection Methodologies, Kluwer Academic Publishers, 2004.
4. Sankar K. Pal, P. Mitra, Data Mining in Soft Computing Framework: A Survey, *IEEE transactions on neural networks*, vol. 13, no. 1, January 2002.
5. Parneeta Dhaliwal, MPS Bhatia and Priti Bansal “A Cluster-based Approach for Outlier Detection in Dynamic Data Streams” Vol. 2, Issue 2, Feb 2010.
6. Ms. S. D. Pachgade, Ms. S. S. Dhande ”Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach” *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 4, June 2012.
7. S.Vijayarani S.Nithya” An Efficient Clustering Algorithm for Outlier Detection” *International Journal of Computer Applications* (0975 – 8887) Volume 32– No.7, October 2011.

8. Dhanashree S. Deshpande “A survey on web Data Mining Application” Emerging Trends in Computer Science and Information Technology -2012.
9. H.S Behera, Rosly Boy,Lingdoh, Diptendra Kodama singh “An Improved hybridized k-means clustering algorithm for high dimensional data set & it’s performance analysis” International Journal on Computer Science and Engineering (IJCSSE).
10. M. Elahi, K. Li, W. Nisar, X. Lv, and H. Wang, “Efficient clustering- based outlier detection algorithm for dynamic data stream,” in Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery – Volume 05, ser. FSKD '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 298–304.
11. Carlos Ordonez, “Clustering Binary Data Streams with K-means”, proceedings of ACM international conference on management of data, sigmod 1998.
12. Madjid Khalilian, Norwati Mustapha , “ Data Stream clustering- Challenges and issues”, Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong ,Vol I, pp.17 - 19,March 2010.
13. Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, “CURE: an efficient clustering algorithm for large databases” ACM LIBRARY, 1999.

HEARTBEAT PULSE DETECTION BY HEARTBEAT SENSING USING INTERNET OF THINGS: IOT

S. Subhasini¹, Dr. V. Kavitha², J. S. Anita Lily³ and B. Satyabama⁴Assistant Professor^{1,3,4} and Associate Professor², Hindusthan College of Arts and Science, Coimbatore

ABSTRACT

This paper evaluates the developing research into utilization of heartbeat facts as a biometric for human identification. Varieties of methods consume to be present proposed for obtaining heartbeat signatures and a choice of handling methods has been inspected. We approach the biometric identification and substantiation problem by describing the three major factors affecting performance: individual alternates, conservational variants, and sensor modifications. The capability to collect and practice the signal, activity the data for individual identification or verification, and circulate the information depends on all three of these factors.

1. INTRODUCTION

In the novel era of information and technology, the volatile growth of electronic devices, smart phones and tablets, which can be transfer substantially or wirelessly has become the essential tool of day-to-day life. Heart, is the most vital component of the circulatory system. It evaluates around 250–350 grams i.e. about the size of a fist. It strokes around 2.5 billion times during as lifetime of 66–68 years. A distinctive unit called Sino atrial node electrically stimulates the heart. This region crops a convinced potential and slowly discharges, thus sending an electrical impulse across the atria. This electrical impulse is sequential in nature and responsible for the systole and diastole in the four slots respectively. Late years have seen a growing interest for wearable sensors and today a few devices are industrially reachable for individual human services, fitness, and movement mindfulness. Notwithstanding the field entertaining ability field taken into account by ebb and flow gadgets, analysts have furthermore considered procedures of such progressions in clinical submissions in remote health inspection frameworks for long haul demo, management and clinical admission to patient's Physiological data. In light of current creative patterns, one can jagged envision a period quicker moderately than later when your routine physical test disappeared before by a two-three-day time of determined physiological observing, consuming reasonable wearable sensors.

This monitoring system satisfies the basic desires of persistent healthcare for heart diseases, also reflects the cost to confirm the universal mode as cost-effective as possible. Furthermore, it can also be collective with real-time analysis algorithms to assess patients' health disorder and give warnings to possible attacks in advance, which can make the common healthcare more intelligent.

2. INTERNET OF THINGS

The Internet of things is the internetworking of objective plans, vehicles (also referred to as "connected devices" and "smart devices"), structures and other items— inserted with electronics, software, sensors, actuators, and network connectivity that permit these objects to accumulate and discussion data. In 2013 the Global Standards proposal on Internet of Things (IoT-GSI) defined the IoT as "the organization of the information society." The IoT allows substance to be sensed and/or arranged remotely across remaining network infrastructure, creating chances for more direct incorporation of the physical world into computer-based systems, and resulting in improved efficiency, precision and economic benefit. When IoT is augmented with sensors and actuators, the technology develops an instance of the more universal class of cyber-physical systems, which also incorporates tools such as smart grids, smart homes, smart transportation and keen cities. Each thing is uniquely identifiable through its implanted calculating system but is able to interoperate within the existing Internet infrastructure. Experts calculate approximately that the IoT will consist of almost 50 billion objects by 2020. Typically, IoT is expected to offer advanced connectivity of devices, systems, and amenities that goes elsewhere machine-to-machine (M2M) infrastructures and covers a diversity of protocols, domains, and applications.

The NSF Industry/University Combined Research Center has demonstrated an IoT-enabled intellectual system of such cases for Intelligent Maintenance Systems (IMS) at University of Cincinnati on a band saw machine in IMTS 2014 in Chicago.

Band saw technologies are not necessarily expensive, but the band saw belt overheads are massive since they destroy much faster. However, without sensing and intelligent analytics, it can only determine by experience when the band saw belt will break. The developed prognostics scheme will be able to identify and monitor the degradation of band saw belts even if the circumstance is changing, so that users will know in near real time when is the best time to substitute band saw. This will knowingly improve user practice, hand safety, and save costs on changing band saw belts before they actually break. The developed analytical algorithm realized on a

cloud server and was made available via the Internet and on mobile devices. The organization will likely be an example of event driven architecture, bottom-up made (based on the context of processes and operations, in real-time) and will consider any additional level. Therefore, model driven, and useful approaches will coexist with new ones capable to treat exceptions and unusual development of processes.

(multi Agent systems, BADSc, etc.). Incorporation with the Internet implies that strategy will use an IP address as a exceptional identifier. However, due to the inadequate address space of IPv4 (which allow billion exceptional addresses), substance in the IoT will have to use IPv6 to contain the particularly great address space essential. Objects in the IoT will not only be strategy with sensory capabilities, but also supply actuation capabilities (e.g., bulbs or locks restricted over the Internet). To a large scope, the future of the Internet of things will not be potential without the support of IPv6; and accordingly the global implementation of IPv6 in the upcoming years will be vital for the successful improvement of the IoT in the future.

3. THE ARDUINO UNO BOARD

The Arduino Uno is a microcontroller enters based on the ATmega328 (datasheet). It has 12 functional input or output pins (of which 5 can be used as PWM outputs), 5 analog inputs, a 12 MHz ceramic resonator, a USB connection, a power jack, an ICSP header, and a reset button. It contains the whole thing needed to sustain the microcontroller; just connect it to a computer with a USB cable or power it with a AC to DC adapter or battery to get started. It's an open source objective computing stage based on a plain microcontroller board, and a development setting for writing software for the floorboard.



FIGURE 1 ARDUINO UNO BOARD

Blood flow through finger at each pulse. The Pulse Sensor is a well-designed plug and play heart rate sensor for Arduino. It also includes an open source monitoring app that graphs your pulse in real time.



Figure-2: Pulse Sensor

5. SYSTEM ARCHITECTURE

The general architecture of IoT applications can be separated into three layers: the sensing layer, the transport layer and the application layer. This kind of arrangement is clear and stretchy sufficient for our monitoring system, thus we design the system architecture based on that universal model. Figure 1 shows the architecture of the IoT-based monitoring system for heart diseases patients.

4. PULSE SENSOR

A Heartbeat sensor is a monitoring device that allows one to assess his or her heart moment in actual time or records the heart rate for later study. It provides a simple way to learn the heart function. This sensor monitors the flow of blood through the International Journal of Engineering Science and Computing, April 2017 6665 <http://ijesc.org/> finger and is designed to give digital output of the heartbeat when a finger is placed on it. When the antenna is working, the beat LED flashes in unis on with each heartbeat. This digital output can be joined to the microcontroller straight to measure the Beats per Minute (BPM) rate. It works on the standard of light modulation by

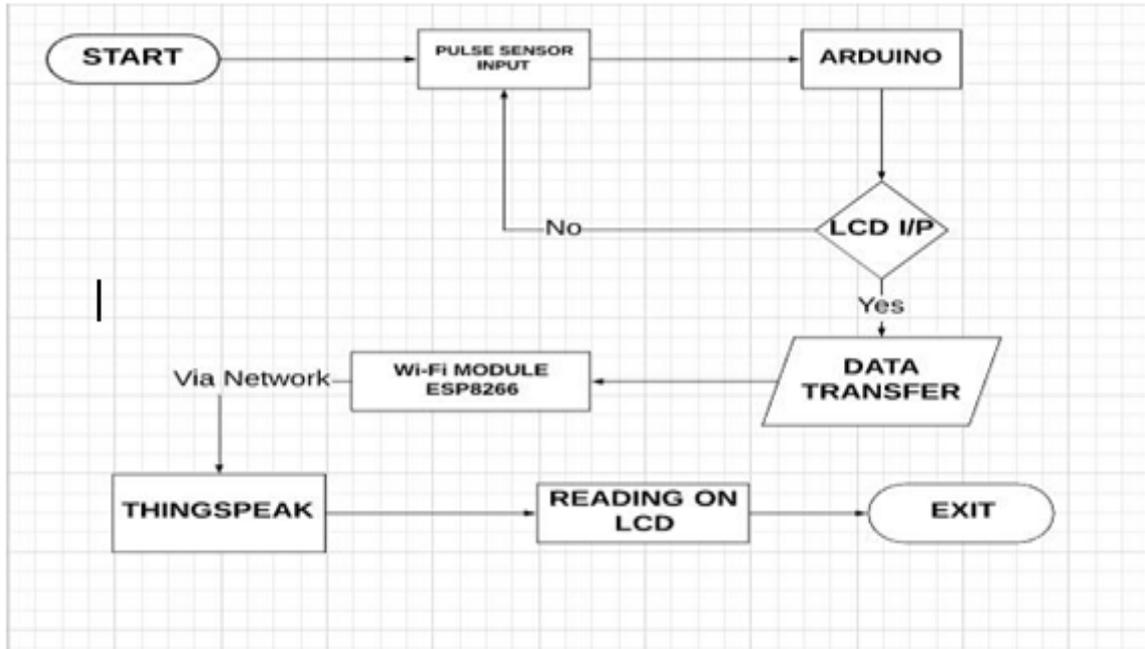


Figure-3: System Architecture

6. SYSTEM IMPLEMENTATION

First, we require attaching the Pulse Sensor to any organ of body where it can discover the pulse easily like finger. Then the Pulse Sensor will evaluate the change in volume of blood, which occurs when every time heart pumps blood in the body. This modify in volume of blood causes a modify in the light strength through that organ. The Arduino will then convert this adjust into the heart beat per minute (BPM). The LED connected at pin 13 will also signal according the Heart Beat. The ESP8266 will then correspond with the Arduino and will send the data to ThingSpeak. The ESP8266 will connect the network to the router and user will offer in the code and will send the data of the sensor online. This facts on the ThingSpeak will be shown in a Graph form performance the past readings too and could be accessed from everywhere over internet. The LCD linked will also give you an idea about the BPM.

7. CONCLUSION

In These days we have an better number of heart diseases including increased risk of heart attacks. Our proposed system user’s sensors that allow for detecting heart rate of a person using heartbeat sensing even if the person is at home. The sensor is then interfaced to a microcontroller that allows inspection heart rate readings and transmitting them over internet. The client might set the elevated as well as low levels of compassion beat edge. After situation these limits, the structure starts monitoring and as soon as patient heart beat goes above a certain limit, the system sends an aware to the organizer which then transmits this over the internet and alerts the health center as well as disturbed users. Also, the system alerts for lesser heartbeats. Whenever the user logs on for monitoring, the structure also displays the live heart rate of the patient. Thus, concerned ones may scrutinize heart rate as well get an paying special attention of heart attack to the person immediately from anywhere and the someone can be saved on time. In our planned research, we tried to counsel a complete paper for finding heart attack using two ways. However, we have some plan about this research.

Time of India, a most important newspaper at India in print that “Researchers in the United States, within the subsequently decade Heart Microeconomic memory chip will be place in blood vessel of human body. The smart phone will collect data and launch the data in series to us”. Researchers are annoying to apply the requirements of Microchip for uses of the knowledge in smart phone. We will try to use this skill in future. If these tools will be built-up, then we can examine heart blockage through this skill by our project.

REFERENCES

1. Internet of Things: internet of things, https://en.wikipedia.org/wiki/Internet_of_things.
2. Effective Ways to Use Internet of Things in the Field of Medical and Smart Health Care: Kaleem ullah, Munam ali Shah, Sijing zhang, IEEE Journal 2016
3. IReHMo: An Efficient IoT-Based Remote Health Monitoring System for Smart Regions, Ngo Manh Kogi, Karan Mitra , 2015 17th International Conference on E-health Networking, Application & Services (HealthCom)
4. Mobile based Home Automation using Internet of Things(IoT), Kumar Mandula, Ram parupalli, E.Magesh , 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)
5. Heartbeat monitoring and alert system using GSM technology, Ufaraoh S.U, Oranugo C.O, International Journal of Engineering Research and General Science Volume 3, Issue 4, July-August, 2015 ISSN 2091-2730
6. Varshney U. Pervasive healthcare and wireless health monitoring. Mobile Networks and Applications 2007; 12(2-3): 113-127.
7. Rofouei, M., Sinclair, M., Bittner, R., Blank, T., Heffron, J. A non-invasive wearable neck-cuff system for real-time sleep monitoring. Proceedings of International Conference on Body Sensor Networks, Dallas, TXm USA, 23-25 May 2011.
8. Majdi Bsoul, Hlaing Minn, Lakshman Tamil. Apnea MedAssist: real-time sleep apnea monitor using single-lead ECG. IEEE Transactions on Information Technology in Biomedicine.

BOUTIQUE MANAGEMENT SYSTEM**Semitha Esthar Rani S¹, Shanmuganathanj², Dr. V. Kavitha³ and G. Subashini⁴**Student^{1,2} and Associate Professor⁴, Department of MCA, Hindustan College of Arts and Science, Coimbatore
Assistant Professor³, Department of BCA, Hindustan College of Arts and Science, Coimbatore

ABSTRACT

The project entitled "BOUTIQUE MANAGEMENT SYSTEM" is a software application, developed for the boutique shops to stay connects with the customers. This software allows organizing the customer, product details in digital form. The main work of this application is to reduce the manual work. Manual work can lead to loss data, as they miss the details. Registration are maintained for the sales and billing purpose of the clothes. By this application, as it is a system work, there won't be loss of data and risk for owner to maintain customer. This Boutique Management System is used to overcome the entire problem which they are facing currently, and making complete atomization of manual system to computerized system.

Keywords: Publication, integration, and execution of applications with Boutiques.

1. INTRODUCTION

The "Boutique Management System" has been developed to override the problems prevailing in the practicing manual system. [1] This software is supported to eliminate and in some cases reduce the hardships faced by this existing system. [2] Moreover this system is designed for the particular need of the company to carry out operations in a smooth and effective manner. [3] The application is reduced as much as possible to avoid errors while entering the data. It also provides error message while entering invalid data. [4] No formal knowledge is needed for the user to use this system. Thus by this all it proves it is user-friendly. Online Boutique Management, as described above, can lead to error free, secure, reliable and fast management system. It can assist the user to concentrate on their other activities rather to concentrate on the record keeping. Thus it will help organization in better utilization of resources.

TEXT IN ROW DATA

In previous versions of SQL Server, text and image data was always stored on a separate page chain from where the actual data row resided. [1] The data row contained only a pointer to the text or image page chain, regardless of the size of the text or image data. [2] SQL Server 2000 provides a new text in row table

option that allows small text and image data values to be placed directly in the data row, instead of requiring a separate data page. [3] This can reduce the amount of space required to store small text and image data values, as well as reduce the amount of I/O required to retrieve rows containing small text and image data values.

CASCADING RI CONSTRAINTS

In previous versions of SQL Server, referential integrity (RI) constraints were restrictive only. If an insert, update, or delete operation violated referential integrity, it was aborted with an error message. SQL Server 2000 provides the ability to specify the action to take when column referenced by a foreign key constraint is updated or deleted.

DISCUSSION

With Boutiques, developers can integrate their applications once and execute them in several platforms. Boutiques removes the technological dependency to a particular platform and facilitates application migration. Although the motivating use cases were taken from neuro informatics, our primary field of interest, nothing prevents the system from being used in other domains.

LOG SHIPPING

The Enterprise Edition of SQL Server 2000 now supports log shipping, which you can use to copy and load transaction log backups from one database to one or more databases on a constant basis. This allows you to have a primary read/write database with one or more read-only copies of the database that are kept synchronized by restoring the logs from the primary database.

METHODOLOGY

Model View Controller or MVC as it is popularly called, is a software design pattern for developing web applications. A Model View Controller pattern is made up of the following three parts:

Model - The lowest level of the pattern which is responsible for maintaining data.

View - This is responsible for displaying all or a portion of the data to the user.

Controller - Software Code that controls the interactions between the Model and View.

MVC is popular as it isolates the application logic from the user interface layer and supports separation of concerns. Here the Controller receives all requests for the application and then works with the Model to prepare any data needed by the View.

2. RELATED WORKS

Table

The database design involves creation of tables that are represented in physical database as stored files. They have their own existence. Each table constitute of rows and columns where each row can be viewed as records that consist of related information and column can be viewed as field of data of same type.

The database design of project is designed in such a way that values are kept without redundancy and with normalized format.

CUSTOMER TABLE

Primary Key: Id

FIELD NAME	DATATYPE	SIZE	DESCRIPTION
Id	Int	5	Customer id
Pname	Varchar	30	Customer name
Psex	Varchar	30	Gender
Adds	Varchar	10	Customer address
Mobile	Int	10	Customer mobileno
Email	Varchar	30	Customer email id
Date	Varchar	30	Customer d.o.b

3. RESULTS AND DISCUSSION

Input Design

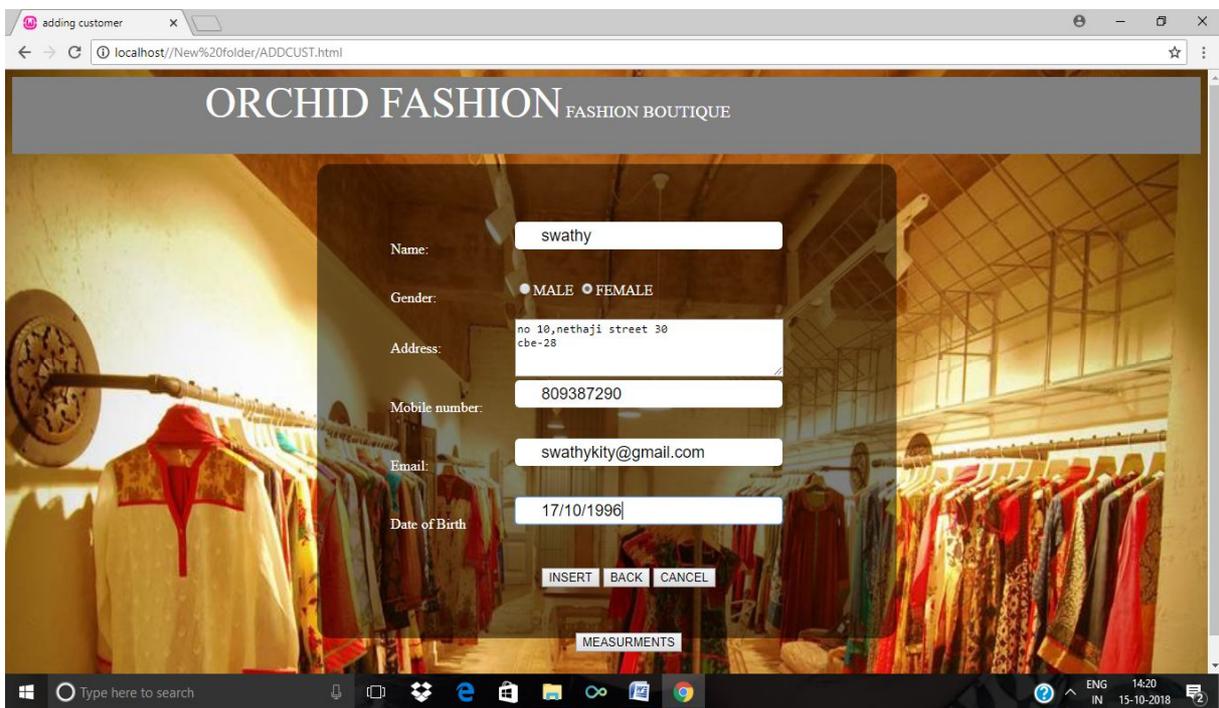


Figure-1: Customer Registration Details

In the input design, user oriented inputs are converted into a computer based system format. It also includes determining the record media, method of input, speed of capture and entry on to the screen. Data entry accepts commands and data through a keyboard. The major approach to input design is the menu and the prompt design. In each alternative user's options are predefined.

Output Design

The output device's capability, print, print capability, response time requirements etc should also be considered. Form design elaborates the way output is presented and layout available for capturing information. It's very helpful to produce the clear, accurate and speed information for the end user.



BILL ID	CUSTOMRE ID	PRODUCT ID	DATE	QUANTITY	Dress	PRICE
12	4	4	15/10/2018	3	M	1200

Total Payment: 1200

BACK CANCEL

Figure-2: Report details

4. CONCLUSION AND FUTURE ENHANCEMENT

Boutique management system with intelligence system is a great improvement over the existing system which has various drawbacks. The computerization of the system with added features has speed up the process. In the current system, the process is slow and time consuming. By the proposed system process has become quick and updates can also be made quickly. Customer and product details can be viewed whenever admin needs. Connecting the system with various branch websites Creating digital catalogue for customer viewing Providing the admin.

REFERENCES

1. Peng RD.Reproducible research in computational science. Science 2011;334(6060);1226-7.
2. Stodden V,McNutt M,Baily DH et al. Enhancing reproducibility for computational methods .Science 2016;354(6317):1240-1.
3. Kurtzer GM,Sochat V,Bauer MW.Singularity:scientific container for mobility of computer. Plos One 2017;12(5):
4. Glatard T,E'tienne Rousseau M,Camarasu-Pop S et al. Software architectures to integrate workflow engines in science gateways.Future Generation computer systems 2017;75(supplement c);239-55.
5. Sherif T,Rioux P,Rousseau ME et al. CBRAIN:a web-based, distributed computing platform for collaborative neuroimaging research. Frontiers in Neuroinformatics 2014;
6. Glatard T,MacIntyre LC et al.The MNI data-sharing and processing ecosystem.Neuroimage 2016;124:1188-95.
7. Geogolewski K,Burns CD,Madison C et al.Nipype:a flexible,lightweight and extensible neuroimaging data processing framework
8. Glatard T,Lartizien C,Gibaud B et al.A virtual imaging platform for multi-modality medical image simulation.IEEE Transactions on medical imaging 2013;32(1):110-8.

CERTAIN INVESTIGATIONS ON INTEGRATED MATERIALIZED VIEW BASED APPROACH IN ETL WITH DSS TO ACHIEVE FAST DATA TRANSFORMATION

A. Uma Maheswari¹ and Dr. N. Revathy²Research Scholar¹ and Associate Professor², Hindusthan College of Arts and Science, Coimbatore

ABSTRACT

A data warehouse can be defined as a subject-oriented, integrated, nonvolatile and time-variant collection of data, which has value and role for decision-making by querying. To avoid accessing base tables and increase the speed of queries posed to a DW, we can use some intermediate results from the query processing stored in the DW called materialized views. Materialized views are physical structures that improve data access time by pre-computing intermediary results. The ETL process bridging the online transaction processing (OLTP) system and the online analytical processing (OLAP) system is often modeled as a separate and independent process. When transformed data are loaded into the data warehouse, the analysis-centric applications take place. To expedite the analysis process, materialized views are often created and used through the query rewrite mechanism in the data warehouse. The proposed approach integrate the ETL process and the data warehouse applications by using views and materialized views to model and perform the ETL process. This process integration has several advantages. First, it achieves fast data transformation and materialized view maintenance through one single materialized view refresh call.

Keywords: Online Transaction Processing (OLTP), Online Analytical Processing (OLAP), Materialized views, ETL.

INTRODUCTION

Data analysis and decision making are ultimate goals in the data warehouse applications. Consider a large data warehouse containing terabytes of data distributed in de-normalized tables each of which has potentially millions or even billions of data rows. Retrieving data from such system is a very resource-intensive and time-consuming process. Queries that retrieve and calculate a large amount of data could require hours to be processed. To speed up the query processing in the OLAP system, a commonly used approach is the use of materialized views.

The materialized view is a database object that contains pre-calculated data typically for highly aggregate queries associated with complex joins of fact and dimension tables. The use of the materialized view in query processing is often transparent through a mechanism called *query rewrite*. In such mechanism, the original query is rewritten to make use of the materialized view instead. The query response time is thus improved. Consequently, the materialized view has become a common but required resident in the OLAP system. Then, consider the subject of data loading to the OLAP system, a well-known process called ETL.

ETL extracts the source data from the operation-centric OLTP system, transforms the data to resolve various kinds of heterogeneity problems and loads the transformed data into the target OLAP system (e.g., data warehouse). The ETL process is usually an independent process that bridges the two systems. It is often achieved by the third party tool which executes the ETL process in a specific window of time. Due to this asynchronous nature, the changes in the OLTP system are not transformed and propagated to the OLAP system in the real time. As a result, real-time data analysis and decision making is difficult to achieve with such process model. On the other hand, since the materialized view has already become a common data warehouse object for improving query performance, it will be beneficial to use the materialized view to model the ETL process so that the ETL process and the data warehouse applications can be seamlessly integrated. This idea motivates us to investigate and design an integration mechanism that is based on the materialized view technology.

In our approach, we combine the ETL process into the OLAP application by conditionally using views and materialized views. The buffering of intermediate results in the ETL process and the fact and dimension tables in the OLAP system are all represented in materialized view objects. The changes propagation and the data transformation are done in one single refresh call that refreshes all the materialized views in the dependency hierarchy. Such design generates an efficient integration and enables near real-time data analysis. Also, the approach can support data cleansing to filter out dirty data and allows corrected data to be re-submitted into the ETL process at the spot where the error is found.

In Section 2, we first review related work in the areas of the ETL modeling, data cleansing and processing performance. Section 3 presents our proposed approach and compares it with the current architecture. We then

illustrate the use of the materialized view to model the ETL process and data cleansing with examples. Later, we discuss the advantages and concerns of our approach. Section 4 describes the conclusion.

This process integration approach has several objectives. First, it should achieve fast data transformation and materialized view maintenance through one single materialized view refresh call. Changes in the OLTP system can be quickly and transparently applied to the materialized views so that a near real-time data analysis can be carried out. Second, the use of views and materialized views to model the ETL process should provide the benefits of encapsulating data transformations in a multi-step SQL process. Third, it should improve the ETL performance and provide better query optimization. Lastly, the use of materialized view in the ETL process should facilitate the data cleansing so that clean data are passed through and processed while dirty data are intercepted and loaded into tables for correction.

VIEWS AND MATERIALIZED VIEWS

A view consists of a stored query accessible as a virtual table in a relational database or a set of documents in a document-oriented database composed of the result set of a query or map and reduce functions. Unlike ordinary tables in a relational database, a view does not form part of the physical schema. It is a dynamic, virtual table computed or collated from data in the database. Changing the data in a table alters the data shown in subsequent invocations of the view.

Materialized views are physical structures that improve data access time by pre-computing intermediary results. At the same time, the use of materialized views requires additional storage space and entails maintenance overhead when refreshing the data warehouse. Data warehouse is capable of answering queries and performing analysis in an efficient and quick manner, in the view of the fact that integrated information is directly available at the warehouse with differences already resolved. Though the data warehouse research community provides effective solutions for the problem of representing data in a form suitable for analytical queries, it has not completely addressed other performance issues such as, query response time for a given aggregated query, view maintenance time, etc.

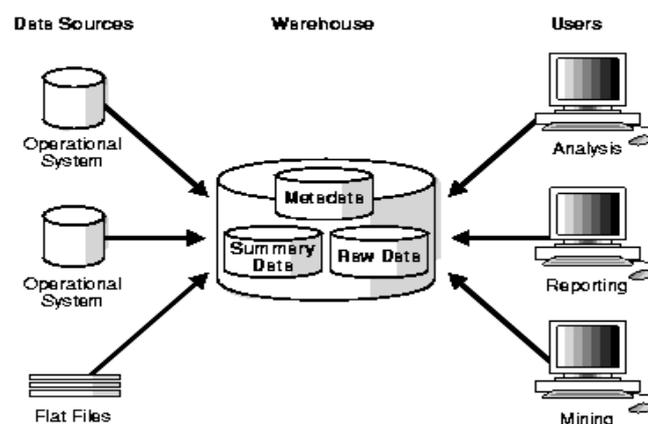
A view takes the output of a query and makes it appear like a virtual table. All operations performed on a view will affect data in the base table. Materialized views are schema objects that can be used to summarize, pre-compute, replicate, and distribute data. A materialized view provides indirect access to table data by storing the results of a query in separate schema object. The existence of a materialized view is transparent to SQL, but when used for query rewrites will improve the performance of SQL execution.

METHODOLOGY

Here, we first show the architecture of the data warehouse, current ETL process model and compare it with the proposed architecture that integrates the ETL and data warehouse applications. Second, we show the use of views and materialized views to achieve the integration and illustrate it using queries.

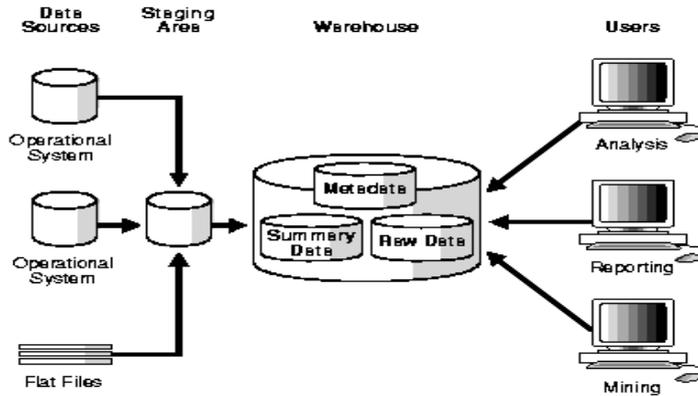
Architecture of a Data Warehouse

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. In addition to a relational database, a data warehouse environment includes an Extraction, Transportation and Loading (ETL) solution, an Online Analytical Processing (OLAP) engine, client analysis tools and other applications that manage the process of gathering data and delivering it to business users.



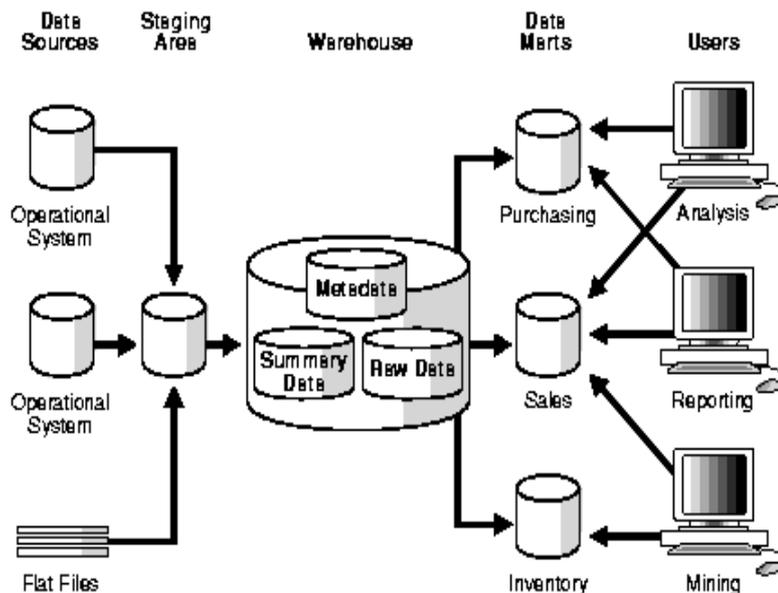
Common Architecture of a Data Warehouse

The above figure shows the common architecture of a Data Warehouse. Here, the source system having any databases or processing systems like Oracle, Teradata, OLTP, Flat Files, Sequential Files etc., and next level is the data warehouse, its consists metadata and raw data. Finally business intelligence process takes place from data warehousing like OLAP, Reporting, and Data Mining etc.



Architecture of a Data Warehouse with Staging Area

The above figure shows the architecture of a data warehouse with staging area. The data warehouse staging area is a temporary location where data from sources systems is copied. A staging area is mainly required in a data warehousing architecture for timing reasons. In short, all required data must be available before data can be integrated into the data warehouse. It composed of the data staging server application and the data store archive of the results of extraction, transformation and loading activity. The staging area can be used for any of the following purposes like, gather data from different sources that will be ready to process at different times or quickly load information from the operational database freeing it up as soon as possible or find changes against current DW/DM values or data cleansing. For example the data staging application server temporarily stores and transforms data extracted from OLTP data sources and the archival repository stores cleaned, transformed records and attributes for later loading into data marts and data warehouses.



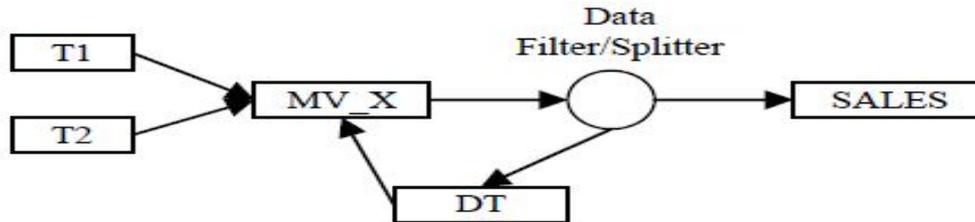
Architecture of a Data Warehouse with Staging Area and Data Marts

The above figure shows the architecture of a data warehouse with staging area and data marts. Data mart is a subset of the data resource, usually oriented to a specific purpose or major data subject that may be distributed to support business needs. The concept of a data mart can apply to any data whether they are operational data, evaluation data, spatial data or metadata. A data mart is a repository of a business organization’s data implemented to answer very specific questions for a specific group of data consumers such as organizational divisions of marketing, sales, operations, collections and others. A data mart is typically established as one dimensional model or star schema which is composed of a fact table and multi-dimensional table.

DESIGN OF EXPERIMENTS AND ANALYSIS OF RESULTS

Example of using materialized views for data cleansing

The use of the materialized views in the ETL process can support data cleansing as well. As previously mentioned automatic conflict resolutions and fixes are nearly impossible due to the complexity of data representations. Part of the conflict resolutions and fixes will have to rely on human experts to examine the dirty data and make corrections. Typically, when transformed data are identified as dirty, those rows are recorded in a separate file or table for the human expert to examine. After corrections, the data can be resubmitted to the transformations process.



Materialized Views for Data Cleansing

The above figure shows an example of using materialized views to support data cleansing. Similar to the example given in Section 3.3, the materialized view MV_X contains intermediate transformation results which are to be passed to SALES. However, the data quality needs to be ensured before the data passing. Therefore, we use a data filter/splitter (as shown in an oval shape in Figure 4) that represents one or more condition checks. The data that pass the condition checks are forwarded to populate SALES. Otherwise, the data are treated as “dirty” and redirected into a table DT for the human expert to examine and correct. The following statement is an example to populate the table DT:

MERGE INTO DT

```

    (SELECT T1RID, T2RID, PRICE_SALES, UNIT_SALES, AMOUNT_SOLD, PROD_ID, C1,
    DECODE (PRICE_SALES < 0, TRUE, 'Invalid PRICE_SALES',
    DECODE (PROD_ID > 250, TRUE, 'Invalid PROD_ID', 'OK')) MSG
  
```

FROM MV_X

WHERE PRICE_SALES < 0 or

```

    T2.PROD_ID > 250)
  
```

WHEN NOT MATCHED

INSERT;

It is noted that the table DT has one extra column, MSG. It is to provide the reason why the data is considered as dirty. The human expert will use the information to correct the data. After the correction, the corrected data in the table DT is resubmitted to the ETL process through MV_X. To support this, the creation statement for MV_X is modified as follows.

CREATE MATERIALIZED VIEW MV_X

FAST REFRESH ON DEMAND

AS

```

    SELECT '1' UMARKER, '0' DTRID
    T1.ROWID T1RID, T2.ROWID T2RID,
    CONV1 (T1.PRICE_SALES) PRICE_SALES,
    CONV2 (T2.UNIT_SALES) UNIT_SALES,
    CONV3 (T1.AMOUNT) AMOUNT_SOLD
    T2.PRODUCT_ID PROD_ID,
    T1.C1
  
```

```
FROM T1, T2
WHERE T1.C1=T2.C1
UNION ALL
    SELECT '2' UMARKER, DT.ROWID DTRID
    T1RID, T2RID, PRICE_SALES,
    UNIT_SALES, AMOUNT_SOLD, PROD_ID, C1
FROM DT;
```

The modified MV_X to support data cleansing has one extra query block connected to the original query through a UNION ALL operator. To support log-based incremental refresh for MV_X, two additional columns (UMARKER AND DTRID) are added. This way, the materialized view MV_X is enable to include corrected data in the ETL process. Another minor modification is needed for SALES whose defining query is added with the filter conditions so that only correct data are passed to SALES.

RESULTS AND ANALYSIS

The above proposed approach that integrates the ETL process and the data warehouse applications has the following Results. It integrates the ETL process and the data warehouse applications in a common framework. Instead of using the third party ETL tool that potentially causes high data transmission overhead, the ETL process is combined into the data warehouse system in a seamless fashion which saves the data transmission overhead from the ETL engine to the target system.

It achieves fast data transformation and materialized view maintenance through one single materialized view refresh call. The use of on-commit nested materialized views enables the changes in the OLTP system to be quickly and transparently applied to the terminal materialized views so that a near real-time data analysis can be realized.

The use of views and materialized views to model the ETL process provides the benefits of encapsulating data transformations in a multi-step SQL process. Compared to the scripting and fine-step GUI-based transformation approaches, this approach offers a neutral representation of the ETL plan which provides better readability and maintainability. The modification of the ETL plan is as simple as the defining query change of the materialized view.

CONCLUSION AND FUTURE WORK

The presented approach provides a seamless integration that can quickly and efficiently propagate the changes from the OLTP source system all the way to the materialized view in the OLAP data warehouse system. Thus, the materialized views used by query rewrite contain up-to-date information and are able to support near real-time data analysis/mining and consequently benefit decision making.

The approach is carried out by a number of intermediate materialized views or views to model the ETL process in the OLAP system. This enables the ELT execution inside the database in which more efficient optimization mechanisms for data processing can be utilized. The ETL process can therefore be more efficient. In addition, the use of on-commit nested materialized views automates the data transformation and the refresh maintenance of terminal materialized views in one single refresh call which greatly eases the process that is currently handled in a complicated way.

Finally, some concerns about this approach such as storage requirements are alleviated by conditionally process. Also, the unnecessary transformation results can be aged out from the intermediate materialized views by range refresh maintenance of the materialized view.

REFERENCES

1. Ashish Gupta and Inderpal Sing Mumick, 1995. Materialized views: techniques, implementations and application, 2nd Edition, University of California.
2. Claudia Imhoff, Nicholas Gallempo and Jonatha G.Geiger, 2001. Mastering Data Warehouse Design, 3rd Edition, Wiley Publishing, Inc.
3. Elzbieta Malinowski and Esteban Zimanyi, 2008. Advanced data warehouse design, 1st Edition, Springer.
4. Gavin Powell, 2005. Oracle Data Warehouse Tuning, Oracle Corporation.
5. Paulraj Ponniah, 2011. Data Warehousing Fundamentals for IT Professionals, 1st Edition, John Wiley & Sons New York.

-
6. Rajiv Parida, 2006. Principles and Implementation of Datawarehousing, 1st Edition, Firewall Media.
 7. Ralph Kimbal, 2002. The Data Warehouse Toolkit, Practical Techniques for Building Dimensional Data Warehouse, 2nd Edition, John Wiley & Sons New York.
 8. Stanislaw Kozielski and Robert Wrembel, 2008. New Trends in Data Warehousing and Data Analysis, Springer.
 9. Students Guide, 2002. Data Warehouse Fundamentals, Oracle Corporation.
 10. W.H.Inmon, 2005. Building the Data Warehouse, 3rd Edition, John Wiley & Sons New York.

STAGING IMPROVEMENT OF ENERGY OPTIMIZATION IN WSN USING BIO-INSPIRED ENVIRONMENT

B. Karthik¹ and Dr. N. Revathy²Research Scholar¹ and Associate Professor², Hindusthan College of Arts and Science, Coimbatore

ABSTRACT

Wireless sensor networks are the sensors which are acclimated to sense the ecological condition like temperature, pressure, sultriness, moisture etc., sensing the environment parameters and sending them to the opportunity and retrieving the aggregated data from the opportunity to the end user. Power is the major constraint in wireless sensor networks. One must need to reduce the power utilization. Wireless sensor networks have sensor nodes in which each node has a processor, antenna and a battery. The batteries use power so that we need to increasing the lifetime of the battery for that some optimization techniques are required to reduce the power utilization. There are some techniques which are stimulated from the lifestyle of animals. In this paper proposing an optimization system which is enthused by the animals behavior which reduces the power consumption of the feeler nodes which is particle swarm optimization (PSO) system. PSO is encouraged by the friendly behavior of birds or schooling of fish. By utilizing this bio-inspired system we can reduce the power inspired by the feeler nodes and at the same time duration of the batteries there in the sensor nodes are increased.

Keywords: Energy Optimization, WSN, Bio-inspired environment.

1. INTRODUCTION

Wireless sensor networks are acclimated to sense the surroundings in immediate and proactive manner. Considering the immediate way it senses the situation whenever the event is occurred or in a random manner. Whereas proactive way senses the surroundings at each and every gap of time. The sensor node senses the situation in two ways mainly, one is immediate and the other is proactive [1]. Immediate be tokens sensing the situation in a normal base. Proactive denotes sensing the situation in an random manner. A sensor network has dissimilar challenges to face in the current trends and the nodes are receiving deployed in different areas to sense dissimilar parameters in the setting. The nodes deployed in undergrounds, and also for conditions monitoring [2] sensing applications, military applications. Wireless sensor networks senses the situation like temperature, pressure, sultriness and so on.., sensing the environment and sending the data to the occasion is the whole method going on First the data is sensed by the nodes in the network and data is aggregated. Then the aggregated data is send to the gate, whenever the end user requires the data at exacting time. Instant then the data is retrieved from the occasion to the end user. The data is send from the starting point to the end user through the sensor nodes. This process requires some amount of energy. The energy consumption is dissolute while transmitting the data from source to destination. Since the sensor nodes deployed in the situation are non-replaceable batteries so if the sensor node is died we require replacing it with other node. Because battery is not replaceable, we cannot power up the batteries. So, what we can do is to reduce the power utilization of the sensor nodes and then by this we can increase the lifetime of the batteries. The sensor network has a processor unit, an antenna and a memory to store the sensed data [3].

The deployed WSN have many problems after deploy ing. WSN consists of different components some of the components may deplete sooner and some may deplete very late depends up on their life time and quality of the network components. Our main motto is to control the power consumed by the network components during their transmission of the data and one more thing one should keep in mind that the lifetime of the batteries used in the network. The batteries which are in the nodes are non-replaceable so we need to increase the lifetime of the batteries in order to get more time working of the nodes without dead.

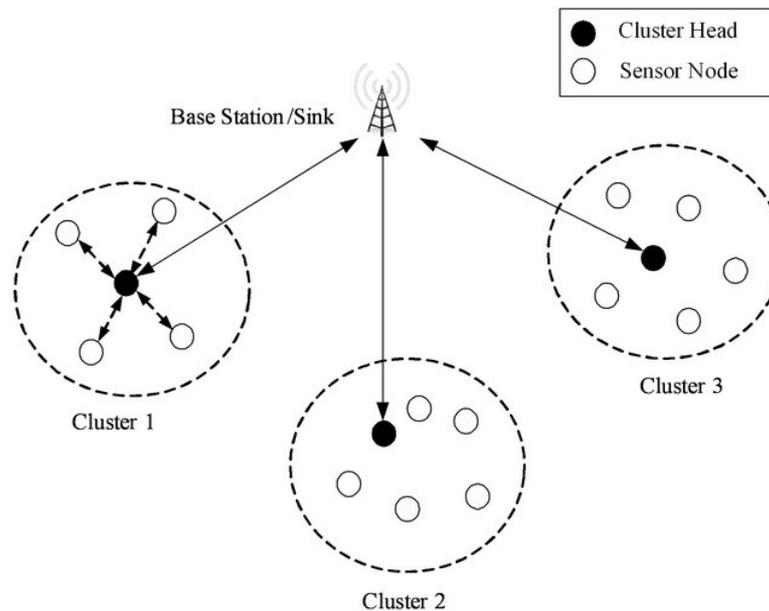
The central unit of WSN is a Node, also called as bit by UC Berkley. Each sensor hub is necessary to be fit for detecting, preparing and imparting the handled data to the neighboring hubs to encircle a system. Sensor hub is now on made out of sensors to detect the physical speculate, simple to computerized converter, microcontroller for calculating and information handling, memory for calculations and information stockpiling, radio unit for short range remote communication and battery unit to control every one of the units. For some applications, if possible, sun based fueled chargeable battery units are utilized to limit the dissatisfaction of sensors and support cost.

Data from early point to the end client There are so many techniques are there to optimize the efficiency. We can optimize the power by making all the sensor nodes through in the network into clusters.

There are some bio-inspired algorithms of optimization techniques depends on the population substructure. Bio-inspired and department population predicated optimization techniques to diminish the puissance utilization in the WSN [4]. The different algorithms which are getting inspired by the life style of animals are chosen to optimize the energy competence in the nodes of the networks.

The objectives of WSN are

1. The vigor and solid quality of the whole communication structure should be improved without growing the energy consumption.
2. No velocity ideas in reunion design will be examined to moreover stretch out the wake-up structure to likewise, impart parameter changes (a couple of byte) to the convenient hubs. For accomplishing this objective, the learning on the example of the arriving association into a scheme is required.
3. The arrangement of the power cautions mechanism and their accurate pre setup.
4. Procedure choices between the moveable hubs and to enhance the unwavering quality of the whole correspondence framework while custody up the legitimately tight energy expenses plan.
5. To create narrative and accurate systematic models to break down the implementation of the proposed conventions and transmission plans. The accuracy of the created models ought to likewise be checked.



One approach can be habituated to control the amount power consumed in WSN is to make the cluster formation of the sensor nodes and each cluster has a head. First, the data is sensed by the sensors and then the sensors transmit the data to its mote then mote is the responsible thing to transmit the data packet to the destination. Interiorly, the data is sensed by the sensors and these sensors transmit the data to their neighboring sensors until it reaches the destination. This whole process consumes more power. Later all the sensor nodes composed different groups called Clusters. Each cluster has some sensors and a single head called cluster head. Now the data is sensed by the sensors in the cluster and the sensed data is transmitted to their heads. These heads transmits the data to the base station and from base station to the destination point.

Power optimization is the process of reducing the required amount of power that a sensor node requires to transmit. In recent years researches are going on in different algorithms based on the different animal's lifestyle and by that inspired life-style. we are going to de-velocity an algorithm and by using that particular algorithm we can reduce the power in WSN by applying such techniques.

In this paper we are proposing an optimization technique which is population predicated convivial department of animal's algorithm. The particle swarm optimization technique is predicated on the swarm astuteness. Flock of birds or schooling of fish. Each bird in the swarm is considered as a particle or solution quandary and each particle has a fitness value. Velocity which shows the over-sight to reach the destination. Position of the particle. Each particle has its personal best value called priest and all the particles in the swarm has an ecumenical best fitness value called globest. This technique is better than the existed algorithm based on ant's life style. The ant's life style gives the shortest path for the destination [5]. But this algorithm gives the optimization solution for the problem.

2. LITERATURE SURVEY

M.Sasikala [6] proposed some clustering algorithms of wireless sensor networks. In this paper the author embodied that wireless sensor networks are the emerging technology, now a days and the data aggregation done through wireless sensor networks is having the quality and the nodes in the network senses and aggregating the data requires power. Power is the major constraint in the wireless sensor networks. By the clustering of the sensor nodes one can reduce the power required for the nodes to transmit the data. By multi hop we can reduce the power utilization. The overview of routing protocols is given in this paper.

Mukul pratap singh and Kunal gupta penned the optimization techniques for reduction the power utilization in [7]. The batteries are non-rechargeable and power is the major constraint in wireless sensor networks and for that to reduce the power utilization optimization techniques are involved in this paper are like Ants Colony Optimization(ACO) and some other techniques like LEACH, HEED also explained in this paper. ACO is the better technique among those which gives us the shortest path from source to destination . Chandni and Anjali Bharti explained the bio inspired algorithms used to reduce the power utilization in [8] now a days research is going on wireless sensor networks. There are some challenges in wireless sensor networks includes cost, power, and life-time. Power is the major constraint to the bio inspired algorithms which reduces the power utilization in networks are brief out in this particular paper.

Salma given a new optimization technique based on grouping of the sensor nodes in [9]. In networks grouping the sensor nodes will reduce the power utilization. The nodes are grouped and form a cluster and each cluster has a head called a mote. The group members sense the data and informed to the mote and then the mote sends the data to the neighbor motes and it takes the responsibility of data aggregation and transmission. By this grouping of sensors optimization can be takes place.

The most important issue in the WSN is optimization of minimum resources. Dehghani in [10] proposed many routing algorithms are penned and described in detail for the optimization purpose. Some genetic algorithms also explained in this paper for the power optimization purpose. Power is the major hottest topic for the researchers in wireless sensor networks and it is the major challenge for them in their research area.

Vimala Rani, Subbrahmanian proposed a bio inspired algorithm for the power optimization in WSN called enhancement of particle swarm optimization technique [11]. This technique is bio inspired technique by the flock of birds and school of fishing. The lifestyle of the birds is inspired to follow up this technique. By this technique power can be optimized.

Debmalya Bhattacharya in [12] explained about the issues related to power in the present networks challenges faced by many researchers in WSN. The major constraint is power. For the optimization purpose some algorithms are explained in the paper.

T. Luo, H.-P. Tan and T. Q. Quek in [13] explained about the WSN how they are going to sense the environment and what the different ways to sense the environment are. The challenges that might have been faced in the WSN.

Sandra Sundra in [14] given distinctive calculations in view of the way of life to the creature conduct and by the conduct we can apply those procedures to the systems and after that we can limit the power utilization in the hubs.

Debmalya Bhattacharya and R Krishna Moorthy in [15] clarified about the WSN and how the different hubs are conveyed in the earth and how they will detect nature they said some unique strategies which limits the cost and builds the lifetime of the hubs.

G.Hemavathy, S.Prabhu in [16] clarified that the WSN comprises of various kinds of various sensors and which are acclaimed to detect the earth of various physical parameters. These creators clarified that there are two calculations which are utilized to allot the vitality to the hubs by figuring the separation and de- pending up on the separation the hubs are getting assigned some vitality and by that we can diminish the power utilization required for transmission of information parcel.

MM Chandane in [17] clarified different issues that are looking in WSN and the difficulties incorporates into WSN how these difficulties are overcome in WSN, he penned a few calculations with the end goal of improvement in WSN hubs.

3. PROPOSED APPROACH

In this paper proposing an optimization technique which is based on bird's life style. This technique is predicated on the swarm intelligence Flock of birds or schooling of fish.[12]. Each bird in the swarm is

considered as a particle or solution quandary and each particle has a fitness value. Velocity which shows the oversight of the path to reach the destination . Position of the particle. Each particle has its personal best value called *pbest* and all the particles in the swarm has an ecumenical best fitness value called *gbest*. [12]



PSO was introduced and de-velocity by Eberhart and Kennedy in 1995. PSO is multi objective and dynamic optimization. It is one of the algorithms taken from inspiration of animals in which flock of animals. In flock of animals find aliment desultorily which is most proximate to victuals position. Animal's up- date to each other about position of food. It will update again and again until food source is found. Particle swarm optimization consists of a group of particles, where particles represent a potential Solution. Every particle is represented by its best position. It calculated on substructure of local best position as *pbest* and global best position *gbest*. PSO calculated on position of particle. Position of the particle is getting updated at every repetition.

By the way of life of feathered creatures this calculation is produced and is utilized for the enhancement reason in the sensor hubs. The accompanying areas give a concise clarification how the flying creature's way of life is taken as a motivation and how it is connected and contrasted with our systems. What are all the conceivable terms in that calculation are clarified unmistakable in following areas.

We as a whole know how flying creatures are flying in the sky day by day we are watching them while they are move starting with one place then onto the next. There is a similitude between the winged creatures relocating from one point to other point and the hubs which are transmitting the information to one point to other point by this comparability.

Birds move in one angle from source point to their destination point for the food particles as appeared in figure 2 every one of the flying creatures in the sky or the gathering doesn't know the area of food however one among all is closer to the food particle.

The bird nearer to that bird started following the path until they reach the food particle. A question is raised to all how we will analyze this feathered creature way of life or the run of winged creatures, way of life of looking for the food particles to our system. Yes we can compare. Presently, let us contrast the flying creature way of life and the system. Each bird is considered as a sensor in our system and group of birds is called a flock a group of sensors is called as a cluster. Each bird having a position and a velocity vector which shows the path for the birds and here in every sensor additionally we will think about the position of the sensor and having a speed vector for heading of the way to transmit the information. At each cycle the birds are changing their positions and those positions are getting refreshed and here in these systems additionally the position of every sensor is updated at each repetition.

Depending up on the fitness, function the best control node is selected. If the fitness function is less than , it is selected as control node. The nodes which are having less fitness function they consume less energy and the nodes which are having more fitness function they consume more energy. Residual energy is calculated for every node after every transmission. By calculating the residual energy we can have clarity of which node consumes less power and which node consumes more power. The node which is having more residual energy consumes less power and the node which is having less residual energy consumes more power.

The residual energy is the remaining energy that is left after the completion of the data transmission from the starting or initial point to the end user. The amount of energy that consumes till the completion of one transmission is called residual energy. We need to calculate the residual energy of each node after completion of the data transmission. By this also we can select the best control node for the cluster.

The above figure shows the flow chart how our proposed system is processed. First the particles are initialized each particle is having some energy and at first transmission of data packet some amount of energy is dissipated and now the update the position of the particle and calculating the amount of energy is consumed for the single transmission. Now for each particle there are two best values called *priest* and *globest* those values plays a major role to finding an optimal solution. Comparing the *globest* with *priest* and if the personal best value is better than the global best then the personal best value is equated to the global best like this the procedure continues until reaches the best solution.

For example, a birds group is there and they need some food. The flock randomly searching for the food. They have no idea about the exact location of the food but, they have a little idea about the distance of the food location. In order to find out the exact location of the food by each bird is impossible or it is too difficult and takes more time. One bird among all knows the location that is the bird is nearer to the food. So, the other birds in the flock or swarm start following the neighbor birds until they reach the food.

In PSO each single bird is considered as a problem solution or a particle in the swarm. Group of birds is considered as a swarm. All of the particles present in the swarm have its own fitness value evaluated from the fitness function. The particle in the swarm follows the other particles to find out the best solution for the optimization problem. Every particle in the swarm has a velocity vector which shows the direction of the path and each particle has a position vector. The position and velocity should be updated at every iteration and each particle has its individual best that is called the personal best value *priest* and all the particles combine and has a global best value called *globest* these values are updated at every iteration until the particles reaches the destination. [12-15] [17-19]. After this the position and velocity of the particles are updated at every iteration as follows

$$\text{velocity}(i,:) = \text{eta} * (\text{w} * \text{velocity}(i,:) + \text{c1} * \text{rand}(\text{size}(\text{globest})) * (\text{globestpop}(i,:) - \text{pop}(i,:))) + \text{c2} * \text{rand}(\text{size}(\text{globest})) * (\text{prbest}(i,:) - \text{pop}(i,:)) \quad (1)$$

$$\text{pop}(I,:) = \text{pop}(I,:) + \text{velocity}(I,:) \quad (2)$$

The equations (1) and (2) are used for the updating the particles position and velocity at the every single iteration.

Velocity (I) is the particle velocity, *Priest* and *globest* are personal best and global best respective velocity as stated before. $\text{Rand}()$ is a random number in between 0 and 1. C_1 C_2 is learning factors and these are kept constant that is 2. First 2 joules of energy i.e., power is given to the sensor nodes. Let us assume 50 particles are present. For each particle in the swarm we are giving 2 joules likewise 50 times of 2 joules and starts transmitting the packet from the starting node to its neighbor. Now the initial power is getting dissipated slowly now we need to calculate the energy of the particle and need to update the position of the particle.

2.1. Proposed Process Arrangements

Initially, we have to choose the particles in the swarm as a system or dimensions. And after that we are giving some energy to the particles called sensor nodes. The packet is broadcast from the source to the finish node through the intra basic hubs in the arrangement. At that point for broadcast of the packet it need some power. After show of the packet some amount of power is degenerate. In the first series this complete process is happened and same method is preceded awaiting the point that it achieves the end node. At each series we will inform the position of the particle and discover out the strength function at every iteration. After manipulative the strength occupation at each iteration we have to look at the strength values of all the client and finding the best robustness value. At this point we select the sensor node which is having the finest strength function as the head of the cluster of the sensor nodes. The principal goal of the proposed approach is to determine the best classify node for the group of nodes. By that we can reduce the power use of the hubs.

The organize nodes consume more power than the ordinary nodes in the network so by the proposed move toward depending up on the strength value we can select the best organize node and by that we can diminish the power.

4. EXPERIMENTAL RESULTS

Here the results are alive using NS simulator as a display place. The X and Y axis are number of iterations and the strength Value correspondingly. Depends up on the strength value we can say that we reach the optimization solution. Here proposed method contains different optimization functions and the graph are given below.

At every iteration the strength value is getting reduced and at some iteration we will get the minimum strength value. The minimum the strength value the highest the optimization solution. By this we can get best optimization solution.

The different Optimization functions comparison which is used for the power optimization. The predictable result is more or less equals to the obtained results.

5. CONCLUSION

By the proposed approach called bio-inspired environment particle is group optimization we can reduce the energy utilization of the networks and the lifetime of the batteries can be increased. Since the batteries are non-rechargeable and by using the proposed approach we can choose the best start for the cluster of nodes called mote and by that we can repeatedly reduce the energy utilization in WSN.

REFERENCES

1. Lisane brisolara, Paulo.r.ferreira, leandrosoares indrusiak, "Ap- plication modelling of performance evaluation of event triggered wireless sensor networks", springer publications 2016.
2. Jennifer Yick, Biswanth Mukherjee, Dipak Ghosal, "Wireless Sensor network survey", International Journal of Computer networks, Vol. 52, pp. 2292- 2330, 2008.
3. Akylidiz, W. Su, Sankarasubramaniam, and E.Cayrici, "A survey on sensor networks", IEEE Communications Magazine, Volume:40 Issue: 8, August 2002, pp.102- 114.
4. Dressler, F, Akan, O.B, "A survey on bio-inspired networking", Computer Network, Vol. 54, pp.881-900, 2010.
5. Clerc, M. and Kennedy, J. "The particle swarm-explosion, stability, and convergence in a multidimensional complex space", IEEE Transactions on Evolutionary Computation, Vol. 6, No. 1, pp. 58-73. 2002.
6. [6] fariha nosheen, sadia bibi, salabath khan,"Ants colony based optimization scheduling algorithm", International conference on open source scheduling algorithm, 2013.
7. M. sasikala, A. Nithya, "Clustering in Wireless Sensor Net works: A Survey, " International journal of computer trends and technology, vol 17 num 3, 2014.
8. Mukul Pratap singh, Kunal gupta, "Techniques of Power Optimization for Wireless Sensor Networks", International Journal of Computer Applications, vol66, num3, 2013.
9. Chandni, anjali bharti, "Optimization through Bio-inspired algorithms in Wireless sensor networks: survey and Future directions ", ISSNIP, 2015.
10. Salma, Mohamed chedly, "Cluster based wireless sensor networks optimization under energy constraints" ISSNIP, 2007.
11. Shahrzad Dehghania , Mohammad Pourzaferanib , Behring Berekatainc , " comparison on energy efficient cluster based routingalgorithms in wireless sensor network" The third infomation systems international conference, 2015.
12. C. Vimalarani, R. Subramanian, S. N. Sivanandam, "An Enhanced PSO Based Clustering Energy Optimization Algorithm for Wireless Sensor Network" scientific world journal, 2016.
13. T. Luo, H.-P. Tan, and T. Q. Quek, "Sensor open flow: Enabling software-defined wireless sensor networks," IEEE Com- munications Letters, vol. 16, no. 11, pp. 1896-1899, Nov. 2012.
14. Sandra sundra, lorena parra, Jaime Lloret,"systems and algorithmsfor wireless sensor networks based on Animal and natural behavior".
15. Debmalya Bhattacharya, R.Krishnamoorthy, "power optimiza tion in wireless sensor networks", International journal of computer science, vol8, 2011.
16. Bonabeau, E., Dorigo, M. and Theraulaz, "Swarm intelligence",Oxford University Press, 1999.
17. R. Storn, K. Price, "Differ- ential evolution - a simple and efficient heuristic for global optimization over continuous spaces", Journal of Global Optimization, Vol. 11, pp. 341-359, 1997.
18. Hongliang Ren, Max Q H Meng, "Bio Inspired approaches for Wireless Sensor Networks", IEEE Conference on Mechatronics and Automation a Survey", pp. 762-768, 2014.
19. MM Chandane, S.G.Bhirud, S.V.Bonde "Energy Optimization inWireless Sensor Network", springer, 2013.

ADVANCEMENT OF CLOUD COMPUTING IN HEALTHCARE SECTOR

S. S. Shanthi

Assistant Professor, Department of Computer Science, NGM College, Pollachi

ABSTRACT

Cloud computing is a recent and fast growing area of development in healthcare. Ubiquitous, on-demand access to virtually endless resources in combination with a pay-per-use model allow for new ways of developing, delivering and using services. Cloud computing is often used in an “OMICS-context”, e.g. for computing in genomics, proteomics and molecular medicine, while other field of application still seem to be underrepresented. Healthcare sector is facing more number of problems like high IT costs, more processing power, scalability and demand for interoperability etc. Present health technology is deficient to address these challenges. Cloud computing having many qualities such as multi-tenancy, flexibility and metered delivery appears a viable approach. The purpose of this document is to provide six areas in which health clouds are resolving key challenges for the healthcare community.”How cloud solutions are applicable and impacting Healthcare” and some of the key principles that healthcare industry need to focus on when building a strategy for their organization’s to adopt cloud computing.

INTRODUCTION

In order to understand Cloud Computing in the healthcare industry, we must understand the basics of cloud computing in general. Figure 1, gives overview of Cloud computing

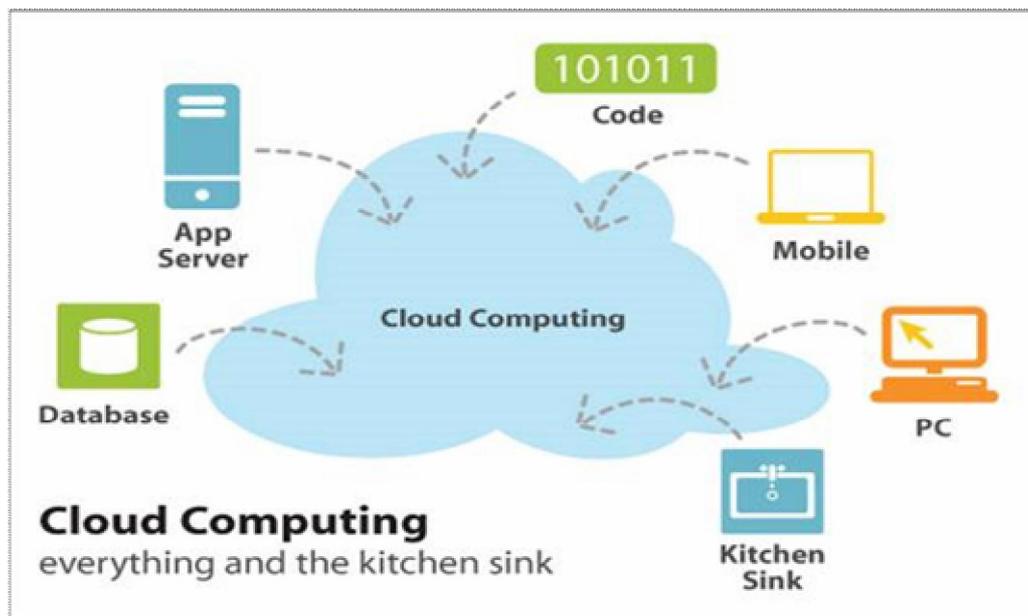


Figure-1: Cloud computing overview

A. Cloud Computing Architecture

To define Cloud computing, “Cloud computing is a new way of delivering computing resources and services”. Cloud Computing can improve health care services, benefit of health care research, and change the face of health information technology. However as with new innovation type, cloud computing should be more rigorously evaluated before its widespread adoption. Healthcare is faster growing its way to adapt cloud computing. This is to use benefits of clouds services at minimum cost, effective use of resources and maximized availability of services. However, like in other fields healthcare is hesitant to embrace the cloud computing environment because of concerns of data privacy, availability and data integrity. Cloud-based services are steadily becoming very large adopted by healthcare organizations.

Cloud computing architecture can be broadly classified into two following parts

1. Front End
2. Back End

Each of the ends is connected through a network, usually Internet. The following Figure 2.ndepicts the graphical view of cloud computing architecture:

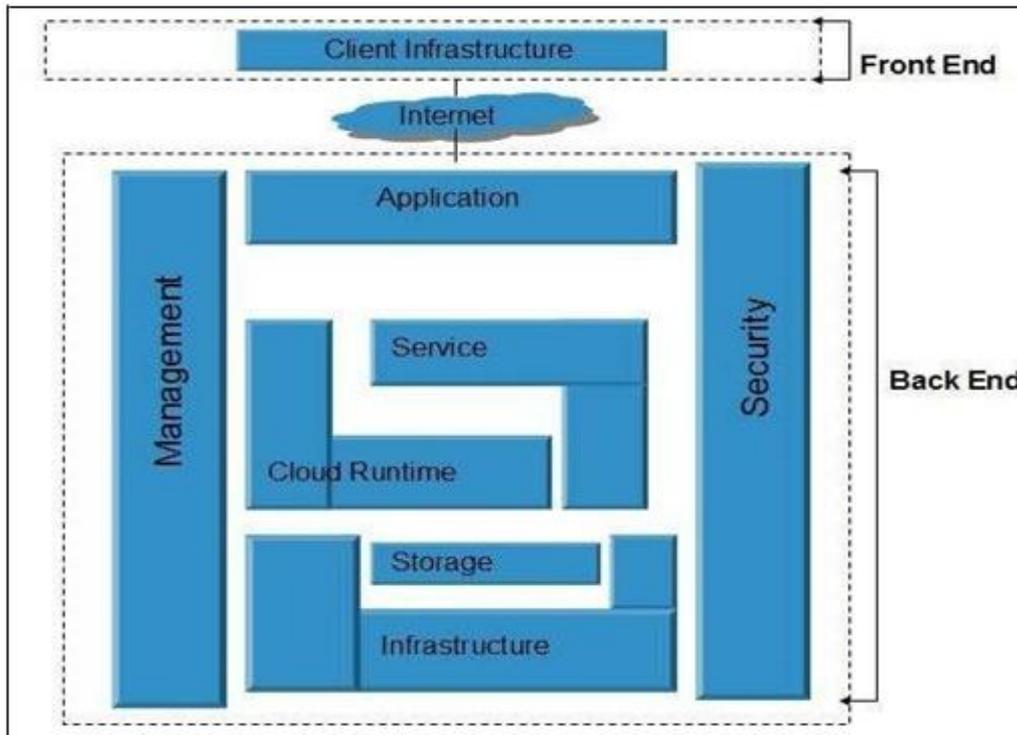


Figure-2: Cloud computing architecture

Front End

It refers to the client part of cloud computing system. As in the above figure, front end consists of interfaces and applications which are used for accessing the cloud computing platforms for example - Web Browser.

Back End

It refers to the cloud itself. The above figure gives overview of back end components; it consists of all the resources, which are required to provide cloud-computing services. Back end comprises of multiple things such as, huge data storage, virtual machines, security mechanism, different services deployment models, different servers, etc.

B. Cloud infrastructure

Cloud infrastructure comprise of many servers, storage devices, cloud management software, network, Deployment software and platform virtualization. The following figure depicts the Cloud Infrastructure Components:

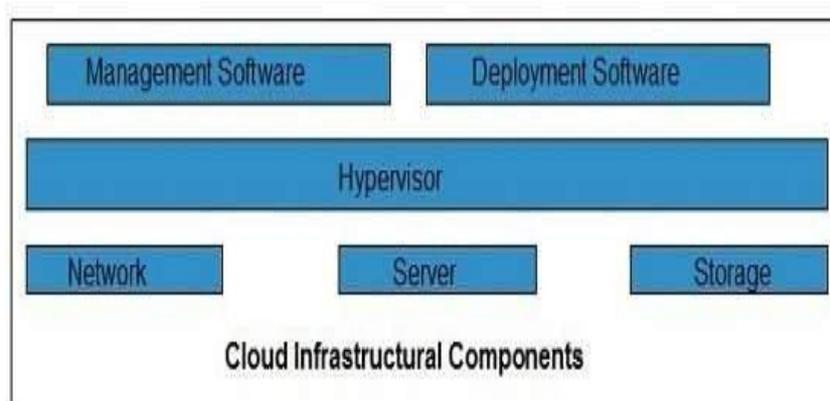


Figure-3: Cloud infrastructure

1. Hypervisor

Hypervisor is one of the infrastructure components; it is a firmware or a low-level program that acts as a Virtual Machine Manager. It allows sharing the single physical instance of cloud resources between many tenants.

2. Management Software

Management Software helps to maintain and configure the infrastructure.

3. Deployment Software

Deployment Software helps to deploy and integrate the application on the cloud.

4. Network

Network is the key component of cloud infrastructure. It allows connecting cloud services over the Internet. Network can be delivered as a utility over the Internet, which means, the customer can customize the network route and protocol.

5. Server

The main role of server is to help to compute the resource sharing and offers other services which include resource allocation, de-allocation, monitoring the resources, providing security etc.

6. Storage

Cloud keeps multiple copies of storage. If a storage resource fails, then it can be extracted from another one, which makes cloud computing more reliable.

Technology is increasing the connectedness of society. People use cell phones, the Internet, and handheld devices to stay in-touch and connected with friends, family, and work. A recent study conducted by the American Life Project found that 74% of American adults use the internet and that 55% of adults connect wirelessly, either using Wi-Fi or WiMAX via their laptops or using handheld devices, such as a smart phone.

There are many businesses and industries affected by this trend in technology and one among them is the health care sector.

C. Impact of Cloud computing to Healthcare industry

Technology is constantly evolving, especially in the highly competitive healthcare industry. Many Hospitals have already undergone changing their legacy systems to include electronic health records (EHRs), a digital format of paper medical records which was mandated by the Health Information Technology for Economic and Clinical Health Act (HITECH) and enforced by the American Recovery and Reinvestment Act of 2009 (ARRA). Such a transformation in healthcare has provided administration personnel, physicians, and nurses with timely access to medical records when needed.

EHRs at many medical facilities are currently housed on traditional client-server architectures. IT technology has already helped simplify operations in this regard, making the process much more efficient and patient-centric than in the olden days. **Adopting cloud computing solutions can make healthcare operations even more convenient and cost effective.**

The cloud offers on-demand computing by using the latest in technology to deploy, access, and use networked information, applications, and resources. Unfortunately, it also has a complex infrastructure that may be challenging to understand.

In most cases, end users are sure to find that cloud computing is the best choice for their healthcare business, as it's often less costly than having multiple computers in various medical rooms — each needing proper hardware, updated software, and network accessibility to upload, store, and retrieve patient or other medical data.

Healthcare IT solutions have offered many worthy benefits to the industry already. And now that enhanced security and safeguards are in place for cloud computing solutions, carriers, and service providers, healthcare organizations can rest easy knowing they're protected from potential loss of control over certain sensitive patient data.

With IT spending on the rise, cloud-based electronic health records (EHRs) is beginning to have an impact on the health industry.

D. Risks of Cloud Computing in Healthcare

Lack of security and privacy are the two primary concerns healthcare providers face when choosing a cloud Solution. In order to overcome these concerns, healthcare businesses must choose a reliable cloud provider Who acts in complete accordance with the provisions set forth in the Health Insurance Portability and Accountability Act (HIPAA) of 1996.

With massive data breaches increasingly reported in recent years, there is a growing uneasiness amongst patients who fear that hospitals and doctors that use a cloud service provider will complicate privacy of their data. There are also concerns of allowing multiple users to share EHRs among facilities.

In addition to patient privacy, data breaches cost healthcare organizations millions and millions of dollars each year. In fact, two of the most recent Cost of a Data Breach Studies from the Ponemon Institute shows that stolen healthcare records cost twice the global average. The average cost per stolen record was \$380 in 2017 (global average was \$141). **Such costs can be devastating for healthcare businesses.**

E. Benefits of Cloud computing in Healthcare Providers

Despite these concerns and risk, cloud computing is continually transforming healthcare in the modern age. First, as a **Software as a Service (SaaS)**, the cloud can offer healthcare organizations on-demand hosted services, providing quick access to business applications and fulfilling customer relationship management (CRM). As an **Infrastructure as a Service (IaaS)**, cloud solutions can offer on-demand computing and large storage for medical facilities.

And lastly, as **Platform as a Service (PaaS)**, the cloud can offer a security-enhanced environment for web-based services and the deployment of cloud applications. Transforming healthcare via the cloud is about more than just the delivery of medical information from multiple computers at anytime, anywhere, and on any mobile device. It's also about the benefits of being able to connect medical centers and cloud users for the purpose of sharing patients' health data over the Internet.

The cloud offers many benefits to a broad range of healthcare stakeholders by virtue of its elastic and virtually unlimited scalability, the high availability and accessibility of data despite volume, velocity, and variety, and a desirable IT budgetary shift from capex to opex. While initial concerns about data privacy and security limited cloud adoptions by the healthcare industry, those fears have been mitigated and cloud adoption has accelerated.

There are numerous advantages and benefits of cloud computing in healthcare. On-demand cloud computing and storage reduces operational costs for healthcare providers such as hospitals and clinics. The cloud can also help support the data-heavy health IT technologies being used in the industry today such as electronic medical records, patient portals, mobile apps, medical devices with IoT technology, and the big data analytics behind improved decision support systems, and therapeutic strategies. The cloud has also changed the face of clinical research, with enhanced support for knowledge-sharing and clinical trial management.

As a highly regulated sector that handles extremely sensitive data, the key health IT requirements of the cloud are

- High data availability, robust backup and disaster recovery capabilities, supporting minimal RTOs and RPOs.
- Effective data security, protecting from unauthorized access and breaches.
- Proven compliance with regulatory frameworks such as the EU's General Data Protection Regulation (GDPR) for the protection of personal data, the US Health Insurance Portability and Accountability Act (HIPAA) for secure data portability, and the HITRUST Alliance's CSF, an industry-mediated certifiable standard for safeguarding sensitive information.

To meet those challenges, some health IT companies need more than their in-house solutions and cloud native resources can provide. That's where NetApp can step in. Cloud Volumes ONTAP is built on AWS storage or Azure storage to meet the cloud requirements of health IT vendors and their end-users. The benefits include:

- **Data Protection:** NetApp Snapshot technology makes instant point-in-time data copies without affecting performance.
- **Disaster Recovery:** NetApp Snap Mirror provides quick and efficient data replication for seamless backup, disaster recovery, and data mobility across hybrid infrastructures.
- **Significantly reducing cloud costs with built-in data storage efficiencies**—compression, deduplications, thin provisioning.
- **Volume cloning technology** instantaneously creates a writable, point-in-time copies of volumes without using additional disk space, useful for dev/test requirements for health IT vendors as well as for data analytics for healthcare researchers and insurers.
- **Automated storage tiering** for cost-effective storage of “cold” data while keeping “hot” data accessible to clinicians, patients, and more.
- **Secure encrypted data** across the entire NetApp data fabric.

F. Examples of How Cloud Solutions Can Transform Healthcare

Healthcare facilities that ultimately decide on a private, public, or hybrid cloud solution can opt for a virtualization platform at VMware or Microsoft. At Innovative Architects, we usually recommend choosing Microsoft's secure cloud platform, which uses Windows Server with Hyper-V and the System Center. This scalable solution is best able to meet most growing business's needs, helping easily power cloud applications and/or supply cloud-based computing and services.

The Microsoft Azure cloud computing system, in particular, can provide on-demand simple access to healthcare applications and data. Using a PaaS environment, Microsoft provides a service to supply providers with networks, servers, and storage.

Microsoft Azure complies with the data protection and privacy laws set forth in HIPAA and the HITECH Act. This system also meets Cloud Security Alliance (CSA) as well as Governance, Risk and Compliance (GRC) criteria. Either implement Azure's .NET Services to integrate public cloud-based applications, or turn to SQL Server-based data services to properly secure the entire infrastructure.

Regardless of what cloud service platform you choose or which provider delivers the best service, the delivery of computing and service must permit sharing of proprietary data resources to help physicians and healthcare providers to do their jobs effectively and efficiently. Both the cloud platform and cloud provider must also ensure all of your digital medical data remains secure and private. So long as these conditions are met, there will be less and less resistance to cloud computing adoption in the healthcare industry.

G. Pros and Cons of Healthcare Systems**The Pros are****1. Low costs**

- a. Cost is lesser with respect to man power ,paper work and many more factors ,but this is applicable to large organizations
- b. Outsourcing information reduce amount spent on new technology
- c. Easier to maintain when compared with existing system. Whether the existing system is On-premise or old traditional technique.
- d. More secure
- e. Companies are hired to watch over the information

2. Interoperability

- a. Access information from anywhere
- b. Can be accessed using different devices
- c. Patient don't need to visit hospital for reports

3. Clinical Research

- a. Data are stored on cloud and are available for clinical research

4. Electronic Medical Records

- a. Patients records are stored electronically

The Cons are**1. Privacy and Security Challenges**

- a. Data maintained in a cloud may contain personal, private or confidential information such as healthcare related information that requires the proper safeguards to prevent disclosure, compromise or misuse.
- b. Globally, concerns related to data jurisdiction, security, privacy and compliance are impacting adoption by healthcare organizations

H. 6 Ways Cloud Computing Is Transforming Healthcare Systems

Here are six areas in which health clouds are resolving key challenges for the healthcare community.

The global healthcare cloud computing market is expected to reach \$9.48 billion in 2020 from \$3.73 billion in 2015 — a 20.5% compound annual growth rate. The market is expected to be dominated by North America, followed by Europe, Asia, and the Rest of the World (RoW). This growth will place new demands on healthcare systems' infrastructure to maintain and improve access to quality care while stabilizing healthcare costs. Enter cloud computing.

As digital transformation ripples across the healthcare industry, advancements in connectivity, security, and cloud services technologies are allowing the healthcare ecosystem to address several of the biggest challenges that this community is facing — leveraging health “clouds” to address key issues, such as healthcare access, personal data privacy, and drug theft and counterfeiting.

Here are six areas in which health clouds are resolving key challenges for the healthcare community.

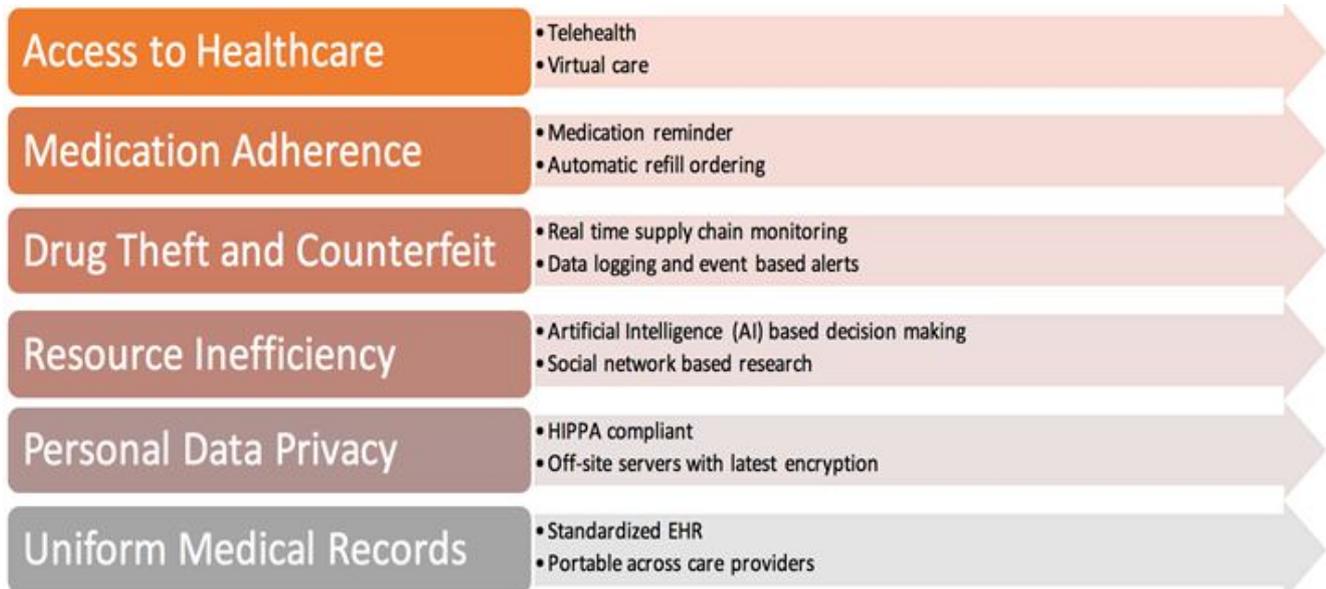


Figure-4: Six ways cloud computing addresses deficiencies in the healthcare system.

Access to healthcare: In remote parts of the country and for patients with busy schedules, accessing healthcare is a major issue. Telehealth or virtual care solutions are gaining popularity, with some states modifying laws to accelerate adoption and address this problem, e.g., Ring Central, Cisco.

Medication adherence: Patients not following the drug regimen prescribed by the doctor, results in avoidable re-admissions to the hospital costing the U.S. healthcare system \$290 billion annually. Market for products that remind the patient when it is time to take the medicine, keep a log and automatically order refills is expanding at a rapid pace, e.g., Omnicell, Medi Pense, Medissimo.

Drug theft and counterfeiting: Theft, counterfeiting, selling expired medicine are some of the problems which can be controlled by monitoring the supply chain. Recently, government has passed the Drug Supply Chain Security Act (DSCSA) which will require all the stakeholders in the pharmaceutical supply chain to monitor and report status on drugs during transportation. This has opened up a market for solutions which monitor and log supply chain events in real-time and report suspicious events, e.g., RxTrace, rfXcel.

Resource inefficiency: Escalating cost of healthcare is always a hot topic among policy makers and no real solution has been implanted to date. One of the major factors adding to the cost of healthcare is inefficiency of resources like medical staff, equipment and easy access to patient resource pool for clinical studies. With the use of artificial intelligence in the healthcare environment, a doctor’s bandwidth can be stretched since data can be augmented with smart machine-based analytics for doctor to evaluate. For clinical trials or scientific studies, a social network-based approach can be used to gain access to a patient pool, e.g., IBM Watson, Apple Research Kit.

Personal data privacy: Each healthcare organization maintaining their own medical records is a nightmare for data security and compliance to Health Insurance Portability and Accountability Act (HIPAA). Not to mention that it adds significant cost for organizations to maintain their own IT infrastructure and be liable for all the data directly. Using cloud-based solutions provides access to latest security technologies and reduces individual liability, e.g., Verizon Thingspace, IBM cloud.

Uniform medical records: Each hospital or care provider using their own Electronic Health Record (EHR) system is not in favor of the consumer. Not only does it add cost to the healthcare system since each hospital has to maintain a different system but it also makes it more resistive for a patient to change care providers. In other words the patient gets tied to a certain care provider and may not always get the optimum care which would be the case if they had easy access to their EHR in a standard format. Industry wide effort is going on to

have one EHR or Personal Medical Record (PMR) tied to the patient so that they get the best care agnostic of the provider network, e.g., Apple PMR, Picnic Health.

Cloud computing and healthcare is a match made in heaven to improve the quality of life for our society. With recent advances in cloud computing — improved performance and storage at better cost, robust security, interoperability, hybrid infrastructure, seamless integration — the healthcare ecosystem, is well positioned to take advantage of connected applications to create and deploy better healthcare solutions.

CONCLUSION

Cloud technology is used to create a network between patients, doctors, and healthcare institutions by providing applications, services and also by keeping all the data in the cloud. There are very wide opportunities and reasons to choose Cloud Technology in healthcare. It is right time to join hands with clinicians and computer specialists to forward in order to implement the large technology which suits small practices to big hospitals. The current technology of adopting Cloud computing in the medical field can improve and solve many collaborative information issues in healthcare organizations and cost optimizations. Standardized cloud-based applications will bring advantages to the patients, physicians, insurance companies, pharmacies, imagining centers, etc. Challenges such as security concerns and interoperability will rise because of the cloud-computing model. Therefore, the adoption of the cloud is progressing slowly. Through the implementation of best practices in the design, deployment and use of it will hopefully generate a Future growth of the cloud-based systems adoption, despite all of the problems.

REFERENCES

- [1] <https://www.wiseguyreports.com/sample-request/2963581-healthcare-cloud-computing-global-markets-to-2022>
- [2] <http://www.healthcareglobal.com/tech/1630/How-Cloud-Computing-is-Changing-the-Health-Care-IT-Industry>
- [3] Scott Good "Why Healthcare Must Embrace Cloud Computing" Forbes May 2013.
- [4] Vishal Gupta, Cloud computing in health care, Express Healthcare, Sept 2011.

**DISCRIMINATIVE PATTERN WITH MULTIPLE SUPPORT VECTOR MACHINE (MSVM)
PREDICTION FRAMEWORK (DPMSvmPF)**

J. Jeyaboopathiraja¹ and Dr. G. Maria Priscilla²Research Scholar¹ and Professor², Sri Ramakrishna College of Arts and Science, Coimbatore

ABSTRACT

In the review, two series of methods have been introduced to solve prediction issues which consist of classification and regression. Straightforward methods such as generalized linear methods have normal accuracy however well-built interpretability on a set of straightforward attributes. The other series, consists of tree-based methods, organize mathematical, unconditional and high dimensional attributes into a complete formation by means of prosperous interpretable information in the data. In this work proposed a new Discriminative Pattern with Multiple Support Vector Machine (MSVM) Prediction Framework (DPMSvmPF) in the direction of achieve the prediction steps with attractive their merits of together efficiency and interpretability. Specifically, DPMSvmPF make use of the brief discriminative patterns with the purpose of are on the prefix hyperplanes in the classifier. Furthermore, DPMSvmPF chooses a restricted no. of the useful discriminative patterns with searching designed for the mainly successful pattern combination in the direction of fit universal linear classifiers. At last compress the no. of discriminative patterns with choosing the mainly effective pattern combinations with the purpose of fit addicted to a widespread classifier model with increased prediction results. This part of fast and successful pattern extraction authorizes the well-built preventability and interpretability of DPMSvmPF.

Keywords: Discriminative Pattern, Generalized Linear Model, Classification, Regression, Multiple Support Vector Machine (MSVM) , Discriminative Pattern with Multiple Support Vector Machine(MSVM) Prediction Framework (DPMSvmPF).

1. INTRODUCTION

For informational indexes with class labels, association patterns [1] with the purpose of happen with irregular recurrence in a few classes versus others can be of significant value. Additionally allude to them as discriminative examples [2-3] in this work, in spite of the fact that these examples have likewise been explored under different names, for example, rising examples [4], differentiate sets [4] and directed descriptive standards [5]. Discriminative examples have been appeared to be valuable for enhancing the prediction results [6] and for circle covering test subgroups [5]. The calculations for finding discriminative examples normally utilize a measure for the discriminative intensity of an example. Such measures are by and large characterized as a component of the example's relative help in the two classes, and can be characterized both essentially as the proportion of the different varieties, for example, information gain, and Gini record and so forth.

Frequent design based characterization has been investigated lately and its capacity was shown by numerous examinations in a few spaces, including (1) association classification [7] on straight out information, where a classifier is fabricated in view of high- support, high- confidence rules; and (2) frequent pattern based classification [8] in light of content or information with complex structures, for example, successions and diagrams, where discriminative frequent examples are taken as highlights to high quality classifiers. A successive itemset (pattern) is an arrangement of things that happen in a dataset no not as much as a user determined least support (minsup). Visit designs have been investigated generally in grouping undertakings. These examinations accomplish promising grouping precision and show the achievement of successive examples in classification.

All the more imperatively, the grouping assignments connect extraordinary significance to the successive itemsets that are profoundly discriminative as for the class labels. Since visit itemsets are produced exclusively in view of help data, not founded on discriminative power, countless itemsets can be created among the mining step. At the point when the total mining comes about are restrictively substantial, yet just the exceedingly discriminative ones are of genuine intrigue, it is wasteful to sit tight perpetually for the mining calculation to complete and after that apply highlight choice to post-process the tremendous measured mining comes about. Notwithstanding for a feature selection with straight difficulty, it could be extremely costly to process a hazardous number, for example, millions, of highlights which is a typical scale in frequent patterns.

In this work proposed a new Discriminative Pattern with Multiple Support Vector Machine (MSVM) Prediction Framework (DPMSvmPF) in the direction of achieve the prediction steps with attractive their merits of together efficiency and interpretability. Furthermore, DPMSvmPF chooses a restricted no. of the useful discriminative

patterns with searching designed for the mainly successful pattern combination in the direction of fit universal linear classifiers.

2. LITERATURE REVIEW

Fang et al [9] introduced a far reaching discourse that characterizes these four example composes and researches their properties and their relationship to each other. Furthermore, these thoughts are investigated for an assortment of datasets (ten UCI datasets, one gene articulation dataset and two hereditary variety datasets). Classify discriminative examples into four gatherings in view of the accompanying kinds of cooperations: (i) driver-traveler, (ii) coherent, (iii) autonomous added substance and (iv) synergistic past independent option. The outcomes show the presence, attributes and measurable noteworthiness of the distinctive kinds of examples. They likewise delineate how design portrayal can give novel bits of knowledge into discriminative example mining and the discriminative structure of various datasets.

Li et al [10] proposed a way to deal with remove the discriminative examples for effective human activity acknowledgment. Each activity is considered to comprise of a progression of unit activities, every one of which is spoken to by an example. Given a skeleton grouping, first naturally separate the key-outlines for unit activities, and afterward mark them as various examples. Likewise further utilize a factual metric to assess the discriminative capacity of each example, and characterize the sack of solid examples as neighborhood highlights for activity acknowledgment. Test comes about demonstrate with the purpose of the separated nearby descriptors could give high exactness in the activity acknowledgment, which show the productivity of strategy in extricating discriminative examples.

He et al [11] considered a novel thought named restrictive discriminative example to address this issue. To mine restrictive discriminative examples, proposed a Conditional Discriminative Patterns Mining (CDPM) calculation to produce an arrangement of non-repetitive discriminative examples. Test comes about on genuine informational indexes show that CDPM has great execution on expelling repetitive examples with the purpose of gotten from noteworthy sub-designs to create a compact arrangement of important discriminative examples.

Cheng et al [12] proposed a few all inclusive models to screen cardiopulmonary conditions, including DPClass, a novel learning approach. Likewise precisely plan movement dataset covering status from GOLD 0 (solid), GOLD 1 (mellow), GOLD 2 (direct), the distance to GOLD 3 (serious). Sixty-six subjects take an interest in this investigation. After de-distinguishing proof, their strolling information is connected to prepare the prescient models. The Radial Bias Function (RBF) with Support Vector Machine (SVM) display yields the most exactness while the DPClass demonstrate gives better understanding of the model systems. Not just give promising answers for screen wellbeing status by essentially conveying a cell phone, yet in addition exhibit how socioeconomics impacts prescient models of cardiopulmonary illness.

Yan et al [13] proposed a novel ordering model in view of discriminative successive structures that are distinguished through a diagram mining process. The proposed approach not just gives an exquisite answer for the graph indexing issue, yet in addition shows how database ordering and inquiry preparing is able to profit by information mining, particularly Frequent Pattern Mining (FPM). Results demonstrate that the minimal list worked under this model can accomplish better execution in preparing graph inquiries. Since discriminative successive structures catch the inherent qualities of the information, they are moderately steady to database refreshes, in this way encouraging examining based element extraction and incremental file upkeep. Besides, the ideas created here can be summed up and connected to ordering groupings, trees, and other muddled structures too.

Fang et al [14] tended to the need of exchanging off the fulfillment of discriminative example disclosure with the effective revelation of low- support discriminative examples from such informational indexes. Likewise proposed a group of antimonotonic measures named SupMaxK with the reason for sort out the arrangement of discriminative examples into settled layers of subsets, which are logically more total in their scope, however require progressively more calculation. Specifically, the individual from SupMaxK with $K = 2$, named SupMaxPair, is appropriate for thick and high-dimensional informational. Examinations on both manufactured informational indexes and a malignancy quality articulation informational collection show that there are low-bolster designs that can be found utilizing SupMaxPair however not by existing methodologies. Besides, we demonstrate that the low- support discriminative examples that are just found utilizing SupMaxPair from the tumor quality articulation informational collection are factually critical and organically important. This delineates the complementarity of SupMaxPairXo existing methodologies for discriminative pattern discovery.

3. PROPOSED WORK

In this work proposed a new Discriminative Pattern with Multiple Support Vector Machine (MSVM) Prediction Framework (DPMSvmPF) in the direction of achieve the prediction steps with attractive their merits of together efficiency and interpretability. Specifically, DPMSvmPF make use of the brief discriminative patterns with the purpose of are on the prefix hyperplanes in the classifier. Furthermore, DPMSvmPF chooses a restricted no. of the useful discriminative patterns with searching designed for the mainly successful pattern combination in the direction of fit universal linear classifiers. At last compress the no. of discriminative patterns with choosing the mainly effective pattern combinations with the purpose of fit addicted to a widespread classifier model with increased prediction results. This part of fast and successful pattern extraction authorizes the well-built preventability and interpretability of DPMSvmPF.

3.1. Problem Formulation

For an prediction task, the information is an arrangement of n cases in a d-dimensional component space together with their labels $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, for $\forall i (1 \leq i \leq n), x_i \in R^d$. It is significant that the qualities in the illustration x_i can be either ceaseless (numerical) or discrete (clear cut). As absolute highlights can be changed into a few double sham pointers are can accept x_i without the loss of consensus. The class y_i is either a class (type) marker or a genuine number contingent upon the particular errand. In past example based techniques, designs are extricated from unmitigated qualities and in this manner they are just ready to deal with the nonstop factors after cautious manual discretization, which is precarious and frequently requires earlier learning about the information. Proposed DPMSvmPF structure is to take in a brief model that comprises of a little arrangement of discriminative examples from the preparation information, which learns and predicts the cases as precisely as would be prudent, i.e., anticipate the right class marker in order assignments and foresee near the genuine number in regression tasks.

Condition is a thresholding boolean capacity on a particular element measurement. The condition is as $(x_{ij} < v)$ or $(x_{ij} \geq v)$, where j shows the particular measurement and v is the limit value. The relational operator in a condition is either < or \geq . For any measurement j in highlights comparing to paired pointers, we limit v to be 0:5. Example is a conjunction clause of conditions on particular element measurements.

Discriminative Patterns allude to those examples which have solid flags on the learning undertakings, given the names of information. For instance, an example with high information gain on the characterization preparing information, or an example with little mean square mistake on the relapse preparing information, is a discriminative example. Discriminative examples have covered prescient impacts. In particular, a couple of discriminative examples are unique instances of different examples. Be that as it may, the second example just encodes a subset of information focuses that the primary example encodes, and along these lines, it doesn't give additional data to the learning procedure. This normal wonder demonstrates that generally taking the best discriminative examples in view of free heuristics squanders the financial plan of the quantity of examples, when the straight blend of these examples are not synergistic. Top-k Patterns are formalized as a size-k subset of discriminative examples, which has the best execution in light of the preparation information.

Discriminative Pattern Generation

The primary segment in the DPMSvmPF system is the age of superb discriminative examples. Utilize hyperplane to elude the arrangement of cases falling into a particular hub in the machine learning classifier. SVM multi-class order is most likely the one-against-all technique. It builds k SVM models where k is the quantity of classes. The mth SVM is prepared with the greater part of the cases in the mth class with positive labels, and every single other case with negative labels. Consequently known l training information $(x_1, y_1), \dots, (x_l, y_l)$, where $x_i \in R^n, i = 1, \dots, l$ and $y_i \in \{1, \dots, k\}$ is the class of x_i , the mth SVM resolve the subsequent issue:

$$\min_{w^m, b^m, \xi^m} \frac{1}{2} (w^m)^T w^m + C \sum_{i=1}^l \xi_i^m$$

$$(w^m)^T \phi(x_i) + b^m \geq 1 - \xi_i^m, \text{ if } y_i = m.$$

$$(w^m)^T \phi(x_i) + b^m \leq -1 + \xi_i^m, \text{ if } y_i \neq m$$

$$\xi_i^m \geq 0, i = 1, \dots, l$$

where the training data x_i are mapped in the direction of a higher dimensional space with the function ϕ and C is the fine factor. Reducing $\frac{1}{2} (w^m)^T w^m$ means with the purpose of we would like in the direction of maximize $2/||w^m||$, the margin among two clusters of data. When data are not linear distinguishable, there is a fine

factor $c \sum_{i=1}^k \xi_i^m$ which be able to decrease the no. of training errors. The basic idea following SVM is in the direction of search meant for a balance among the regularization period $\frac{1}{2} (w^m)^T w^m$ and the training errors. Following solving (1), there are k conclusion task:

$$\begin{aligned} &(w^1)^T \phi(x) + b^1, \\ &\vdots \\ &(w^k)^T \phi(x) + b^k \end{aligned}$$

Articulate x is in the group which have the main value of the choice function:

$$class\ of\ x \equiv \arg\ max_{m=1, \dots, k} ((w^m)^T \phi(x) + b^m)$$

For all intents and purposes we take care of the double issue of (1) whose number of factors is the same as the quantity of information in (1). Consequently k l-variable quadratic programming issues are solved. Another significant strategy is known as the one-against-one technique. It was presented in [16], and the main utilization of this procedure on SVM was in [17]. This technique develops $k(k - 1)/2$ classifiers where everyone is prepared on information from two classes. For preparing information from the i^{th} and the j^{th} classes, we fathom the accompanying double characterization problem:

$$\begin{aligned} &\min_{w^{ij}, b^{ij}, \xi^{ij}} \frac{1}{2} (w^{ij})^T w^{ij} + c \sum_{i=1}^l \xi_i^{ij} \\ &(w^{ij})^T \phi(x_t) + b^{ij} \geq 1 - \xi_t^{ij},\ if\ y_t = i, \\ &(w^{ij})^T \phi(x_t) + b^{ij} \leq -1 + \xi_t^{ij},\ if\ y_t = j \\ &\xi_t^{ij} \geq 0, i = 1, \dots, l \end{aligned}$$

There are distinctive techniques for doing the future testing after all $k(k-1)/2$ classifiers are built. After a few tests, we choose to utilize the accompanying voting procedure : if $sign((w^{ij})^T \phi(x) + b^{ij})$ says x is in the i^{th} class, at that point the vote in favor of the i^{th} class is included by one. Something else, the j^{th} is expanded by one. At that point foresee x is in the class with the biggest vote. The voting approach portrayed above is likewise called the "Maximum Wins" technique. In the event that those two classes have indistinguishable votes, figured it may not be a decent technique, now we essentially select the one with the smaller index. For each discriminative example, there is one comparing twofold measurement depicting whether the occurrences fulfill the example or not. Since the measurement of the example space is equivalent to the quantity of discriminative examples which is an expansive number after the age stage, we have to additionally choose a set number of examples and therefore influence the example to space little and productive. It is additionally justified regardless of a say that this mapping procedure can be completely parallelized for speedup.

Top-k Pattern Selection

After a substantial pool of discriminative examples is created, additionally top-k determination should be done to recognize the most informative and interpretable examples. A guileless path is to utilize heuristic capacities, for example, information gain and gini index, to assess the noteworthiness of various examples on the prediction and pick the best positioned designs. In any case, the impacts of best ranked patterns in view of the basic heuristic scores may have a huge segment of covers and in this manner their blend does not work ideally. Consequently, to accomplish the best execution and discover integral examples, propose two effective arrangements: forward choice and LASSO, which settle on choices in light of the impacts of the example mixes as opposed to considering distinctive pattern autonomously.

4. EXPERIMENTATION RESULTS

In this area, direct broad trials to exhibit the interpretability, proficiency and adequacy of proposed DPMSvmPF and Discriminative Pattern-based Prediction structure (DPPred) [18]. Additionally first present test settings, examine the effectiveness and interpretability, and afterward give the outcomes on grouping and relapse undertakings and also parameter examination. A few characterization datasets from University of California, Irvine (UCI) Machine Learning Repository are utilized as a part of the analyses, as appeared in Table 1 with measurements of the quantity of cases and the quantity of highlights. In the datasets adult, hypo and sick, the proportion of standard prepare/test part is 2:1. Along these lines, for the other grouping and relapse datasets, we separate the datasets into prepare/test (2: 1) by unprejudiced examining as preprocessing.

For classification tasks, to contrast and DDPMine, utilize the same datasets including adult, hypo, sick, crx, sonar, chess, waveform, and mushroom.. Since both DDPMine and DPPred accomplish relatively close

precision on the datasets waveform and mushroom, these two datasets are overlooked. What's more, the execution of DPPred and DPMSvmPF on high-dimensional datasets (nomao, musk and madelon datasets) is additionally explored, since DDPMine performs ineffectively on high-dimensional information. The metric is the precision on the testing information: higher exactness implies better execution. For relapse datasets, we pick general datasets, for example, bike and crime, and additionally clinical datasets where designs will probably be available, for example, parkinsons. Moreover, to make the mistakes in various datasets practically identical, min-max standardization is embraced to scale the consistent names into [0; 1]. The metric is the Rooted Mean Square Error (RMSE) on the testing information: a lower RMSE implies better execution.

Table-1: The statistics of real-world datasets from UCI Machine Learning Repository

Type	Dataset	# instances	# dimensions	Variable type
Classification	Adult	45222	14	Mixed
	Hypo	3772	19	Mixed
	Sick	3772	19	Mixed
	Chess	28056	6	Mixed
	Crx	690	15	Mixed
	Sonar	208	60	Numeric
High dimension	Nomao	29104	120	Mixed
	Musk	7074	166	Numeric
	Madelon	1300	500	Numeric
Regression	Bike	17379	10	Mixed
	Parkinsons	5875	16	Numeric
	Crime	1994	99	Numeric

Think about proposed DPMSvmPF classifier on the same datasets utilized as a part of DPPred-F and DPPred-L are appeared in the figure 1 and figure 2. The outcomes are appeared in Table 2 and table 3. DPPred-F and DPPred-L dependably have higher precision over other existing techniques. An essential reason of this preferred standpoint is that the applicant designs created by tree-based models in DPPred and DPMSvmPF are significantly more discriminative and hence more compelling on the particular characterization undertaking than those successive however less valuable examples. All the more shockingly, DPMSvmPF shows far better execution than the complex DPPred display on a few datasets, while its correctnesses on different datasets are as yet similar with RF, which is because of the viability of the example determination module where we select the ideal example mix as opposed to choosing designs autonomously. This demonstrates the proposed show is extremely compelling in arrangement undertakings while it is profoundly brief and interpretable.

Table-2: Test accuracy on classification datasets

Dataset /classifiers	Accuracy (%)					
	Adult	Hypo	Sick	Crx	Sonar	Chess
DPMSvmPF	87.52	99.25	98.63	89.93	88.36	93.51
DPPred-F	85.93	99.06	98.12	88.21	86.14	91.54
DPPred-L	84.02	99.21	97.53	88.12	84.41	91.91
DT	83.69	92.05	92.36	76.69	72.69	89.05
DDPMine	81.25	90.56	90.58	86.36	75.79	90.08

Table-3: Test error rate on classification datasets

Dataset /classifiers	Error Rate (%)					
	Adult	Hypo	Sick	Crx	Sonar	Chess
DPMSvmPF	12.48	0.75	1.37	10.07	11.64	6.49
DPPred-F	14.07	0.94	1.88	11.79	13.86	8.46
DPPred-L	15.98	0.79	2.47	11.88	15.59	8.09
DT	16.31	7.95	7.64	23.31	27.31	10.95
DDPMine	18.75	9.44	9.42	13.64	24.21	9.92

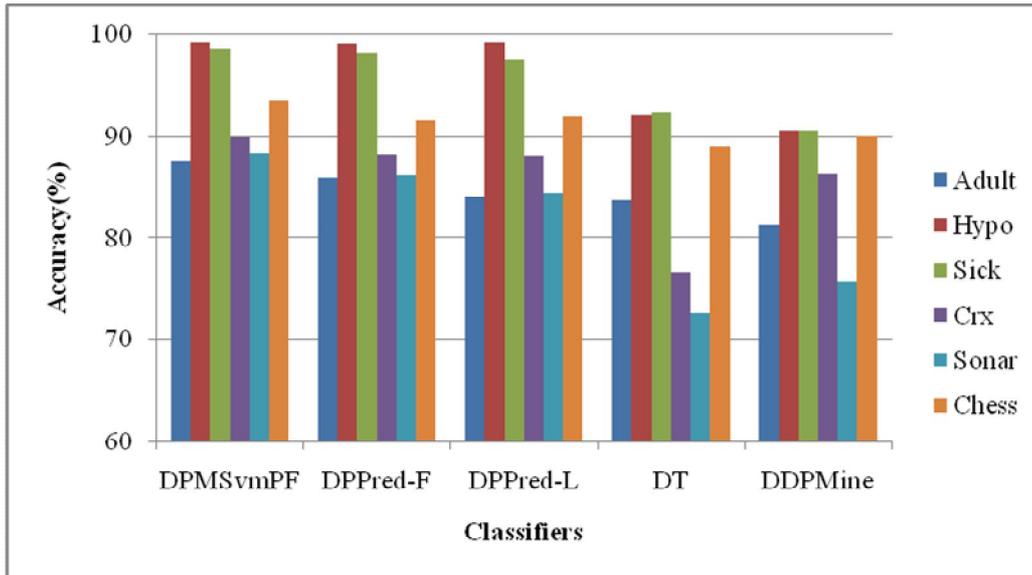


Figure-1: Classification accuracy results comparison

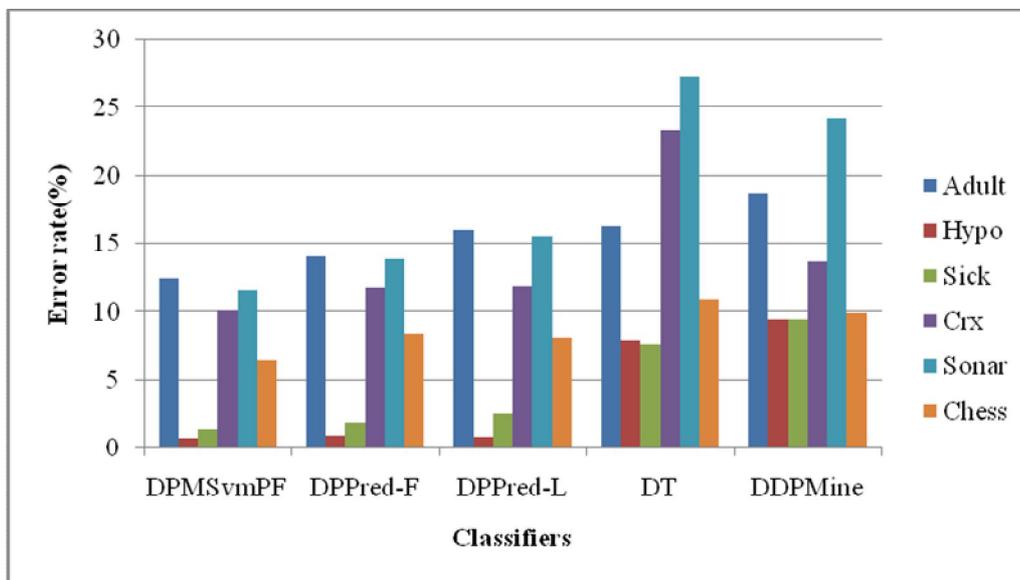


Figure-2: Error rate results comparison

5. CONCLUSION

In this paper, we propose a new Discriminative Pattern with Multiple Support Vector Machine (MSVM) Prediction Framework (DPMSvmPF) to address the classification issues and give high results with a lesser amount of discriminative patterns. Furthermore, DPMSvmPF selects a restricted no. of the useful discriminative patterns with searching designed for the mainly successful pattern combination in the direction of fit universal linear classifiers. The size of discriminative patterns is compacted by choosing the generally successful pattern combinations related in the direction of their analytical results in a widespread classifier. Extensive results show with the purpose of DPMSvmPF is capable in the direction of classifier high-order interactions and presented a small number of interpretable patterns in the direction of assist human experts recognize the data. DPMSvmPF gives comparable or even better performance than the state-of-the-art methods DPMSvmPF and random forest model in classification.

REFERENCES

1. P.-N. Tan, M. Steinbach, and V. Kumar. Introduction to data mining. Addison-Wesley, 2005.
2. G. Fang, R. Kuang, G. Pandey, M. Steinbach, C. Myers, and V. Kumar. Subspace differential coexpression analysis: problem definition and a general approach. Pacific Symposium on Biocomputing, 15:145–56, 2010.
3. G. Fang, G. Pandey, W. Wang, M. Gupta, M. Steinbach, and V. Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. IEEE TKDE, 2011.

4. G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In Proc. of SIGKDD , pp. 43–52, 1999.
5. P. Novak, N. Lavrac, and G. Webb. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *Journal of Machine Learning Research*, 10:377–403, 2009.
6. H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In *ICDE* , pp. 716–725, 2007.
7. Veloso, W. M. Jr., and M. Zaki, “Lazy associative classification,” in Proc. of *ICDM* , 2006, pp. 645–654.
8. M. Deshpande, M. Kuramochi, N. Wale, and G. Karypis, “Frequent substructure-based approaches for classifying chemical compounds,” *IEEE Trans. Knowl. and Data Eng.* , vol. 17, no. 8, pp. 1036–1050, 2005.
9. Fang, G., Wang, W., Oatley, B., Van Ness, B., Steinbach, M. and Kumar, V., 2011. Characterizing discriminative patterns. arXiv preprint arXiv:1102.4104.
10. Li, J., Liu, G. and Wong, L., 2007, Mining statistically important equivalence classes and delta-discriminative emerging patterns. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 430-439). ACM.
11. He, Z., Gu, F., Zhao, C., Liu, X., Wu, J. and Wang, J., 2017. Conditional discriminative pattern mining: Concepts and algorithms. *Information Sciences*, 375, pp.1-15.
12. Cheng, Q., Shang, J., Juen, J., Han, J. and Schatz, B., 2016, Mining discriminative patterns to predict health status for cardiopulmonary patients. In *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 41-49.
13. Yan, X., Yu, P.S. and Han, J., 2005. Graph indexing based on discriminative frequent structure analysis. *ACM Transactions on Database Systems (TODS)*, 30(4), pp.960-993.
14. Fang, G., Pandey, G., Wang, W., Gupta, M., Steinbach, M. and Kumar, V., 2012. Mining low-support discriminative patterns from dense and high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 24(2), pp.279-294.
15. Lo, D., Cheng, H., Han, J., Khoo, S.C. and Sun, C., 2009, Classification of software behaviors for failure detection: a discriminative pattern mining approach. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 557-566.
16. S. Knerr, L. Personnaz, and G. Dreyfus. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In J. Fogelman, editor, *Neurocomputing: Algorithms, Architectures and Applications*. Springer-Verlag, 1990.
17. U. H.-G. Kressel. Pairwise classification and support vector machines. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods – Support Vector Learning*, pp. 255–268, Cambridge, MA, 1998. MIT Press.
18. Shang, J., Jiang, M., Tong, W., Xiao, J., Peng, J. and Han, J., 2018. DPPred: An Effective Prediction Framework with Concise Discriminative Patterns. *IEEE Transactions on Knowledge and Data Engineering*, 30(7), pp.1226-1239.

DIGITAL MARKETING IN INTERNATIONAL ARENA**P. Chandradevi¹, A. Priya² and M. Subhashini³**

Assistant Professor, Department of Commerce, Sri Krishna Arts and Science College, Coimbatore

^{2,3}Department of Commerce, Sri Krishna Arts and Science College, Coimbatore

ABSTRACT

The world is transforming, and technology is taking the lead. The whole business world is in and around Digital Marketing. Today, everything is going digital – entertainment, health, real estate, banking, etc.

The digital performance of the market has accelerated to an exponential space. Digital marketing matters to heterogeneous generation as everyone is dependent on it.

The current generation is living in a competitive environment where money is considered as priority meanwhile they lack time.

The digital marketing is one of those solution for the problem of time management. As we all know that time and tide waits for none, the consumer expects to get wide range of market spaces from a single place.

Everyone in one or the other way is related in digital marketing, not just as products but the ideas, awareness and even the way to find things are being digitalised.

Digital marketing makes the people to be more socialised to the environment. Globalization is made easier due to various market and trade policy.

Interactive Chabot's, voice search, influencer marketing are few digital marketing trends.

Numerous challenges, scope and major benefits of digital marketing would be showcased in forth coming paper. Digital marketing has not just shaken nationally whereas it took over international arena.

People cannot survive without digital marketing, for everything that we use, update which is most important part. So if we cannot use a basic mobile without update its considered as out dated.

Hence here we are to make the spectators aware of the hidden truth, tricks and innovations which have taken around the world.

Keywords: Digital Marketing, Technology, Globalization, Social Media, Business

INTRODUCTION

Technology and marketing are inseparable. There are approx. 1,876 marketing technology vendors out there today, and 60% of marketers presume their companies to increase investments in technology this year. The early period of digital marketing technology can be traced back to the 1980s, when computers became sophisticated enough to store huge volumes of customer information. The digital databases of the 1980s transformed buyer-seller relationships, allowing brands to track their consumers like never before. But the process was still a manual one. Digital marketing became more sophisticated in the 2000s and the 2010s, when the proliferation of devices' capable of accessing digital media led to sudden growth. Statistics produced in 2012 and 2013 showed that digital marketing was still growing. With the development of social media in the 2000s, such as LinkedIn, Facebook, YouTube and Twitter, consumers became highly dependent on digital electronics in daily lives. Therefore, they expected a seamless user experience across different channels for searching product's information. The change of customer behaviour improved the diversification of marketing technology.

Digital marketing is also mentioned to as 'online marketing', 'internet marketing' or 'web marketing'. The term digital marketing has grown in popularity over time. In the USA online marketing is still a popular term. In Italy, digital marketing is referred to as web marketing. Worldwide digital marketing has become the most common term, especially after the year 2013. Digital media growth was estimated at 4.5 trillion online ads served annually with digital media spend at 48% growth in 2010. An increasing portion of advertising stems from businesses employing Online Behavioural Advertising (OBA) to tailor advertising for internet users, but OBA raises concern of consumer privacy and data protection.

BRAND IDENTITY

Digital marketing provides a wide range of different brands to the customers who at the loop and corners of the world. Digital marketing plays a major role as it reaches all kinds of people irrespective of their age, place, position, etc., to gain the goodwill. It would reach out more influencers in the market. They as the whole serve

as a promoting agent of the market. You can choose what gender, age group, geographic area and even interests that your audience has, and that can mean much more eligible eyes are seeing your post. And in the end, that means more brand awareness and more potential sales down the line.

Every page, every post and every link on your social media pages is an opportunity to be seen and noticed by web users. They help with SEO, and they have the power to improve or hinder your search engine rankings.

THE BIG 3

There’s a growing arsenal of digital marketing tools to help you gain better position for your website, services or products. If you look at all the varieties listed on other websites, you’ll probably be a bit overwhelmed. Basically, there are 3 categories you should know and consider:

1. SEO: Search Engine Optimization
2. SEM: Search Engine Marketing
3. SMM: Social Media Marketing

SEO: Search Engine Optimization

It helps place your website in the top search engine page results for relevant keywords. SEO uses a variety of methods such as content marketing, link building, directory population, social marketing and a ton of other tech-tricks to help you get better positioning. Website or on-page SEO refers to the process of making your website appeal to both search engines and users. Online directories are an important part of the SEO ranking system. While most of your potential customer’s search Google or Bing for information, the search engines reference these databases as the modern equivalent of a phonebook, only on a much larger scale.

SEM: Search Engine Marketing

This is digital advertising. When your customers are surfing and researching, we’ll put keyword-relevant text, graphic or video ads in front of them while they search, and on their favourite websites. We can target so that:

- Your ads are only presented to your best prospects and customers.
- Your ads are seen nationally or locally in as small as a five-mile radius.
- Your ads will only be displayed at times when your customers are most likely to be looking.

SMM: Social Media Marketing.

Digital marketing through social media channels.

Pew Research Centre reports 67% of internet users now use social media sites. Yet, many businesses still have trouble reaching the social crowd. It provides cost effective digital marketing solutions to strategically manage your blog, Facebook, Twitter and many of the other popular social sites. Whether you just need a strategy, or want full content creation and management, certain marketing has a social solution to fit your digital marketing needs. AS

STATISTICS OF CONSUMER ADOPTION:

Annual growth continues apace, especially in active mobile social users - 39% penetration up 5% from 2017.

Share of web traffic by device highly favours mobile at 52% (+4% year-on-year change), whilst Desktop remains in second place with only 43% of device share to all web pages, down by 3% year-on-year.

DIGITAL IN 2018 INTERNET PENETRATION MAP

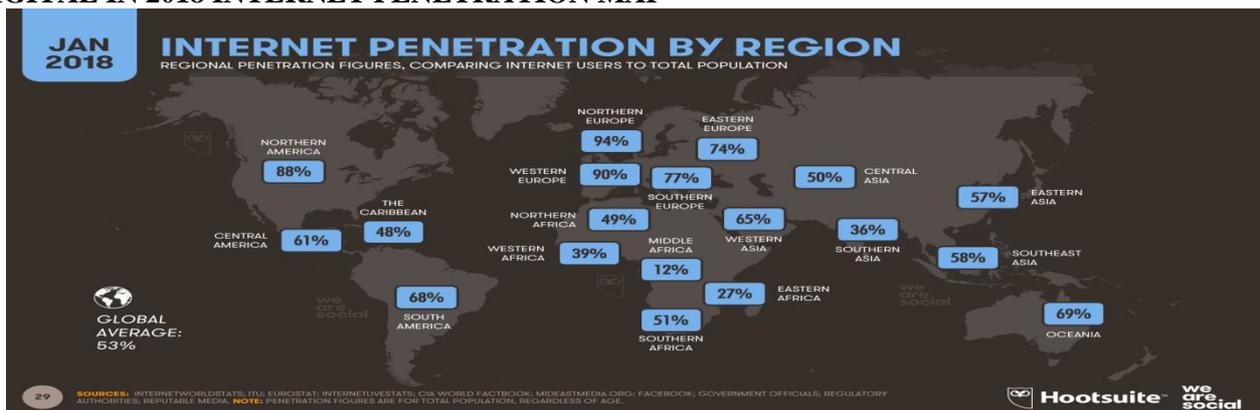


Fig-1

Source: https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjAre7JyYHgAhXaEHIKHemVCH0QjRx6BAgBEAU&url=https%3A%2F%2Fwearesocial.com%2Fus%2Fblog%2F2018%2F01%2Fglobal-digital-report-2018&psig=AOvVaw2IAm47_TVVRJvXIX9JI7jJ-&ust=1548252880393229

The above given map clearly shows the recent study of how much the whole population is penetrated to the social media which is termed to be the main platform for the digital marketing.

Here by it gives a global average of 83%, meanwhile northern Europe with the highest of 94% and the least is South Asia with 36%

Social users interact with brands on social

Knowing how to behaviour / post on social media is also important as this is reflective of your brand personality and subsequently influencers social uses to buy or unfollow! Social media is increasingly being used as a customer service platform where customers and potential customers want answers quickly and in real-time. Only 27% of social media user respondents will be prompted to purchase products after seeing behind the scenes content shared on brand social pages.

51% said that they would unfollow brands on social media if they posted irritating posts and 27% said they would mark/report the brand and page as spam and block them. This is why posting relevant, interesting content that resonates with your target audience and social users is important in retaining your Reach and Engagement with potential customers.

BRAND ACTIONS ON SOCIAL THAT PROMPT CONSUMERS TO PURCHASE

Brand Actions on Social That Prompt Consumers to Purchase
Q2 2017

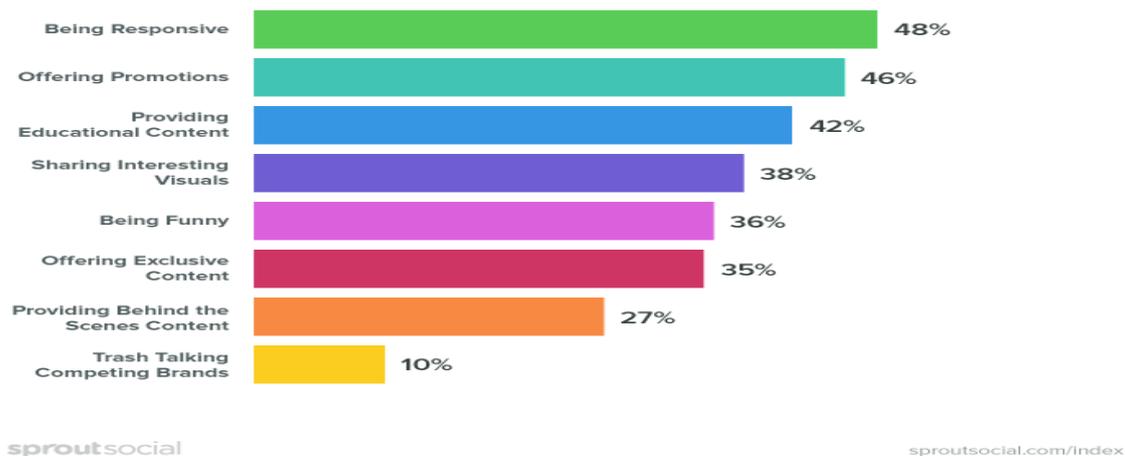


Fig-2

Source: <https://www.google.com/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwik6NaRyoHgAhWOT30KHVvECRgQjRx6BAgBEAU&url=https%3A%2F%2Fsproutsocial.com%2Finsights%2Fdata%2Fq2-2017%2F&psig=AOvVaw0t3H2pu7FA-ra9tuMw8Ukz&ust=1548253019402026>

The social media acts as major place through which marketing could reach out at different extent by targeting a certain group of customers.

BENEFITS OF DIGITAL MARKETING

It may come as no surprise that the marketplace has become increasingly more digital as technology continues to evolve. The benefits of digital marketing are becoming more prevalent every day. More and more consumers are researching and buying products online. According to Forbes, 82% of consumers conduct research online. And, Tech Crunch reports that 79% of people shop online.

So how do you reach these consumers? Here’s the answer: digital marketing.

More and more small businesses are implementing digital marketing tactics to effectively reach and engage their target consumers online. In fact, U.S. digital marketing spend will rise to about \$332 billion by 2021. Digital marketing tactics has proven to be the most cost-effective way to reach potential customers.

WHAT'S AD GROWTH GOING TO LOOK LIKE IN THE DIGITAL ARENA?

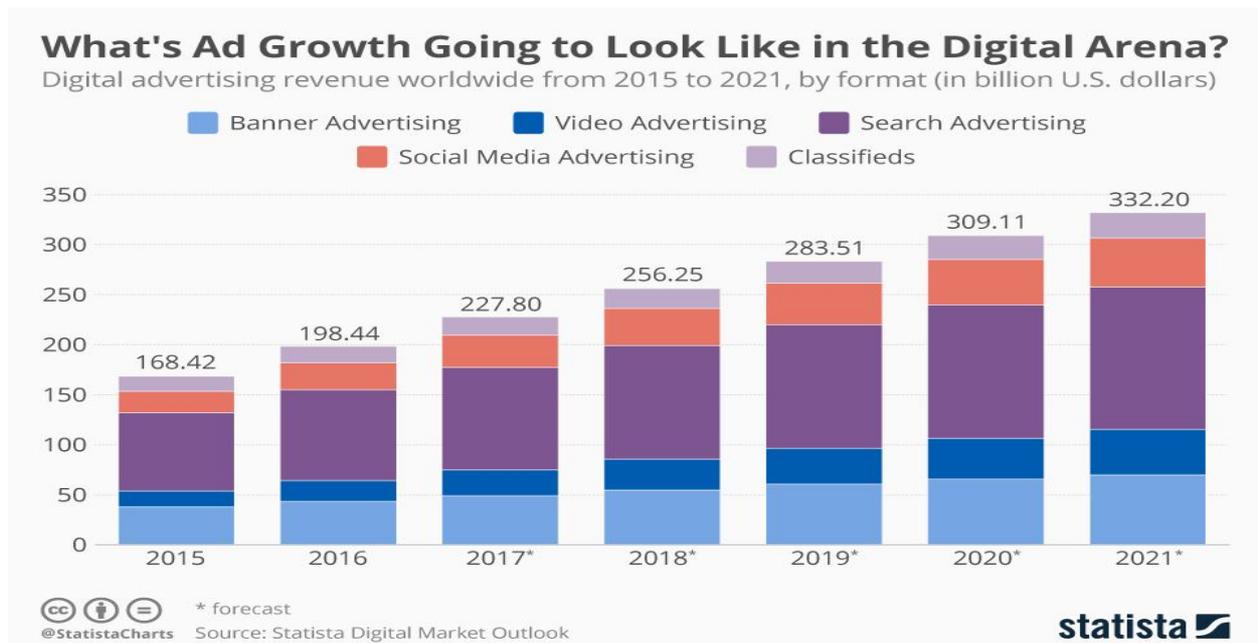


Fig-3

Source: https://www.google.com/url?sa=i&source=images&cd=&ved=&url=https%3A%2F%2Fnmc-mic.ca%2Fnews%2Fresearch%2Fdigital-advertising-canada-continues-grow-data-shows%2Fattachment%2Fstatista_digital_advertising_revenue_worldwide_n%2F&psig=AOvVaw1-wtjU42ZY7dAaGw11RIk4&ust=1548253100995725

➤ **Digital marketing is the most powerful form of marketing**

Digital marketing has the potential to transform the way that you reach and engage your customers.

➤ **Using digital marketing tactics is the most cost-effective way to market your business**

One of the most important benefits of digital marketing is that it is the most cost-effective ways to market your business. When it comes to traditional marketing, it's very difficult for small businesses with limited budgets to compete with larger businesses for ad space.

➤ **Digital marketing is the most measurable form of marketing**

The only way to know for sure is to measure your success over time. While it can be difficult to track the success of a traditional marketing crusade like a radio advertisement or mailer, every digital marketing tactic that you use is measurable. This benefit is every reason why you need to invest into digital marketing.

➤ **One of the greatest benefits of digital marketing is that it allows you to target your ideal buyers**

When you purchase poster space or place an ad in a magazine, you're taking a shot in the dark that this messaging will reach its intended audience. Though there are certainly ways to increase your chances of success with traditional marketing tactics, it just doesn't offer the same targeting skills as digital marketing.

By improving targeting, you can work to get more for your marketing budget and resources. With digital marketing, you no longer have to worry that you are spending money on ads that won't reach those who are likely to be interested in your products. With sophisticated targeting abilities, digital marketing tactics allow you to take comfort in knowing that you are focusing your marketing efforts on plans that actually work.

➤ **Most people are starting their buyer's journey online**

Think about how often you turn to Google or another search engine to find the information you essential. Your customers are no different when they start to research the products or services that can help them solve their biggest problems.

➤ **Your customers are on social media, and digital marketing helps you reach them**

The most important on our list of benefits of digital marketing is this. No matter what industry your business is in, there's a good chance that your buyers are spending their time on social media channels like Facebook or Twitter. According to e - Marketer, 2.34 billion people, or about one-third of the earth's population, uses social media platforms regularly.

➤ SEO and local SEO helps you reach more qualified buyers online

SEO is another powerful digital marketing tactic that offers a variety of benefits. By optimizing your site content for the search engines, you can work to reach more buyers online. Using relevant keywords that help call your merchandise or service offering can help you bring more targeted traffic to your site, which increases conversions over time.

SCOPE OF DIGITAL MARKETING

Hasn't the internet driven all of us crazy? well, it sure has. There was a time when a new serial on the TV used to be the hot topic whereas today, the online posts or a new music video on YouTube grabs our attention. What is this? this is a shift in the choice and preferences. Digital media is gaining mass attention because of the fresh air it has got with itself. It's like living in a new era. We are experiencing a revolution, while we are shifting from the traditional to the Digital media.

- It is flexible
- It is easy
- Eco – friendly
- Fastest Reach
- High engagement
- Influential
- Small investments and big returns
- Measure immediate result

The Indian City, Pune is reported to emerge as a hub for Digital Marketing companies and agencies. Hence, the Digital Marketing scope in Pune is recorded to be relatively high.

The United States has a broad scope of Digital Marketing. The United States is one of the most powerful countries in the world because of its technological advancements. The technological advancements have the digital domain as the most integral component of the entire structure.

When we see Dubai, we know that Dubai has always stood ahead in the technological aspect. the companies in Dubai are going digital and thus there are plenty of job opportunities. The month of August 2017 itself marked for as many as 3650 jobs. Well, that's a good enough number. This is surely indicative that there is a remarkable scope of Digital Marketing in Dubai.

The huge land mass of Canada, loosely populated has marked its global presence with a good number of Digital Marketing job opportunities in Canada. An average of 2275 job vacancies are available in Canada under Digital Marketing domain on a daily basis. The wildfire of Digital Marketing has wrapped this country in its arms as well.

DIGITAL MARKETING TRENDS FOR 2019**1. Artificial Intelligence**

Artificial intelligence will take over the world! Or at least the world's simpler jobs.

AI can analyse consumer behaviour and search patterns, utilizing data from social media platforms and blog posts to help businesses understand how users and customers find their products and services. For example, Facebook messenger bots (which we'll talk more about later) can help you automate and optimize your customer service. Businesses adopting AI in 2K19 will be able to save costs and accelerate growth, getting an edge over their competitors

2. Programmatic Advertising

Programmatic advertising means using AI to automate ad buying so you can target more specific audiences. Real-time bidding or auctions, for example, is a type of programmatic ad buying. This automation is much more efficient and fast, which means higher conversions and lower customer gaining costs.

3. Chatbots

Chatbots will continue to be an important section of digital marketing in 2019.

This AI-based technology uses instant messaging to chat in real-time, day or night, with your customers or site visitors. With 1.4 billion people interacting with chatbots, 80% of savvy businesses are already using or plan to

use chatbots by 2020. And by 2022, chatbots will help businesses save over \$8 billion per annum, especially in the banking and healthcare industries. Many customers prefer interacting with chatbots as they are responsive, give answers promptly, accurately recall your entire buying history, and never lose patience. These virtual assistants offer outstanding customer service, meeting customers' expectations and automating repetitive tasks – which mean you can focus on more important work.

4. Personalization

If you want to stand out in 2019, you need to personalize your marketing – and that means personalized content, products, emails, and more. With the availability of data like purchase history, consumer behaviour and links clicked, custom content has never been easier. In fact, 96% of marketers believe that personalization advances customer relationships. Businesses like Netflix and Amazon are already leveraging the power of personalization. Logging on to your Netflix account, for example, immediately shows you the evidence of this: the banner, carousels, order, artwork, text and search are all personalized for you.

Personalization = Maximize enjoyment + Minimize search time

5. Influencer Marketing

Using influencers is a very effective marketing tool that works to attract customers. Influencers can be anyone from celebrities and Instagram or YouTube stars to well-known bloggers and journalists who help spread the word about your business or product through their social channels.

CONCLUSION

Digital Marketing in International Arena plays a prominent role in the global business as the whole world is running out of time and there is easy availability of internet, results in the tremendous growth of Digital Marketing globally. Artificial Intelligence have been a major space in the creation and development of global market. Digital Marketing is not just an emerging sector but it's the future of the business all over the world. Digitalise or die, because if we does not develop to the changing world we could not cope up with the competitive world.

Digital marketing a unique way of international integration.

REFERENCES

- https://en.wikipedia.org/wiki/Digital_marketing
- <https://steadfastcreative.com/digital-marketing-success-7-ways-to-improve-brand-awareness/>
- <https://www.investopedia.com/terms/d/digital-marketing.asp>
- <https://www.redcrowmarketing.com/digital-marketing/>
- https://www.sas.com/en_us/insights/marketing/digital-marketing.html
- <https://www.singlegrain.com/digital-marketing/11-digital-marketing-trends-you-can-no-longer-ignore-in-2018/>
- <https://www.smartinsights.com/marketplace-analysis/customer-analysis/digital-marketing-statistics-sources/>
- <https://www.smartinsights.com/online-brand-strategy/international-marketing/develop-international-digital-marketing-strategy/>
- <https://www.techmoweb.com/blog/know-the-importance-of-digital-marketing-and-its-benefits>

AN INTRODUCTION OF EDGE DETECTORS FOR BRAIN TUMOR

B. Chitradevi¹, N. Thinaharan² and G. Suganya³

Assistant Professor^{1,2} and Research Scholar³, Thanthai Hans Roever College (Autonomous), Perambalur

ABSTRACT

Edge detection is one of the fundamental approaches in image processing which aim at identifying points in a digital image. Proposed technique is used for computing an optimal threshold value used by adaptive threshold for edge detection. The various edge detection algorithms are compared with the proposed algorithm and their performance are evaluated using the evaluation metrics. From the experimental results, the proposed algorithm was better than the existing methods.

Keywords: Medical image processing, Brain tumor, Edge detection

INTRODUCTION

Magnetic resonance imaging (MRI) is a non-invasive imaging technique which is used in medical applications to produce high quality images of the inside of the human body. Recently, this technique is successfully applied to obtain detailed pictures of organs, soft tissues, bones and virtually all other internal body structures. After MRI image acquisition various post processing algorithms are applied to magnetic resonance images in order to extract more information or enable better visualization of information in magnetic resonance images. Especially, detection of tissue borders is performed, as a process of a great importance while recognition of pathological alternations in MRI images.

In medical image processing, edge detection is a fundamental task that needs to point out the exact edges to get the best results. The aim of this process is to find points in an image where discontinuities or sharp changes in intensity arise. This process is vital to understanding the content of an image and has its applications in image analysis. The edge detection aims to localize the boundaries of objects in an image and it is a basis for many image processing applications. Usual approaches to edge detection are computationally expensive because each set of operations has conducted for each pixel. In usual approaches, the computation time rapidly rise with the size of the image.

EDGE DETECTION METHODS

Edge detection is a fundamental method used in most image processing applications to get information from images as a ancestor step to feature extraction and object segmentation. This process detects boundaries between objects and the background in the image at which the image brightness changes sharply or more formally has discontinuities. The image containing these boundaries is known as edge map. In this paper, the various edge detection techniques are applied for obtaining the fine edges as follows:

- i. Canny edge detector
- ii. Sobel edge detector
- iii. Proposed System

i. Canny Edge Detector

The Canny edge detection operator was developed by John F. Canny in 1986 and uses a multi-stage algorithm to detect a wide range of edges in images. An optimal edge detector is based on the following three criteria:

1. *Good detection:* The algorithm should mark as many real edges in the image as possible.
2. *Good localization:* Marked edges should be close as possible as to the edge in the real scene.
3. *Minimal response:* A given edge in the image should only be marked once, and where possible, image noises should not create false edges [5].

This edge detection method has following in five important steps

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +1 \\ -1 & 0 & +1 \end{bmatrix} \quad G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

- Smoothing: The image is smoothened using the Gaussian filter, to reduce noise.
- Detecting gradients: The edges ought to be checked where the angles of the image have vast extents. The local gradient $G(x,y)$ is calculated by using the following equation,

$$G(x,y) = [G_x^2 + G_y^2]^{1/2}$$

- The edge direction is computed at each point with the values of G_x and G_y by using the following equation,

$$\alpha(x,y) = \tan^{-1}(G_y/G_x)$$

- The strength of edge point is non-maximum suppression. In this method, only local maxima should be marked as edges. The edge points detected in step 2 give rise to ridges in the gradient magnitude image. The technique then follows beside the top of these points as well as sets to zero every pixel which are not really on the ridge top so as to provide a thin line in the output, a process which is known as non maximal suppression.
- Double thresholding: Potential edges are determined by thresholding. The ridge pixels are then thresholded using two thresholds, T_1 T_2 , with $T_1 < T_2$. Ridge pixels with values more important than T_2 are said to be solid edge pixels. Edge pixels with qualities somewhere in the range of T_1 and T_2 are said to be frail edge pixels [6].
- Edge tracking by hysteresis: In the last step the algorithm performs edge linking by incorporating the weak pixels that are 8-connected to the strong pixels. Last edges are controlled by smothering all edges that are not associated with an extremely certain (solid) edge.

$$G_x = \begin{matrix} & \begin{matrix} -1 & 0 & +1 \end{matrix} \\ \begin{matrix} -2 & 0 & +1 \end{matrix} & & \\ \begin{matrix} -1 & 0 & +1 \end{matrix} & & \end{matrix} \quad G_y = \begin{matrix} \begin{matrix} +1 & +2 & +1 \end{matrix} \\ \begin{matrix} 0 & 0 & 0 \end{matrix} \\ \begin{matrix} -1 & -2 & -1 \end{matrix} \end{matrix}$$

The canny edge detector has smoothing effect to remove noise and good localization response. It enhances signal noise ratio as well as immune to noisy Environment but difficult to implement to real time response and also time consuming [7].

ii. Sobel Edge Detector

The Sobel operator is particularly used for edge detection algorithms. Technically, it is a discrete differentiation operator, computing an image intensity function. At each point in the image, the result of the Sobel operator is either the corresponding opposite of the gradient vector or the norm of this vector. The Sobel operator is based on convolving the image with a small, separable, and integer valued filter in horizontal and vertical direction and is therefore relatively inexpensive in terms of computations [8]. On the other hand, the opposite of the gradient approximation that it produces is relatively crude, in particular for high frequency variations in the image. Here, G_x and G_y are masks represented as follows,

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ 2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

$$|G| = \sqrt{G_x^2 + G_y^2}$$

The point of introduction of the edge (with respect to the pixel framework) offering ascend to the spatial angle is given by:

$$\theta = \arctan \left(\frac{G_y}{G_x} \right) - \frac{3\pi}{4}$$

In this system, orientation θ is taken to mean that the direction of maximum contrast from black to white runs from left to right on the image, and other angles are measured anticlockwise from this. Often, this absolute magnitude is the only output the user observes the two components of the gradient are conveniently computed and added in a single pass over the input image [9].

iii. Proposed system

Thresholding is called adaptive threshold when a different type of threshold is used for different regions in the image. This is known as local or dynamic Thresholding. Thresholding assumes that the image have pixel values generally different from the background. In the proposed approach, initially edges are extracted using an adaptive variation threshold. The connectivity of the edges is increased and then obtained using ACO. The threshold is used to segment an image by setting all pixels whose intensity values are above a threshold to a foreground value and all the remaining pixels to a background value. Whereas the conventional threshold operator uses a global threshold for all pixels, adaptive threshold changes the threshold dynamically over the image. This more sophisticated version of threshold can accommodate changing lighting conditions in the image, e.g. those occurring as a result of a strong illumination gradient or shadows. The threshold value is calculated for each pixel in the image. If the pixel value is below the threshold value it is set to the background value, otherwise it assumes the foreground value.

CONCLUSION

Performance Metrics for various edge detection methods

	Sobel	Canny	Proposed Method
SNR	23.032	23.144	24.184
MS	5.8077	7.5106	4.9085
SSIM	0.8625	0.9853	0.9983

This paper has presented comparison of various edge detection algorithms by measuring their performance using evaluation metrics. Here, the adaptive threshold is proposed for edge detection. From the experimental results, it is inferred that the proposed method shows good result when compared to other methods.

REFERENCES

1. N. Senthilkumaran and R. Rajesh," Edge Detection Techniques for Image Segmentation – A Survey of Soft Computing Approaches",International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009 250 .
2. Harelimana Joyeuse, et.al, "Implementation of Edge Detection for a Digital Image", KIST-AO Electronics and telecommunication engineering 2008.
3. Guennadi (Henry) Levkine,"Prewitt, Sobel and Scharr gradient 5x5 convolution matrices", Image Processing Articles, Second draft, June 2012.
4. Wenshuo Gao; Xiaoguang Zhang; Lei Yang; Huizhong Liu, "An improved Sobel edge detection," Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on , vol.5, no 1.
5. Harelimana Joyeuse, et.al, "Implementation of Edge Detection for a Digital Image", KIST-AO Electronics and telecommunication engineering 2008.
6. Vishwadeep Garg et.al.," Image Texture Enhancement through an Improved Grunwald-Letnikov Fractional Differential Mask, International Journal of Advanced Computer Science and Applications, Vol.3, No.3, Dec2011.

AN INTRODUCTORY STUDY ON WEB BASED SUPPORT SYSTEM

V. Rengaraj¹ and V. Thiagarasu²Research Scholar¹, Manonmaniam Sundaranar University, Tirunelveli
Principal², Gopi Arts and Science, Gopichettipalam**ABSTRACT**

Data mining deals with finding patterns in data that are by user-definition, interesting and valid. It is an interdisciplinary area involving databases, machine learning, pattern recognition, statistics, and visualization. The Combination between Web services and software agents provides a promising computing paradigm for efficient service selection and integration of inter-organizational business processes. Data Mining and Decision Support: Integration and Collaboration presents a conceptual framework, plus the methods and tools for integrating the two disciplines and for applying this technology to business problems in a collaborative setting. The decision-making is not only guided by the information provided by DSS but rather than the Web technology, the process is entirely based on communication between ISP Agents and Web agent.

Keywords: Data Mining, Web based System, Decision Support System.

INTRODUCTION

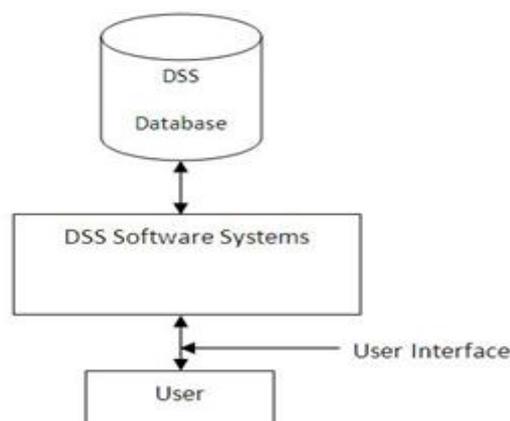
Data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Decision support systems are the core of business IT infrastructures to translate a wealth of business information into tangible and lucrative results. Web technology has provided an effective means of delivering and sharing decision support functionalities. Decision support systems are the core of business IT infrastructures to translate a wealth of business information into tangible and lucrative results. Collecting, maintaining, and analyzing large amounts of data, however, are mammoth tasks that involve significant technical challenges, expenses, and organizational commitment.

Decision Support System: A decision support system (DSS) is an information system that supports business or organizational decision-making activities. DSSs serve the management, operations and planning levels of an organization (usually mid and higher management) and help people make decisions about problems that may be rapidly changing and not easily specified in advance—i.e. unstructured and semi-structured decision problems. Decision support systems can be either fully computerized or human-powered, or a combination of both.

Components of Decision Support Systems (DSS): A decision support system consists of three main components, namely database, software system and user interface.

1. DSS Database: It contains data from various sources, including internal data from the organization, the data generated by different applications, and the external data mined from the Internet, etc. The decision support systems database can be a small database or a standalone system or a huge data warehouse supporting the information needs of an organization. To avoid the interference of decision support system with the working of operational systems, the DSS database usually contains a copy of the production database.

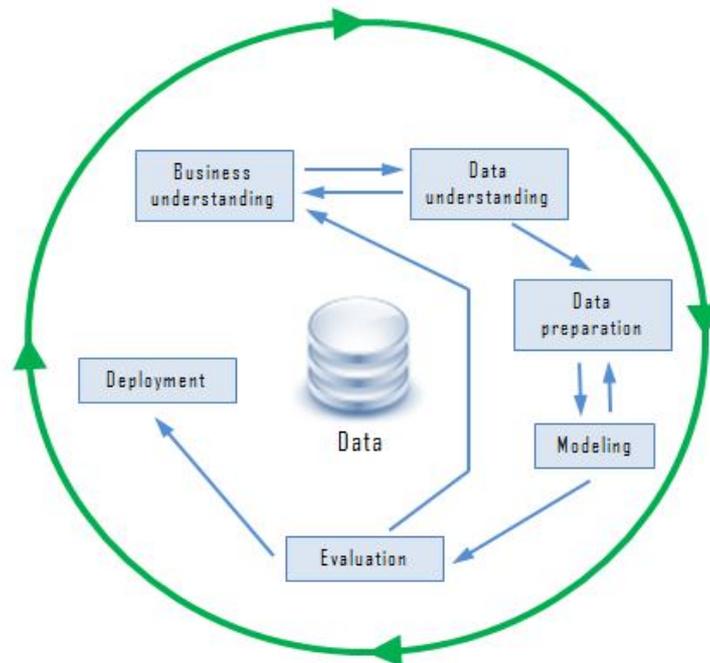
2. DSS Software System: It consists of various mathematical and analytical models that are used to analyze the complex data, thereby producing the required information. A model predicts the output in the basis of different inputs or different conditions, or finds out the combination of conditions and input that is required to produce the desired output.



3. DSS User Interface: It is an interactive graphical interface which makes the interaction easier between the DSS and its users. It displays the results (output) of the analysis in various forms, such as text, table, charts or graphics. The user can select the appropriate option to view the output according to his requirement.

DATA MINING

Data mining is a promising and relatively new technology. Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence(AI) and statistical.



DATA MINING TECHNIQUES AND THEIR APPLICATION

The three main areas of data mining are

- (a) Classification,
- (b) Clustering and
- (c) Association rule mining.

**Classification assigns items to appropriate classes by using the attributes of each item. The k-nearest neighbour (k-NN) method uses a training set, and a new item is placed in the set, whose entries appear most among the k-NNs of the target item. K-NN queries are also used in similarity search, content based image retrieval – (CBIR).

**Clustering methods group items, but unlike classification, the groups are not predefined.

**Association rule mining – (ARM) considers market basket or shopping-cart data, that is, the items purchased on a particular visit to the supermarket. ARM first determines the frequent sets, which have to meet a certain support level.

Market segmentation - Identify the common characteristics of customers.

Customer churn - Predict which customers are likely to leave some x-company and go to a competitor.

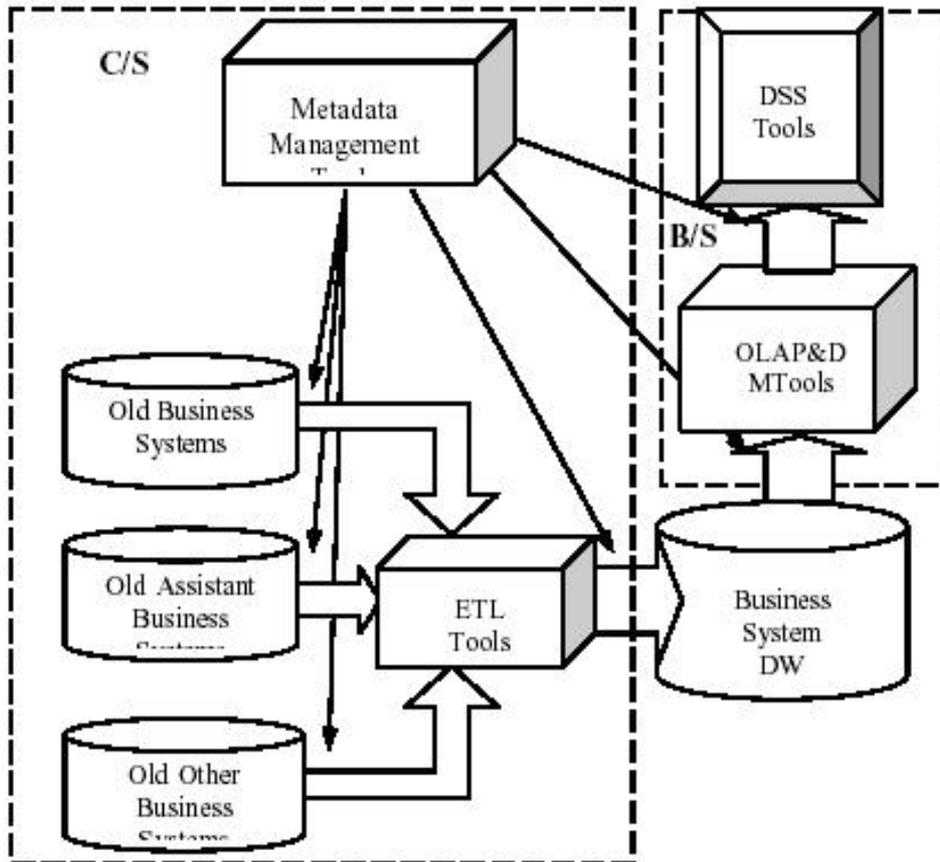
Fraud detection - Identify which transactions are most likely to be fraudulent.

Direct marketing - Identify which prospects should be included in a mailing list to obtain the highest response rate.

Interactive marketing - Predict what each individual accessing a Web site is most likely interested in seeing.

Market basket analysis - Understand what products or services are commonly purchased together; e.g., bread and milk.

Trend analysis - Reveal the difference between a typical customer present and future.



WEB-BASED SUPPORT SYSTEMS

The WSS is a natural evolution of studies on various computerized support systems. An ultimate goal of computer scientists is to build fully automated computer systems that have the same or even a higher level of intelligence as human beings. It is hoped that these systems can replace human beings to perform various activities, either simple or complex.

BENEFITS OF WEB TECHNOLOGY

- The Web provides a distributed infrastructure for information processing.
- The Web delivers timely, secure information and tools with a user friendly interface.
- The Web has no time or geographic restrictions. Users can access systems at any time and any place.

Web used technologies are employed to improve the capacity of decision support systems through decision models, On-line Analysis Processing (OLAP) and data mining tools that allow "standardized" publishing and sharing of decision resources on the Internet.

APPLICATION OF BUSINESS AND ECONOMICS

- Financial Planning
- Risk Analysis
- Supply Chain Planning
- Marketing Plans
- Text and Video Mining
- Handwriting/Speech
- Recognition
- Image and Pattern Recognition
- Long-range Economic Planning
- Homeland Security

CONCLUSION

Data mining in decision support system provides a selection of data analysis, simulation, visualization and modeling techniques, and software tools. Data mining is a powerful new technology with great potential to help companies focus on the most important information in the data they have collected about the behavior of their customers and potential customers. The evolution of the application dimension is the extension of decision support systems to computerized support systems. A Web-based DSS uses the Web as a portal to the underlying DSS. It lets interested users access and make use of the underlying DSS through the Web.

REFERENCES

1. Michael J. Shaw, Chandrasekar Subramaniam, Gek Woo Tan, Michael E. Welge; Decision Support Systems 31 (2001) 127–137: Knowledge management and data mining for marketing.
2. Salvatore T. March, Alan R. Hevner: Integrated Decision Support: A Data Warehousing Perspective.
3. Sundarraj, R.P.: A Web-based AHP approach to standardize the process of managing service-contracts. Decision Support Systems 3, 343--365, (2004).

AN EFFICIENT FUZZY TECHNIQUE INCORPORATED WITH MAMMOGRAM IMAGE ENHANCEMENT**Dr. T. A. Sangeetha¹ and Dr. N. Revathy²**Assistant Professor¹, Department of Computer Science, Kongu Arts and Science College (Autonomous), Erode
Associate Professor², Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore

ABSTRACT

One of the most significant causes of increased women death rate in the world is due to Breast cancer. Mammography is the most effective method for early detection of breast diseases. The main aim of mammography is to detecting, non-palpable cancers during its premature stage. Conversely, mammograms are extremely complex to deduce being the fact that the pathological transformations of the breast are slight and their visibility is very poor with low contrast and noise. Mammograms have the significant information such as microcalcifications and masses, which are extremely complicated to detect because mammograms are of low-contrast. Cancer detection using mammography mainly concentrates on features of tiny microcalcifications, together with the number, size and spatial arrangement of microcalcification clusters and morphological features of individual microcalcifications. In the current scenario, Content-Based Image Retrieval (CBIR) techniques have gained considerable attention for medical images analysis. It is necessary to enhance the mammogram images because the mammogram images are very noisy, low-contrast, blur and fuzzy, for accurate identification and early diagnosis of breast cancer. In this paper, proposed an efficient techniques to enhance the mammogram image using various transforms and fuzzy enhancement. The various transforms are Curvelet transform, Contourlet Transform, Non-subsampled contourlet transform (NSCT). K-means algorithm and Expectation-maximization (EM) algorithm are used for segmentation process. The results proved that the fuzzy enhancement achieves better results in enhancing the mammogram image. The innovative technique of Fuzzy Concept is included in the mammography images which is used to produce the purified images. That images is utilized to take a clear decision in the medical field. Moreover some of the performance factors are discussed in this research work and proofed with the efficiency.

Keywords: Breast Cancer, Mammography, Image Enhancement, K-means algorithm, EMz algorithm.

INTRODUCTION

BREAST cancer is one of the main reason for death among women. Early detection and treatment are regarded as the most suggested approach to reduce breast cancer mortality [1]. Breast cancer is a type of cancer caused by breast tissue, which occurs mostly in the inner lining of milk ducts or the lobules that supply the ducts with milk [2]. Cancers originates from ducts are known as ductal carcinomas, while those originates from lobules are known as lobular carcinomas. Breast cancer usually occurs in humans and it occurs in other mammals too. Majority of human cases occur in women, male breast cancer can also occur [3].

A. Mammogram

Mammogram is a medical test that uses x-rays to take pictures of the internal structure of the breast. The testing is also known as "mammography." A mammogram is a radiograph of the breast tissue. It is an effective method of investigating the breast, typically for the diagnosis of breast cancer [4]. Mammography is a radiographic examination of the breast and the most significant investigation to identify early stages of breast cancer.

But the digital mammography is very noisy, low-contrast, blur and fuzzy and hence there is a requirement for enhancing images [5]. This is essential for enhancing the Peak Signal-to-Noise Ratio (PSNR) and reducing the Mean Squared Error (MSE) for accurate identification.

Mammograms are done for two reasons

- i. **Screening:** When women participate on a routine basis to have mammograms done to find breast cancer at an premature stage. This type of mammogram looks for the indications that breast cancer may be spreading, even though no symptoms are there.
- ii. **Diagnostic:** This is typically done to check for breast cancer after a lump or any other sign/symptom has been found such as pain, nipple discharge, skin thickening, or a change in breast size or shape. It will be used as a second test if a screening mammogram finds something that is not normal.

One of the most important objectives of mammogram image enhancement is to enhance the contrast between regions of interest and the background. Also the medical images fluctuate extensively in terms of acquisition, noise characteristics and quality [6]. Thus, there is a requirement to process an image on the image basis. This

motivates the design and construction of effective mammogram image enhancement techniques using various transforms and fuzzy enhancement method.

The paper is organized as follows. Section II describes the related works, Section III deals with various methodology used for image enhancement and Section IV describes about the experimental results and Section V represents the conclusion of the proposed method.

RELATED WORK

Digital mammography is one of the most suitable methods for early detection of breast cancer. However, the visual clues are faint and vary in appearance which makes diagnosis difficult and challenging [7]. There is a significant requirement for developing methods for automatic classification of irregular areas in mammograms for aiding radiologists to improve the efficiency of screening programs and avoid unnecessary biopsies. Micro calcifications occurs in mammogram image as small localized granular points with high brightness. It cannot be detected easily by naked eye because of its miniaturized dimension. Due to its small size, about 10-40% of microcalcification clusters are missed by radiologists[8][9].

Curvelet was developed by Candes and Donoho, for providing effective indication of smooth objects with discontinuities along curves. Detecting and enhancing the boundaries between different structures is important in image processing, especially in the field of medical imaging. In many important applications, images exhibit edge discontinuities across curves. Some studies have been presented with curvelet in image processing. Dettori and Semler presented a comparative study between ridgelet, wavelet and curvelet transform on some computed tomography (CT) scans. The relative study indicated that curvelet yields better results than wavelet or ridgelet.

METHODOLOGY

The proposed method uses various transforms and fuzzy enhancement for mammogram image. K-means algorithm is used for image segmentation.

Mammogram image enhancement

Mammography is a specific kind of imaging that utilizes a low-dose x-ray system to check breasts. A mammography exam is called as mammogram, used to assist in the premature detection and early diagnosis of breast cancer and related diseases in women. An x-ray is noninvasive medical tests that assist physicians in diagnosing the disease. Imaging with x-rays involves exposing a part of the body to a tiny amount of ionizing radiation to generate pictures of the inside of the body. X-rays are the traditional and most commonly used form of medical imaging.

Fuzzy Enhancement

In image denoising, detecting major image details and change the degree of noise smoothing appropriately. Noise is uncorrelated in the spatial domain and the wavelet domain. Fuzzy feature for single channel image denoising is used to enhance image information in wavelet sub-bands and then using a fuzzy membership function to shrink wavelet coefficients, correspondingly. This fuzzy feature space helps to distinguish between important coefficients, which belongs to image discontinuity and noisy coefficient.

Large weights should be given to neighboring coefficients with similar magnitude, and a small weight is given to neighboring coefficients with dissimilar magnitude. The larger coefficients are produced by noise which becomes always isolated or unconnected, but their edge coefficients are clustered and persistent. The adjacent points are more similar in magnitude. Hence a fuzzy function $m(l, k)$ of magnitude similarity and a fuzzy function $s(l, k)$ of spatial similarity [25] is defined as follows:

$$m(l, k) = \exp \left[- \left(\frac{y_{s,d}(l, j) - y_{s,d}(l + l, j + k)}{Thr} \right)^2 \right]$$

$$s(l, k) = \exp \left(- \left(\frac{l^2 + k^2}{n} \right) \right)$$

where $y_{s,d}(l, j)$ and $y_{s,d}(l + l, j + k)$ are central coefficient and neighbor coefficients in the wavelet sub-bands respectively. $Thr = c \times \hat{\sigma}_n$, $3 \leq c \leq 4$, $\hat{\sigma}_n$ is estimated noise variance, and N is the number of coefficients in the local window $k \in [-K \dots K]$, and $l \in [-L \dots L]$.

Based on the fuzzy functions, adaptive weight $w(l, k)$ for each neighboring coefficient is determined as follows:

$$w(l, k) = m(l, k) * s(l, k)$$

Using the adaptive weights $w(l, k)$, the fuzzy feature for each coefficient in the wavelet sub-bands are obtained as follows:

$$f(i, j) = \frac{\sum_{l=-L}^L \sum_{k=-K}^K w(l, k) * |v_{i,j}(i+l, j+k)|}{\sum_{l=-L}^L \sum_{k=-K}^K w(l, k)}$$

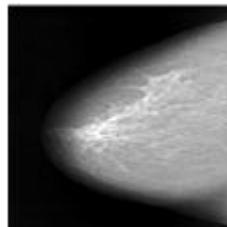
Fuzzy feature can distinguish well between edge structure and noise when compared to the local mean.

Experimental Results for Fuzzy enhancement

Besides, the quality of the images are evaluated using the traditional distortion measurements such as

- Mean Squared Error (MSE)
- Peak Signal-to-Noise Ratio (PSNR)

For these images the mean square error and peak signal to noise ratio is calculated to find the better transformation. The input mammographic denoised image is shown below



Input Denoised image

To calculate the mean square error and peak signal to noise ratio, the following formula is used.

Mean Squared Error

MSE of the output image is defined as

$$MSE = \frac{\sum_{i=1}^M \sum_{j=1}^N |x(i, j) - \hat{x}(i, j)|^2}{MN}$$

where $x(i, j)$ is the original image, $\hat{x}(i, j)$ is the output image, and MN is the size of the image.

Peak Signal to Noise Ratio (PSNR)

PSNR is defined as

$$PSNR = 20 \text{Log}_{10} \left[\frac{(2^n - 1)}{\sqrt{MSE}} \right] \text{ (dB)}$$

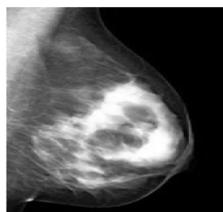
where n is the number of 8bits/pixel used in representing the pixel of the image.

Breast cancer data sets are incorporated to Fuzzy Enhancement techniques. Hence the Fuzzy Enhancement technique is compared with Non-Fuzzy Enhancement technique to produce the comparative result values.

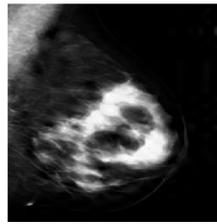
Fuzzy, Non- Enhancement and their PSNR and MSE values are calculated and tabulated as given below.

Image Enhancement Techniques	Database
	WDBC
Non Fuzzy Enhancement	1.958480
Fuzzy Enhancement	1.763289

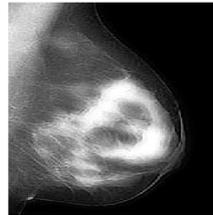
Enhancement Techniques	Database(dB)
	WDBC
Non Fuzzy Enhancement	46.360221
Fuzzy Enhancement	47.167234



(a) Mammogram Image



(b) Non-Fuzzy Enhancement



(c) Fuzzy Enhancement

OVERALL PERFORMANCE RANKING

Methodology used	Rank
Enhanced image using various transforms	6
Mammogram image enhancement using Fuzzy Enhancement	5

Thus the mammogram image enhancement using fuzzy enhancement achieves better results since its PSNR value is high with very low MSE value and it is used for efficient image enhancement.

The resultant images are taken for opinion from Dr. Anandhakumar, Radiologist in Erode Town has stated that, “By looking the results of the given sample images, it gives its view on various techniques used in this research. The techniques for processing the images shows visual perception of the doubtful area (i.e.,) whether it is cancerous or Non-cancerous region. A normal person cannot know about the processing of images as this can be known by the radiologists perception. The final transform for enhancing images shows efficient and clear way of detection. It is suggested that the images finally concluded are said to be reliable.”

CONCLUSION AND SCOPE

At present, one of the most important causes of cancer death among middle aged women is breast cancer. Mammography is technique used by radiologists for early detection and diagnosis of cancer in breast images. Digital mammogram has turned out to be the most effective technique for premature breast cancer detection. Digital mammogram captures an electronic image of the breast and accumulates it in a computer. Mammogram images are very noisy, low-contrast, blur and fuzzy and hence the mammogram images are enhanced for accurate identification of breast cancer. Processing these images require high computational abilities. In this paper, the mammogram images are enhanced using various transforms and fuzzy enhancement. K-means algorithm and EM algorithms used for efficient image enhancement process. It is found from the experiments that Fuzzy image enhancement method using EM algorithm is efficient and useful in capturing relevant clinical information since its PSNR value is high with very low MSE value. As a future work , various effective transforms with efficient algorithms are used to detect the mammogram images accurately.

REFERENCES

1. H.D.Cheng, X. Cai, X. Chen, L. Hu, X. Lou, “Computer aided detection and classification of micro calcification in mammograms: a survey”, Pattern recognition vol. 36,pp. 2967-2991, 2003.
2. D.NarainPonraj, M.Evangelin Jenifer, P. Poongodi, J.Samuel Manoharan, “A Survey on the Preprocessing Techniques of Mammogram for the Detection of Breast Cancer”, Journal of Emerging Trends in Computing and Information Sciences, vol. 2, no. 12, pp 656-664, December 2011.
3. http://en.wikipedia.org/wiki/Breast_cancer.
4. T.S.Subashini, V. Ramalingam and S. Palanivel, “Pectoral Muscle removal and Detection of masses in Digital Mammogram using CCL”, International Journal of Computer Applications, vol. 1, no. 6, pp. 66-70, 2010.
5. Keir Bovis and Sameer Singh, “Enhancement Technique Evaluation using Quantitative Measures on Digital Mammograms”, Proc. 5th International Workshop on Digital Mammography, Toronto, Canada, M.J. Yaffe (ed.), Medical Physics Publishing, pp. 547-553, 2000.

6. R. Krishnamoorthy, N. Amudhavalli and M.K. Sivakkolunthu, "An Adaptive Mammographic Image Enhancement in Orthogonal Polynomials Domain", *International Journal of Computer and Information Engineering*, vol. 4, no. 2, pp. 120-128, 2010.
7. B. Verma, P. Zhang, "A novel neural-genetic algorithm to find the most significant combination of features in digital mammograms", *Applied Soft Computing* vol. 7, pp. 612-625. 2007.
8. R.G. Bird, T.W. Wallace, B.C. Yankaskas, "Analysis of cancers missed at screening mammography", *Radiology*, vol. 184, pp. 613-617,1992.
9. H. Burhenne, L. Burhenne, F. Goldberg, T. Hislop, A.J. Worth, "Interval breast cancers in the screening mammography program of British Columbia: Analysis and classification," vol. 162, pp.1067–1071,1994.
10. J. K. Romberg, M. B. Wakin, and R. G. Baraniuk, "Multi-scale geometric image processing", In *Proceedings of the SPIE: Visual Communications and Image Processing*, pp. 1265-1272, 2003.
11. M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Trans. Image Process*, vol. 14, pp. 2091-2106, 2005.
12. D.L. Donoho, I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage", *J. Amer. Statist. Assoc.* vol. 90 , no.432,1200–1224, 1995.

SURVEY ON MACHINE LEARNING WITH BIG DATA SOFTWARE DEVELOPMENT

Dr. M. Thangamani, K. PavithKumar, T. PreethiKongu Engineering College, Tamilnadu

ABSTRACT

Huge gatherings of data are valuable resource from which potentially new and useful knowledge can be discovered through data mining. Data mining is an increasingly admired field in visualization, machine learning and other data operation and knowledge mining techniques aimed at gain an insight into the relationships and patterns hidden in the data. Machine learning is concerned with the design and development of algorithms and techniques that allow computers to "learn". Machine learning has a wide spectrum of applications. Machine learning made decision based on past history. Every day social media generates millions of millions data. The increase in data volume causes difficulties in extortion useful information for decision support. These types of data cannot access through traditional tools. This survey provides new ideas about machine learning with Big Data.

1. INTRODUCTION**A. Machine Learning**

Machine learning is used in different domains [1]. All machine-learning approaches lie on a spectrum based on the ease of interpreting their results. For example, data classification trees produce rules that classify data. By reading those rules, you can easily understand how a classification tree classifies data. The security heuristics that distills attack patterns to protect, for instance, ports or networks. In image analysis machine learning is used to identify distinct forms and shapes, such as for medical analyses or face and fingerprint recognition. Deep learning to generate rules for data analytics and big data handling, such as are used in marketing and sales promotions. The object recognition and predictions from combined video streams and multi sensor fusion for autonomous driving and pattern recognition to analyze code for weaknesses such as criticality and code smells are machine learning intelligence

B. Learning Algorithm

Supervised learning means the algorithm generates a function that maps inputs to desired outputs. One standard formulation of the supervised learning task is the classification problem: the learner is required to learn the behavior of a function which maps a vector into one of several classes by looking at several input-output examples of the function. Unsupervised learning is an agent which models a set of inputs: labeled examples are not available. Semi-supervised learning, which combines both labeled and unlabeled examples to generate an appropriate function or classifier. Reinforcement learning is an algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm [2, 3].

In supervised learning a set of input pattern and the target output are presented to the network repeatedly till the difference between the target output and the actual output of the network reaches a certain predetermined value. During the training process the difference between the actual output and the target output is compared. This difference is used to adjust the connection weights to the neurons in such away that the output of the network matches closely to the target output. Perception uses supervised training. In unsupervised training output pattern for given input patterns is not required. The neural networks construct internal models that capture regularities in input pattern. The process of training the network consists of letting it discover salient features of the training set and using these features to group the inputs in to classes that it finds distinct. Difference between supervised and unsupervised learning depends on whether the learning algorithm uses pattern-class information. Supervised learning assumes the availability of a teacher or supervisor who classifies the training examples into classes, where as unsupervised learning does not. Unsupervised learning must identify the pattern class information as a part of the learning process.

3. RELATED WORK

When using machine-learning techniques, the big challenges include the identification of suitable attributes to represent learning style and the lack of prior data about the student. Building an accurate model of learning style is dependent on using appropriate features. Substantial information about the student's behaviour can be collected, but only some features are indicative of learning style. Predominantly the information comes from the navigation profile, as students will move frequently to and spend more time on resources. Declan Kelly et.al. planned [4] a novel adaptive educational system that dynamically determines learning style using machine learning techniques. It describes how it uses the Felder & Solomon Index of Learning Style to design an

environment for different learning styles. It also describes how, as the student interacts with the learning environment, it uses the Naïve Bayes algorithm to predict the student's preferred learning style and adaptively customize the learning environment.

Huge mining have various task. Data preparation, all activities aimed to prepare data for analytics. Data representation, how data are represented and representation choices for each analysis process. Data analytics, the analytics to be computed. Data processing, how data are routed and parallelized. Data visualization and reporting, an abstract representation of how the results of analytics are organized for display and reporting. Many companies and business not have IT expertise and sufficient budget to have benefits from them. In order to fill this break, a model-based representation for Big Data Analytics-as-a-service can be used. Domenico Redavid [5] model, composed by declarative, procedural and deployment models, can be used to select a deployable set of services based on a set of user preferences shaping a Big Data Campaign. The deployment of a BDC requires that the selection of services has to be carried out on the basis of coherent and non conflictual user preferences. The ontology can be exploited to enhance the selection and composition operations on a given set of annotated web services.

Big data analytics are compute-intensive and offer unique scalability disputes to existing homomorphic computing method [6, 7, 8] which can initiate too much overhead. Existing techniques cannot adequately protect big data analytics from faults or malicious tampering. It prevalent approaches contain isolation, hardening of software and hardware, and active monitoring. However, computing infrastructure such as cloud computing resources are generally built to serve many applications, so isolation is not feasible. Additionally, it is unrealistic to eliminate every software and hardware bug and security vulnerability for any reasonably complex real world systems even with the best possible engineering practices. In many fields, much useful patterns can be extracted by applying machine learning techniques to big data. Jian yin [9], address the problem of data confidentiality in big data analytics. In many fields, much useful patterns can be extracted by applying machine learning techniques to big data. They also presented a scheme to provide provable secure data confidentiality and discuss various techniques to optimize performance of such a system. The model provide a "big picture" of the Requirements Engineering work products created and used in Big Data software development projects [10].

Justin McHugh et.al [11] proposed the Semantics Toolkit, a framework that enables access to polyglot persistent Big Data stores while giving the appearance that all data is fully captured within a knowledge graph. It allows data to be stored across multiple storage platforms. Big Data stores such as Hadoop, graph databases, and semantic triple stores with the best-suited platform adopted for each data type, while maintaining a single logical interface and point of access, thereby giving users a knowledge-driven veneer across their data.

Chenn-Jung Huang[12] proposed an application of machine learning techniques to web-based learning diagnosis system that supports a Web-based thematic learning model, enhance the learners' ability of knowledge integration by giving the learners the opportunities to select the learning topics that they are interested, and gain knowledge on the specific topics by surfing on the Internet to search related learning courseware and discussing what they have learned with their colleagues. Based on the log files that record the learners' past online learning behavior, an intelligent diagnosis system is used to give appropriate learning guidance to assist the learners in improving their study behaviors and grade online class participation for the instructor. The achievement of the learners' final reports can also be predicted by the diagnosis system accurately.

Yi wan [13] investigated, the new and more people realize fatigue crack propagation rate curve is more basic curve than S-N curve to describe material fatigue specialty in structure reliability analysis. Establishing fatigue crack propagation rate is the key to forecasting structure fatigue lifetime and fatigue crack propagation rate is usually calculated by some estimating model, In the paper, we presented optimal common machine learning algorithm as least squares support vector machine-LSSVM method for fatigue crack propagation rate forecast.

Semi-supervised learning has attracted a significant amount of attention in pattern recognition and machine learning. Most previous studies have focused on designing special algorithms to effectively exploit the unlabeled data in conjunction with labeled data. They [14] build a model to improve the classification accuracy of any given supervised learning algorithm by using the available unlabeled

Now a day clinical data mining is recent research fields which utilize data mining and machine learning capabilities for revealing the biological patterns. Furthermore oncogenomics research domain endeavors to identifying and analyzing cancer related genes and thus helps in diagnosis at genotype level [15]. Various biological and computational techniques have been used by scientists to early detect cancer type. Collection of

large cancer data repositories has hiked the research in this domain. Various machine learning approaches have been used to predict if tumor is malignant or not.

4. CONCLUSION

This paper provides survey about machine learning with big data. In future, this approach can be extended to implement classifier in big data analytics.

REFERENCE

1. Panos Louridas and Christof Ebert, (2016), "Machine learning", IEEE software, pp.110-115.
2. Jyothi Bellary, BhargaviPeyakunta and SekharKonetigari, (2010), "Hybrid Machine Learning Approach In Data Mining", Second International Conference on Machine Learning and Computing, pp.305-308..
3. Jiang Hua,"Study on the Application of Rough Sets Theory in Machine Learning", (2008), Second International Symposium on Intelligent Information Technology Application, pp.193-196, IEEE.
4. Declan Kelly1 & Brendan Tangney, (2005), "First Aid for You': Getting to know your Learning Style using MachineLearning", Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05), IEEE.
5. DomenicoRedavid, Roberto Corizzo and DonatoMalerba, (2018), "An OWL Ontology for supporting Semantic Services in Big Data platforms", IEEE International Congress on Big Data, pp.228-231.
6. Brakerski, Zvika, Craig Gentry, and Vinod Vaikuntanathan, (2014), "Fully Homomorphic Encryption without Bootstrapping." ACM Transactions on Computation Theory (TOCT) , Vol.6, No. 3 , pp.13.
7. Craig Gentry, (2009), "Fully Homomorphic Encryption Using Ideal Lattices", 41st ACM Symposium on Theory of Computing (STOC).
8. Halevi, Shai, and Victor Shoup, (2014), "Algorithms in helib." In Advances in Cryptology-CRYPTO 2014, pp. 554-571. Springer Berlin Heidelberg.
9. Jianyin, Dongfungzhao, (2015),"Data confidentiality challenges in big data applications", IEEE International conference on big data, pp.2886-2888.
10. Darlan Arruda,Nazim H. Madhavji, (2017), "Towards a Big Data Requirements Engineering Artifact Modelin the Context of Big Data Software Development Projects", IEEE International Conference on Big Data ,4725-4726 .
11. Justin McHugh, Paul E. Cuddihy, Jenny Weisenberg, Williams, Kareem S. Aggour, Vijay S.Kumar, (2017), "Integrated Access to Big Data Polystores through a Knowledge-driven Framework", IEEE International Conference on Big Data (BIGDATA), pp.1494-1503.
12. Chenn-Jung Huang, Ming-Chou Liu, San-Shine Chu, Chin-Lun Cheng, (2014) "Application of Machine Learning Techniques to Web-Based Intelligent Learning Diagnosis System", Proceedings of the Fourth International Conference on Hybrid Intelligent Systems (HIS'04), IEEE.
13. Yi WAN, Chengwen WU, (2009), "Fitting and prediction for crack propagation rate based on machinelearning optimal algorithm", International Conference on E-Learning, E-Business, Enterprise Information Systems, and E-Government, pp.94-96.
14. Pavan Kumar Mallapragada, Rong Jin, Anil K. Jain, and Yi Liu, (2009), " SemiBoost:Boosting for Semi-Supervised Learning", IEEE transactions on pattern analysis and machine intelligence, Vol. 31, No. 11, pp.2000-2014.
15. Beibei Zhu, Lei Yang, Xiaoyu Wu and TianchuGuo, (2015), "Automatic Recognition of Books Based on Machine Learning", 3rd International Symposium on Computational and Business Intelligence, pp.74-78.

E-COMMERCE

Subha S¹, Blessy Ebanazar. D² and Carolin Joyce Paiva³

Assistant professor¹, Department of Commerce, Sri Krishna Arts and Science College, Coimbatore
^{1,2}Department of Commerce, Sri Krishna Arts and Science College, Coimbatore

ABSTRACT

E-Commerce or Electronics Commerce is a way of modern business, which understands the need of business organizations, manufacturers and customers to reduce cost and improve the quality of goods and services while increasing the delivery on time. Ecommerce refers to the paperless exchange of business information. In today's world everything is available online. From books and newspapers we read to the daily things we use everything is available on a single click. This is of all because of development of e-commerce. It is e-commerce which have made buying simple. Just a simple click it brings world to our doorstep. It not only increases the number of buyers but also enables the manufacturers to produce quality goods. It made everything very simple. There are different types of e-commerce. The impacts of e-commerce can not be said in short listed. Now it plays a major role in the life of every people in the world.

Keywords: Mobile commerce, Outline transaction, Remarket, Seasonal demand, Standard of living.

INTRODUCTION

E- Commerce or electronic commerce is platform where buying and selling of goods or services takes place over online. It removes the gap between the buyers and sellers. E-commerce is a very easiest and most convenient way of shopping. Where in a single click it brings world to its doorstep. Customers need not go to shop physically instead they can access through their computers, mobile phone and through TV and could save their time. They even don't need to use cash physically .they can pay online itself. This type of convenience has increased the number of buyers easily. You can shop anything, anywhere and at anytime through e commerce.

HISTORY OF E-COMMERCE

Rome was not built in a day in a same way e-commerce did not develop just like that it has a huge and vast history for its transformation. From barter system to online business it might taken so many years and a great evolution to take place. The development of e-commerce was because of existence of computers, technologies and internet. In 1991 when internet was opened for commercial use many business showed their identity over internet. In 2000 lots of companies in from United States and Western Europe countries started using internet to develop their business. But it did not give up any results so many of the companies disappeared. It was a company called brick and mortar who used the e-commerce in a useful and successful way. Now the online vendors knew the usage of e-commerce and started using it very well. They even gave advertisements in various search engines with low cost. This made small and emerging vendors to use e-commerce in effective way. Amazon and eBay were the first online companies to give online payment. The most frequently sold products in internet are music, books, computers, and office supplies. The most popular online shopping website is Amazon was founded by Jeff bezos.and it has all categories of products selling.

EVOLUTION OF E-COMMERCE

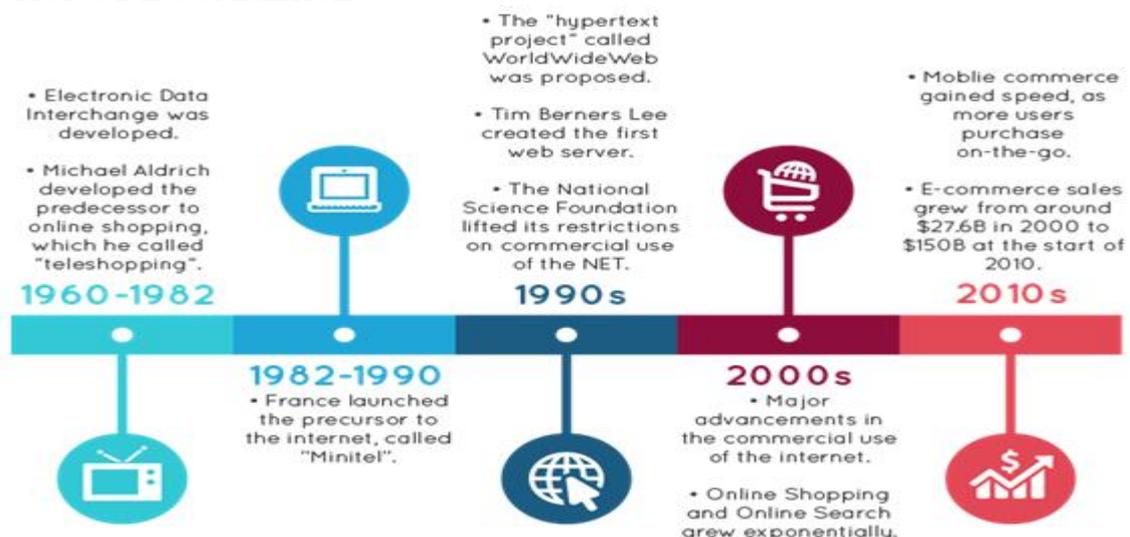


Fig-1.1

https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=2ahUKEwj3w6O03YHgAhUdk3AKHeC_BVkJhx6BAGBEAM&url=https%3A%2F%2Fwww.cheqsystems.com%2Findustries%2Fcommerce%2F&psig=AOvVaw1LIt2R2A9uDeB9sqr4yvo0&ust=1548258199551594

Types of E - Commerce

Basically there are six types of e- commerce and they are given below

1. Business to business (b2b)
2. Business to consumer (b2c)
3. consumer to consumer (c2c)
4. Consumer to business (c2b)
5. Business to administration (b2a)
6. Consumer to administration (c2a)

1. Business to business (b2b)

This type of business is where selling of goods takes place within the companies i.e. it sells goods from one company to other without giving it to public. For example:- Skype from Microsoft company is b2b service where Skype could be accessed in any mobiles, computers or TV etc.

2. Business to consumer (b2c)

In this type of business anybody can have access through i.e. anybody can buy goods and services through online. It offers goods and services directly to customers this is known as business to consumer. For example:- Amazon is a online business where it offers great variety of products and anybody can buy it.

3. Consumer to consumer (c2c)

In this case the business here is the goods is sold by consumer to consumer through third parties.

For example:- eBay is good example where the one consumers sells the used product to other consumers. It is a way of second hand buying and selling.

4. Consumer to business (c2b)

Here it is opposite way of doing business where here business does not provide goods and services instead customers will do this to business. For example: - when you buy a product and give review about it. It helps other people to buy this product

5. Business to administration (b2a)

This type of business engages in the activity or transactions between businesses to administration. These services are developed in connection with e- government.

This includes employment, social security, legal documents etc..

6. Consumer to administration (c2a)

Here it simply refers an electronic transaction between individual customers and administration. For example:- filing tax returns, making online payments etc..

Features of E-Commerce

An E-commerce websites is essential for running an online business, however, it is equally important to have one with all the important features to assure success. Some of the features are:-

FEATURES OF E-COMMERCE



Fig-1.2

<https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjQnYbO3oHgAhUPX30KHAlpAzUQjhx6BAgBEAM&url=https%3A%2F%2Fskytechicon.com%2Fecommerce-images-editing%2F&psig=AOvVaw37WW1bZcxakhp9hZqkiAmj&ust=1548258465913575>

➤ **Payment Gateway Integrations**

A good e-commerce websites gives you an option of integrating with diverse payment gateways by not limiting your choices to a selected few.

➤ **Shopping cart**

This is an integral part of any e-commerce store, the shopping cart. This is where your end consumers store their products to continue with the checkout process.

➤ **Communication**

Provision to send and receive timely notifications regarding your orders are available on the panel. This way a merchant can keep his customers informed about the status of the order and receive the same via the system. Examples: EDI

➤ **Relationship marketing**

E-commerce builds long term mutually satisfying relations. This leads to life time loyal customers.

➤ **Online selling**

E-commerce uses online selling. It has revolutionized selling through E-tailing specially for:

- Shares and financial services
- Books and music
- Airlines tickets and hotel bookings etc.,

BENEFITS OF E-COMMERCE

❖ **Low financial cost**

- One of the ecommerce benefits is that it has a lower start up cost. They have several upfront costs such as store signs, store design, buying inventory etc.

❖ **Potential income**

- Online stores are always open for business. For a customer to order at night, you don't need to have employees working at night shift to ensure all orders get processed.

❖ **Sell internationally**

- A new brand can sell to customers around the world easily. We may have the ability to discover your new audiences all over the world.

❖ **Easy to showcase best-sellers**

- E-commerce benefits like being able to easily display best-sellers makes it easier to show off products to customers. The reason why the customers want to buy your best sellers is because they are proven.

❖ **Personalized online experiences**

- Website personalization, one of the online business advantages, can enhance the online shopping experiences.

❖ **Remarket to customer**

- It's easy to create retargeting ads to retarget customers in your area when running an online business making it one of the most profitable E-commerce benefits.

❖ **Keep eye on consumers buying habit**

- Through e-commerce retailers can easily keep eye on consumers buying interest and habits to provide them their expected product.

❖ **Stay open 24*7/365**

- The most important benefits that e-commerce merchants can enjoy is store timings are now 24*7/365 as they run e-commerce websites all the time.

BENEFITS OF E-COMMERCE



Fig-1.3

<https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwi72ev53YHgAhVIknAKHRmIBTUQjhx6BAgBEAM&url=https%3A%2F%2Fwww.360psg.com%2Fblog%2Fhow-e-commerce-offers-impact-user-experience&psig=AOvVaw34v6njEL5rikouW25bQMsk&ust=1548258177704901>

IMPACTS OF E-COMMERCE

The impact of E-commerce on business is immense and cannot be exhaustively elucidated in short article like this. The impact is expected to increase as internet penetration in emerging markets increases.

- a) Marketing strategy of differentiation increasing effective.
- b) Product life cycles are shortened
- c) Greater use of digital promotion
- d) Workforce planning to support highly seasonal demand.
- e) Increased use of automation in E-commerce
- f) E-commerce likely to involve greater use of multi-currency transactions.
- g) It is relatively easy for smaller firms to sell online.
- h) Economics of scale are becoming increasingly important.
- i) Need for employees to have a broader range to digital skills.
- j) Concerns over the working conditions of staff working in e-commerce warehouse.

EXAMPLES OF E-COMMERCE:

❖ **Amazon**

Amazon is one of the famous e-commerce example.

It was started in Seattle Washington U.S.A and was founded in 1994 by Jeff Bezos.

It was just American company to sell goods online.

He was entitled as person of the years for the company’s success.

Although the company’s headquarters is at U.S.A, it has separate websites in other economically developed countries like U.K, Japan, France etc.

It also has retailers websites dealing for popular brands. According to the reports in 2008 it attracts 615 million customers every year.

It is because of the development of e-commerce.

❖ **Make my trip**

Indian online travel agency was founded in the year 2000 by Deep Kalar.

It has its headquarters in Gurugram, Haryana.

The speciality of this website is it give online travel services like air tickets for within India and international vacations packages, resort reservation, railways and bus tickets.

Nearly 7 million mobile users Make My Trip websites or applications.

It has got the world travel awards in the year 2013 and 2014.

In 2016 Make My Trip and IBIBO groups merged together.

Thus these developments were happened due to the mass development of e-commerce.

EXAMPLES FOR E-COMMERCE



Fig-1.4

https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwjKsLz47oHgAhUIX30KHYY1yD_0Qjhx6BAGBEAM&url=https%3A%2F%2Fwww.goodworklabs.com%2Ftop-five-e-commerce-websites-in-india%2F&psig=AOvVaw2iK2W5KD4ArI8DbjXV-eLy&ust=1548262801307058

CONCLUSION

E-commerce is not only buying and selling of goods over internet but it has also increased our standard of living of people. Thanks to development of internet without it we wouldn't have got accessed to different types of goods around the world. it is not only beneficiary to customers but also to the manufacturers and emerging vendors around the world. So, therefore I conclude that the future of e-commerce will lead to globalization of business and develop the economy highly with the advanced technology.

REFERENCES

1. <https://www.disruptiveadvertising.com/business/business-to-business-examples/>
2. https://www.ecommerce-land.com/history_ecommerce.html
3. <https://emtv.com.pg/six-different-types-of-e-commerce/>
4. <https://www.oberlo.in/blog/20-ecommerce-advantages-and-disadvantages>
5. <https://www.techbullion.com/impact-e-commerce-empowered-customers/>
6. https://www.tutorialspoint.com/e_commerce/e_commerce_overview.htm
7. <https://en.wikipedia.org/wiki/E-commerce>

DATA MINING APPLICATION IN CREDIT CARD FRAUD DETECTION

J. Karunanithi¹ and S. Jagadesh Kumar² and V.Ponmagal³Associate Professor¹, Thavathiru Santhalinga Adigalar Arts Science and Tamil College
Assistant Professor^{2,3}, Thavathiru Santhalinga Adigalar Arts Science and Tamil College**ABSTRACT**

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

INTRODUCTION

Credit Card Transactions continue to grow in number, taking an ever-larger share of the US payment system and leading to a higher rate of stolen account numbers and subsequent losses by banks. Improved fraud detection thus has become essential to maintain the viability of the US payment system.

Fraud is a million dollar business and it is increasing every year. The PwC global economic survey 2019 suggests that close to 50% of companies worldwide reported fallen victim to fraud in the past two years. Banks have used early fraud warning systems for some years. Large-scale data-mining techniques can improve on the state of the art in commercial practice. Scalable techniques to analyze massive amounts of transaction data that efficiently compute fraud detectors in a timely manner is an important problem especially for e-commerce. Besides scalability and efficiency, the fraud-detection task exhibits technical problems that include skewed distributions of training data and non uniform cost per error, both of which have not been widely studied in the knowledge-discovery and data mining community. Data mining is a relatively new technique that is proving to be extremely effective in detecting fraud, and it offers insurers new opportunities to reduce losses. Although heavily used in the finance and telecommunications industries, data mining has not yet achieved widespread use in the insurance industry, perhaps because it is viewed as too great a technological step, requiring a high level of statistical expertise.

Credit Card Transaction: In today's increasingly electronic society and with the rapid advances of electronic commerce on the Internet, the use of credit cards for purchases has become convenient and necessary. Credit card transactions have become the de facto standard for Internet and Web based e-commerce. The US government estimates that credit cards accounted for approximately US \$13 billion in Internet sales during 1998, taking a larger share of the US payment system, and have led to a higher rate of stolen account numbers and subsequent losses by banks. However, the growing number of credit card transactions provides more opportunity for thieves to steal credit card numbers and subsequently commit fraud.

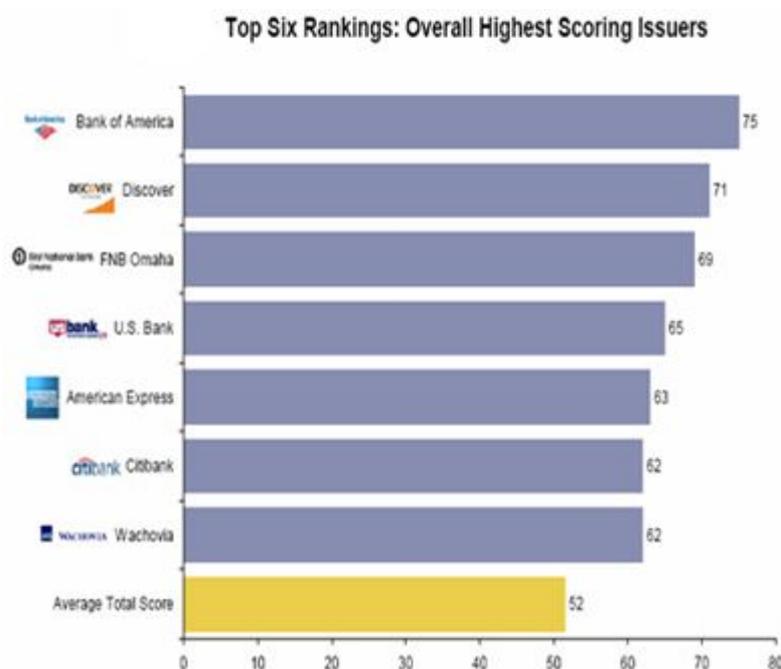
Detecting Frauds: Techniques used for fraud detection fall into two primary classes: statistical techniques and artificial intelligence. Examples of statistical data analysis techniques are:

- Data preprocessing techniques for detection, validation, error correction, and filling up of missing or incorrect data.
- Calculation of various statistical parameters such as averages, quantiles, performance metrics, probability distributions, and so on. For example, the averages may include average length of call, average number of calls per month and average delays in bill payment.
- Models and probability distributions of various business activities either in terms of various parameters or probability distributions.
- Computing user profiles.
- Time-series analysis of time-dependent data.
- Clustering and classification to find patterns and associations among groups of data.
- Matching algorithms to detect anomalies in the behavior of transactions or users as compared to previously known models and profiles. Techniques are also needed to eliminate false alarms, estimate risks, and predict future of current transactions or users.

Fraud management is a knowledge-intensive activity. The main AI techniques used for fraud management include

- Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud.
- Expert systems to encode expertise for detecting fraud in the form of rules.
- Pattern recognition to detect approximate classes, clusters, or patterns of suspicious behavior either automatically (unsupervised) or to match given inputs.
- Machine learning techniques to automatically identify characteristics of fraud.
- Neural networks that can learn suspicious patterns from samples and used later to detect them. Other techniques such as link analysis, Bayesian networks, decision theory, and land sequence matching are also used for fraud detection.

The top six ranking banks in the credit card forgery are listed in the below graph



MACHINE LEARNING AND DATA MINING

Early data analysis techniques were oriented toward extracting quantitative and statistical data characteristics. These techniques facilitate useful data interpretations and can help to get better insights into the processes behind the data. Although the traditional data analysis techniques can indirectly lead us to knowledge, it is still created by human analysts.

To go beyond, a data analysis system has to be equipped with a substantial amount of background knowledge, and be able to perform reasoning tasks involving that knowledge and the data provided. In effort to meet this goal, researchers have turned to ideas from the machine learning field. This is a natural source of ideas, since the machine learning task can be described as turning background knowledge and examples (input) into knowledge (output).

If data mining results in discovering meaningful patterns, data turns into information. Information or patterns that are novel, valid and potentially useful are not merely information, but knowledge. One speaks of discovering knowledge, before hidden in the huge amount of data, but now revealed.

Credit card data and cost models : Chase Bank and First Union Bank, members of the Financial Services Technology Consortium (FSTC), provided us with real credit card data for this study. The two data sets contain credit card transactions labeled as fraudulent or legitimate. Each bank supplied 500,000 records spanning one year with 20% fraud and 80% non fraud distribution for Chase Bank and 15% versus 85% for First Union Bank.

In practice, fraudulent transactions are much less frequent than the 15% to 20% observed in the data given to us. These data might have been cases where the banks have difficulty in determining legitimacy correctly. In some of our experiments, we deliberately create more skewed distributions to evaluate the effectiveness of our

techniques under more extreme conditions. Bank personnel developed the schemata of the databases over years of experience and continuous analysis to capture important information for fraud detection. We cannot reveal the details of the schema beyond what we have described elsewhere.

The records of one schema have a fixed length of 137 bytes each and about 30 attributes, including the binary class label (fraudulent/legitimate transaction). Some fields are numeric and the rest categorical. Because account identification is not present in the data, we cannot group transactions into accounts. Therefore, instead of learning behavior models of individual customer accounts, we build overall models that try to differentiate legitimate transactions from fraudulent ones. Our models are customer-independent and can serve as a second line of defense, the first being customer-dependent models.

Most machine-learning literature concentrates on model accuracy (either training error or generalization error on hold-out test data computed as overall accuracy, true-positive or false-positive rates, or return-on-cost analysis).

This domain provides a considerably different metric to evaluate the learned models' performance—models are evaluated and rated by a cost model. Due to the different dollar amount of each credit card transaction and other factors, the cost of failing to detect a fraud varies with each transaction. n Cumulative cost = \sum cost (i)

Average cost = Cumulative cost / n

Hence, the cost model for this domain relies on the sum and average of loss caused by fraud. We define and where Cost (i) is the cost associated with transactions i, and n is the total number of transactions.

After consulting with a bank representative, we jointly settled on a simplified cost model that closely reflects reality. Because it takes time and personnel to investigate a potentially fraudulent transaction, each investigation incurs an overhead.

Other related costs—for example, the operational resources needed for the fraud-detection system—are consolidated into overhead. So, if the amount of a transaction is smaller than the overhead, investigating the transaction is not worthwhile even if it is suspicious. For example, if it takes \$10 to investigate a potential loss of \$1, it is more economical not to investigate it.

OUTCOME COST

Miss (false negative—FN)

Tranamt False alarm (false positive—FP)

Overhead if tranamt > overhead or 0 if

Tranamt \leq overhead Hit (true positive—TP)

Overhead if tranamt > overhead or

Tranamt if tranamt \leq overhead

Normal (true negative—TN) 0

Therefore, assuming a fixed overhead, we devised the cost model shown in Table 1 for each transaction, where tranamt is the amount of a credit card transaction. The overhead threshold, for obvious reasons, is a closely guarded secret and varies over time. The range of values used here are probably reasonable levels as bounds for this data set, but are probably significantly lower. We evaluated all our empirical studies using this cost model.

DISADVANTAGES OF DATA MINING

Privacy Issues it has always been a major concern in this country. In recent years, with the widespread use of Internet, the concerns about privacy have increase tremendously. Because of the privacy issues, some people do not shop on Internet. They are afraid that somebody may have access to their personal information and then use that information in an unethical way; thus causing them harm. Although it is against the law to sell or trade personal information between different organizations, selling personal information have occurred. For example, according to Washing Post, in 1998, CVS had sold their patient's prescription purchases to a different company. In addition, American Express also sold their customers' credit care purchases to another company. What CVS and American Express did clearly violate privacy law because they were selling personal information without the consent of their customers. The selling of personal information may also bring harm to these customers because you do not know what the other companies are planning to do with the personal information that they have purchased.

Security issues: Although companies have a lot of personal information about us available online, they do not have sufficient security systems in place to protect that information. For example, recently the Ford Motor credit company had to inform 13,000 of the consumers that their personal information including Social Security number, address, account number and payment history were accessed by hackers who broke into a database belonging to the Experian credit reporting agency. This incidence illustrated that companies are willing to disclose and share your personal information, but they are not taking care of the information properly. With so much personal information available, identity theft could become a real problem.

Misuse of information/inaccurate information:- Trends obtain through data mining intended to be used for marketing purpose or for some other ethical purposes, may be misused. Unethical businesses or people may used the information obtained through data mining to take advantage of vulnerable people or discriminated against a certain group of people. In addition, data mining technique is not a 100 percent accurate; thus mistakes do happen which can have serious consequence.

CONCLUSION

Presently, virtual reality represents a broad range of techniques that are rapidly evolving the credit card detection using Data mining. Advanced data mining techniques and neural network algorithm can be combined successfully to obtain a high fraud coverage combined with a low false alarm rate.

REFERENCES

1. Data mining “Han Kamber pei”
2. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, Advances in Knowledge Discovery & Data Mining.
3. Ronald J. Brachman and Tej Anand, "The Process of Knowledge Discovery in Databases," in Advances in Knowledge.

WEB SERVICES

Sindhu T

Student, Computer Science, Sri Ramakrishna Arts and Science College for Womens, Coimbatore

ABSTRACT

Web service meant by interaction between two machines over a network. service also uses SOAP, REST, and XML-RPC as a means of communication. A web service represents a collection of open protocols and standards used for exchanging data between the applications. Software applications written in various programming languages and running on various platforms can use web services to exchange data over computer networks same as the Internet in a manner similar to inter-process communication on a single computer. are but all are not Web services. A Web service uses only three types such as SOAP, REST and XML-RPC for communication.

INTRODUCTION

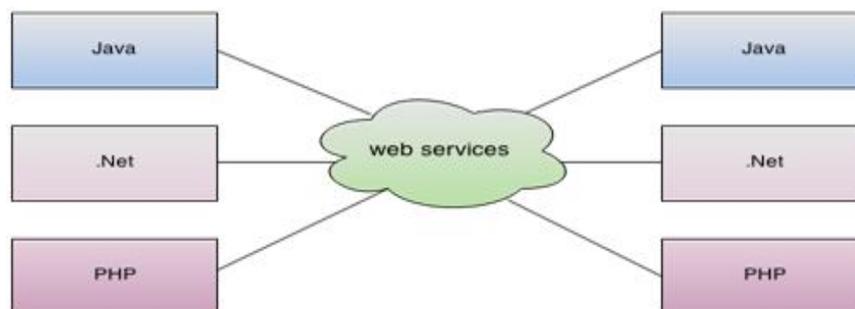
A web service is a network accessible interface to the application functionality that is built using the standard Internet technologies. The web service has some special characteristics as follows.

XML-based: Web services uses the XML as the data transport protocols. XML eliminates the platform dependencies and other protocols may have.

Loosely coupled: A web service and the consumer of the web service are not close to one another; the web service interface can change without affecting the client's interaction with the services

Ability to be synchronous or asynchronous in the synchronous scenario, the client blockwaits for the service to complete. the asynchronous scenario, the client invokes a services and then executes other operations, and the result could be received at a later point.

The simplest Web services system has two participants: A service producer (provider) A service consumer (requester). The provider presents the interface and implementation ofservices, and the requester uses the Web services.



For ex: Java can interact with Java and application. is a language independent way of communication.

HISTORY

The history starts from inter-process communication in which two processes running on a system exchange information between each other. Inter-process communication is achieved using pipes and shared memory and com (Component object model). Afterit extended to LAN where two processes in separate machine in a network exchange data with help of DCOM, RPC and sockets. But it is not possible to exchange data between processes if architecture of system is different because of the different data representation techniques. To solve this issues inXDR(External Data Representation) format evolved which specified the common standards for data representation. a lap now we are exchanging data in the processes separated by internet using protocol. The LAN evolved to internet and web the XDR is evolved into the XML. XDR was binary format and now XML is text format but both are same by nature of working. And the SOAP protocol evolves to achieve communication between applications on internet. And now REST architecture isthe top of every technology.

ADVANTAGES OF WEBSERVICES

There are many advantages of using web services

Interoperability: One of the main advantages of web service is interoperability. Web services allow applications to communicate and exchange data and to share services among themselves. The common standard-based communications methods have been developed and these make it possible for web service to be the platform-independent.

Usability: Web services are designed to be used like a webpage request and receive data. Web services are the same. The capability of web services may differ from simple information lookup to complex algorithmic computations.

Reusability: Web Services are designed to combined to deliver more added-value services. Web services are the building blocks to makes it easy and reuse the Web Service components in other services also legacy applications can be wrapped into web services to be used by others.

Deployability: Web Services are deployed over the Internet standards such as standard Apache, Axis2 to provide HTTP,WSDL driven services. This makes easy to deploy.

Cost: The cost is reduced due to new systems are assembled from packaged webservice. The saved cost can be a benefit to both the solution provider and the customers. Moreover, efficiency is also achieved at the same time.

Differences between Web Services vs API

Both API and Web services are used for communication or interchange of data. Web service is Network based it is over a HTTP or FTP. Web services are used to exchange data in a specified format(JSON,XML). Where as API is super set of Webservices. API Communications may or may not occur over network and can be in any data format.

Ex: A typical e-Commerce application is made up of number of categories which are connected via Webservices. A High end Refrigerator are connected to an application will use APIs to monitor performance. (No network is required).

WEB SERVICES

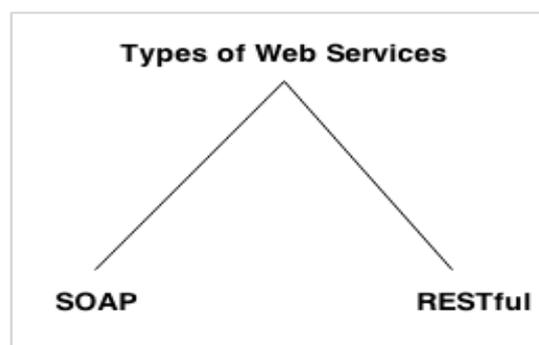
- 1) It is SOAP based service and returns data as XML.
- 2) It only supports the HTTP protocol.
- 3) It can only be hosted on IIS.
- 4) It is not open source but can be used by any client that understands the XML.
- 5) It needed SOAP protocol to receive and send data over the network so it is not a light weight architecture.

Web API

- 1) A Web API is a HTTP based service and returns JSON/XML data.
- 2) It supports HTTP protocol.
- 3) It can be hosted an application.
- 4) It is an open source and it is used by any client that understands JSON / XML.
- 5) It is a light weight architected and good for device ,which have limited bandwidth like mobile device.

TYPES OF WEBSERVICES

There are Mainly two types of web services. These are SOAP webservice and Rest Webservice.



SOAP WEBSERVICE

SOAP stands for Simple Object Access Protocol. It is a XML based protocol for accessing web service. SOAP is a W3C recommendation for communication between the two applications.

SOAP is XML based protocol. It is platform independent and language independent. By using SOAP, it will be able to interact with other programming language applications.

ADVANTAGES OF SOAP WEBSERVICES**WS Security**

SOAP defines its own security known as WS Security. Language and Platform independent: SOAP web services can be written in any programming languages and executed in any platform.

SOAP is platform and language independent.

SOAP provides simplified communications through proxies and firewalls.

SOAP has the ability to leverage different transport protocol including HTTP and SMTP.

Disadvantages of Soap Web Services

Slow: SOAP uses the XML format that must be parsed to be read. It defines many standards that must be followed while developing the SOAP application. So its slow and consumes more bandwidth and resources.

WSDL dependent: SOAP uses WSDL and doesn't have any other mechanism to discover the services.

RESTful WEB SERVICES

REST stands for Representational State Transfer. REST is a architectural style not a protocol.

Advantages of RESTful Web Services

Fast: RESTful WebServices are fast because there is no strict specification like SOAP. It consumes less bandwidth and resource.

Language and Platform independent: RESTful web services can be written in any programming languages and executed in any platform.

Can use SOAP: RESTful web services can use SOAP web services for implementation.

Permits different data format: RESTful web service permits different data format such as Plain Text, XML and JSON.

Restful Web Services is a stateless client/server architecture where webservices are resources and can be identified by their URIs.

REST Client applications can be use HTTP GET/POST methods to invoke Restful webservices. REST does not specify any specific protocol to use but in almost all cases it's used over HTTP/HTTPS. When compared to SOAP webservices are light weight and does not follow any standard. We can use XML, JSON, text or any type of data for request and response.

Representational State Transfer (REST) is an architectural style that specifies constraints such as the uniform interface that is applied to a web service induce desirable properties such as performance, scalability that enables the services to work best on the web. In the REST architectural style, data and functionality are considered resources and are accessed by Uniform Resource Identifiers (URIs), typically links on the web. The resources are acted upon by using a set of simple defined operations. The REST architectural style constrains an architecture to a client/server architecture and is designed to use a stateless communication protocol, typically called as HTTP. In the REST architecture style the clients and servers exchange representations of resources by using a standardized interface and protocol.

The following principle encourage RESTful applications to be simple lightweight and fast:

Resource identification through URI: A RESTful web service exposes a set of resources that identify the targets of the interaction with its client. Resources are identified by URI which provide a global addressing space for resource and service discovery

Uniform interface: Resources are manipulated using a fixed set of four create, read, update, delete operations such as PUT, GET, POST, and DELETE. PUT creates a new resource which can be then deleted by using DELETE method. GET method retrieves the current state of a resource. POST method transfers a new state onto a resource.

Self descriptive messages: Resources are decoupled from their representations so that their content can be accessed in a variety of formats such as HTML, XML, plain text, PDF, JPEG, JSON, and other format. Metadata is about the resource is available and used for example to control caching, detect transmission errors, negotiate the appropriate representation format, and perform authentication or access control.

Stateful interactions through hyperlinks: Every interaction with the resource is stateless that is request messages are self contained. Stateful interaction are based on the concept of explicit state transfer. Several techniques exist to exchange state, such as URI rewriting, cookies, and hidden form field. State can be embedded in the response messages to point to valid future states of the interaction. REST is an architectural style for designing distributed system. It is not a standard but set of constraints such as being stateless having a client server relationship and a uniform interface. REST is not strictly related to HTTP but its most commonly associated with it.

Principles of REST

Resource expose easily understood directory structure URIs.

Representations transfer JSON or XML to represent data object and attributes.

Messages use HTTP methods explicitly (for ex, GET, POST, PUT, and DELETE).

Stateless interactions store no client context on the server between the requests. State dependencies can limit and restrict scalability. The client holds session state.

HTTP methods

HTTP methods used to map CRUD (create, retrieve, update, delete) operations to HTTP requests.

GET

Retrieve information. GET requests must be safe how many times it repeats with the same parameters the results are the same. They can have side effects but the user does not expect them, so they cannot be critical to the operation of the system. Requests can also be partial or conditional.

Retrieve an address with an ID of 1:

GET /addresses/1

POST

Request that the resource at the URI can do something with the provided entity. POST method is used to create a new entity but it can also be used to update an entity.

Create a new address:

POST /addresses

PUT

Stores an entity at a URI. PUT can create a new entity or update an existing the old one. PUT request is idempotent. Idempotency is the main difference between the expectations of PUT vs a POST request.

Modify the address with an ID of 1:

PUT /addresses/1

Note: PUT replaces an existing entity. If only subset of data elements are provided, the rest will be replaced with empty or null.

PATCH

Update only the specified fields of an entity at URI. A PATCH request is neither safe nor idempotent (RFC 5789) Because a PATCH operation cannot ensure the entire resource has been updated.

PATCH /addresses/1

DELETE

Request that resource be removed the resource does not have to be removed immediately. It could be an asynchronous or long-running request.

Delete an address with an ID of 1:

DELETE /addresses/1

HTTP status codes

Status codes indicates the result of the HTTP request

1XX - informational

2XX - success

3XX - redirection

4XX - client error

5XX - server error

REFERENCE

- <https://www.javapedia.net/Web-services/1325>
- <https://hostpresto.com>
- <https://www.pyarb.com/developer/java/web-services-history-introduction-soap-vs-rest.html>
- <http://sfdcsrini.blogspot.com/2015/09/what-is-web-service-and-what-are.html>
- <https://www.quora.com/What-is-the-difference-between-web-services-and-API>
- <https://spring.io/understanding/REST>

CLOUD COMPUTING SECURITY ISSUES & CHALLENGES..,

Chandra Devi P¹, Afritha A. Nazar² and Anjana A³

Assistant Professor¹ and Student^{2,3}, Department of Commerce, Sri Krishna Arts and Science College, Coimbatore

ABSTRACT

Cloud Computing is an emerging paradigm which has become today's hottest research area due to its ability to reduce the costs associated with computing. Since Cloud computing stores the data and its disseminated resources in the environment, security has become the main obstacle of environment. There are number of users used cloud to store their personal data, so that data storage security is required on the storage media. The major concern of cloud environment is security during upload the data on cloud server. Cloud computing, undoubtedly, has become the buzzword in the IT industry today. In this paper we gonna discuss about a short preview and security issues of cloud computing on storage, data and network and their virtualization. It also makes an attempt to describe the security challenges in Software as a Service (SaaS) model of cloud computing and also endeavors to provide future security research directions.

Keywords: Cloud computing, Cloud for everyone, Private vs. Public cloud, developing cloud services, models of cloud, Security issues, Encrypted Data storage, security challenges, specific challenges, Related Technologies, Future Trends, Conclusion, References.

INTRODUCTION

Cloud Computing: Cloud Computing is rapidly being accepted as a universal access appliance on the Internet. But here we have considered the standard definition which was given by the National Institute of Standards and Technology (NIST): "Cloud Computing is model for enabling ubiquitous, convenient, on demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".

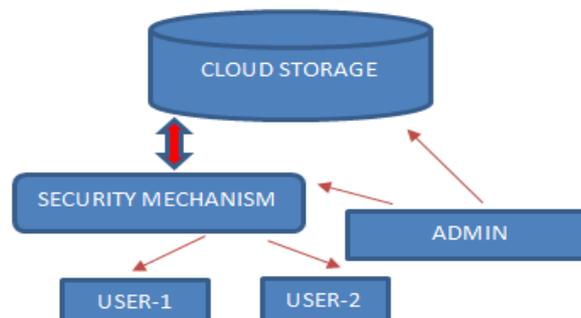


Fig: Cloud Storage Mechanism

CLOUD FOR EVERYONE**Cloud Computing For Family**

- Centralization of E.MAILs.
- Collaboration of schedules.
- Collaborating house-hold budget .
- Collaboration on contact list.
- Collaboration on school /college projects.
- Sharing family photos. etc..,

Cloud For Community

- Collaborating Event Management.
- Community group schedules.
- Collaborating group projects & events.
- Collaborating task management.
- Collaborating on budgets.

Cloud For Cooperation

- Managing schedules & contact list.
- Managing projects, marketing materials.
- Collaborating expense and budgets report.

PRIVATE CLOUD VS. PUBLIC CLOUD

A Detailed comparison and economic model of using public cloud verses private cloud for database workloads is presented by tak et al .The author consider the intensity of workload ,burstiness , as well as the growth rate of the workload in their evaluation.

DEVELOPING CLOUD SERVICES

Cloud computing offers a lot to software developers , who can now develop web-based applications that take advantage of the power and reach of cloud computing.

ADVANTAGES OF CLOUDE DEVELOPMENT

One of the underlying advantages of cloud development is that of economy of scale.

- Cloud infrastructures.
- Low & reliable cost.
- Rapid provisioning & automatic scaling.
- Easy development.

DISADVANTAGES OF CLOUD DEVELOPMENT :

- Security risk.
- Third -party involvement.
- Another important problem is that if the host goes offline, all process will become risky.

MODELS OF CLOUD COMPUTING :

- Private cloud
- Community cloud
- Public cloud
- Hybrid cloud

Private Cloud

The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers.

Community Cloud

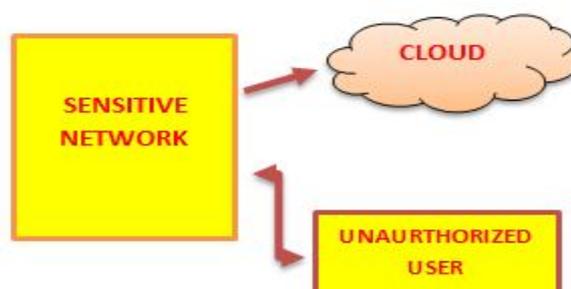
The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns.

Private Cloud

The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization.

Hybrid Cloud

The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities.



SECURITY ISSUES IN CLOUD

There are numerous security issues for cloud computing as it work with numerous fields like network ,Database ,Operating System ,Virtualization and Concurrency control and Memory Management.

We will focus only on some aspects of secure cloud computing.one is to store data efficiently in foreign machines . Another one is to query encrypted data as much of data on cloud may be encrypted.

DATA ENCRYPTION

Encryption is a key technology for data security. Understand data in motion and data at rest encryption. Remember, security can range from simple (easy to manage, low cost and quite frankly, not very secure) all the way to highly secure (very complex, expensive to manage, and quite limiting in terms of access).

ENCRYPTED DATA STORAGE FOR CLOUD

Since the data will be placed anywhere in cloud it is important that the data must be encrypted . We are using secure co-processor parts cloud infrastructure to enable efficient encrypted storage of encrypted storage of sensitive data.

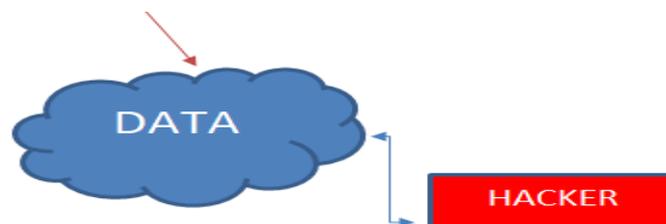
Security Feature in Cloud Computing: There are several main challenges for building a secure and trustworthy cloud system. they are,

Outsourcing: Outsourcing means that customers physically lose control on their data and tasks. The loss of control problem has become one of the root causes of cloud insecurity. To address outsourcing security issues, first, the cloud provider shall be trustworthy by providing trust and secure computing and data storage; second, outsourced data and computation shall be verifiable to customers in terms of confidentiality, integrity, and other security services. In addition, outsourcing will potentially incur privacy violations, due to the fact that sensitive data is out of the owner's control.

Massive data and intense computation

Cloud computing is capable of handling mass data storage and intense computing tasks. Therefore, traditional security mechanisms may not suffice due to unbearable computation or communication overhead. For example, to verify the integrity of data that is remotely stored, it is impractical to hash the entire data set. To this end, new strategies and protocols are expected.

Connected Devices



SECURITY ISSUES OF CLOUD

The security of corporate data in the cloud is difficult, as they provide different services like **Network as a service (NaaS)**, **Platform as a service (PaaS)**, **Software as a service (SaaS)**, **Infrastructure as a service (IaaS)**. Each service has their own security issues . They are,

Data Security: Data Security refers as a confidentiality, integrity and availability. These are the major issues for cloud vendors. Confidentiality is defined as a privacy of data. Confidentiality are designed to prevent the sensitive information from unauthorized or wrong people.

Regulatory Complaint: Customers are eventually accountable when the security and completeness of their own data is taken by a service provider.

Data Locations: When users use, they probably won't know exactly where their data will hosted and which location it will stored in. In fact, they might not even know what country it will be stored in.

Privileged user access: Outside the resource data that is processed contains major risk, as deploy services, avoid the mortal, consistent and human resource manage IT shops works on the house programs.

Trust Issue: Trust is also a major issue in cloud computing. Trust can be in between human to machine, machine to human, human to human, machine to human. Trust is revolving around assurance and confidence. In cloud computing, user stores their data on cloud storage because of trust on cloud.

Data Recovery: It is defined as the process of restoring data that has been lost, corrupted or accident.

CLOUD SECURITY CHALLENGES

There are some key security challenges. They are

Authentication: Throughout the internet data stored by cloud user is available to all unauthorized people. Henceforth the certified user and assistance cloud must have interchangeability administration entity.

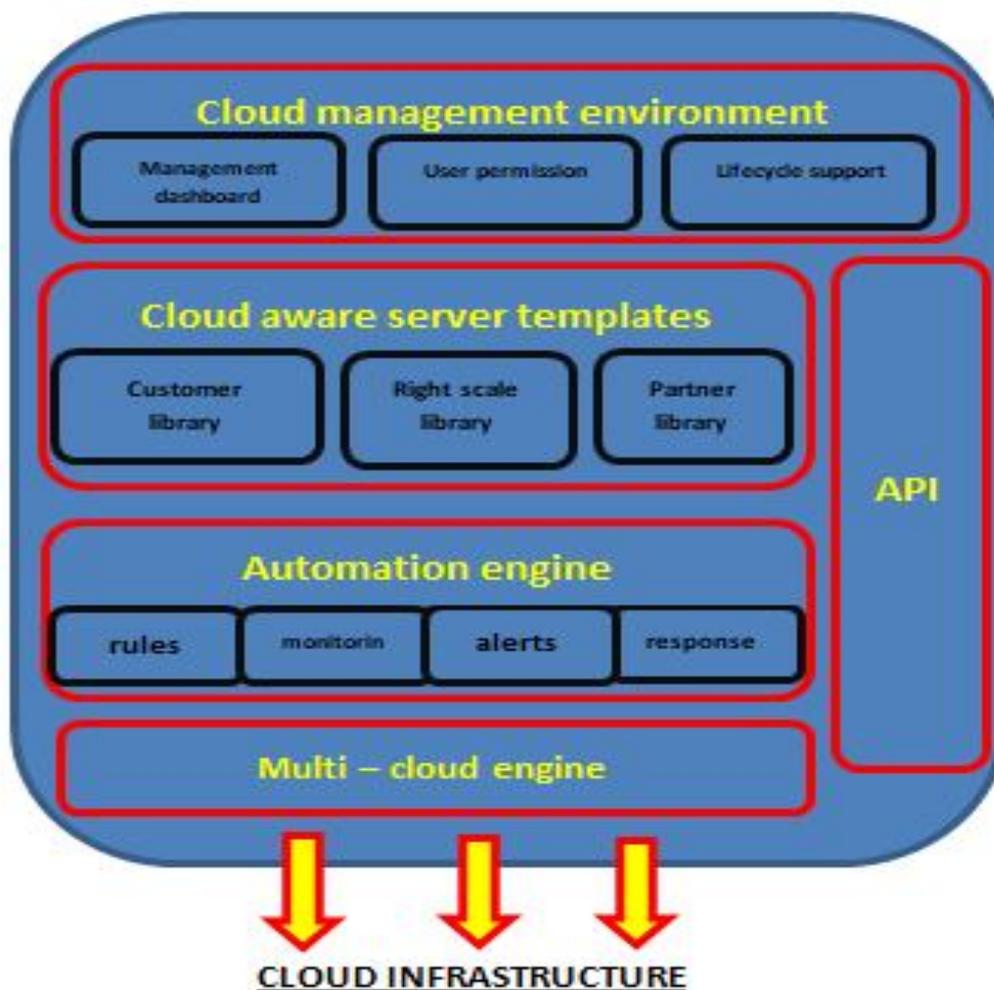
Access Control: To check and promote only legalized users, cloud must have right access control policies. Such services must be adjustable, well planned, and their allocation is overseeing conveniently.

Policy Integration: There are many cloud providers such as Amazon, Google which are accessed by end users. Minimum number of conflicts between their policies because they use their own policies and approaches.

Service Management: In this different cloud providers such as Amazon, Google, comprise together to build a new composed services to meet their customers need.

Trust Management: The trust management approach must be developed as cloud environment is service provider and it should include trust negotiation factor between both parties such as user and provider.

MANAGING CLOUD



CLOUD SPECIFIC SECURITY CHALLENGES

Information Security

In a traditional on - premise application deployment model, the sensitive data of each enterprise continues to reside within the enterprise boundary and is subject to its physical, logical and personnel security and access control policies. This involves the use of strong encryption techniques for data security and fine-grained authorization to control access to data.

Network Security

In a SaaS deployment model, sensitive data is obtained from the enterprises, processed by the SaaS application and stored at the SaaS vendor end. All data flow over the network needs to be secured in order to prevent leakage of sensitive information. This involves the use of strong network traffic encryption techniques such as Secure Socket Layer (SSL) and the Transport Layer Security (TLS) for secure.

Web application security

- SaaS is software deployed over the internet and/or is deployed to run behind a firewall in local area network or personal computer.
- The key characteristics include Network-based access to, and management of, commercially available software and managing activities from central locations rather than at each customer's site, enabling customers to access application remotely via the Web.
- SaaS application development may use various types of software components and frameworks. These tools can reduce time-to-market and the cost of converting a traditional on premise software product or building and deploying a new SaaS solution.

Common Cloud Standards

Security based accreditation for Cloud Computing would cover three main areas which are technology, personnel and operations. Technical standards are likely to be driven by organizations

Server consolidation

The increased resource utilization and reduction in power and cooling requirements achieved by server consolidation are now being expanded into the cloud. Server consolidation is an effective approach to maximize resource utilization while minimizing energy consumption in a cloud computing environment.

Reliability & Availability of Service

The challenge of reliability comes into the picture when a cloud provider delivers on-demand software as a service. The software needs to have a reliability quality factor so that users can access it under any network conditions (**such as during slow network connections**).

Access Controls

Authentication and identity management is more important than ever. And, it is not really all that different. This is not all that different from how you secure your internal systems and data, and it works the same way

CURRENT SECURITY SOLUTIONS

There are several research works happening in the area of cloud security. Several groups and organization are interested in developing security solutions and standards for the cloud. The Cloud Security Alliance (CSA) is gathering solution providers, non- profits and individuals to enter into discussion about the current and future best practices for information assurance in the cloud. The Open Grid Forum publishes documents to containing security and infrastructural specifications and information for grid computing developers and researchers

RELATED TECHNOLOGIES**Server Virtualization**

Server virtualization is a key technology that enables **Infrastructure as a service (IaaS)**. It uses software modules instead of hardware devices.

Storage virtualization

It uses hardware entities instead of software modules in processing. There are two types in storage virtualization. They are,

- File virtualization.
- Block virtualization.

FUTURE TRENDS

- Cloud computing is latest technology that is being widely used all over the world. Once the organization takes the decision to move to the cloud, it loses control over the data.
- Thus, the amount of protection needed to secure data is directly proportional to the value of the data. Number of cloud platforms are available now in educational as well as in enterprises circle .
- New security techniques need to be developed and older security techniques needed to be radically tweaked to be able to work with the clouds architecture.
- As the development of cloud computing technology is still at an early stage, we hope our work will provide a better understanding of the design challenges of cloud computing.
- Resource Locality and Data Segregation are the two key security challenges on which not much information is available in the existing literature, which necessitates that this can be further taken up for research.

CONCLUSION

Though there are numerous advantages in using a cloud-based system, there are yet many practical issues which have to be sorted.

As described in the paper, currently security has lot of loose ends which scares away several potential users. Until a proper security module is not in place, potential users will not be able to leverage the true benefits of this technology.

One of the biggest security worries with the cloud computing model is the sharing of resources. Cloud service providers need to inform their customers on the level of security that they provide on their cloud.

In this paper, we first discussed various models of cloud computing, security issues. Data security is major issue for Cloud Computing. There are several other security challenges including security aspects of network and virtualization.

This paper has highlighted all these issues of cloud computing. We believe that due to the complexity of the cloud, it will be difficult to achieve end-to-end security.

As the development of cloud computing technology is still at an early stage, we hope our work will provide a better understanding of the design challenges of cloud computing, and pave the way for further research in this area.

These issues mentioned above will be the research hotspot of cloud computing. There is no doubt that cloud computing has bright future.

REFERENCES

1. IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS) Vol. 1, No. 2, December 2011
2. International Journal of Advances in Engineering & Technology, June, 2015. ©IJAET
3. International Journal on Cloud Computing: Services and Architecture (IJCCSA) ,Vol.3, No.4, August 2013
4. Michael Miller . "CLOUD COMPUTING" vol-2. 2016 edition.
5. Ronald L. Krutz and Russell Dear Vines. "cloud security:A comprehensive Guide to secure computing"
6. Dinakar Sitaram and Geetha Manjunath. "moving to the cloud:Developing app in the new world of cloud computing"

DIABETES MELLITUS (DM) PREDICTION AND CLASSIFICATION USING BAT OPTIMIZATION BASED ON ADAPTIVE NEURO-FUZZY**R. S. Padma Priya¹ and Dr.P.Senthilvadivu²**Research Scholar¹ and Head², Hindusthan College of Arts and Science, Coimbatore**ABSTRACT**

Information mining approach analyzes patient's maladies. Diabetes Mellitus is an unending problem to influence different organs of the human body. Early expectation can spare human life and can take command over the infections. In this investigation, we have proposed a novel model dependent on information digging strategies for anticipating type 2 diabetes mellitus (T2DM). The primary issues that we are endeavoring to fathom are to enhance the exactness of the forecast display. This should be possible utilizing any of the element choice calculations accessible in information mining. This paper proposes a Modified Bat Algorithm (MBA) for highlight choice to wipe out insignificant highlights from a unique dataset. The Bat calculation was adjusted utilizing straightforward arbitrary inspecting to choose the irregular examples from the dataset. Positioning was with the worldwide best highlights to perceive the transcendent highlights accessible in the dataset. The chose highlights are utilized to prepare a KNN and ANFIS order calculation. Here proposed improvement based characterization technique ANFIS is order effective precision and brief day and age for contrasted with existing enhancement based grouping strategies.

Keywords: ANFIS, Bat Algorithm, Diabetes Mellitus, Information Mining, KNN.

I. INTRODUCTION

Diabetes is a constant sickness with the possibility to cause an overall Health Care emergency. As indicated by International Diabetes Federation 382 million individuals are living with diabetes around the world. By 2035, this will be served as 592 million. Early forecast of diabetes is very testing undertaking for restorative experts because of complex relationship on different elements. Diabetes influences human organs, for example, kidney, eye, heart, nerves, foot and so on. Information mining is a procedure to remove valuable data from huge database. It is a multidisciplinary field of software engineering which includes computational process, machine learning, measurable procedures, arrangement, grouping and finding designs.

Diabetes mellitus (DM) is an incessant ailment that is described by high blood glucose. Almost 50% of all diabetics have family heredity factors, which is a standout amongst the most essential highlights of DM. Disappointment of the pancreas to create enough insulin and the body's wasteful use insulin are both pathologic reasons for DM. There are two sorts of DM. The pathogenesis of sort 1 diabetes mellitus (T1DM) is that the pancreas secretes harmed β -cells, keeping it from bringing down blood glucose level in time. Insulin opposition and insulin emission inadequacy are the pathogeneses of sort 2 diabetes mellitus (T2DM), which is additionally, called noninsulin subordinate DM. In the previous 30 years of advancement in China, with rising number of diabetics, individuals have begun to understand that this incessant sickness has profoundly affected each family and everybody's day by day life. There is a climbing pattern in the extent of diabetics in the overall public, and the development rate of male diabetics is higher than that of female diabetics, The International Diabetes Federation (IDF) displays the most recent information on DM in the Diabetes Atlas (Seventh Edition) [1]. It demonstrates that in 2015, the quantity of diabetics worldwide was near 415 million. Regarding the populace development pattern of diabetics, it predicts that the number wills way to deal with 642 million or one of every ten grown-ups.

Diabetes

Diabetes Mellitus (DM) is generally alluded as Diabetes; it is the condition in which the body does not appropriately process sustenance for use as vitality. The majority of the sustenance we eat is transformed into glucose or sugar for vitality. The pancreas, an organ makes a hormone called insulin to enable glucose to get into the phones of our bodies. At the point when a body is influenced with diabetes, it couldn't make enough insulin or couldn't utilize its very own insulin. This makes sugar incorporate up with blood. A few pathogenic procedures are associated with the improvement of diabetes. These range from immune system obliteration of the β -cells of the pancreas with subsequent insulin insufficiency to irregularities that outcome in protection from insulin activity. The regular side effects for the diabetic patients are visit pee, expanded thirst, weight reduction, moderate mending in wound, happiness, expanded yearning and so forth. Diabetes can cause genuine wellbeing confusions including coronary illness, visual impairment, and kidney disappointment and low-furthest point removals.

II. RELATED WORKS

So as to acquire increasingly helpful and important information, we understood that the preprocessing techniques and parameters ought to be picked objectively. B.J.Oommen. [2] proposed the advantages of Bat algorithm. The convergence speed and algorithm searching precision are determined by the pulse of loudness and pulse rate. X.-S. Yang et al. [3] proposed the problem formulation and solution search for various nature inspired optimization techniques. L. Wang et al [4] has proposed an online dictionary learning for improving the performance of PSO. Y. Gheraibia et al[5] introduces a metaheuristic Ant Colony algorithm for search space reduction and improved results. Vijayan V. [9] audited the advantages of various preprocessing procedures for foreseeing Diabetes Mellitus. The preprocessing techniques were essential part investigation (PCA) and discretization. It reasoned that the preprocessing strategies enhanced the precision of the gullible Bayes classifier and choice tree (DT), while the help vector machine (SVM) exactness diminished. Wei [10] broke down hazard elements of type 2 diabetes dependent on the FP-development and Apriori calculations. Guo [11] proposed the recipient working trademark (ROC) territory, the affectability, and the particularity prescient qualities to approve and check the trial results. Based on a compelling expectation calculation, we require a suitable method to make the model advantageous for everybody [12]. We found that Sowjanya [13] had built up an android application-based answer for conquer the inadequacy of mindfulness about diabetes in his paper. The application utilized the DecisionTree classifier to anticipate diabetes levels for clients. The framework likewise given data and proposals about diabetes. It utilized a genuine world dataset gathered from a healing center in the Chhattisgarh province of India. Shi et al. [14] thought about that avoiding type 2 diabetes mellitus (T2DM) ought to be coordinated toward people. In this manner, they concentrated on building up a diabetes hazard evaluation demonstrate and built up a diabetes chance score framework dependent on cell phones.

The highlights are picked by various combinatorial streamlining issue dependent on the conduct of ants. ACO Inspecting technique thinks about Ant Colony Optimization (ACO) to address the issue of class anomaly that happens regularly is proposed. The above investigation shows that there is the need of cross breed technique as a singular methodology is certainly not sufficient or inducing for beginning period area (Yu, Ni, and Zhao, 2013). The bosom malignant growth acknowledgment is finished utilizing the neighborhood straight wavelet neural system. The execution of the preparation parameters is enhanced utilizing Recursive Least Square (RLS) approach. The recommended model uncovers the association loads of the neurons in the covered up and yield layer. This strategy is observed to be powerful on examination with the regular techniques (Senapati, Mohanty, Dash, and Dash, 2013).

III. METHODOLOGY

3.1 Modified Bat Algorithm

Bat algorithm is a knowledge enhancement calculation roused by the echolocation conduct of bats [9]. Echolocation functions as a kind of sonar: bats, primarily microbats, discharge a noisy and short solid heartbeat. When they hit an article, after a small amount of time, the reverberate will return back to their ears. Te bat gets and recognizes the area of the prey along these lines. Furthermore, this astonishing introduction component makes bats ready to recognize the contrast between an impediment and a prey and enables them to chase even in total murkiness. So as to mimic the searching conduct of the bats, the natural system of the bat calculation to mimic the bats' scavenging conduct complies with the accompanying glorified presumptions:

- (1) All bats use echolocation to detect separation, and they likewise know the contrast between nourishment/prey and the foundation boundaries in some supernatural way.
- (2) A bat flies arbitrarily with speed V_i at position X_i with a fixed recurrence f_{min} , changing wavelength λ , and tumult A_0 to look for prey. They can consequently change the wavelength (or recurrence) of their produced heartbeats and modify the rate of heartbeat discharge $r \in (0, 1)$ contingent upon the vicinity of their objectives.
- (3) Although the tumult can differ from numerous points of view, Yang expect that the uproar changes from a huge positive A_0 to a base consistent esteem A_{min} .

Based on three romanticized suppositions, the algorithm creates a lot of arrangements haphazardly and after that looks through the ideal arrangement by cycle and fortifies the neighborhood seek during the time spent looking. By creating the neighborhood arrangement close to the ideal arrangement by arbitrary battle, BA finally finds the worldwide ideal arrangement.

The foraging space of bats is the d dimension. At time $t-1$, the location and the flight velocity of the i th bat are X_i^{t-1} and V_i^{t-1} respectively, and X_* is the current global optimal location. At time t , the velocity and position of the i th bat are updated by using the following equations:

$$f_i = f_{min} + (f_{max} - f_{min}) \beta$$

$$V_i^t = V_i^{t-1} + (X_i^{t-1} - X_s) f_i$$

$$X_i^t = X_i^{t-1} + V_i^t$$

Where f_{min} and f_{max} are the minimum and maximum frequency of the sound waves emitted by bats, respectively β is a random number obeying the uniform distribution in $[0, 1]$. When setting up the initial values, the frequency of launch sound waves of each bat obeys the uniform distribution in $[f_{min}, f_{max}]$. First of all, according to, obtain their frequencies, and then, according to update the velocities and positions.

For the local search, each bat carries out the random walk based on the optimal solution. The following equation is used to produce a new solution:

$$X_{new} = X_{old} + \varepsilon \bar{A}(t)$$

Where ε is a random number obeying the uniform distribution in $[-1, 1]$, X_{old} is a solution randomly selected from the current optimal solution $\bar{A}(t)$, and shows the average loudness for all bats at the t th iteration. The update rules of the loudness of pulse emission of the bat A_i and the velocity r_i are described as follows: if a bat finds prey, it will reduce its impulse response and increase the velocity of its pulse emission. In BA, the loudness of pulse emission of the bat A_i and r_i velocity is adjusted by the following equations:

$$A_i^{t+1} = \alpha A_i^t$$

$$r_i^{t+1} = r_i^t [1 + \exp \gamma t],$$

Where r_i^0 is the initial velocity and A_i^0 is the initial loudness, which are selected randomly. α and γ are constant ($0 < \alpha < 1 ; \gamma > 0$).

4. K - NEAREST NEIGHBOR ALGORITHM

KNN is a technique which is utilized for ordering objects dependent on nearest preparing models in the element space. KNN is the most fundamental kind of case based learning or sluggish learning. It expects all examples are focuses in n -dimensional space. A separation measure is expected to decide the "closeness" of examples. KNN arranges an example by discovering its closest neighbors and picking the most mainstream class among the neighbors.

Highlights of KNN

- a) All cases of the information compare to the focuses in a n -dimensional Euclidean space
- b) Classification is deferred till another occasion arrives
- c) In KNN, the Classification is finished by contrasting component vectors of the distinctive focuses in a space locale.
- d) The objective capacity might be discrete or genuine esteemed.

In KNN, the preparation tests are for the most part portrayed by n -dimensional numeric qualities. The preparation tests are put away in a dimensional space. At the point when a test (obscure class mark) is given, k -closest neighbor classifier begins looking through the ' k ' preparing tests which are nearest to the obscure example or test. Closeness is chiefly characterized regarding Euclidean separation. The Euclidean separation between two P and Q i.e. $P (p_1, p_2, \dots, p_n)$ and $Q (q_1, q_2, \dots, q_n)$ is characterized by the accompanying condition:

$$D(P, Q) = \sum_{i=1}^n (P_i - Q_i)^2$$

5. ADAPTIVE NETWORK-BASED FUZZY INFERENCE SYSTEM (ANFIS)

A versatile neuro-fluffy illation framework or versatile system based fluffy interface framework (ANFIS) could be a sensibly fake neural system that is bolstered Takagi– Sugeno fluffy illation framework. The strategy was produced inside the mid Nineties. Since it incorporates each neural systems and scientific rationale standards, it's capability to catch the upsides of each in an exceptionally single structure. Its illation framework relates to a gathering of fluffy IF– THEN decides that have learning ability to rough nonlinear capacities. Thus, ANFIS is considered to be a widespread reckoner for abuse the ANFIS in an extremely extra practical and ideal way, one will utilize the best parameters acquired by hereditary guideline.

A versatile neuro-fluffy interface framework (ANFIS) could be a blend of ANN and Fuzzy illation System (FIS) in such the least complex way that neural system learning calculations region unit wont to check the parameters of FIS. A reasonable extra essential side is that the framework should dependably be logical regarding fluffy on the off chance that rules; because of it's upheld the fluffy framework intelligent cloud data. We've utilized first– arrange Sugeno fluffy model among a few FIS Models. The Sugeno fluffy model is most commonly connected one for its high interpretability and machine power and indispensable ideal and versatile procedures.

The Sugeno fluffy model gives a logical way to deal with get fluffy standards from a gathering of input– yield data sets. Further the ideal estimations of the following (parameters in versatile neuro-fluffy illation framework) might be found by abuse the littlest sum sq. technique (LSM). When the preface parameters don't appear to be mounted, the inquiry region winds up bigger and along these lines the union of instructing turns out to be slower. The half and half learning (HL) rule consolidating LSM and BP calculations might be wont to unravel this drawback. it completely was appeared in this the hectoliter rule is exceptionally practical in Training the ANFIS. This standard unites a great deal of snappier since it diminishes the component of the hunt zone of the BP rule. All through the preparation strategy, the introduce parameters and thusly the sequent parameters zone unit tuned till the predefined reaction of the FIS is accomplished. The hectolitre rule contains a traditional dancing strategy. To begin with, the following parameters zone unit known abuse LSM once the estimations of the commence parameters territory unit mounted. At that point, the following parameters region unit direction mounted while the blunder is engendered from the yield complete to the info complete, and in this way the start parameters territory unit refreshed by the BP rule [13].

A. Neuron-Fuzzy Approach

The neural-fluffy model is a successful technique for displaying nonlinear frameworks, for example, climate information because of the mix of focal points of neural frameworks and fluffy rationale frameworks. The Neuro-versatile learning strategies give a technique to the fluffy displaying method to pick up data about a dataset, so as to figure the participation work parameters which enable the related fluffy induction framework to follow the given information/yield information (Jang 1993).Each fluffy framework contains three fundamental parts: fuzzification, derivation and defuzzification.

B. Design of ANFIS

The ANFIS approach takes in the tenets and participation capacities from information. The ANFIS design is exhibited in figure 2. The round hubs speak to hubs that are settled. While the square hubs are hubs that have parameters to be educated.

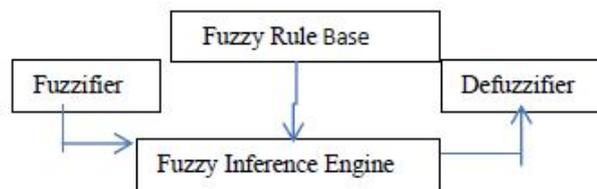


Fig. 1 Fuzzy interface system
Fig-1: Fuzzy Interface System

A two Rule Sugeno ANFIS has rules of the form

A two Rule Sugeno ANFIS has rules of the form:
 If x is A_1 and y is B_1 THEN $f_1 = p_1x + q_1y + r_1$ 1
 If x is A_2 and y is B_2 THEN $f_2 = p_2x + q_2y + r_2$ 2

When Training the Network there is a forward pass and a backward pass. The the input vector is propagated through the network layer by layer using forward pass. In the backward pass, the error is sent back through the network in a way similar to back propagation [1].

In Layer 1, the output of each node is:

$$O_{1,i} = \mu_{B_{i-2}}(y) \quad \text{for } i = 3,4$$

An is membership functions grade for x and y .The membership functions could be any shape. Using the Gaussian membership function given by:

$$\mu_{\bar{A}}(x) = \frac{1}{(1+x)^2} \dots \dots \dots$$

In layer 2, every node in this layer is fixed. Here used ‘AND’, ‘OR’ membership grades for example the product:

$$O_{2,i} = w_i = \mu_{A_i}(x)\mu_{B_i}(y), \quad i = 1,2 \dots$$

Layer3 contains the fixed node which calculates the ratio of the firing strengths of the rules:

$$O_{3,i} = \bar{w}_i = \frac{w_i}{w_1+w_2}$$

The nodes in layer 4 are adaptive and perform the consequent of the rules:

$$O_{4,i} = \bar{w}_i f_i = \bar{w}_i (p_i x + q_i y + r_i) \dots \dots$$

The parameters in this layer are to be determined and are referred to as the consequent parameters,

In layer 5 there is a single node that computes the overall output:

$$O_{5,i} = \sum \bar{w}_i f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \dots \dots \dots$$

Based on each layer data is applied and update the weight and bias values and predicted the bestvalue .

6. RESULT

This proposed optimized ANFIS method based classification is given high accuracy and less time period diabetic data analysis.

Table-1: Time period analysis of feature extraction methods

ACO	BAT	MBAT
3.5	3.2	2.8

Fig-2: Time period calculation for optimization algorithm
time period

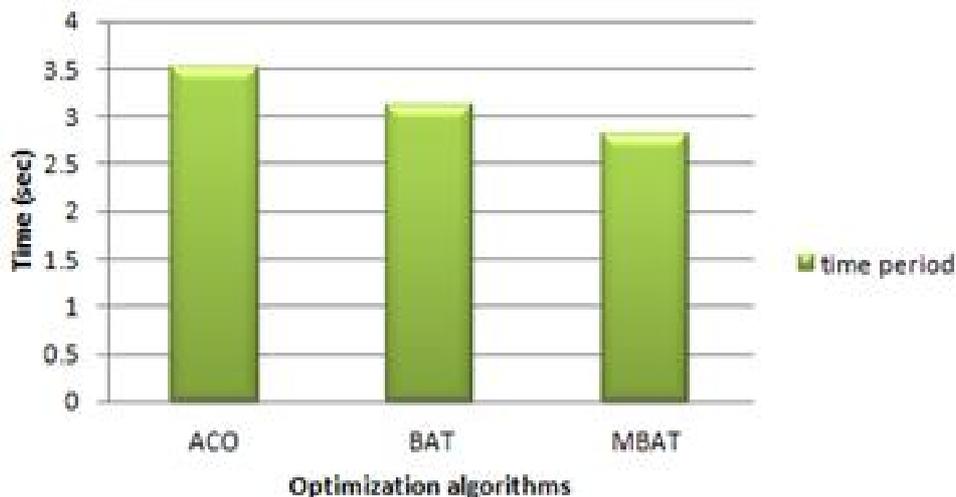


Table-2: Parameter analysis for existing and proposed classification methods

Algorithm	Accuracy	Precision	recall
KNN	89.3	74.3	76.4
ANN	92.4	76.2	75.2
OANFIS	97.5	78.4	79.5

Table 2 show the parameters (accuracy, precision, recall) analysis for existing (KNN, ANN) and proposed (OANFIS) classification methods.

Fig-3: Parameter analysis chart for existing (KNN, ANN) and proposed (OANFIS) classification methods performance

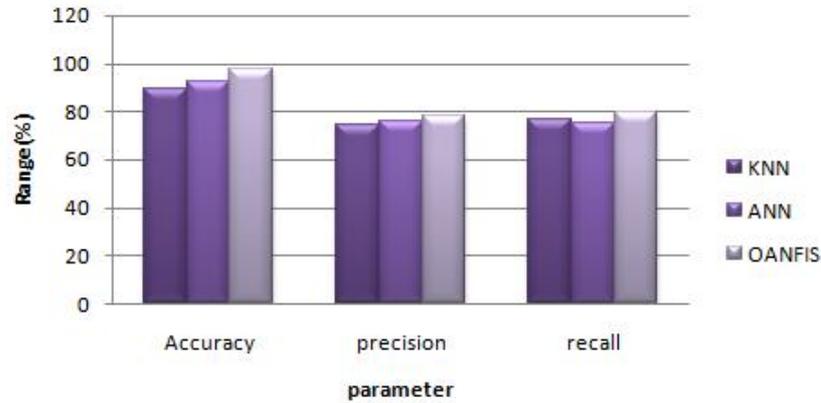


Table-3: Time Period analysis for existing and proposed classification methods

Algorithm	Time period	RMES
KNN	3.56	0.046
ANN	3.21	0.037
OANFIS	2.86	0.312

Fig-4: Time Period analysis chart for existing and proposed classification methods

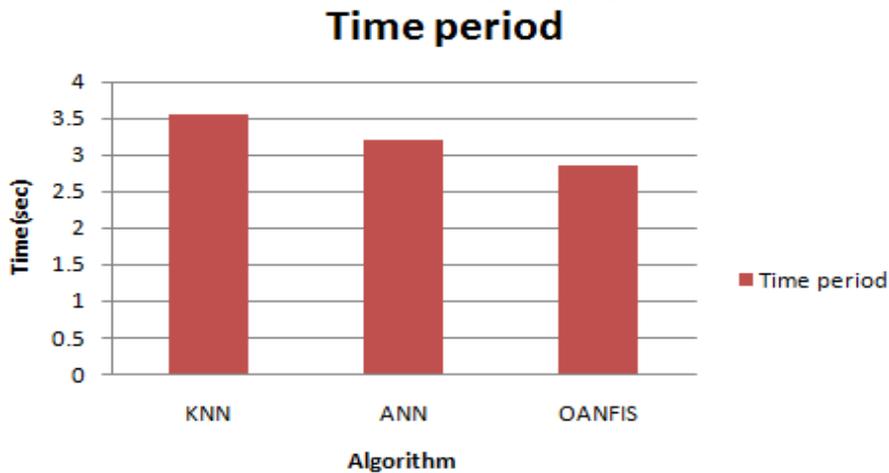


Figure-4 show the parameters (time period) analysis for existing (KNN, ANN) and proposed (OANFIS) classification methods

Fig-5: RMSE analysis for existing and proposed classification methods

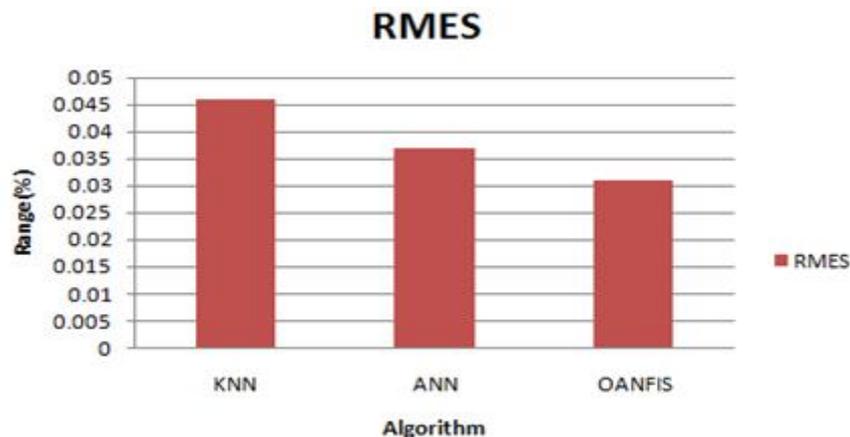


Figure.5 shows the parameters (RMSE) analysis for existing (KNN, ANN) and proposed (OANFIS) classification methods

7. CONCLUSION

In this paper, we have proposed Diabetes-Finder for determination of Type-2 Diabetes. The proposed model built the OANFIS grouping framework by acquiring the class shrewd bunch focuses i.e., class by class approach. The proposed model is probed PID dataset. Because of class by class approach the computational time is diminished (because of less number of examples) and exactness is expanded (by maintaining a strategic distance from lopsided order). Exploratory outcomes on PID informational index demonstrated that proposed model with couple of neurons in concealed layer beat as far as exactness contrasted with KNN, SVM, ANN, and regular OANFIS.

8. REFERENCES

1. M. M. Noack and S. W. Funke, "Hybrid genetic defaced Newton method for global optimization," *Journal of Computational and Applied Mathematics*, vol. 325, pp. 97–112, 2017.
2. N. Bell and B. J. Oommen, "A novel abstraction for swarm intelligence: particle field optimization," *Autonomous Agents and Multi-Agent Systems*, vol. 31, no. 2, pp. 362–385, 2017.
3. X.-S. Yang, S. Deb, S. Fong, X. He, and Y.-X. Zhao, "From Swarm Intelligence to Metaheuristics: Nature-Inspired Optimization Algorithms," *The Computer Journal*, vol. 49, no. 9, Article ID 7562319, pp. 52–59, 2016.
4. L. Wang, H. Geng, P. Liu et al., "Particle Swarm Optimization based dictionary learning for remote sensing big data," *Knowledge-Based Systems*, vol. 79, pp. 43–50, 2015.
5. Y. Gheraibia, K. Djafri, and H. Krimou, "Ant colony algorithm for automotive safety integrity level allocation," *Applied Intelligence*, pp. 1–15, 2017.
6. B. R. Adarsh, T. Raghunathan, T. Jayabarathi, and X.-S. Yang, "Economic dispatch using chaotic bat algorithm," *Energy*, vol. 96, pp. 666–675, 2016.
7. Z.-Z. Liu, D.-H. Chu, C. Song, X. Xue, and B.-Y. Lu, "Social learning optimization (SLO) algorithm paradigm and its application in QoS-aware cloud service composition," *Information Sciences*, vol. 326, pp. 315–333, 2016.
8. D. Wu, S. Xu, and F. Kong, "Convergence Analysis and Improvement of the Chicken Swarm Optimization Algorithm," *IEEE Access*, vol. 4, pp. 9400–9412, 2016.
9. X. S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*, vol. 284, pp. 65–74, Springer, 2010.
10. S. Mirjalili, S. M. Mirjalili, and X.-S. Yang, "Binary bat algorithm," *Neural Computing and Applications*, vol. 25, no. 3-4, pp. 663–681, 2013.
11. Siddheswar Ray and Rose H Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143. Citeseer, 1999.
12. Judong Shen, Shing I Chang, E Stanley Lee, Youping Deng, and Susan J Brown. Determination of cluster number in clustering microarray data. *Applied Mathematics and Computation*, 169(2):1172–1185, 2005.
13. M. Lichman. UCI machine learning repository, 2013.
14. Mehmet Recep Bozkurt, N'ILUFER YURTAY, Ziyet Yilmaz, and Cengiz Sertkaya. Comparison of different methods for determining diabetes. *Turkish Journal of Electrical Engineering & Computer Sciences*, 22(4):1044–1055, 2014.
15. Duen-Yian Y., Ching-Hsue C. and Yen-Wen c., A predictive model for cerebrovascular disease using data mining. *Expert Syst. Appl.*, Vol 38(7), pp. 8970-8977, 2011. doi: 10. 101 6/j.eswa.201 1.01.1 14

CLASSIFICATION PROBLEMS IN SUPERVISED MACHINE LEARNING

Karpagadeepa S, Theinmozhi M. and Vasuki S

Assistant Professor, Thavathiru Santhalinga Adigalar Arts Science and Tamil College, Coimbatore

ABSTRACT

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. Supervised ML is the most developed and popular branch of Machine Learning. The main idea of Supervised ML is that you provide a set of input parameters and the expected result. In this way you “teach” the ML algorithm with correct answers, thus the name “supervised”. Classification problems are very common in practice. Such algorithms provide an answer to the question of whether something is an instance of some limited set of answers or not.

Keywords: Artificial intelligence, Machine learning, SML.

1. INTRODUCTION

Artificial Intelligence is a way of making a computer, a computer-controlled robot, or a software think intelligently, in the similar manner the intelligent humans think.

AI is accomplished by studying how human brain thinks, and how humans learn, decide, and work while trying to solve a problem, and then using the outcomes of this study as a basis of developing intelligent software and systems.

In the real world, the knowledge has some unwelcomed properties

- Its volume is huge, next to unimaginable.
- It is not well-organized or well-formatted.
- It keeps changing constantly.

AI Technique is a manner to organize and use the knowledge efficiently in such a way that

- It should be perceivable by the people who provide it.
- It should be easily modifiable to correct errors.
- It should be useful in many situations though it is incomplete or inaccurate.

AI techniques elevate the speed of execution of the complex program it is equipped with. The ability of a system to calculate, reason, perceive relationships and analogies, learn from experience, store and retrieve information from memory, solve problems, comprehend complex ideas, use natural language fluently, classify, generalize, and adapt new situations. The intelligence is intangible. It is composed of –

- Reasoning
- Learning
- Problem Solving
- Perception
- Linguistic Intelligence

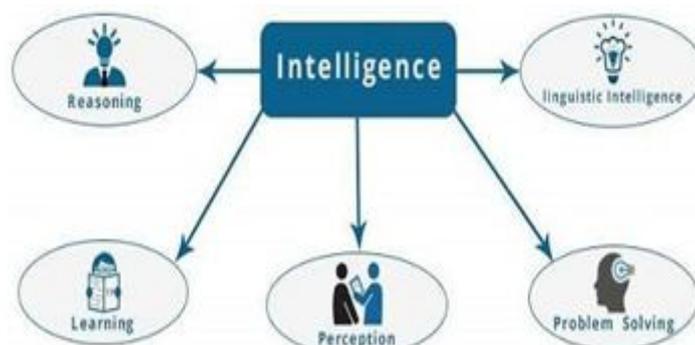


Figure-1.1: Types of intelligence

2.1 MACHINE LEARNING APPROACH

Learning is defined as the acquiring of knowledge or skill, in a particular domain. This definition is related to human beings. In psychology, various generalized definitions of learning have been proposed.

Many of them interpret learning as the change in behavior of a being, subject to a given situation, or as a sequence of his or her repeated experiences in that situation. It is an idea to learn from examples and experience, without being explicitly programmed.

When applying Machine Learning to the same problem, a data scientist takes a totally different approach. Instead of devising an algorithm himself, he needs to obtain some historical data which will be used for semi-automated model creation.

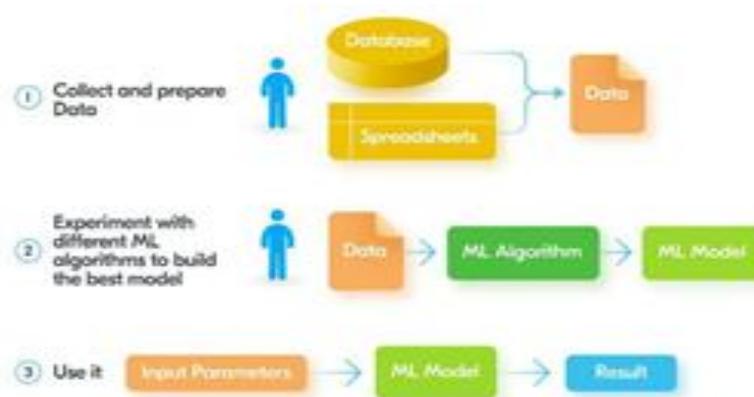


Figure-2.1: Machine learning approach

After obtaining a decent set of data, a data scientist feeds the data into various ML algorithms. The output of any ML algorithm is a model, which can predict new results.

A data engineer can use different knobs to fine-tune the learning algorithm and obtain different models. The model which produces the best results is used in production.

The usage of the ready model is akin to what we have in the traditional programming solution. It gets input parameters and produces a result. The full flow is depicted below:

2.2 EXAMPLES OF MACHINE LEARNING

- Face detection: Identify faces in images (or indicate if a face is present).
- Email filtering: Classify emails into spam and not-spam.
- Medical diagnosis: Diagnose a patient as a sufferer or non-sufferer of some disease.
- Weather prediction: Predict, for instance, whether or not it will rain tomorrow.

2.3 TYPES OF MACHINE LEARNING

Machine learning uses two types of techniques: supervised learning, which trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data.

2.3.1 SUPERVISED MACHINE LEARNING

Supervised ML is the most developed and popular branch of Machine Learning. The main idea of Supervised ML is that you provide a set of input parameters and the expected result. In this way you “teach” the ML algorithm with correct answers, thus the name “supervised”.

To use Supervised Learning your data needs to be “labeled”. That means along with input parameters, the data should contain answers or labels in ML terminology.



Figure-2.3: Building a model with supervised ML

For currency exchange prediction problem the label is the value of the currency exchange rate. In other words, with Supervised ML we provide questions and answers to build a model. After the model is built, we can ask for answers to new questions.

3.1 CLASSIFICATION PROBLEMS IN SUPERVISED ML

Classification problems are very common in practice. Such algorithms provide an answer to the question of whether something is an instance of some limited set of answers or not.

For example, we might have an image and need to “classify” an object on it. Does it have a cat? Does it have a dog? Or for medical diagnostics, the classification determines whether a patient has a certain disease or not.

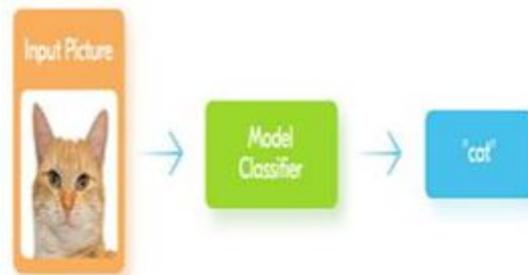


Figure-3.1: Using a model classifier for object detection

A set of images/texts/data and correct set of answers for each of them. The ML algorithm gets those questions and answers and produces a model. Later this trained model is used for making new guesses.

The usual restriction of classification algorithms is that they can provide an answer only to those instances they were trained for. For example, if you provided a lot of images with cats and “labeled” those images as having them, the final model will be able to detect cats on new images. But it will not be able to detect dogs.

3.2 CLASSIFICATION TECHNIQUES:

Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring. Use classification if your data can be tagged, categorized, or separated into specific groups or classes. For example, applications for hand-writing recognition use classification to recognize letters and numbers.

Notwithstanding, the arrangement refreshing has dependably been a testing issue when ABE is utilized to build get to control plans. In this paper, we propose a novel plan that empowering effective access control with dynamic arrangement refreshing for enormous information in the cloud.

We focus on working up an outsourced methodology invigorating strategy for ABE structures. Our strategy can dodge the transmission of scrambled information and limit the calculation work of information proprietors, by making utilization of the beforehand encoded information with old access arrangements.

In image processing and computer vision, unsupervised pattern recognition techniques are used for object detection and image segmentation. Common algorithms for performing classification include support vector machine

(SVM), boosted and bagged decision trees, k-nearest neighbor, Naïve Bayes, discriminant analysis, logistic regression, and neural networks.

3.3 REGRESSION ANALYSIS PROBLEMS IN SUPERVISED LEARNING:

Classification algorithms work only when we have a limited set of possible results. But they can’t help when the output has to be a numerical value we are trying to predict. Consider our currency exchange rate example. We have a set of input parameters and the requirement to predict the numerical value of the exchange rate. So, the exchange rate has an unlimited set of possible answers.

To deal with such problems there are regression analysis algorithms. To apply a regression analysis algorithm data scientist should go through the same process as we described before. It should collect data which contains input parameters and correct answers (labels).

This data is fed to a regression analysis algorithm and it produces a trained model. After getting the model, we can use it for predicting new values using only input parameters.

So, from a high-level perspective, classification and regression analysis algorithms are very similar and differ only in the possible results they can produce.

3.4 THE MOST COMMON USE-CASES FOR REGRESSION ANALYSIS ARE:

- Forecasting stock prices
- Forecasting currency exchange rates
- Estimating real estate prices
- Estimating used car prices
- Predicting energy consumption for buildings
- Retail store sales forecasting

UNSUPERVISED ML

Unsupervised Machine Learning tries to find hidden insights in raw, unlabeled data. In other words, we provide some data, but the data has no answers. This is why it's called "unsupervised" – the unsupervised algorithm should figure something out without being trained like in Supervised Learning.



Figure-3.4: Unsupervised ML working pipeline

4.1 CONCLUSION

In this three directions of ML: Supervised, Unsupervised, and Reinforcement Learning. All of them have some practical application in the real world and each has its own distinct features. Supervised ML is currently the most evolved and practically applicable subset of ML.

To put it in practice you need to have a task which can be formulated as a classification or regression analysis problem and a decent set of labeled data.

There are dozens of ready to use classical ML algorithms and also Deep Learning algorithms to tackle more challenging stuff like image, text, and voice processing.

On the contrary, Unsupervised ML has much less applicability in production. While associative algorithms help in data analysis for retail and online stores, clustering and PCA are more applicable to support Supervised ML algorithms.

5.1 FUTURE SCOPE

Advance we consider the best execution in other instances include

- What kind of flower is displayed in the image?
- What are emotions present in the text message?
- Is it a positive or a negative review?
- Is there a celebrity in the photo?
- is the email a spam or not?

The basic algorithm is the same. We need to have a set of images/texts/data and correct set of answers for each of them. The ML algorithm gets those questions and answers and produces a model. Later this trained model is used for making new guesses.

REFERENCES

1. Saloky, T.: Applications of Artificial Intelligence Techniques. Košice: Elfa, Košice, 1993. ISBN 80-7079-227-1
2. Góts, I.- Saloky, T.: Scientific Research and Pedagogical Aspects of the Doctoral Study at the Faculty of Mechanical Engineering. In Proceedings of AE in ACT'94. International Workshop on Advanced Education and Control Technology. Prague: CTU, Prague, 1994, pp. 31-34
3. Saloky, T.: Applications of Machine Learning Techniques. Košice: Elfa, Košice, 1998. ISBN 80-88786-73-8
4. Sergii boiko: Practical Machine Learning for Solving Real World Problems, 2018.
5. Susankahler: sas data science, 2018
6. Mandeep Sidana, Technology Consultant, Sifium Technologies, 2016

DATA MINING AND ITS IMPLICATIONS**B. Saravana Kumar¹ and Dr. P. Senthil Vadivu²**Assistant Professor, K. S. G College of Arts & Science, Coimbatore
Professor & Head, Hindusthan College of Arts & Science, Coimbatore

ABSTRACT

Data plays an important role in today's business world. Database maintenance and relevant usage of data for appropriate tasks are vital factors for running a successful business. Data mining helps in supporting business objectives to the modern corporate world.

Keywords: Data mining tools, techniques, algorithms in data mining

INTRODUCTION

Data mining is about processing the data and identifying patterns and trends in that information which helps in reliable judgment of business objectives. Data mining tools, techniques and algorithm procedures helps in implementing prompt database protocols for successful business. Data mining analysis is performed by using properties of the focus of analysis. It is mainly used in business strategies. Data analysis helps in decision making in the ongoing business and it also helps for prediction of future. It is used in the betterment of business solutions.

DATA MINING

Data mining is the method of extracting potentially useful information from the data sets. It uses the suite of methods to organize, examine and combine large data sets, including machine learning, visualization methods and statistical analysis. The final goal of data mining is to produce concrete information from data and interpret information which is practically applied. The factors influencing data mining includes managing databases and data pre-processing.

WORKING OF DATA MINING PROCESS

Modeling is an important step in the data mining process. A model which could possibly be a mathematical relationship or a certain sampling of data- is created from the data in question, and this model can be applied to situations related to those which produced the original data in which the desired data has not yet been discovered. Although modeling has been known to mathematicians for centuries, its potential was previously limited by the absence of adequate computational resources. The techniques that have been grasped or created for data mining purposes include a variety of different tools. These tools include the genetic algorithm, artificial neural networks, decision trees, nearest neighbor, and the idea of rule induction.

THE PRODUCTS OF DATA MINING

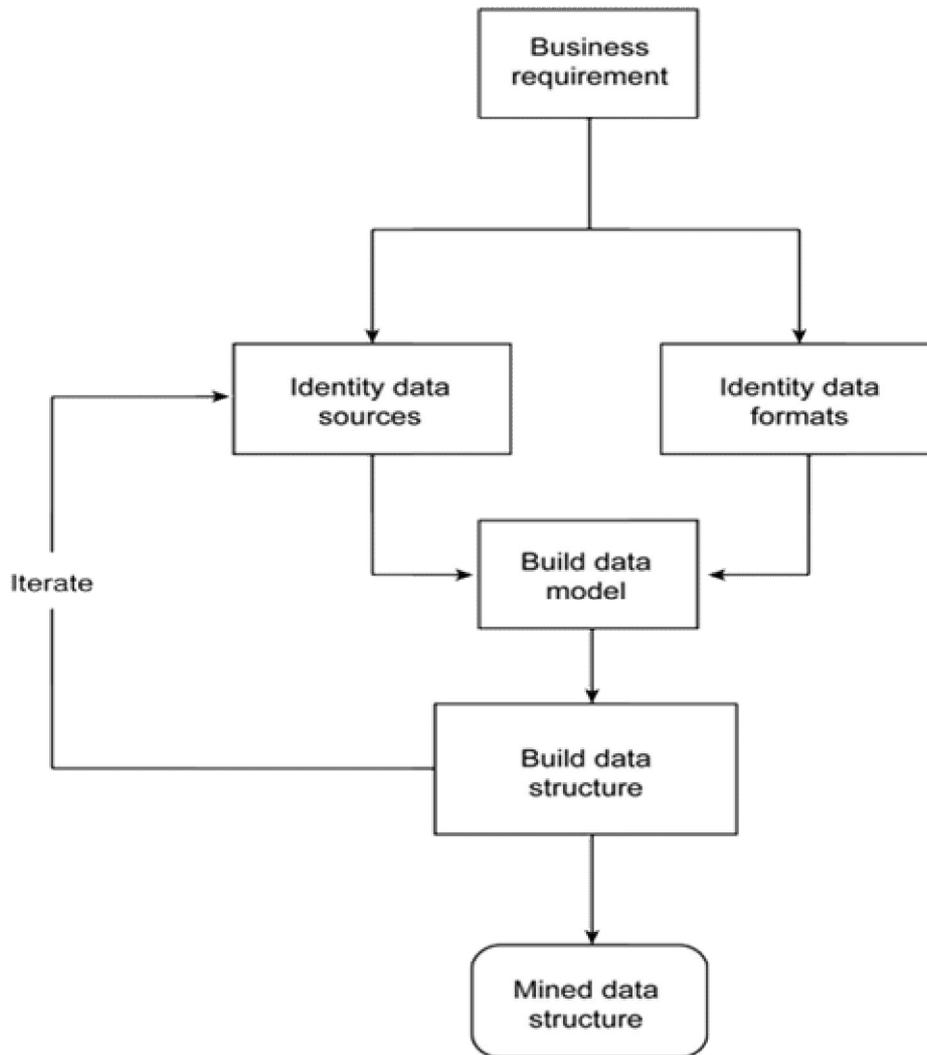
Data mining is still being developed, but its accomplishments to date are already impressive. Companies and institutions working in a wide range of fields can use this technique to find ways to put their data to use to make themselves more profitable and successful. Industries as diverse as retail, health care, aerospace and finance have already begun to pick up on and grasp the potentials of data mining. The following are the features of data mining which it can already do:

- Finding the characteristics of customers who are likely to buy certain products.
- Predict which customers will cut ties with a given company and begin working with a competitor.
- Warn of potentially fraudulent transactions.

GLOBAL DATA MINING

Global data mining is used to examine data from the world to create a wider range of new and useful content depending on the situation. It helps the companies to develop plans and actions on a global scale. It helps to save time, money and other resources. There is a price difference depending on how much data a company wants and for how long they are in need of such information. Mining helps the companies to save money and makes their customers to save more money as they can drop product prices. In some countries it is illegal to use data mining for anything other than "scientific research." Many European countries have this law in effect. There are many different ways to use global data mining such as for games, business, science and engineering, human rights and medical research.

OUTLINE OF DATA MINING PROCESS



DATA MINING TOOLS

- Rapid miner
- Orange
- Weka
- Knime
- Sisense
- SSDT (SQL server data tools)
- Apache Mahout
- Oracle Datamining
- Rattle
- DataMelt
- IBM Cognos
- IBM SPSS Modeler
- SAS Data Mining
- Teradata
- Board
- Dundas BI

ADVANTAGES OF DATA MINING TOOLS

Data mining is performed with comparatively modest database systems and simple tools and software packages. The process of extract, transformation and loading (ETL) of data is performed using data mining tools. The ETL process eases complex data transactions which are involved in business.

KEY TECHNIQUES OF DATA MINING

The techniques and technologies from the database management, statistics and machine learning have focused better understanding as how to process and draw conclusions from the vast amount of information. The following techniques are used in data mining:

➤ Association

Association (or relation) is probably the better known and most familiar and straightforward data mining technique. Here, we can make simple correlation between two or more items, often of the same type to identify patterns. For example, when tracking people's buying habits, we might identify that a customer always buys cream when they buy strawberries, and therefore suggest that the next time that they buy strawberries they might also want to buy cream.

➤ Classification

Classification is a more complex data mining technique that helps the user to collect various attributes together into discernable categories, which can be used to draw further conclusions or serve some function. For example, if you're evaluating data on individual customers' financial backgrounds and purchase histories, you might be able to classify them as "low," "medium," or "high" credit risks. You could then use these classifications to learn even more about those customers.

➤ Clustering

Clustering is very similar to classification, but involves grouping chunks of data together based on their similarities. For example, a data analyst can choose to cluster different demographics of the audience into different packets based on how much disposable income they have, or how often they tend to shop in a particular store.

➤ Prediction

Prediction is one of the most valuable data mining techniques, since it's used to project the types of data in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, a data analyst can review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future.

➤ Sequential patterns

Often used over longer-term data, sequential patterns are a useful method for identifying trends, or regular occurrences of similar events. For example, with customer data an analyst can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, analyst can use the information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history.

➤ Decision trees

Related to most of the other techniques (primarily classification and prediction), the decision tree can be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer.

➤ Combinations

In practice, it's very rare that we would use one of these exclusively. Classification and clustering are similar techniques. By using clustering to identify nearest neighbors, we can further refine his classifications. Often, we use decision trees to help build and identify classifications that we can track for a longer period to identify sequences and patterns.

➤ Tracking patterns

One of the most basic techniques in data mining is learning to recognize patterns in data sets. This is usually recognition of some aberration in data happening at regular intervals, or an ebb and flow of a certain variable over time. For example, you might see that your sales of a certain product seem to spike just before the holidays, or notice that warmer weather drives more people to your website.

➤ Regression

Regression is used primarily for planning and modeling. It is used to identify the likelihood of a certain variable, given the presence of other variables.

For example, an analyst could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression's main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

➤ Outlier detection

In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data. For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process.

ALGORITHMS IN DATA MINING

An *algorithm* in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends. The algorithm uses the results of this analysis over various iterations to find the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics.

➤ Choosing the Right Algorithm

Choosing the best algorithm to use for a specific analytical task can be a challenge. While we can use different algorithms to perform the same business task, each algorithm produces a different result, and some algorithms can produce more than one type of result. For example, you can use the Microsoft Decision Trees algorithm not only for prediction, but also as a way to reduce the number of columns in a dataset, because the decision tree can identify columns that do not affect the final mining model.

➤ Choosing an Algorithm by Type

SQL Server Data Mining includes the following algorithm types

- **Classification algorithms** predict one or more discrete variables, based on the other attributes in the dataset.
- **Regression algorithms** predict one or more continuous numeric variables, such as profit or loss, based on other attributes in the dataset.
- **Segmentation algorithms** divide data into groups, or clusters, of items that have similar properties.
- **Association algorithms** find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis.
- **Sequence analysis algorithms** summarize frequent sequences or episodes in data, such as a series of clicks in a web site, or a series of log events preceding machine maintenance.

However, there is no reason that we should be limited to one algorithm to obtain solutions. Experienced analysts will sometimes use one algorithm to determine the most effective inputs (that is, variables), and then apply a different algorithm to predict a specific outcome based on that data. SQL Server Data Mining lets us to build multiple models on a single mining structure, so within a single data mining solution an analyst can use a clustering algorithm, a decision trees model, and a Naïve Bayes model to get different views on your data. Analyst might also use multiple algorithms within a single solution to perform separate tasks: for example, we could use regression to obtain financial forecasts, and use a neural network algorithm to perform an analysis of factors that influence forecasts.

DATA IMPLEMENTATIONS AND PREPARATION

Data mining itself relies upon building a suitable data model and structure that can be used to process, identify, and build the information that we need. Regardless of the source data form and structure, analyst should organize the information in a format that allows the data mining to take place in as efficient a model as possible.

CONCLUSION

Thus data mining plays an important role in the modern day scenario, whereby large and voluminous data are handled in the best possible manner, which helps to develop the business sector and other areas whereby people have to work with the data, reformat it, or restructure it, regardless of the tools and techniques.

REFERENCES

1. Kamber Pei, Jaiwei Micheline Jian, “Data Mining: Concepts and Techniques”, 3 rd edition, June 2011.
2. Kobielus, James, “The Forrester Wave: Predictive Analytics and Data Mining Solutions”, Forrester Research, July 2008.

ESSENTIALS OF CLOUD COMPUTING WITH ITS CHALLENGES

S. Kowsalya¹, B. Reethu² and S. Suprraja³

Assistant Professor¹ and Student^{2,3}, Sri Krishna Arts and Science College, Coimbatore

ABSTRACT

Cloud computing is an advanced development in the recent days. It has formed the conceptual and infrastructural basics for tomorrow's computing. The technology has brought a new change in the field of education, business and in IT industries. Now a days most of the industries store information and data in the cloud because the data size is huge and maintenance of such large amount of data is very difficult, so we store our data in the cloud. Even though the technology is developed we come across some privacy issues and data security problems during the storage of data into the cloud. In this paper, we will discuss about the concepts as well as current privacy issues, security problems and the solution methodology to sort out the current problems within the context of cloud computing which are undergone by the industries.

Keywords: Cloud Architecture, Cloud Computing, Cloud Service Providers, Data Security, Privacy Issues.

IV. INTRODUCTION

Cloud computing has become the emerging issues in the recent years. The IT infrastructure has completely changed. In the past, an expensive infrastructure has to be deployed in order to store the operational data of the organization. In last decades, they used Relational Databases to store the data and two or more servers are located inside the organization and the client needed to request data from these server machines. This was very costly since the organization was indeed of hiring many personnel for deploying, maintaining and managing the cloud infrastructure. From this cloud computing environment, on demand services such as resources and applications are provided through the internet. Software and hardware are the two cloud environment resources in the data centre.

Organizations are under the need of cloud due to the abundant growth in the resources, Data retrieval and ability is also offered by cloud computing. Due to the demand for the cloud the storage of data and information is less safe and cannot be accessed by different user. The purpose of the paper will bring out the solutions for the privacy issues and security constraints in cloud.



Fig-1: Cloud Computing Infrastructure

V. CLOUD COMPUTING INFRASTRUCTURE

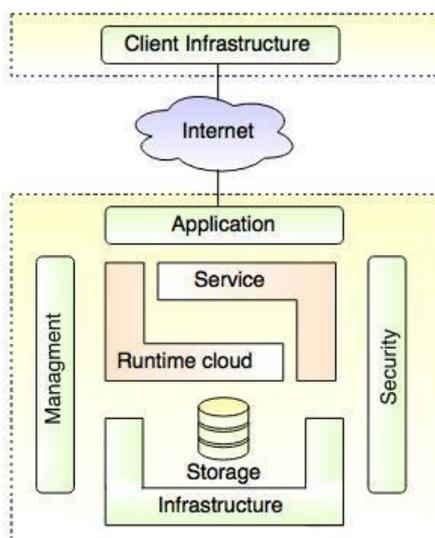


Fig-2: Cloud Computing Architecture

Cloud infrastructure refers to the hardware and software components such as servers, network and virtualization software that are needed to support the computing requirements of a cloud computing model. Service providers or IT departments host their resources virtually and deliver to the user through network or internet. The term cloud computing is evaluated from the distributed computing and grid computing. In recent years, the cloud has divided into two types – to rent the infrastructures in cloud, or to rent any specific service in the cloud. Hardware and software usage where the former one who deals with the cloud directly, then the later one is confined with the ‘soft’ products or services from the cloud service and infrastructure.

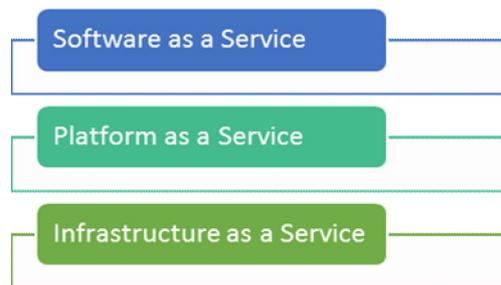


Fig-3: Cloud Services

There are three types of terminology in the computing world they are

- Software as a Service (SaaS)
- Platform as a Service (PaaS)
- Infrastructure as a Service (IaaS)

A. Software as a Service (SaaS)

The cloud service providers will provide some software to the client. So that, the client can use all the resources provided in the cloud service providers. Through web browser the client can access resources. The client can only receive the information it do not have any control over the setting of the cloud, network and storage.

Major benefits of SaaS are

- Rapid Scalability
- Accessibility from any location with Internet
- Eliminates infrastructure concerns
- Custom levels of service offerings
- Bundled maintenance and Support

B. Platform as a Service (PaaS):

The client will provide us the software tool to display and install the customization applications which are provided by the cloud service provides. Physical settings are controlled it is restricted by the user or client server. The client can have the rights of changing the application setting provided by the cloud service provider.

Major Benefits of PaaS are

- Community – Most of the time, many people are involved in building cloud applications in PaaS environments. This creates a strong supportive community that can help your development team along the way.
- No more upgrades – Companies are not required to update or upgrade the infrastructure software. Instead, the PaaS provider handles all upgrades, patches and routine software maintenance.
- Lower cost – Companies face lower risk since they do not have to make upfront investment in hardware and software.
- Simplified deployment – The development team can concentrate on developing the cloud application without having to worry about the testing and deployment infrastructure.

C. Infrastructure as a Service (IaaS)

Cloud providers run the cloud software which is necessary to make computing resources. Cloud services such as Amazon EC2 use this kind of model and change the resources according to the utility.

Major benefits of IaaS are

- Reduces total cost of ownership and capital expenditures
- Users pay for the service that they want, on the go
- Access to enterprise-grade IT resources and infrastructure
- Users can scale up and down based on their requirements at any time

VI. ESSENTIAL CHARACTERISTICS OF CLOUD

A. On Demand Self Services

Customers are able to get resources without any interaction with the cloud providers. Example: email account.

B. Broad Network Access

Customers are able to access the resources from anywhere at any time over networks such as the internet using a client such as mobile phone, laptop from a range of limit devices. Example: Mobile Phones, Laptop, and Desktop.

C. Resource Polling

The cloud provider resources are polled to some multiple customers. Typically virtualization technology are used to facilitate multi-tenancy [4] and enable computing resources to be dynamically assigned and reallocated based on customer demand.

D. Rapid Elasticity

Resources can be quickly provisioned and related, sometimes automatically based on demand. Customers can easily increase or decrease their use of a cloud service to meet the current needs.

E. Measured Services

Customers pay only for the resources they actually used within the service. Typically the service providers will supply customers with a dashboard so they can track their usage.



Fig-4: Cloud Characteristics

VII. DEPLOYMENT MODEL

A. Public Cloud

It is open to the public and everyone can use it. It is developed for the public to use it and satisfy their needs. It could be owned by government or private providers.

B. Private Cloud

This organization is exclusively used for the organizations needs we must pay to the particular organization for what we have used. It can be either on premises and off-premises.

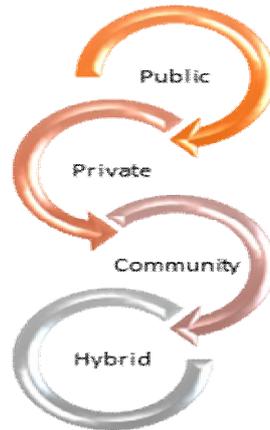


Fig-5: Deployment of Cloud

C. Community Cloud

It is an infrastructure which is stored between the government and organization. This community cloud could be on premises or off-premises

D. Hybrid Cloud

This model is a combination of two or more deployment models such as private, community or public. A cloud can be considered as hybrid cloud if the data moves from a data centre to a private cloud or public cloud and vice versa.

VIII. ADVANTAGES OF CLOUD SERVICES

It only takes you a few minutes or hours to set up a cloud service application with robust features. One can access the cloud service from any computing device attached to the internet including smartphones, tablets and laptops. If we have to access a service, we can do it from anywhere; home, at the airport, at the office, etc.

Cloud services are also scalable. As a company grows, one can increase the subscription to the resources need to go in line with the growth. The same is the case when the company’s operations shrink. Most cloud providers have packages that allow customers to choose the level of service they need.



Fig-6: Advantages of cloud services

Finally, companies that adopt cloud services usually benefit from improved efficiency and lower costs. This gives them an edge over their competitors and can help them to quickly gain market share. For start-ups, cloud services can help them to become profitably quickly without risking investments in robust in-house IT infrastructure, hardware or software.

Businesses should research well to understand their exact needs and find providers who offer the solutions that will work for them.

IX. DATA SECURITY

Every technology has both advantage and disadvantage. All the advantages are listed above and here are the drawbacks in the cloud security and privacy. The first and most thing is loss of data, the companies store the data on to the cloud. The cloud service providers must keep the data in the safe manner but due to the low cost the cloud providers are enabled to create a secured storage of the cloud in this case the important data in the cloud can be lost. Cloud migration is an another major drawback in the cloud, all the business data, important files and other elements are transferred by the desktop or server to the cloud through the internet.

X. PRIVACY ISSUES

Long days back the cloud service providers used the traditional way of storing the data as the technology has grown the cloud services has also developed. The data which are stored in the cloud must be secured from the data encryption process, the cloud service providers must create a tool to enhance the data security

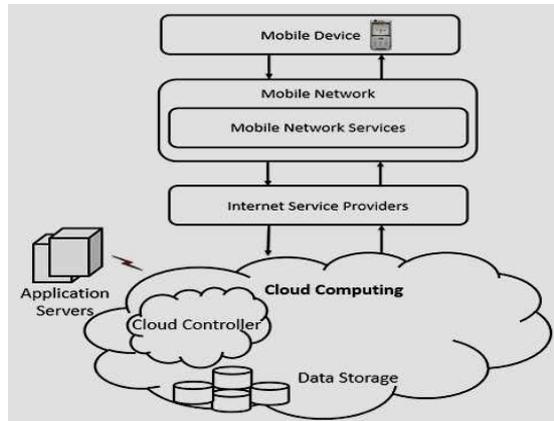


Fig-7: Security and Privacy Issues of Cloud

XI. CHALLENGES IN CLOUD

Cloud computing challenges are numerous and thorny, to be sure. These days, everyone is in the cloud — but that doesn't mean that they've figured out how to overcome all the challenges of cloud computing.

Cybersecurity experts are even more concerned about cloud security than other IT staffers are. A 2018 Crowd Research Partners survey found that 90 percent of security professionals are concerned about cloud security. More specifically, they have fears about data loss and leakage (67 percent), data privacy (61 percent) and breaches of confidentiality (53 percent).

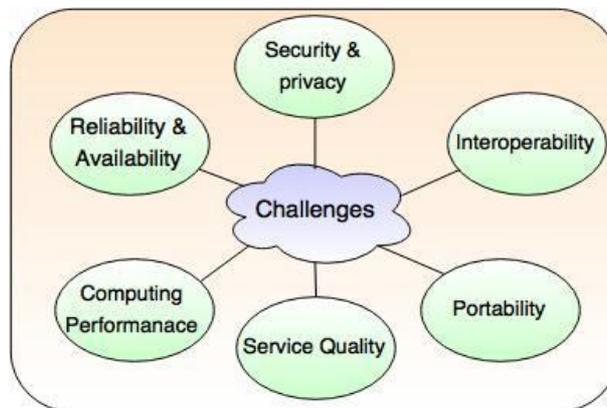


Fig-8: Cloud Computing Challenges

Most organizations aren't using just one cloud. According to the RightScale findings, 81 percent of enterprises are pursuing a multi-cloud strategy, and 51 percent have a hybrid cloud strategy (public and private clouds integrated together). In fact, on average, companies are using 4.8 different public and private clouds.

Multi-cloud environments add to the complexity faced by the IT team. To overcome this challenge, experts recommend best practices like doing research, training employees, actively managing vendor relationships and re-thinking processes and tooling.

While launching a new application in the cloud is a fairly straightforward process, moving an existing application to a cloud computing environment is far more difficult.

More specifically, many of the companies migrating applications to the cloud reported time-consuming troubleshooting (47 percent), difficulty configuring security (46 percent), and slow data migration (44 percent), trouble getting migration tools to work properly (40 percent), and difficulty syncing data before cutover (38 percent) and downtime during migration (37 percent).

To overcome those challenges the IT leaders surveyed said they wished they had performed more pre-migration testing (56 percent), set a longer project timeline (50 percent), hired an in-house expert (45 percent) and increased their budgets (42 percent).

XII. FUTURE WORKS

Cloud computing is a new and widely emerging domain and it must have to overcome all the security issues and privacy problems in order to make it as an important technology for the future use even though there are many problems and issues which are unseen, it could be a gateway for the future developers and the door for the future research are always open.

Cloud Technologies are mostly virtualization environments. Aspirants seeking to make good in the Cloud need to be adept in networking and virtualization and gain hands-on exposure with live deployments. The roles in Cloud Computing might range from cloud developers to operators. Every role comprises of the knowledge of the cloud computing basics and certain domain specific skills.

XIII. CONCLUSION

The user must have knowledge about the cloud service provider; they must look across the reviews and comments.

Available space: First and most things is, the user must check the available space in the cloud to store the data.

If in case the data is lost cloud service provider must have a backup of all the data and retention strategy.

This paper is about the data security and privacy issues on the cloud. It is noted that, the cloud computing technology has rapidly grown and the usage of the cloud is increased. Due to the insufficient storage space and security problem the organization are facing many problems it is mainly on the cloud service providers hand to bring out a better change in the cloud services. It can be concluded that the data security and privacy issues are the two major problems.

REFERENCES

- A. Karnik, "Performance of Cloud based data retrieval" M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
- FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
- J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Cloud Computing*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
- M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.
- M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: <http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/>
- R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
- S. M. Metev and V. P. Veiko, *Cloud Computing and its technology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
- S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *Intrinsic complexity in cloud.*, vol. 20, pp. 569–571, Nov. 1999.

AUTOMATIC TIMETABLE GENERATOR**Aiswarya Lekshmi¹ and Dr. A. V. Senthil Kumar²**Student¹ and Director², Hindusthan College of Arts & Science, Coimbatore

ABSTRACT

In this system we intend to display the Timetable as per the faculty is been allotted for the respective subject in accordance to that rooms will be available. The manual system of preparing time table in colleges with large number of students is very time consuming and usually ends up with various classes clashing either at same room or with same teachers having more than one class at a time. These are just due to common human errors which are very difficult to prevent in processes such as these. To overcome these problems people usually taking the previous years' timetable and modifying it but still it is a tedious job to incorporate changes. To overcome all these problems we propose to make an automated system. The system will take various inputs like details of students, subject sand class rooms and teachers available, depending upon these inputs it will generate a possible time table, making optimal utilization of all resources in a way that will best suit any of constraints or college rules. List of subjects may include electives as well as core subjects. The case is similar to schools and other educational institutions. So our aim is to develop a general purpose which can efficiently generate optimal solutions.

Keywords: Automatic Time Table, Assign Staffs for classes, View Time Table.

1. INTRODUCTION

Automatic Timetable Generator is a .NET based software used to generate timetable automatically. Currently timetable is managed manually. It will help to manage all the periods automatically. Automatic time table generator is a .NET based software used to generate timetable automatically. Currently time table scheduling was done manually with a single person or some group involved in task of scheduling it manually, which takes a lot of effort and time. Planning timetables is one of the most complex and error-prone applications. It will help to manage all the periods automatically. Maximum and minimum work load for a Faculty for a day, week and month will be specified for the efficient generation of table. It is a comprehensive timetable management solutions for colleges which help to overcome the challenges in manually setting the timetable.

2. OVERVIEW

Genetic algorithms are general search and optimization algorithms inspired by processes and normally associated with natural world. Genetic algorithm mimics the process of natural selection and can be used as a technique for solving complex optimization problems which have large spaces They can be used as techniques for solving complex problems and for searching of large problem spaces. Unlike many heuristic schemes, which have only one optimal solution at any time, Genetic algorithms maintain many individual solutions in the form of population. Individuals (parents) are chosen from the population and are then mated to form a new individual (child). The child is further mutated to introduce diversity into the population [10]. Rather than starting from a single point within the search space.

3. AUTOMATIC TIMETABLE GENERATOR

1. Admin
2. Staff

The Admin contains

- Add subject
- Add staff

The staff contains

- Register
- Login
- View timetable

1. Admin

An admin can create the timetables and update timetables. The Admin will finalize the timetable for respective classes. He will make sure that how many periods will be there in a day, what subjects are to be covered in a day, how many periods should be there for a particular subject in a week. Admin contains:

- Add subject

The admin can add subjects to the automatic timetable generator.

- Add staff

The admin can add staffs to the automatic time table generator.

2. Staff

The faculty are log into the application by providing respective login credentials (user name & password). On submitting they are re-directed to the faculty home page. The faculty members were given rights to view the existing pre generated time table. The staff contains:

- Register

Faculties can register to the application.

- Login

Faculties can login to the application using their respective login id and password.

- View timetable

Faculties can view the timetable after login to it.

3.EXISTING SYSTEM

The time table management system is a very old process. In every schools and colleges, the generation of the time table is required. Till now, all the processes are manually done in most of the colleges because no one wants to spend the money on such process. Some time table management system are available but some cost are charged for using them. The existing systems are not user friendly because they require lots of data to generate. It is very hard to handle the database in the presently available system because it is computer based applications. There is no web based time table management system available till now. So, all the features of the web based applications are missing. Security is main factor for any application.

In the presently available system, the security level is very low because there is no criterion of login id and password. Database management is also very difficult in the desktop based application. To use this application the expert is requires. In our system we have tried to solve all these problems.

DISADVANTAGES

1. Inconsistency in data entry, room for errors, miskeying information.
2. Large ongoing staff training cost.
3. System is dependent on good individuals.
4. Reduction in sharing information and customer services.
5. Time consuming and costly to produce reports.
6. Lack of security.
7. Duplication of data entry.

4. PROPOSED SYSTEM

The objective of the project is to create a web based time table management system to be used by any college but basically it is created according to the needs of ou. Its main is to perform the basic requirements of my department and to maintain the databases of subjects and staffs more efficiently. As it is a web based system, the maintenance and working is somehow difficult from the manual mode of approach. Till now, the time table making process is manual work. So because of this the teacher has some difficulty to work on this. They have to first login through their id and then they have to fill all the details then time table will be generated. But we have tried hard to make this system very user friendly and easy to understand.

As this is web based project, it is easier to fetch the data from the database and remove the unwanted data by just clicking the delete button. This project will limit the time and money factor involve in the time table management system. The maintenance of this web based is much easier and accurate than the existing manual system. As this web based system, the security features are somehow higher than the manual system. In this system, there is less chance of mishandling of the data because only the administrator will login through their login id and password and upload all the related data and generate the time table. So there is no misuse of the data.

ADVANTAGES

1. Eliminate paper-based process
2. Intuitive & user-friendly
3. Automatic Timetable Scheduling
4. Generate multiple timetables
5. Customization and flexibility
6. Optimal resource allocation
7. Highly secure

5. CONCLUSION

Automatic timetable generator is a web based application for generating timetable automatically. It is a great difficult task that to manage many faculties and allocating subjects for them at a time manually & proposed system will help to overcome this disadvantage. Thus we can generate timetable for any number of courses and multiple semesters. This system will help to create dynamic pages so that for implementing such an system we can make use of the different tools are widely applicable and free to use also.

6. FUTURE SCOPE

FUTURE SCOPE To generate timetable for the institute which will be less time consuming and free of human errors along with high level of efficiency and precision. Moreover improve the overall process of timetable generation with help of genetics algorithm along with the assistance of technology

CREDIT CARD FRAUD DETECTION SYSTEM**Abidh Ali A¹ and Dr. A. V. Senthil Kumar²**Student¹ and Director², Hindusthan College of Arts & Science, Coimbatore

ABSTRACT

Credit-card-based purchases can be categorized into two types: 1) physical card and 2) virtual card. In a physical-card based purchase, the cardholder presents his card physically to a merchant for making a payment. To carry out fraudulent transactions in this kind of purchase, an attacker has to steal the credit card. If the cardholder does not realize the loss of card, it can lead to a substantial financial loss to the credit card company. In the second kind of purchase, only some important information about a card (card number, expiration date, secure code) is required to make the payment. Such purchases are normally done on the Internet or over the telephone. To commit fraud in these types of purchases, a fraudster simply needs to know the card details. Most of the time, the genuine cardholder is not aware that someone else has seen or stolen his card information. The only way to detect this kind of fraud is to analyze the spending patterns on every card and to figure out any inconsistency with respect to the "usual" spending patterns. Fraud detection based on the analysis of existing purchase data of cardholder is a promising way to reduce the rate of successful credit card frauds. Since humans tend to exhibit specific behaviorist profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system.

Keyword: physical card, virtual card, financial, behaviorist profiles

1. INTRODUCTION

System analysis is concerned with the comparison study about the existing system and proposed system. The system analysis is essential when the software must interface with other elements such as software, people and other resources. The essential purpose of the place is to find the need and to define the problem that needs to be solved

Core Features

- The system stores previous transaction patterns for each user.
- Based upon the user spending ability and even country, it calculates user's characteristics.
- More than 20 -30 % deviation of users transaction (spending history and operating country) is considered as an invalid attempt and system takes action.

Along with increasing credit cards and growing trade volume in China, credit card fraud rises sharply. How to enhance the detection and prevention of credit card fraud becomes the focus of risk control of banks. This paper proposes a credit card fraud detection model using outlier detection based on distance sum according to the infrequency and unconventionality of fraud in credit card transaction data, applying outlier mining into credit card fraud detection. Experiments show that this model is feasible and accurate in detecting credit card fraud.

The importance of Machine Learning and Data Science cannot be overstated. If you are interested in studying past trends and training machines to learn with time how to define scenarios, identify and label events, or predict a value in the present or future, data science is of the essence. It is essential to study the underlying data and model it by selecting an appropriate algorithm to approach any such use case. The various control parameters of the algorithm need to be tweaked to fit the data set. As a result, the developed application improves and becomes more efficient in solving the problem.

2. BACK GROUND**2.1 Description of the Case**

In our partner bank, a rule based fraud detection system was in place. These rules were generated from a list provided by the credit cards software vendor (who has many 163 international customers) and from the own experience of bank's Customer Security and Fraud Department. The rules were basically of two types: online and offline. Online rules produce alerts to be inspected by inspection officers who work 7 days and 24 hours. Suspicious transactions appear on their screen as a list where the most suspicious ones appear at the top. The inspector can reach the other details and previous transactions of the cardholder and then decide either the transaction is legitimate or check the transaction with the customer (by sending SMS or by calling). Offline rules are also known as hidden rules and they automatically send SMSs to customers that basically include the message "from your card such and such transaction has been made, if it is out of your consent please call your

bank.” Most of the fraud cases are identified when the customers make a reply to these SMSs. For offline rules and also for most online rules, the bank first gives provision to the transaction then, if a fraud is captured, the card is blocked to prevent any further frauds. The partner bank in this project has been using data analysis/mining solutions for some years and a lot of historical information regarding customers and their credit card usages had been collected. Unfortunately, however, these data marts had been developed mostly for marketing and risk purposes, and they provided only limited information for fraud detection. We thus had to invent and introduce new variables that would be directly beneficial in the fraud detection problem. The number of transactions for each card differs from one to other; however, each transaction record is of the same fixed length and includes the same fields. These fields include transaction date and hour, transaction amount, transaction type, transaction channel, MCC code, merchant address, and the currency used in the transaction [15]. While transaction type shows whether this transaction is a purchase or a cash advance, transaction channel depicts the medium of the transaction (POS, web, ATM, etc.), and MCC code displays the type of the merchant store where the transaction takes place. Each merchant type has a specific four-digit number assigned, for example, all pharmacies have one common MCC code. Sometimes larger merchants can be assigned a unique code, for example, while all small hospitals are assigned the same code, a large and well-known hospital can have its own code. As we anticipated, the information associated with a transaction is quite limited and thus more information (variables) is needed to build a better model. As such, the variables that form the card usage profile and the methods used to build the model make the

difference in the fraud detection systems. Our aim in defining the variables used to form the data mart is to discriminate the profile of the fraudulent card usage by the fraudsters from the profile of legitimate (normal) card usage by the actual card holders. The variables are one of the six main types: statistics on all transactions of cardholders, statistics on regions of transactions, MCC statistics, daily amount statistics, channel usage statistics, and daily number of transactions statistics. The existing fraud detection solution was a quite successful one as compared to other banks in Turkey (this, we know from our experiences in other banks). We would like to note that Turkey is well known with its advanced banking IT infrastructures and software systems in Europe (as was stated many times during the bank acquisitions and mergers in the last decade). This fraud solution was composed of some expert rules based on the experiences of two credit card software vendors and on the own experiences of the bank. With these rules some groups of transactions that are more likely to include suspicious ones are determined (according to expert opinions or information obtained from the official organization Inter Banks Card Center). For example, since internet transactions are statistically (and logically) more risky, they were sending out SMSs to all internet transactions. Later, by an improvement study they wanted to reduce the number of SMSs since although one individual SMS is very cheap (about two dollar cents), millions of SMSs sum up to considerable amounts. In this regard, they decided not to send SMSs to transactions whose amounts are less than 50 dollars. From our perspective such a reduction in SMSs can be quite risky in that (1) the real risk the bank is faced to is the remaining available limit on the card rather than the amount of current transaction, and (2) sometimes fraudsters test counterfeit cards by smaller transactions and if everything is okay they continue with larger ones. Thus, one needs to track the remaining card limits to prevent more risk as much as possible.

3. EXISTING SYSTEM

In existing scheme the validation of the card was verified at the point terminal of purchase. The card verifier terminal was not directly integrated to banking /financial institution transaction process. The point terminal system verify only card number validity. The user perspective details such as transactional limit of the customer, customer's frequency in using the credit cards, etc were not revealed in the existing scheme. The credit card validity process of the existing scheme was more time consuming. The process operation of the existing system was more complex. In order to migrate to smart card based payment systems, Banks will have to make a number of changes to their existing systems. Among these are:

DRAWBACKS

The drawbacks of the existing system are

In existing scheme the validation of the card was verified at the point terminal of purchase.

The card verifier terminal was not directly integrated to banking /financial institution transaction process.

The user perspective details such as transactional limit of the customer, customer's frequency in using the credit cards, etc were not revealed in the existing scheme.

4. PROPOSED SYSTEM

1. Due to Behavior and location analysis approach like IP address 192.168.1.1 tracing, there is a drastic reduction in the number of False Positives transactions identified as malicious by an FDS although they are actually genuine.
2. The system stores previous transaction patterns for each user.
3. Based upon previous data of that user the system recognizes unusual patterns in the payment procedure.
4. The System will block the user for more than 3 invalid attempts

5. CONCLUSION

Conclusion Clearly, credit card fraud is an act of criminal dishonesty. This article has reviewed recent findings in the credit card field. This paper has identified the different types of fraud, such as bankruptcy fraud, counterfeit fraud, theft fraud, application fraud and behavioral fraud, and discussed measures to detect them. Such measures have included pair-wise matching, decision trees, clustering techniques, neural networks, and genetic algorithms. From an ethical perspective, it can be argued that banks and credit card companies should attempt to detect all fraudulent cases. Yet, the unprofessional fraudster is unlikely to operate on the scale of the professional fraudster and so the costs to the bank of their detection may be uneconomic. The bank would then be faced with an ethical dilemma. Should they try to detect such fraudulent cases or should they act in shareholder interests and avoid uneconomic costs? As the next step in this research program, the focus will be upon the implementation of a 'suspicious' scorecard on a real data-set and its evaluation. The main tasks will be to build scoring models to predict fraudulent behavior, taking into account the fields of behavior that relate to the different types of credit card fraud identified in this paper, and to evaluate the associated ethical implications. The plan is to take one of the European countries, probably Germany, and then to extend the research to other EU countries.

PUBLIC DISTRIBUTION SYSTEM**Viswapal Venugopal¹ and Dr. A. V. Senthil Kumar²**Student¹ and Director², Hindusthan College of Arts & Science, Coimbatore

ABSTRACT

Public distribution System in the country facilitates the supply of food grains to the poor at a subsidized price. Essential items such as selected cereals, sugar and kerosene at subsidized prices to holders of ration cards is the objective of efficient Public Distribution System. The PDS also helps to modulate open - market prices for commodities that are distributed through the system. Government accords great importance to the objective of measuring outcomes of PDS so as to ensure that equal distribution system serves up the purpose for which it was set up. The Authorized Wholesale Distributors are appointed by the District Collectors while the Retail Ration Distributors are appointed by the District Supply Officers. The Authorized Retail Distributors distribute rationed articles to the cardholders on production of the ration cards issued by the taluk Supply Officers/City Rationing Officers at the scale and price fixed by the Government from time to time. PDS is operated under the joint responsibility of the Central and the State Governments. The Central Government has taken the responsibility for procurement, storage, transportation and bulk allocation of food grains, etc. The responsibility for distributing the same to the consumers through the network of Fair Price Shops (FPSs) rests with the State Governments.

Keywords: Public distribution system, Fair Price Shops, commodities, Retail Ration Distributors

1. INTRODUCTION

Public distribution system is a government-sponsored chain of shops entrusted with the work of distributing basic food and non-food commodities to the needy sections of the society at very cheap prices. Wheat, rice, kerosene, sugar, etc. are a few major commodities distributed by the public distribution system. The system is often blamed for its inefficiency and rural-urban bias. It has not been able to fulfill the objective for which it was formed. Moreover, it has frequently been criticized for instances of corruption and black marketing. In coverage and public expenditure, it is considered to be the most important food security network.

However, the food grains supplied by the ration shops are not enough to meet the consumption needs of the poor or are of inferior quality. The average level of consumption of PDS seeds in India is only 1 kg per person / month. The PDS has been criticised for its urban bias and its failure to serve the poorer sections of the population effectively. The targeted PDS is costly and gives rise to much corruption in the process of extricating the poor from those who are less needy. Today, India has the largest stock of grain in the world besides China, the government spends Rs. 750 billion (\$13.6 billion) per year, almost 1 percent of GDP, yet 21% remain undernourished. Distribution of food grains to poor people throughout the country is managed by state governments. As of date there are about 500,000 Fair Price Shops (FPS) across India.

1.1 Overview

The central and state governments shared the responsibility of regulating the PDS. While the central government is responsible for procurement, storage, transportation, and bulk allocation of food grains, state governments hold the responsibility for distributing the same to the consumers through the established network of Fair Price Shops (FPSs). State governments are also responsible for operational responsibilities including allocation and identification of families below poverty line, issue of ration cards, supervision and monitoring the functioning of FPSs (Fair Price Shops) Under PDS scheme, each family below the poverty line is eligible for 35 kg of rice or wheat every month, while a household above the poverty line is entitled to 15 kg of food grain on a monthly basis.

A below poverty line card holder should be given 35 kg of food grain and the card holder above the poverty line should be given 15 kg of food grain as per the norms of PDS. However, there are concerns about the efficiency of the distribution process.

1.2 Public distribution system in India-evolution

Evolution of public distribution of grains in India had its origin in the 'rationing' system introduced by the British during the World War II. In view of the fact that the rationing system and its successor, the public distribution system (PDS) has played an important role in attaining higher levels of the household food security and completely eliminating the threats of famines from the face of the country, it will be in the fitness of things that its evolution, working and efficacy are examined in some details.

It was really the generation of World War's own compulsions that forced the then British Government to introduce the first structured public distribution of cereals in India through the rationing system-sale of a fixed quantity of ration (rice or wheat) to entitled families (ration card holders) in specified cities/towns. The system was started in 1939 in Bombay and subsequently extended to other cities and towns. By the end of 1943, 13 cities had been brought under the coverage of rationing and by 1946, as many as 771 cities/towns were covered. Some rural areas, suffering from chronic shortage were also covered. The Department of Food under the Government of India was created in 1942, which helped in food matters getting the serious attention of the government. When the War ended, India, like many other countries, decided to abolish the rationing system. This was in 1943. However, on attaining Independence, India was forced to reintroduce it in 1950 in the face of renewed inflationary pressures in the economy immediately after independence "which were accentuated by the already prevailing high global prices of foodgrains at the end of the War, which were around four times higher than the prewar prices".

Public distribution of foodgrains was retained as a deliberate social policy by India, when it embarked on the path of a planned economic development in 1951. It was, in fact, an important component of the policy of growth with justice. In the first five year plan, the system, which was essentially urban based till then was extended to all such rural areas which suffered from chronic food shortages. It was also decided to have two variations of the system, Statutory Rationing Areas, where foodgrains availability was supposed to be only through the Ration Shops and Non-Statutory Rationing Areas, where such shops would only supplement the open market availability. The system, however, continued to remain an essentially urban oriented activity. In fact, towards the end of the first five year plan (1956), the system was losing its relevance due to comfortable foodgrains availability. The net (gross minus 12.5 per cent for seed, feed and wastage) retail level availability of foodgrains had jumped from 54.0 million tonnes in 1953 to 63.3 million tonnes in 1954 and remained at 63 plus million tonnes up to end of the first five year plan. This situation even prompted the government to abandon procurement of foodgrains and remove all controls on the private trading in foodgrains. However, true to its cyclic nature, the production dropped to 58.3 million tonnes in 1958, when the second five year plan had just started and forced the government to not only restart the procurement of cereals and put control on trading of foodgrains but re-examine the need for public distribution system (PDS). It was decided to re-introduce PDS. Other essential commodities like sugar, cooking coal, kerosene oil were added to the commodity basket of PDS. There was also a rapid increase in the Ration Shops (now being increasingly called the fair price shops-FPSs) and their number went up from 18000 in 1957 to 51000 in 1961. Moreover, quantity of foodgrains distributed through PDS started getting increased with PL- 480 availability. Thus, by the end of the Second Five Year Plan, PDS had changed from the typical rationing system to a social safety system, making available foodgrains at a 'fair price' so that access of households to foodgrain could be improved and such distribution could keep a check on the speculative tendencies in the market. The concept of buffer stocks was also incorporated in the overall food policy, although no buffer worth the name was required to be created in view of easy and continuous availability of PL-480 grains.

Creation of Food Corporation of India and Agricultural Prices Commission in 1965 consolidated the position of PDS. Government was now committed to announce a minimum support price for wheat and paddy and procure quantities that could not fetch even such minimum prices in the market. The resultant stocks were to be utilized for maintaining distribution through the PDS and a portion of these were used to create and maintain buffer stocks. In fact, if stocks happened to be inadequate for maintaining a certain level of distribution through PDS, government had to resort to imports to honour its charge to PDS consumers. All through the ups and downs of Indian agriculture, PDS was continued as a deliberate social policy of the government with the objectives of:

- i) Providing foodgrains and other essential items to vulnerable sections of the society at reasonable (subsidized) prices;
- ii) To have a moderating influence on the open market prices of cereals, the distribution of which constitutes a fairly big share of the total marketable surplus; and
- iii) To attempt socialisation in the matter of distribution of essential commodities.

Public Distribution System in the country facilitates the supply of food grains to the poor at a subsidized price. Essential items such as selected cereals, sugar and kerosene at subsidized prices to holders of ration cards is the objective of efficient Public Distribution System. The PDS also helps to modulate open - market prices for commodities that are distributed through the system. Government accords great importance to the objective of measuring outcomes of PDS so as to ensure that equal distribution system serves up the purpose for which it

was set up. The Authorized Wholesale Distributors are appointed by the District Collectors while the Retail Ration Distributors are appointed by the District Supply Officers. The Authorized Retail Distributors distribute rationed articles to the cardholders on production of the ration cards issued by the taluk Supply Officers/City Rationing Officers at the scale and price fixed by the Government from time to time. PDS is operated under the joint responsibility of the Central and the State Governments. The Central Government has taken the responsibility for procurement, storage, transportation and bulk allocation of food grains, etc. The responsibility for distributing the same to the consumers through the network of Fair Price Shops (FPSs) rests with the State Governments

2. EXISTING SYSTEM

The manual system is error prone. It is time consuming. It is very difficult for a person to produce report. There are chances for manipulating the sales reports and changing it. Public Distributing system involves a lot of manual entries with the applications to perform the desired task. It had its origin with focus on distribution of food grains in urban scarcity areas when the country passed through critical food shortages. It had its origin with focus on distribution of food grains in urban scarcity areas when the country passed through critical food shortages. The size of the family decides the quantum of essential and nonessential commodities required for consumption. Based on the size of the family the entitlement is fixed and allotment of rice, kerosene, sugar and other controlled articles is made to FPS. It was found in the study that Family cards are not only viewed as essential requirement for purchase of essential articles but also to avail benefits under various Government schemes and programmes. Even the respondents with joint family had family cards separately for each family. Both the Centre and State Governments have been taking serious efforts in providing essential commodities in time to the customers.

DRAWBACKS

- Inaccurate result in case of duplicating, delay and inconsistency in reporting.
- No chance of knowing other members sales daily, It is not user friendly.
- It was found that all sections of people have used PDS articles and the lower price was the motivating factor. Non-controlled articles sold currently at FPS were not attractive to Rural consumers.

3. PROPOSED SYSTEM

Public Distributing System is a specific requirement of the consumers, Employee, Distributors that reduce the overburden specifically to the administrators, need for the new system is due to major drawbacks of existing system. Reports can be generated at any time within few seconds and also analysis can be performed much more frequently which helps in taking a particular consumers, Employee, Distributors details bending. The details regarding the entire consumer, employee, distributor can also be maintained as their information is very helpful and sometimes becomes a critical requirement. To overcome these problems in existing system we develop "Online Distributing Systems".

FEATURES

- Flexible and easy to access
- Less time consuming process
- Report are maintained properly
- It reduces the overburden for the administrator.
- Fast, flexible, less prone to errors and save time.

4. CONCLUSION

Public Distributing System would be very effective for counseling system. This project is quite simple and easy to understand even for the end users. This project is very compatible in nature. Further enhancements can affect the flow of the project. The main motto behind this is the accuracy and this factor is considered and implemented to the fullest possible extent. This project has been very helpful in knowing the software deeply and has good idea over the databases. This project includes reports that would be helpful in knowing the sketch in general. It has helped to have a systematic way of approach on tasks. Normally applications fail not because of wear and tear but by eventually failing to perform because of cumulative maintenance.

PERFORMANCE EVALUATION OF MOBILE ADHOC NETWORK USING DYNAMIC SOURCE ROUTING PROTOCOL

S. SasikalaResearch Scholar, Department of Computer Science, Sri Saraswathi Thyagaraja College, Pollachi

ABSTRACT

Mobile Ad-hoc Network (MANET) is an infrastructureless wireless network has a routable networking environment. It consists of several gadget devices. The movement of packets in a routing protocol facilitates from source to destination via some intermediate nodes. Every mobile node act as a router in the network. The node work as a router and it will expend some amount of its energy. Energy consumption is the main challenging issues due to the limited battery energy of mobile devices. This paper focus on a number of energy efficiency metrics to reduce energy consumption when transmitting and receiving packets during communication. The main aim is to improve network life time due to limited power supply. To increase the energy efficieny ,we examined the variants of energy efficient DSR routing protocols to improve the life time of network.

Keywords: DSR, Energy Efficiency, MANET

I. INTRODUCTION

In wireless technology, a mobile Adhoc network(MANETs), is one of the emerging types of network in which mobile nodes or hosts establish a connection with other devices on an ad-hoc basis. MANETs are a self-managing, self-configuring and self-administrating where devices are connected with each other through a wireless link. MANET works without any centralized management or fixed infrastructure[1]. The applications of mobile ad hoc networks are defense, disaster recovery, transportation, mining, heavy construction and event management.

Most of the today's available networks are based on wireless technology where the establishment of network is on demand.. Mobile phones play an vital role in everyday life. Laptops, game consoles, pagers, PDAs and other similar devices have already adopted the wireless technology[2]. With increasing number of communication in wireless technology a need of establishing connection devices has arisen. The main challenging task to communicate and establish a communication for a different purpose. Devices provide different services and it has to know in advance which service and device required to fulfill the requirements. The wireless technology is capable of supporting a device with its own configuration and manual configuration.

The Major challenges with MANETs are dynamic topology, bandwidth, multi hop routing and limited battery resources. Due to limited battery power to transmit a packet from one node to another node the battery power consumes more.MANET has a frequent topology change so an extra effort done by the nodes make to reduce the network life time.[4]

The network protocols can be categorized in two ways

1. Maximum network lifetime protocol: The maximum network lifetime protocol networks are considered which will work for a long time [5].
2. Minimum energy protocol - In minimum energy protocols where nodes consume less energy.

In MANET energy is consumed during data communication in four ways: Receiving, Sleeping, Idle and Residual energy. If there is no path found in the network a new path is discovered that will consume more energy[6]. If a node is idle there will be power drain leads to overhearing . MANET is multi hop, where node acts as an intermediate , it consumes energy. and the higher battery consumption will result lesser network lifetime[7,8]. Many energy efficient routing protocols are being investigated where on demand routing protocols are found better performance to increase network life time. In Energy efficient performance analysis, DSR routing protocol performs well as compared to other routing protocol[9].

IIDSR PROTOCOL

The Dynamic Source Routing is an on-demand routing protocol.The protocol is composed of two mechanism:Route discovery and Route maintenance[11].By these two mechanism it allow to discover route from source node to destination node in adhoc networks.

1. Route Discovery: Route discovery consist of two mechanism that is route request and route reply[12]. It is the method where the source node receives the end node destination path. In order to reduce the cost of

route discovery, DSR RREQs broadcast the packets to entire network even if no reply are received. When an intermediate node forward packets along the route and found that the packet is broken, the node chooses the another route to send a packet without discard it.

Algorithm: Route discovery in DSR

1. Start
2. Initialize Source(S) to Destination (D)
3. Source node s flood RREQ message to all nodes in network
4. If route exists, Then Broadcast RREQ message to all intermediate nodes
Else
Start route discovery and send message End
5. If RREQ is repeated, Then Discard route
Else
a. Add address to route record
b. Send request to adjacent node c. Receive RREQ to D
d. Copy route record and send response to S request RREP
End
6. End

2. Route Maintenance

Route maintenance is used to identify link whether it is able to carry packet on it or not. A routing entry contains an information about intermediate node, whether the node has entire routing path to send that packet throughout from source to destination[13,14].

Algorithm: Route maintenance in DSR

1. Start
2. Initialize S, D
3. Start route discovery module and send RREQ
4. Receive requests from intermediate nodes
5. If link is disconnected between nodes
Then
a. Send error message to S
b. Delete all routes from buffer c. Start route discovery again
End
6. End

III ENERGY EFFICIENT DSR ROUTING

Protocol

DSR routing protocol is on demand routing protocol. The nodes can participate in routing data packets because they are dynamic in nature [15]. The major role of DSR routing protocol is to establish an efficient path and to increase the network life time. [14]. MANET is infrastructure less network where energy conservation is very essential for the functioning of entire network[16]. During active and inactive mode the node energy can be achieved by limiting mobile node's energy. The energy consumption can be minimized by the following approaches:

1. Transmission power control
2. Load distribution
3. Power down or sleep

a) Transmission power control approach

The transmission power control approach choose the best possible routing path to minimizes transmission energy[17]. The routing path able to deliver data packet from a source node to a destination node. MANET discover a route and broadcasts flooding of packets in the network. The nodes contain various transmitting power. The transmission range able to transmit any pair of node should be efficient[18]. The node can transmit according to the requirement to a destination node to save energy, interference and congestion in the network[19].

b) Load distribution approach

The load distribution approach focuses on energy usage among the node. It is used to avoid over utilized among the nodes while selecting a routing path[20]. In this approach it helps to optimize battery power. By choosing the shortest path over utilized nodes may go out of network which will affect the network lifetime[21]. So load distributed approach prevent a node from overloading instead of choosing a shortest path. It increases the network life time and underutilized the path in an effective way.

c) Power down/sleep approach

If a node is not in active communication then to save energy it is in a sleep state.

IV ENERGY CONSUMPTION MODES

The wireless mobile adhoc network are connected to other mobile nodes. Mobile nodes transmit and receive packets from other nodes and require energy while communicate[22]. In order to reduce the power consumption among other nodes, the wireless transmitter of a sender, receivers, forwarders does not involve in this communication. The total energy of nodes is spent in following modes. They are a Transmission Mode, Reception Mode, Idle Mode and Overhearing Mode

a. Transmission Mode

A node is said to be in transmission mode where it sends data packet to other nodes in network. The nodes require to transmit data packet is called as Transmission Energy(T_e). Transmission energy(T_e) depends on size of data packet (in Bits), means while the size of a data packet is increased the required Transmission energy (T_e) is also increased [23]. The transmission energy can be formulated as:

$$T_e = (330 * \text{Plength}) / 2 * 10^6 \text{--- (1)}$$

And

$$P_T = T_e / T_p \text{--- (2)}$$

Where T_e is Transmission Energy, T_p is Transmission Power, T_p is time taken to transmit data packet and Plength is length of data packet in Bits.

b. Reception Mode

If a node receives a data packet from other nodes then it is said to be reception mode. The energy take to receive packet is called as Reception Energy (R_e). The reception energy can be calculated as:

$$R_e = (230 * \text{Plength}) / 2 * 10^6 \text{--- (3) And}$$

$$R_P = R_e / T_r \text{--- (4)}$$

Where R_e is a Reception Energy, R_P is a Reception Power, T_r is a time taken to receive data packet and Plength is length of data packet in Bits.

c. Idle Mode

The node is neither generally transmits nor receives any data packets. This idle mode consumes power because it have to listen to wireless medium continuously to detect a packet [24]. In order to receive a packet the node can then switch to receive mode from idle mode. Idle mode does not handle data communication operations, where the wireless interface consumes a approaches the amount that is consumed in the receive operation. Idle mode defines that wasted energy should be eliminated or reduced. The power consumed in Idle Mode is:

$$P_I = R_P \text{--- (5)}$$

Where P_I is power consumed in Idle Mode and R_P is power consumed in Reception Mode[25].

d. Overhearing Mode

When a node receives data packets that are not intended for it, then it said to be in overhearing mode. It consumes energy while unnecessarily receiving packets.

The power consumed in overhearing mode is:

$$P_{over} = P_R \dots (6)$$

Where P_{over} is power consumed in Overhearing Mode and P_R is power consumed in Reception Mode[26,27].

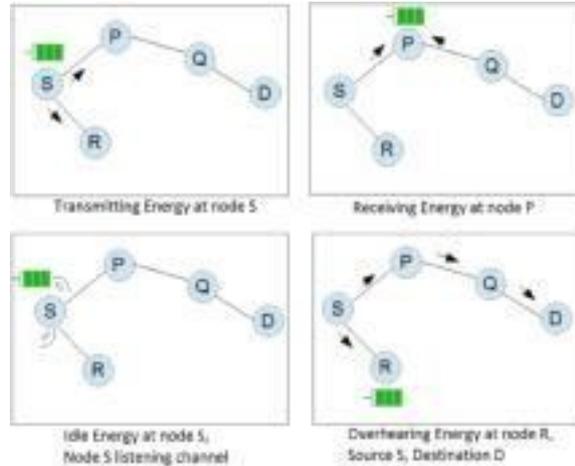


Fig-1: Energy Consumption Modes

IV EFFICIENCY EVALUATION IN MANET USING DSR

The Energy Efficiency Consumes less power consumption, and also increase the network and maintain the performance level.

DSR perform quick adaptation to frequent changes in routing and host movement. The results include the following metrics:

- Network Life time
- Packet Delivery Ratio
- Energy Consumption
- Throughput
- End to End Delay
- Routing Overhead

Simulator	NS-2
Protocol	Dynamic Source Routing
Node movement model	Two Ray Ground
Simulation Area	500m x 500m
Simulation time	150 seconds
Node Speed	20 m/sec
Pause time	0 to 100 sec in steps of 20s
Packet sending rate	8 packets / sec
Traffic type	CBR
No. of Nodes	50
No. of Sources	5

The simulation results were shown in the following figures

1. NETWORK LIFETIME

The network lifetime an estimation process schedule performed before the network becomes nonfunctional.

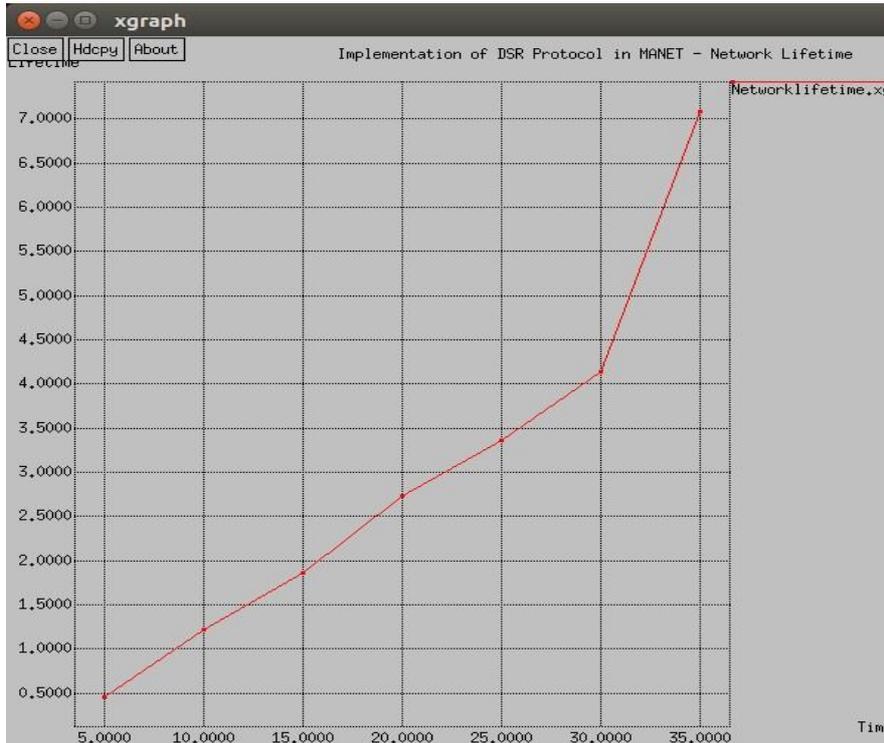


Fig-2: Network Lifetime in DSR

2. PACKET DELIVERY RATIO Packet Delivery is the ratio of number of the data packets successfully delivered to the destination.



Fig-3: Packet Delivery Ratio in DSR

3. NODE ENERGY

A node in a network can spend an amount of energy while transmitting packet and it can be finalized by adding the total energy utilization of the network

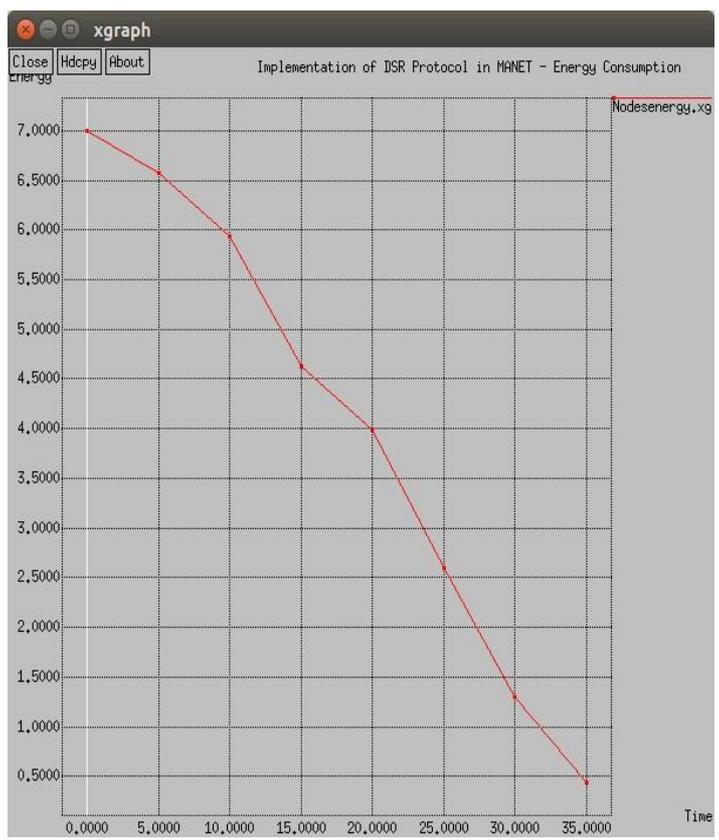


Fig-4: Node Energy in DSR

4. THROUGHPUT

Throughput is the average rate of successful message delivery over a communication channel. Throughput refers that data can be transferred from one location to another in a given amount of time.

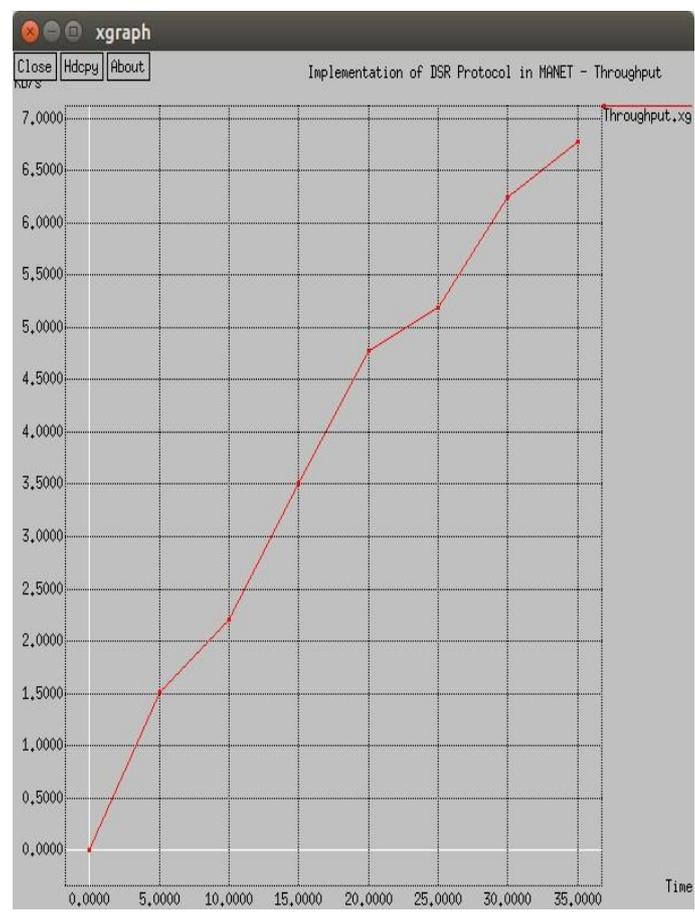


Fig-5: Throughput in DSR

5. END TO END DELAY

Routing Delay is an average delay of data packets from the source to the destination.

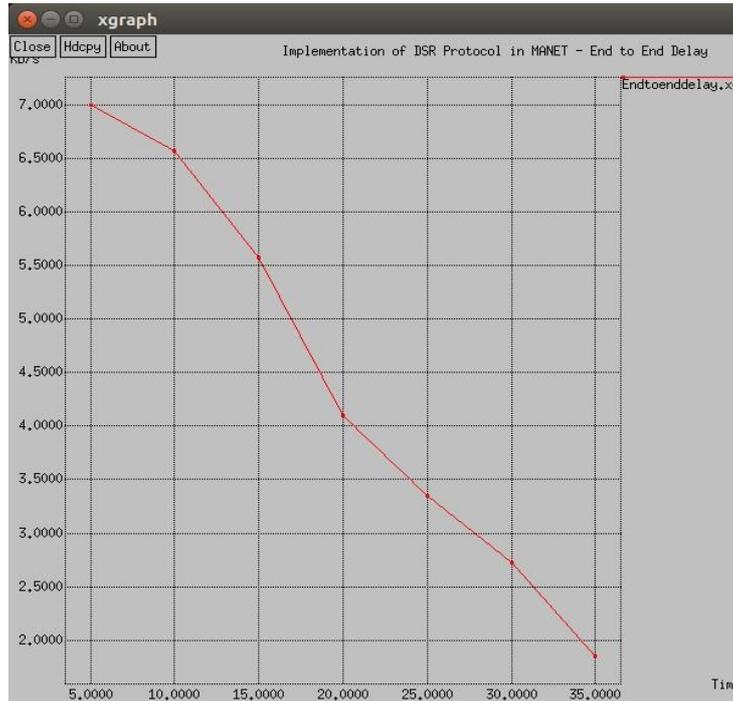


Fig-6: End to End Delay in DSR

6. ROUTING OVERHEAD

Routing Overhead is the number of routing packets required for network communication.



Fig-7: Routing Overhead in DSR

CONCLUSION

This implementation of DSR protocol using Mobile adhoc network is to evaluate the energy level utilization in a network since there is a Routing delay, packet delivery ratio and Throughput of a network has to be improved in MANET. Therefore the Future work with a Branch and Bound, A discrete optimization problem solving technique is combined with DSR to prolong the Network Lifetime and extend the Energy Efficiency in MANET.

REFERENCES

1. D. B. Johnson, D. A. Maltz, Yih Chun Hu and Jorjeta G. Jetcheva, "The Dynamic Source Routing for Mobile Ad Hoc Wireless Networks", May 17, 2007.
2. C. Perkins, P. Bhagwat, "Highly dynamic Destination Sequenced Distance Vector routing (DSDV) for mobile computers", ACM SIGCOMM, PP: 234-244, August 1994.
3. Chansu Yu, Ben Lee Hee and Yong Youn, "Energy Efficient Routing Protocols for Mobile Ad hoc Networks", June 4, 2007.
4. A. P. Chandrakasan, S. Sheng and R. W. Brodersen, "Low-power CMOS Digital Design, Solid State Circuits", IEEE Journal, Vol. 27, PP:473-84, 1992.
5. A. R. Lebeck, X. Fan, H. Zeng and C. S. Ellis, "Power Aware Page Allocation, in Architectural Support for Programming Languages and Operating Systems", PP: 105-116, 2000.
6. L. Benini, A. Bogliolo and G. De Micheli, "A Survey of Design Techniques for System Level Dynamic Power Management, Very Large Scale Integration (VLSI) Systems", IEEE Transactions on, Vol. 8, PP: 299-316, 2000.
7. Q. Li, J. Aslam and D. Rus, "Online Power-Aware Routing in Wireless Ad-Hoc Networks", MOBICOM, July 2001.
8. Sunsook Jung, Nisar Hundewale and Alex Zelikovsky, "Energy Efficiency of Load Balancing in MANET Routing Protocols", Sixth International Conference on Vol. 23, Issue 25, PP: 476 – 483, May 2005.
9. Tanu Preet Singh, Shivani Dua, Vikrant Das, "Energy Efficient Routing Protocols in Mobile Ad hoc Networks", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 1, PP: 1-7, Jan 2012.
10. Ramanathan R, Rosales Hain R, "Topology Control of Multi hop Wireless Networks using Transmit Power Adjustment", The Conference on Computer Communications, PP: 404-413, 2000.
11. Wattenhofer R, Li L, Bahl P and Wang Y-M, "Distributed Topology Control for Power Efficient Operation in Multihop Wireless Ad Hoc Networks", Conference on Computer Communications, 2001.
12. Pushparaj, R. and M. Dinakaran, 2014. Energy Efficient Issues and Routing Challenges in MANET", Current Trends in Engineering and Technology(ICCTET), DOI: 10.1109/ICCTET.2014.6966258: 26-
13. Narayanaswamy, S., V. Kawadia, R.S. Sreenivas .R. Kumar, 2002. Power Control in ad hoc networks theory, architecture, algorithm and implementation of the COMPOW protocol, in Proc. European Wireless 2002 Next Generation Wireless Networks: Technologies, Protocols, Services and Applications, Florence, Italy, pp: 156-162.
14. Elbatt, T. and A. Ephremides, 2004. Joint Scheduling and Power Control for wireless adhoc networks, IEEE Trans. Wireless Commun., 3(1): 74-85.
15. Chiasserini, C.F. and R.R. Rao, 2001. Energy Efficient Battery Management, IEEE J. Select. Areas Commun., 19(7): 1235-1245.
16. Chiasserini, C.F. and R.R. Rao, 2003. Improving Energy Saving in Wireless Systems by using dynamic power management, IEEE Trans. Wireless Commun., 2(5): 1090-1100.
17. Tonguz Ozan, K. and Gianluigi Ferrari, 2010. Adhoc Wireless Networks communication theoretic perspective, John Wiley & Sons Ltd, U.K.
18. Ramachandran Pushparaj and M. Dinakaran., 2016. Signal Strength and Residual Power Based Optimum Transmission Power Routing for Mobile Ad hoc Networks, Elsevier, Procedia Computer Science, 92: 168-174.
19. Douros, Vaggelis, G. and George C. Polyzos, 2011. Review of some fundamental approaches for power control in wireless networks, Elsevier Computer Communications, 34: 1580-1592.
20. Krunz Marwan, Alaa Muqattash and Sung-Ju Lee, 2004. Transmission Power Control in Wireless AdHoc Networks: Challenges, Solutions and Open Issues, IEEE Network, 18(5): 8-14.

21. Fotino Marco and Floriano De Rango, 2011. Energy Issues and Energy Aware Routing in Wireless AdHoc Networks, Mobile Ad Hoc Networks: Protocol Design, Prof. Xin Wang (Ed.), ISBN:978-953-307-402-3, InTech.
22. Kim, D., J. Garcia-Luna-Aceves, K. Obraczka, J.C. Cano and P. Manzoni, 2003. Routing Mechanisms for Mobile Ad Hoc Networks based on the Energy Drain Rate, IEEE Transactions on Mobile Computing, 2(2): 161-173.
23. Kirousis, L., E. Kranakis, D. Krizane and A. Pele, 1997. Power Consumption in Packet Radio Network, In: Symposium on theoretical aspects of computer science (STACS).
24. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficient optimization with informed transmitted", IEEE Trans. Wireless Commun., vol. 11, no. 8, pp. 2946-2957, Aug. 2012.
25. Y. Tao and Wei Luo, "Modified Energy-Aware DSR Routing for Ad hoc Network", International Conference on Wireless Communications, Networking and Mobile Computing, WiCom, pp: 1601-1603, 2007.
26. D.B. Johnson, D.A. Maltz, and Y.-C. Hu, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks (DSR)", Internet-Draft, draft-IETF-manet-dsr-09.txt, Apr. 2003.
27. C.E. Perkins, E.M. Belding-Royer, and I.D. Chakeres, "Ad Hoc On-Demand Distance Vector (AODV) Routing", IETF Internet draft, Oct. 2003.
28. J. Kuruvila, A. Nayak, and I. Stojmenovic, "Progress and Location Based Localized Power Aware Routing for Ad Hoc and Sensor Wireless Networks", Int'l J. Distributed Sensor Networks, vol. 2, pp. 147

CRIME RECORDS MANAGEMENT SYSTEM**Aathinamarun¹, Dr. A. V. Senthilkumar²**Student¹ and Director², Department of Computer Applications, Hindusthan College of Arts and Science, Coimbatore

ABSTRACT

The proposed system applies to all Police stations across the country and specifically looks into the subject of Crime Records Management. It is well understood that Crime Prevention, Detection and Conviction of criminals depend on a highly responsive backbone of Information Management. The efficiency of the police function and the effectiveness with which it tackles crime depend on what quality of information it can to it. Initially, the system will be implemented across Cities and Towns and later on, be interlinked so that a Police detective can access information across all records in the state thus helping speedy and successful completion to cases. The project has been planned to be having the view of distributed architecture, with centralized storage of the database. The application for the storage of the data has been planned. Using the constructs of SQL server and all the user interfaces have been designed using the DOT Net technologies.

Keywords: Crime, Criminals, Investigators, Lawyers.

I. INTRODUCTION

The Crime Records Management System applies to Police Stations all across the country and specifically looks into the subject crime prevention, detection, conviction of criminals depending on a highly responsive backbone of Information Management [1]. The efficiency of the Police and the effectiveness with which it tackles crime depend on what quality of information it can derive from its existing records and how fast it can have access to it.

A. Modules**1) Station module**

Each of the station must first register with the Software. For the registration part each station enter their details like station name, address, phone no, station in charge etc. and get a User Id from the Software. Once the prospective station registers with the software they can avail the existing records.

2) Citizen module

Each of the citizens, who has a complaint to register, must first register with the Software. For the registration part each person enter their details like name, address, phone no., E-Mail ID etc. and get a User Id & password from the Software. Once the registration is complete, the citizen can sign-in to the website & register their complaint.

3) Crime module

This module is used for entering all details about the crime. It contains the date and time, police station where it is recorded, place, Nature of Crime, Location of the Crime etc.

4) Search module

In this module we can search the crime in station wise, nature of crime.

5) Administrators Module

The module will be protected by user ID and password. Ordinary users of the software will not be permitted to enter into this area of the software.

The module will be focusing on the maintenance like Master Data Maintenance, Removal of old and outdated data from the software etc.

6) Avocation Module

This Module deals with the Law part of every Crime. [3]

II. SYSTEM ANALYSIS

System Analysis works with users to identify goals and build systems to achieve them. System analysis is an important phase of any system development process. The system is studied to the minutest details and analyzed. The system analyst plays the role of an interrogator and dwells deep in to the working of the present system. In analysis, a detailed study of these operations performed by a system and their relationships within and outside of the system is done. A key question considered here is, "what must be done to solve the problem?" One aspect of analysis is defining the boundaries of the system and determining whether or not the candidate system should

be considered. The system is viewed as a whole and the inputs to the system are identified. The outputs from the system are traced through the various processing that the input phases through in the organization. During analysis, data are collected on available files, decision points, and transaction handled by present system. Once analysis is completed the analyst has a firm understanding of what is to be done.

III. EXISTING SYSTEM

In the existing crime management system, most of the operations are done manually like send complaints, taking actions against crimes, view status etc. So with the existing system if anybody wants to complaint against crimes he must do it through the police. If we are doing the system manually, so many minor errors will occur. Error detection in the previous entries made and data cross verification is another important function. These are done manually, and it would take time.

1) Drawbacks of the existing system can be concluded as follows

- * The existing system is time consuming and not very user friendly.
- * The officer dealing with a particular case cannot take decision by himself even when he is having the first hand knowledge/information about the case and he can expect obstructions from higher authorities/officials.
- * Even an efficient officer cannot/may not able to handle more than one case at a time.
- * In most of the cases, the innocent are accused in the existing system.
- * As per our jurisdiction, "Let thousand criminals escape not a single innocent be punished ". As a result of this and other factors that influence investigation, such as bribery, the innocent becomes accused in several situations in the eyes of Justice.
- * The existing system could provide only investigation and there is no Advocating, Counseling facilities etc. hence no contact or control or co-ordination on these cell.

V. PROPOSED SYSTEM

Our mission statement directs us to promote the public interest while balancing the Law. Obviously we'll not rush to judgment, but will look at the facts and render a thoughtful decision based on these facts. Organization is an independent statutory body which was created to maintain an effective and efficient police service for the public. Its primary task includes securing continuous needs of the public in an efficient manner. As part of the responsibility, 'our committee' is responsible for performance monitoring (performance planning and review formerly professional standards and performance monitoring) decided to look at the contribution and effectiveness of the organization within the public. We need the whole hearted support of each and every individual member of the site and cooperation of the users. The administrator, along with the investigators, lawyers, counselors, and other authorities statutory partners in our organization. Given their key role in these partnerships, we felt that an overview of strength and Weakness of the organization is needed.

The aim of the project is to bring about improvement to the organization's contributions; this report necessarily concentrates on overcoming weakness and raising standards. To avoid conducting and unfair evaluation efforts which are made to put this contribution into this context. It believes that partnership work is highly beneficial to the organization and that partnership work is the way forward to reduce crime and disorder.

1) Advantages of the Proposed System:

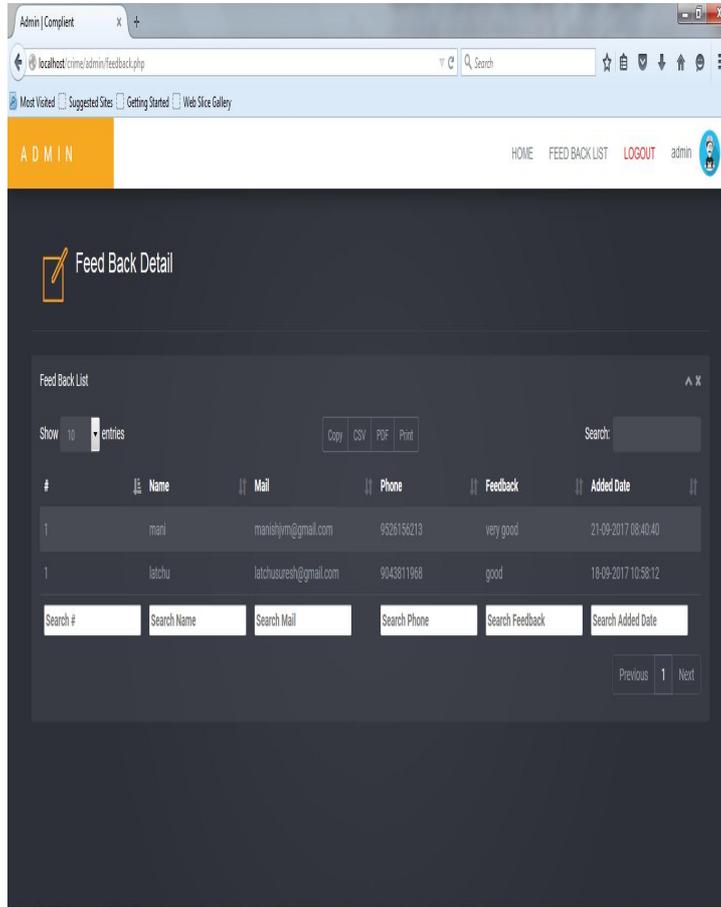
- * Better communication, better leadership, better training, more defined roles in terms of contribution to contribution to the society.
- * Working with partners.
- * Reducing crime and disorder.
- * Confidentially and anonymity issues. The proposed system has control over all its cells and is perfectly coordinated. Also the cells can act individually. To take our organization to heights we need the wholehearted co-operation of the public.

2) Objectives of the Proposed System

In the fast moving world, if people lack something, it is time. All are busy in their world. So the main objective of our product is better communication, better leadership, reducing crime and disorder etc. The product provides a framework within which a user can easily work with. We know users are of many categories, like users from who know working with computers very well to users who didn't know about computers. So all the category can use the software. So it should be user friendly.

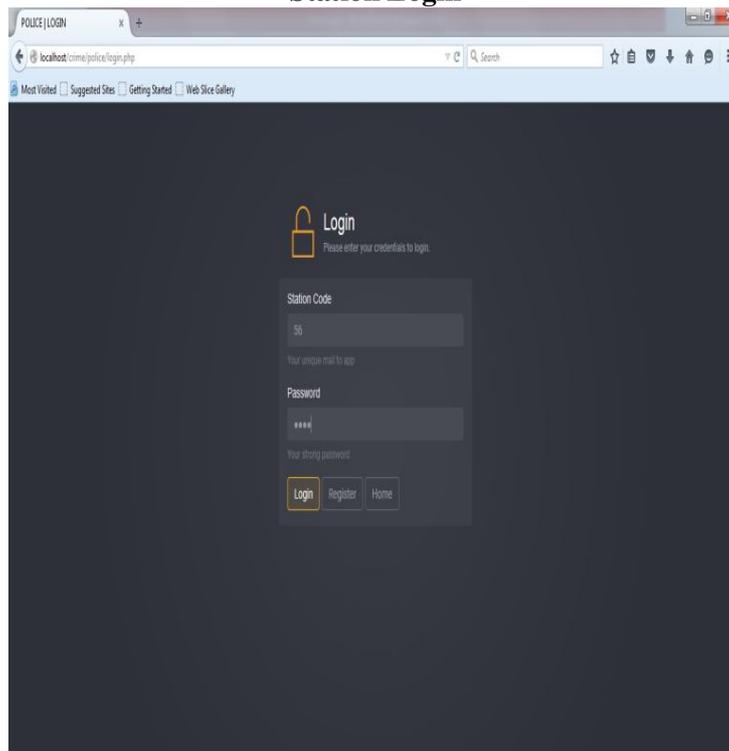
The product provides a framework, which is error free. We know a crime management system is actually a critical process having many calculations and operations. So each simple error laid to big problem. So it should be error free and our objective is to build error free software.

Admin Feedback View



Admin feedback view on the website

Station Login



Station login to the website

Station Registration

The screenshot shows a web browser window titled 'POLICE REGISTER'. The address bar shows 'localhost:8080/police/register.php'. The page has a dark theme. At the top left is a 'Login' button. Below it is a 'Register' section with a person icon and the text 'Please enter your data to register:'. The registration form consists of several input fields arranged in two columns:

- Station Code: 56
- Station ID: A45
- Station Incharge: mani
- Station Division: kottakulam
- Phone Number: 9876543210
- Address: 56, colony, kottakulam
- Password: ****
- Repeat Password: ****

A 'Register' button is located at the bottom right of the form.

New station registration

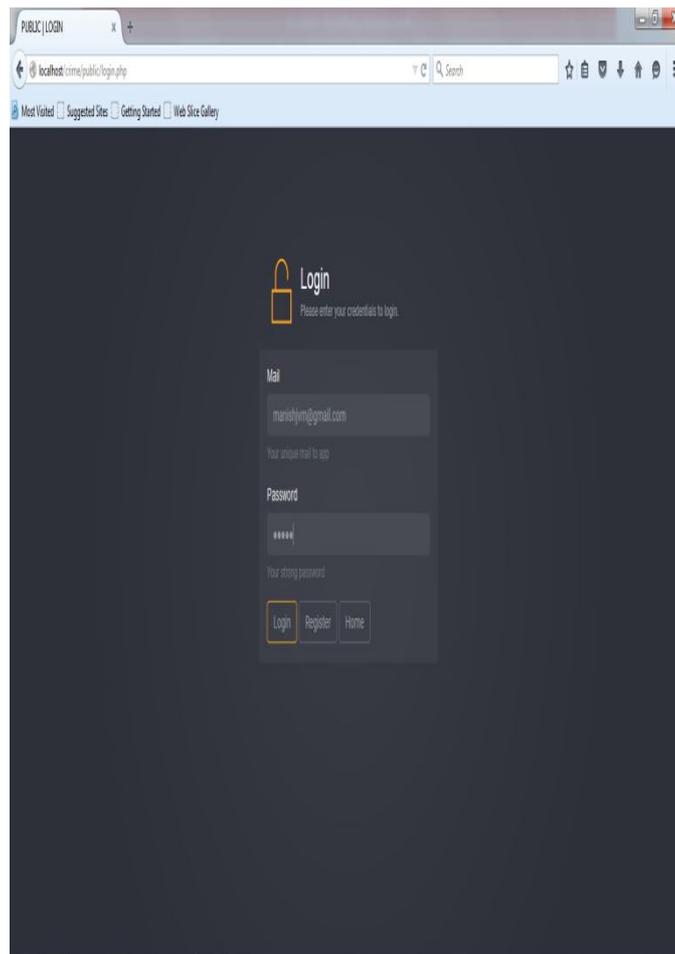
Station Compliant Details

The screenshot shows a web browser window titled 'POLICE | Compliant'. The address bar shows 'localhost:8080/police/welcome.php'. The page has a dark theme. At the top left is a 'POLICE' logo. At the top right is a 'LOGOUT' button and a user profile icon. Below the header is a 'Compliant Detail' section with a list icon. The main content is a 'Compliant List' table with the following data:

#	Name	Age	Address	Phone	Aq.Name	Aq.A
1	ramesh	22	56, bajanai madum street, pullingudi.	9876543210	mani	20

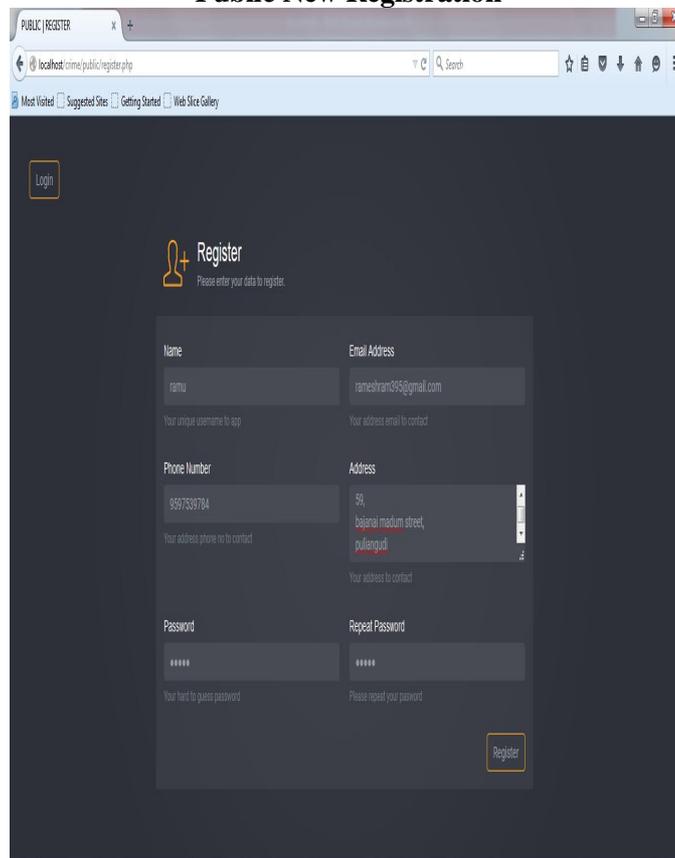
Below the table are search filters for #, Name, Age, Address, Phone, and Aq.Name. There are also 'Previous' and 'Next' pagination buttons.

Public Login



Public login to the website

Public New Registration



Public new registration

New Compliant Registration

USER FEEDBACK LOGOUT rameshram395@gmail.com

New Compliant

New Compliant

Complainer Name	Age	Phone Number
ramesh	22	99759784

Address

59, bajajal madam street,
pullangudi

Acquest Name	Acquest Age	Case
mani	20	Civil Criminal

Acquest Address Case Description

35/ship road, pullangudi land cheating problem

Register

New compliant registration

Public Feedback

USER FEEDBACK LOGOUT rameshram395@gmail.com

Feedback Form

Type Your Feedback

very nice

Close Save changes

New Compliant

Complainer Name	Age	Phone Number

Address

Acquest Name	Acquest Age	Case
		Civil Criminal

Acquest Address Case Description

Register

Public feedback

VI. CONCLUSION

The Software developed is found to be working efficiently and effectively. It results in regular and timely action against crime reported. It can be observed that the information can be obtained easily and accurately. The Software is made user friendly to the maximum so that any lay man can run the software provided he could access to the system via the login password. It believes that partnership work is highly beneficial to the organization and that partnership work is the way forward to reduce crime and disorder. Hence, we wish to remind that its culture should fully endorse partnership work; we urge to ensure the attention they require.

REFERENCES

1. http://en.wikipedia.org/wiki/Criminal_record
2. <http://thinknext.co.in/rupali/homepage.aspx>
3. Anil Jaiswal, Neeta Gunjal, PoojaLondhe, Shikha Singh, Ramesh
4. Solanki, "Crime Automation & Reporting System" ,International Journal of Science and Modern Engineering (IJISME),Volume-1, Issue-11, October 2013
5. <http://topicideas.org/how-to-online-crime-reporting-and- management system>

MANUSCRIPT SUBMISSION

GUIDELINES FOR CONTRIBUTORS

1. Manuscripts should be submitted preferably through email and the research article / paper should preferably not exceed 8 – 10 pages in all.
2. Book review must contain the name of the author and the book reviewed, the place of publication and publisher, date of publication, number of pages and price.
3. Manuscripts should be typed in 12 font-size, Times New Roman, single spaced with 1” margin on a standard A4 size paper. Manuscripts should be organized in the following order: title, name(s) of author(s) and his/her (their) complete affiliation(s) including zip code(s), Abstract (not exceeding 350 words), Introduction, Main body of paper, Conclusion and References.
4. The title of the paper should be in capital letters, bold, size 16” and centered at the top of the first page. The author(s) and affiliations(s) should be centered, bold, size 14” and single-spaced, beginning from the second line below the title.

First Author Name₁, Second Author Name₂, Third Author Name₃

1 Author Designation, Department, Organization, City, email id

2 Author Designation, Department, Organization, City, email id

3 Author Designation, Department, Organization, City, email id

5. The abstract should summarize the context, content and conclusions of the paper in less than 350 words in 12 points italic Times New Roman. The abstract should have about five key words in alphabetical order separated by comma of 12 points italic Times New Roman.
6. Figures and tables should be centered, separately numbered, self explained. Please note that table titles must be above the table and sources of data should be mentioned below the table. The authors should ensure that tables and figures are referred to from the main text.

EXAMPLES OF REFERENCES

All references must be arranged first alphabetically and then it may be further sorted chronologically also.

• **Single author journal article:**

Fox, S. (1984). Empowerment as a catalyst for change: an example for the food industry. *Supply Chain Management*, 2(3), 29–33.

Bateson, C. D.,(2006), ‘Doing Business after the Fall: The Virtue of Moral Hypocrisy’, *Journal of Business Ethics*, 66: 321 – 335

• **Multiple author journal article:**

Khan, M. R., Islam, A. F. M. M., & Das, D. (1886). A Factor Analytic Study on the Validity of a Union Commitment Scale. *Journal of Applied Psychology*, 12(1), 129-136.

Liu, W.B, Wongcha A, & Peng, K.C. (2012), “Adopting Super-Efficiency And Tobit Model On Analyzing the Efficiency of Teacher’s Colleges In Thailand”, *International Journal on New Trends In Education and Their Implications*, Vol.3.3, 108 – 114.

- **Text Book:**

Simchi-Levi, D., Kaminsky, P., & Simchi-Levi, E. (2007). *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies* (3rd ed.). New York: McGraw-Hill.

S. Neelamegham," Marketing in India, Cases and Reading, Vikas Publishing House Pvt. Ltd, III Edition, 2000.

- **Edited book having one editor:**

Raine, A. (Ed.). (2006). *Crime and schizophrenia: Causes and cures*. New York: Nova Science.

- **Edited book having more than one editor:**

Greenspan, E. L., & Rosenberg, M. (Eds.). (2009). *Martin's annual criminal code: Student edition 2010*. Aurora, ON: Canada Law Book.

- **Chapter in edited book having one editor:**

Bessley, M., & Wilson, P. (1984). Public policy and small firms in Britain. In Levicki, C. (Ed.), *Small Business Theory and Policy* (pp. 111–126). London: Croom Helm.

- **Chapter in edited book having more than one editor:**

Young, M. E., & Wasserman, E. A. (2005). Theories of learning. In K. Lamberts, & R. L. Goldstone (Eds.), *Handbook of cognition* (pp. 161-182). Thousand Oaks, CA: Sage.

- **Electronic sources should include the URL of the website at which they may be found, as shown:**

Sillick, T. J., & Schutte, N. S. (2006). Emotional intelligence and self-esteem mediate between perceived early parental love and adult happiness. *E-Journal of Applied Psychology*, 2(2), 38-48. Retrieved from <http://ojs.lib.swin.edu.au/index.php/ejap>

- **Unpublished dissertation/ paper:**

Uddin, K. (2000). A Study of Corporate Governance in a Developing Country: A Case of Bangladesh (Unpublished Dissertation). Lingnan University, Hong Kong.

- **Article in newspaper:**

Yunus, M. (2005, March 23). Micro Credit and Poverty Alleviation in Bangladesh. *The Bangladesh Observer*, p. 9.

- **Article in magazine:**

Holloway, M. (2005, August 6). When extinct isn't. *Scientific American*, 293, 22-23.

- **Website of any institution:**

Central Bank of India (2005). *Income Recognition Norms Definition of NPA*. Retrieved August 10, 2005, from <http://www.centralbankofindia.co.in/home/index1.htm>, viewed on

7. The submission implies that the work has not been published earlier elsewhere and is not under consideration to be published anywhere else if selected for publication in the journal of Indian Academicians and Researchers Association.

8. Decision of the Editorial Board regarding selection/rejection of the articles will be final.



INDIAN ACADEMICIANS & RESEARCHERS ASSOCIATION

Major Objectives

- To encourage scholarly work in research
- To provide a forum for discussion of problems related to educational research
- To conduct workshops, seminars, conferences etc. on educational research
- To provide financial assistance to the research scholars
- To encourage Researcher to become involved in systematic research activities
- To foster the exchange of ideas and knowledge across the globe

Services Offered

- Free Membership with certificate
- Publication of Conference Proceeding
- Organize Joint Conference / FDP
- Outsource Survey for Research Project
- Outsource Journal Publication for Institute
- Information on job vacancies

Indian Academicians and Researchers Association

Shanti Path ,Opp. Darwin Campus II, Zoo Road Tiniali, Guwahati, Assam

Mobile : +919999817591, email : info@iaraedu.com www.iaraedu.com



EMPYREAL PUBLISHING HOUSE

- Assistant in Synopsis & Thesis writing
- Assistant in Research paper writing
- Publish Thesis into Book with ISBN
- Publish Edited Book with ISBN
- Outsource Journal Publication with ISSN for Institute and private universities.
- Publish Conference Proceeding with ISBN
- Booking of ISBN
- Outsource Survey for Research Project

Publish Your Thesis into Book with ISBN “Become An Author”

EMPYREAL PUBLISHING HOUSE

Zoo Road Tiniali, Guwahati, Assam

Mobile : +919999817591, email : info@editedbook.in, www.editedbook.in