URL. : http://ijrar.com **ISSN 2349 - 5138** **E - ISSN 2348 - 1269**

An open Access, peer reviewed, refereed, online and print research journal

International Journal of Research and Analytical Reviews

UGC Approved Research Journal

Periodicity - Quarterly

Atman Publishing Academy

International Journal of Research and Analytical Reviews
Atman Publishing Academy
2061-C/2/B, Nr. Adhyatma Vidya Mandir, Sanskar Mandal, Bhavnagar-364002.
Contact : 9427903033 E mail : **editorsijrar@gmail.com**, **ijrar1@gmail.com**

International Journal of Research and Analytical Reviews

ijrar.com

Rs. 900

Subscription	1 year	2 years	5 years
Individual	3600	7200	18000
Institutional	4000	8000	20,000
Advertisement	1000/Black	3000/colour	Per Page

Send your Paper(s)/Article(s) and Contact us on any one of following

E mail: (1) editorsijrar@gmail.com (2) ijrar1@gmail.com (3) drrbjoshi@ijrar.com

Contect No.: +91 9427903033

1. Thoughts, language vision and example in published research paper are entirely of author of research paper. It is not necessary that both editor and editorial board are satisfied by the research paper. The responsibility of the matter of research paper/article is entirely of author.
2. Editing of the IJRAR is processed without any remittance. The selection and publication is done after recommendations of at least two subject expert referees.
3. In any condition if any National/International University denies accepting the research paper/article published in IJRAR, than it is not the responsibility of Editor, Publisher and Management.
4. Only the first author is entitle to receive the copies of all co-author.
5. Before re-use of published research paper in any manner, it is compulsory to take written permission from the Editor – IJRAR, unless it will be assumed as disobedience of copyright rules.
6. All the legal undertakings related to IJRAR is subject to Bhavnagar Jurisdiction.

Editor
International Journal of Research and Analytical Reviews
Atman Publishing Academy
2061-C/2/B, Nr. Adhyatma Vidya Mandir, Sanskar Mandal, Bhavnagar-364002.
Contact : 9427903033 E mail : **editorsijrar@gmail.com, ijrar1@gmail.com**

URL. : http://ijrar.com **ISSN 2349-5138** **E-ISSN 2348-1269**

An open Access, peer reviewed, refereed, online and print research journal

International Journal of Research and Analytical Reviews

URL. : http://ijrar.com **ISSN 2349-5138** **E-ISSN 2348-1269**

An open Access, peer reviewed, refereed, online and print research journal

Review Committee

International Journal of Research and Analytical Reviews

Special Issue July 2018

Contents

International Journal of Research and Analytical Reviews

INTERNATIONAL CONFERENCE ON BIG DATA ANALYTICS AND INTELLIGENT TECHNOLOGIES

MARCH 1 & 2 , 2018

ICBAIT - 2018

JOINTLY ORGANIZED BY

MADURAI KAMARAJ UIVERSITY,
TAMIL NADU, INDIA

UNIVERSITY OF MARIBOR, SLOVENIA

DEPARTMENT OF COMPUTER SCIENCE
MADURAI KAMARAJ UNIVERSITY
University with Potential for Excellence
Re-accredited by NAAC with 'A' Grade in the 3rd cycle
Madurai 625 021,Tamil Nadu, India.
MADURAI-625 021.TAMIL NADU, INDIA

Preface

The First International Conference on Big Data Analytics and Intelligent Technologies held during March 1& 2, 2018, at the Madurai Kamaraj University, Madurai, Tamil Nadu, India.

Madurai Kamaraj University is on its persistent journey for the past 51 years and is accorded with the status of "University with Potential for Excellence" - A status conferred by the University Grants Commission .

In this world of digital era we all know that the Data are being generated every moment at a phenomenal and ever growing rate. Social media plays a vital role in generating volume of images and text documents. This has created an avalanche effect in the digital media and we all know that these data need to be stored and analyzed effectively and efficiently.

Big data analytics is the process of examining large and varied data sets -- i.e., data with properties of Volume, Variety, Velocity and Veracity- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions. Big data analytics applications enable data scientists, predictive modelers, statisticians and other analytics professionals to analyze growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs.

Research in the area of Big Data Analytics focus on the theory development, novel and innovative techniques and suitable solutions of big data analytics in broad domains as the traditional approaches are not powerful enough to handle the issues. Research in the area of Intelligent Techniques focus on the theories and t techniques of building computer systems, which capture the intelligent behaviors in complex environments.

The papers in this proceeding address challenges and best practices related to Big data Analytics and Intelligent Technologies. All the papers published in the proceedings have undergone a plagiarism check and a blind review process .

The papers are organized as :

- Session 1 - Big Data Analytics
- Session 2 - Internet of Things
- Session 3 - Image Processing and Biometrics
- Session 4 - Smart Systems

We would like to thank our Resource Persons, Reviewers, Committee Members for their contribution for organizing the Conference. Our special thanks to the authors and participants of the conference. We would like to extend our thanks to Dr. Marjan Hericko and Dr. Muhamed Turkanovic, University of Maribor, Slovenia for their support. The relentless work of editorial team members is highly appreciated and acknowledged. The hard work contributed by the research scholars of the Department of Computer Science P. Aruna Saraswathy and M.Sivakami made the papers worthy of publication.

The Proceedings will be beneficial for the researchers in the domain of Big Data Analytics and Intelligent Technologies.

G. Arumugam
M. Thangaraj

Novel Comparative Analysis of Performance Metric in OSPF and EIGRP in a Real Secured Network Environment

[1]Prasanya Devi .P, [2]Ravindran.M
[1,2]Department of Computer Science, Government Arts College,Melur,
Tamil Nadu State,India.

ABSTRACT — *In today's world, with increasing usage of computer networks and internet, the importance of Network, computers, information security and data security is obvious. Usual intrusion detection and prevention systems are hasty in the sense that they use a set of signatures, which lift at the same rate as new techniques are exposed, to identify malicious traffic patterns. Wireshark is chosen as a tool for traffic analysis in our local network real-time environment. We Find that TCP control and data traffic have high correlation levels during benign normal application. We find entropy calculation, port probability, flow distribution for our real-time network datasets and compare OSPF routing protocol and EIGRP routing protocol for achieving performance metrics in the Network environment. Data centers are experiencing an increase in network security threats resulting in the loss of revenue, productivity, and business opportunity. Comprehensive security policies and architectures that include network-based intrusion detection systems (NIDS) are a means to combat this expanding threat. NIDS perform analysis of all traffic passing on a network segment or subnet.*

Keywords: — OSPF, EIGRP, Network Intrusion Detection (NID), Network Vulnerability (NV).

Introduction

The world is becoming more inter connected with the advent of the Internet and new working Technology, Network Security (NS) is becoming a great importance because of intellectual property that can be easily acquired through the internet. An effective network security plan is developed with the understanding of security issues.

Potential attackers, needed level of security, and factors that make a network vulnerable to attack. To lessen the vulnerability of the computer and the network there are many products available. These tools are encryption, authentication mechanisms, intrusion detection, security management and firewalls Information security is the process of protecting information. It protects its availability, privacy and integrity. Access to stored information on computer databases has increased greatly. More and More computers get connected to both private and public networks whether most of traffic is generated by traditional data transfer applications such as HTTP, NNTP or FTP. This consequently made TCP/IP the most widely used protocol for computer network and accounts for vast majority of the Traffic over wide area network particularly the Internet.1

EIGRP and OSPF are dynamic routing protocols used in practical networks to disseminate network topology to the adjacent routers. There are various numbers of static and dynamic routing protocols available but the selection of appropriate routing protocol is most important for routing performance. The right choice of routing protocol is dependent on several parameters. In this paper, we implement two routing protocols, namely EIGRP and OSPF, and further do performance evaluation for real-time applications.

The emergence of high speed internet access and government plans to push the broadband to homes and universities has increased the importance of IP networks. The type of traffic on networks changes rapidly with the development of new technologies such as network APPS and pear-to-peer Network setups etc. Such new applications have caused rapidly increased the traffic burden on core Internet routers and the need to monitor the traffic types, which had not been fully considered in development of protocols, has play an important role of network architecture design.

The network intrusion detection system(IDS) can be placed at a choke point such as the company's connection to a trunk line, or it should be placed on each of the hosts that are being monitored to protect from intrusion. Intrusion, incident and attack are three terms that we frequently come across while discussing Intrusion Detection System

Aims and Objectives

The performance of each routing protocol is different from each other. Among all routing protocols, we choose EIGRP and OSPF routing protocols for doing performance evaluation for real-time traffics. The main aim of this work is to evaluate which protocol, EIGRP or OSPF, is most suitable to route in real-time traffic:

- To discuss about different features of the routing protocols.
- To implement the proposed routing protocols in IP networks.
- To select the quantitative metrics such as convergence activity, end-to-end delay, packet delay variation, flow size distribution.
- To analyze the protocol performance theoretically and by simulation

Packets Information

A. Packet Filtering (Pf)

When a packet is received, it is first decoded to extract the information list used for filtering **[1].** At each node of the search tree, the packet is checked against the Boolean expression of the children of this node. The packet is directed to the node whose Boolean expression is satisfied .This process goes on till a leaf is reached. The action associated with this segment is updated as well.

B. Packets Information Log

Consider a queue Q. The neighbor routers rb, rc feed data in queue Q. Each Q has the order by which the packets should enter along with its associated information. Each router maintains a log record. Let Q_{in} be the traffic before entering the queue and Q_{out} be the traffic after leaving the queue. At any instant time, the traffic is represented as

R(Q,qp(t),I,F)

where

1. qp(t) is the predicted state of queue at any time't'.
2. I is the traffic before entering the queue by the information collected from neighbouring routers (rb, rc).
3. D is the traffic after leaving the queue, collected at router rd.

If R(Q,qp(t),I,F) is false and the routers are not protocol faulty, then the packets are dropped maliciously at time 't'. Each packet forwarded maintains a log record which includes:

- Header name P
- IP address (from where it was forwarded)
- Packet size(no. of routers it should traverse) ps
- Time at which it arrived at the router

Three criteria could be used for predicting the state of the Packet P:

- If P came from F, then the packet is leaving Q
- If P came from I (P traversed D), the packet is entering the Q and will exit at qp+ps
- If P came from I (P hasn't traversed D), the packet is entering the final Queue and is received at the destination.

To detect how the attack occurred, two conditions have to be satisfied :

- Buffer limit(B) is maintained at each router. If B<qp+ps, then the packet P is dropped due to congestion.
- Otherwise, the packet P is dropped due to malicious access.

C.Confidence Value (Cv) Test

We introduce a term Cv which is the probability of an attack to occur. If a packet P is dropped at time t at queue length qp, then Cv is raised. This is suitable only for a single packet is lost.

We use the following terms :

qs(t) – size of the queue for packet P
ps – size of the packet P
qlim – max size of queue
X – new malicious packet inserted
Cv is calculated as below:

$Cv = Prob(\text{Packet P to be dropped})$
$= Prob(\text{ more space in the queue})$
$= Prob(qs(t)+ps \le qlim)$
$= Prob(X+qp(t)+ps \le qlim)$
$= Prob(X \le qlim-qp(t)-ps)$
$= Prob(Y \le (qlim-qp(t)-ps-\mu)/\sigma)$
Random variable $Y= (X- \mu)/\sigma$
$=Prob(Y \le y)$
$y = (qlim-qp(t)-ps- \mu)/ \sigma$
$= (1+erf(y/\sqrt{2}))/2$

Intrusion Detection System (IDS)

Intrusion Detection Systems (IDS) are established based on the types of attack [2] that are most commonly used. Network intrusions consist of packets that are introduced cause problems for the following reasons.

* To consume resources uselessly
* To gain system knowledge that can be exploited in later attacks.
* To interfere with any system resources intended function.

A. Classification Of Intrusion Detection Systems (IDS)

Intrusion Detection System briefly classified into two main categories:

1. ***Host Based Intrusion Detection***: HIDSs evaluate information found on a single or multiple host systems, including contents of operating systems, system and application files.
2. ***Network Based Intrusion Detection***: NIDSs evaluate information captured from network communications, analysing the stream of packets which travel across thenetwork .

B. Vulnerability-Assessment (IDS)

There are two basic models used to analyze the events and discover attacks:

- ***Misuse detection model*** – Intrusion Detection System detect intrusions by looking for similar activities such as vulnerabilities or known intrusion signatures.
- ***Anomaly detection model*** - IDS detect intrusions by searching « abnormal » network traffic.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

The misuse detection model is commonly referred as IDS commercial tool; always Vendors must update intrusion signatures. Anomaly detection based IDS model have the capability to detect attack symptoms without specifying attack models, but these models are very sensitive to false alarms. In the present study we have utilized the proposed IDS approach's based on the anomaly detection model.

Network Vulnerablity (NV)

An attack has connectivity with will have some level of vulnerability in network. Network Vulnerability (NV) [3] is impossible to eliminate entirely. One of the faction is, an attacker who knows about a security flow in the software that the network relies on that is unknown to the network administration. On being the network s are secured, in the mean time vulnerability occurs because of many hosts are administer by primary users of the system, who may lack the proper training to configure a secure computer system.

For the question-How the hosts are damaged by the network attacker, This gives the answer as Network attackers normally start their work by searching for vulnerability on the host they can communicate with on the target's network. When they use to increase the vulnerability on level of the host vulnerability in discovered. Thus the attacker can damage the communication channel those are in secure.

INTERIOR VERSES EXTERIOR ROUTING PROTOCOLS

Some routing protocols are designed to be used within an organization, while other routing protocols are designed for use between organizations. The current lead Interior Gateway Protocol (IGP) is OSPF. Other Interior Gateway Protocols include IS-IS, RIP, and EIGRP.

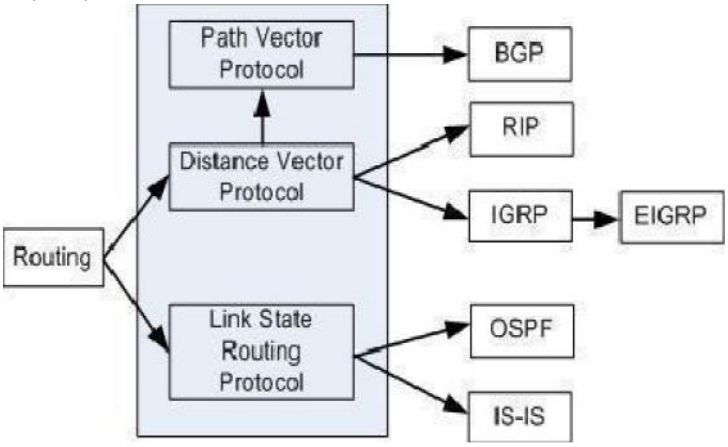


FIG 1: Interior Vs. Exterior Routing Protocols

A.DISTANCE VECTOR PROTOCOL

Distance vector protocols are based on two algorithms that are Bellman-Ford or Ford Fulkerson. The distance vector protocols choose the best path to a remote network by judge the distance. To be the best route each time least number of hops (routers) is determined.

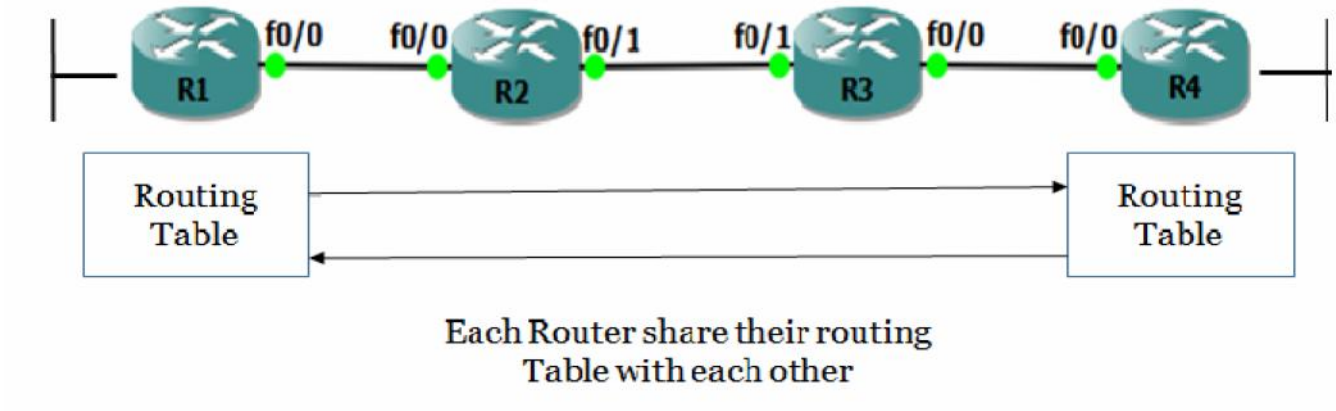


FIG 2: Distance vector routing protocol

B. Link State Routing Protocols

Link State protocols are also called shortest path first (SPF) or distributed database protocols, are build approximately a well-known algorithm of graph theory, E.W. Dijkstra's shortest path first algorithm. In the form of Link State Advertisement (LSA) each router shares its link information. Link state information is used by a link state router to generate a topology map and in the topology to choose the finest path to the destination. SPF tree is then applied to the LSDB to reach the destination to find the best path and the best path is then added to the routing table.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

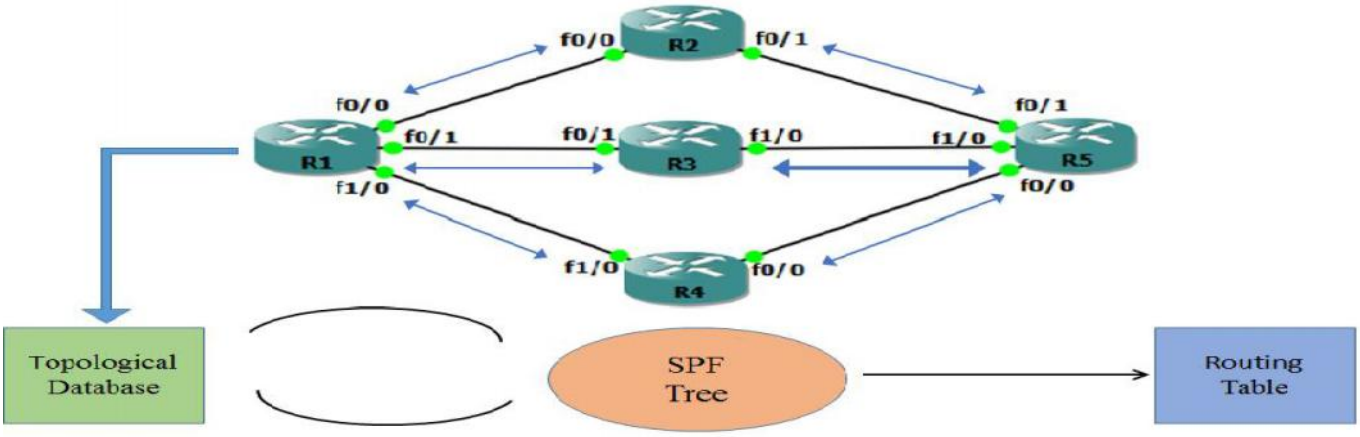


FIG 3: Example of SPF within Link State Protocols

C.Difference Between Distance Vector Protocol (DVP) And Link State Protocol(LSP)

In Table 1 we differentiate DVP and LSP as follows

DISTANCE VECTOR PROTOCOL(DVP)	***LINK STATE PROTOCOL(LSP)***
DSP is used in small networks, and it has a limited number of hops	LSP is used in larger networks, and it has unlimited number of hops.
DSP has a high convergence time	In LSP convergence time is low.
In DSP periodically advertise updates	In LSP advertises only new changes in a network.
advertises only the directly connected routers and full routing tables	Advertise the updates, and flood the advertisement.
Loop is a problem, and it uses split horizon, route poisoning and hold down as loop preventing techniques,	Link state has no loop problems.

TABLE1: Difference between Distance Vector Protocol And Link State Protocol

D.Open Shortest Path First (OSPF)

Open shortest path first [4]is a routing protocol that was developed try the interior Gateway protocol(IGP)working group of the Internet Engineering task firce for Internet protocol(IP)networks.OSPF is a link state routing protocol that is used to distribute information within a Single Autonomous System(AS).

- *Area(A*)-is a collection of networks ,hosts and routers all contained within an autonomous system.
- *Border Router(BR*)-At the border of an autonomous system special routers summarize the information about the area and send it to other ares.
- *BackBone(BB)*-Among the areas inside an autonomous system is a special area called BackBone; all the areas inside an autonomous system must be connected to the backbone.
- *Backbone Router(BBR*)-The router inside the backbone.

Each OSPF router stores the local network connection state with Link State Advertisement (LSA) and advertises to the entire AS. LSA is a packet that contains all relevant information regarding a router's links and the state of those links. Each router receives the LSA generated by all routers within the AS. The LSA collection then forms Link State Database (LSDB). Each LSA is the description of the surrounding network topology of a router. Hence, the LSDB reflects the AS network topology.Based on the link-state database, each router or system calculates a shortest-path spanning tree, with itself as the root, using the SPF algorithm. OSPF has five different packet types. Each packet has a specific purpose in OSPF route.

1. Hello packet.
2. Database description.
3. Link state request packet.
4. Link state update.
5. Link state acknowledgment packet.

E. Enhanced Interior Gateway Routing Protocol (EIGRP)

EIGRP [5]is a CISCO proprietary protocol, which is an improved version of the interior gateway routing protocol (IGRP).EIGRP is being used as a more scalable protocol in both medium and large scale network. EIGRP is said to be an extensively used IGRP where route computation is done through Diffusion update Algorithm (DUAL)[6]. However, EIGRP can also be considered as hybrid protocol because of having link state protocol properties

EIGRP uses the following four key technologies that combine to differentiate it from other routing technologies:

1) Neighbour discovery/recovery mechanism: Enables routers to dynamically learn about other routers on their directly attached networks
2) Reliable transport Protocol: It is responsible for guaranteed, ordered delivery of EIGRP packets to all neighbours

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

3) DUAL Finite State Machine: DUAL embodies the decision process for all route computations
4) Protocol Dependent Modules: EIGRP's protocol-dependent modules are responsible for network layer protocol-specific requirements

Implementation

A. Transmission Time (Tt)

In fig 4. We illustrate the transmission time(Tt) difference between EIGRP and OSPF

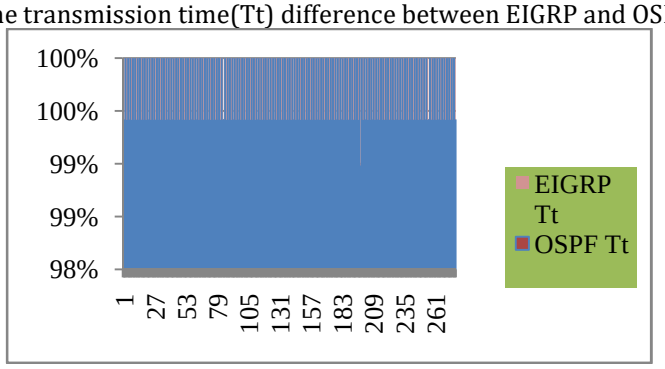


FIG 4: *Transmission Time (Tt)*

B. Variable Bit Rate(VBR) For Source To Destination

In fig 5, We illustrate the Variable Bit Rate (VBR) for source to destination packet transfer with time duration in microseconds(ms) .

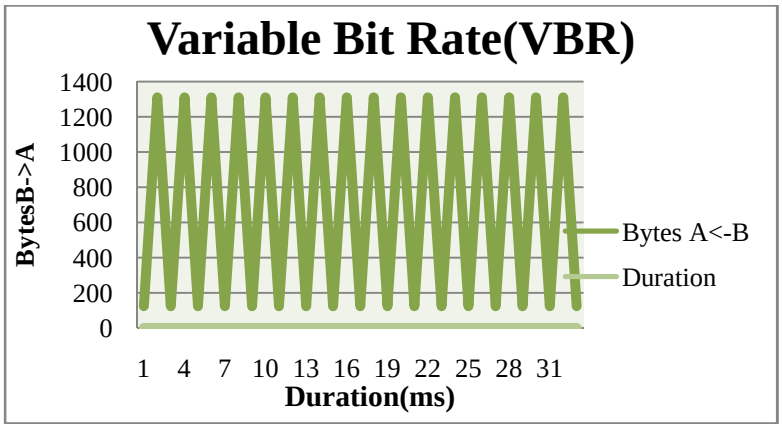


FIG 5: Variable Bit Rate(VBR)

C. Bandwidth (BW) for OSPF&EIGRP

OSPF sends partial updates when a link-state change occurs. The updates are flooded to all routers in the area. In a quiet network, OSPF is a quiet protocol. In a network with substantial topology changes, OSPF minimizes the amount of bandwidth used..Enhanced IGRP uses partial updates. Partial updates are generated only when a change occurs; only the changed information is sent, and this changed information is sent only to the routers affected. The following fig 6.Shows the comparison of Bandwidth (BW) [7]calculation in OSPF and Bandwidth (BW) calculation in EIGRP for our given experimental setup real time Network environment. This shows the end result as Enhanced IGRP is very efficient in its usage of bandwidth.

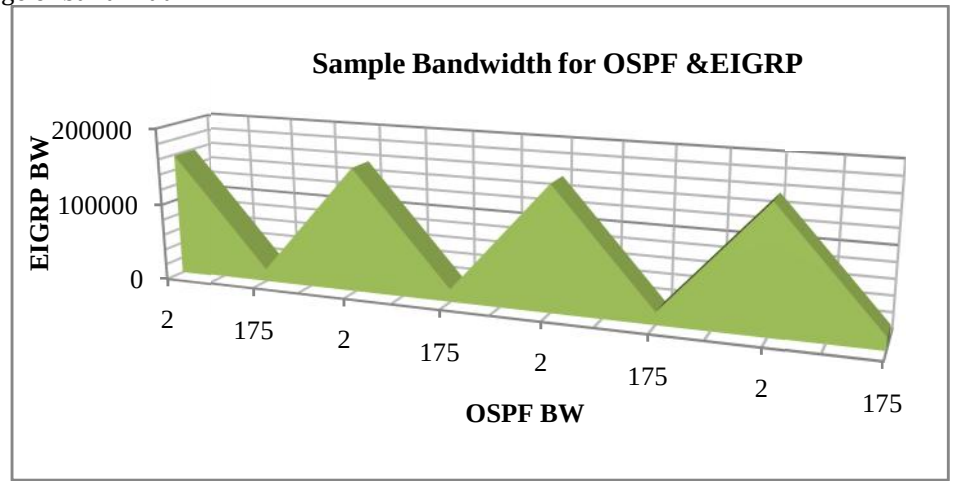


FIG 6: Sample Bandwidth For Ospf &Eigrp

E. Latency(Delay)

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

The following fig. 6 shows the delay time for our given experimental setup real time Network environment

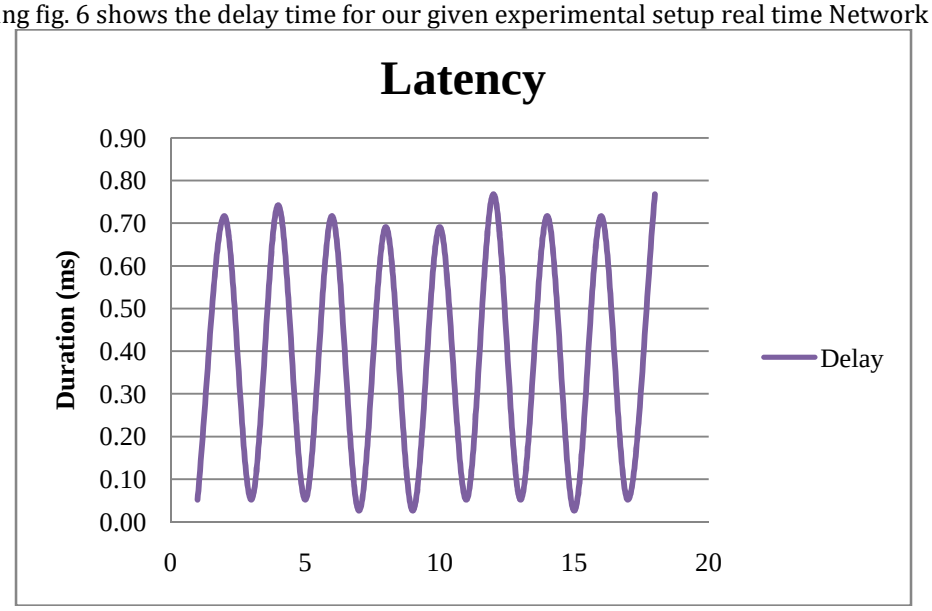


FIG7:*Latency(Delay)*

F. Flow Size Distribution(fs)

The following fig 7.Shows the Flow Size Distribution (fs) for our given experimental setup real time Network environment

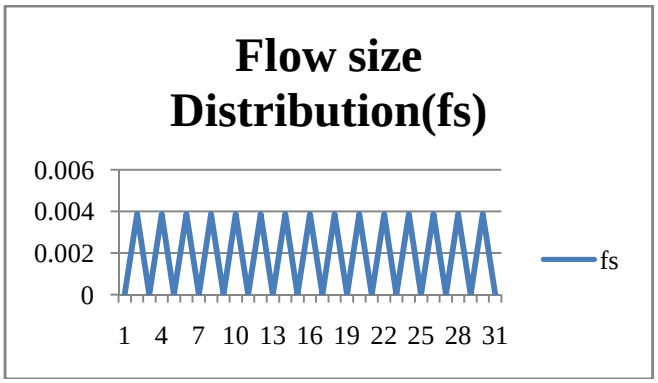


FIG 8: *Flow Size Distribution(FS)*

G. Entropy Calculation

The following fig 8.Shows the Entropy calculation for our given experimental setup real time Network environment.

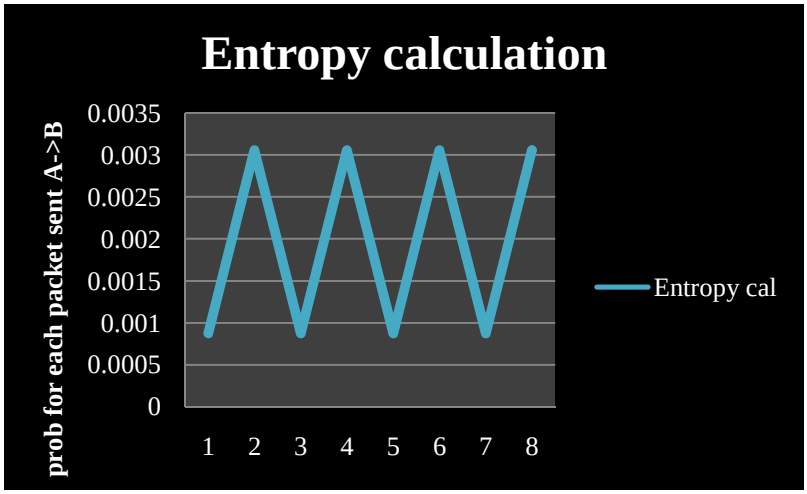


FIG 9: *ENTROPY CALCULATION*

H. Entropy Table Based On Autonomous System (AS) Network

The following Table 2we calculate the Entropy for the dataset in our real time network environment and achieve the probability result ***0.993453*** and normalization factor results ***2.7007*** on large data of about ***1,250,322*** .

Address A	Address B	Entropy calculation	Latency(ms)
192.168.xx.149	192.168.xx.253	0.000873	0.05
192.168.xx.4	192.168.xx.253	0.003055	0.72
192.168.xx.124	192.168.xx.978	0.000873	0.05
192.168.xx.149	192.168.xx.253	0.003055	0.74
192.168.xx.149	192.168.xx.12	0.000873	0.05
192.168.xx.12	192.168.xx.189	0.003055	0.72
192.168.xx.149	192.168.xx.103	0.000873	0.03
192.168.xx.4	192.168.xx.217	0.003055	0.69
	probability entropy	0.993453	

TABLE 2: Entropy Table Based On Autonomous System (AS) Network

Future Work and Conclusion

The performance of packet delay variation for EIGRP is better than for OSPF. We observed that the packet delay variation of OSPF network is high while the one of EIGRP network is low with available network load. In future Security analysis for both OSPF and EIGRP is also done.

We have implemented a compromised router detection protocol that dynamically infers, based on measured traffic rates and buffer sizes, the number of congestive packet losses that will occur. Subsequent packet losses can be attributed to malicious actions. Our protocol maintains log record and helps the user to know where the packet loss happened in the topology of the network. It also tells us whether it is due to malicious access or traffic congestion.

Many continuous-time and discrete-time traffic models have been developed based on traffic measurement data. The choice of traffic models involves at least two major considerations. The first consideration is accuracy. Accurate traffic models should include not only source behavior but also possible policing or congestion avoidance. The second consideration is ease of queuing analysis. Traffic models are useful only if network performance can be evaluated. It is always possible to evaluate network performance via computer simulations, but analysis is preferred whenever analysis is tractable

Our analysis on the packet capture and filter (Pf) is the initial attempt on various routing protocol to enhance bandwidth, to reduce delay and throughput analysis. Our study may be enhanced to different graph models for easy analysis of network vulnerability and security within a human – computer community using the idea of association and connectedness.

References

1. P. Prasanya Devi and M .Ravindran ,“ *A Study of packet capture and filtering on a Network Environment*” , NCCIS2012 , national conference at Dwaraka Doss GoverdanVaishnav College, Chennai , ISBN : 978-81-905709-4-0 , PP. 10-18.
2. M .Ravindran and P .Prasanyadevi ,“ *Risk Assessment Model (RAM) with Graph Based (GB)On vulnerable TCP/IP Network traffic* , International conference @ Mysore Universiy , Mysoru.
3. Behrouroz A . Forourzan , “ *Data communications and networking* “ , Fourth edition , Tata Mcgraw Hill Education Private Limited , New Delhi ISBN : 978-0-07-063415 ISBN 10: 0-07-063414-9.
4. Cisco “*OSPF Design Guide*” http://www.cisco.com/en/US/tech/tk365/technologies_White_paper09186a0080094e9e.shtml
5. Cisco, “Internet Technology Handbook” http://www.cisco.com/en/US/docs/internetworking/technology/handbook/ Enhanced_IGRP.html
6. Cisco, *“IP Routing, Introduction to EIGRP”* Document ID: 13669. http://www.cisco.com/en/US/tech/tk365/technology_tech_note09186a0080093f07.shtml
7. Amanpreet Kaur, Dinesh Kumar,“*Comparative Analysis of Link State Routing Protocols OPSF and IS-IS”*,IJCST, vol-3, issue-4, july-aug 2015.

Detection of Gait Abnormality with EELM

[1]M.Pushpa Rani, [2]S.A.Vajiha Begum
[1,2]Department of Computer Science, Mother Teresa Women's University, TamilNadu, Kodaikanal

ABSTRACT *rain diseases are characterized by progressive nervous system dysfunction which create severe gait abnormalities. Improper diagnosis of brain disorder can lead to inappropriate treatment and serious consequence on patient health. Diagnosis of diseases is presently based on neurologist observation is tough for early detection. Hence Gait analysis is an effectual tool for the early analysis and identification of brain disorder by computational techniques. Hence for the effective identification of neurological disorder, machine intelligence-based gait analysis technique is proposed in this work to gain the advantages of fast and accurate identification of the gait bound diseases.*
Keywords: *Gait Analysis; Enhanced Extreme Learning Machine; Brain Disorder.*

I. INTRODUCTION

Gait is basically the pattern of movement, means how we walk and the studies stating that every person has a unique gait pattern[14]. Examination of human gait patterns can afford useful information related to the physical and neural functions, and it may pay way for the analysis of brain disorders in extreme conditions [10,11]. Thus, the analysis of human gait patterns is essential to recognize and classify the type of brain disorder of the patients.

Alzheimer's Disease(AD), Huntington's Disease(HD), Parkinson's Disease(PD), Amyotrophic Lateral Sclerosis(ALS), stroke, paraneoplastic disorders and multiple sclerosis are some of the neurological disorders [1,15]. The nervous system controls the process of walking. If a portion of neural network that controls this process is damaged, can cause to produce abnormal movement in the person, is the primary feature of neurodegenerative disease [2].

Accurate diagnosis of these diseases in fast way help the patient to receive the appropriate care as soon as possible. Additionally, identifying the disorder in a more simplified way will help the doctors in the early analysis and treatment process. To achieve this objective machine intelligence-based gait analysis technique is proposed, shows how the human gait is related to neurological disorder classification. The proposed work is divided into Feature extraction phase, Classification phase and Identification phase.

Initially the gait video is preprocessed and the gait features are extracted with suitable feature selection technique. Then the extracted features are ranked to avoid redundant features. The ranked features are used for the training and testing process with the Enhanced Extreme Learning Machine (EELM) technique for the classification of the brain disorders. Then intelligent based gait analysis technique is used for the perfect identification of particular neurological disorders by comparing with the standard gait patterns in the database. For comparison of the gait pattern of patients suffering from brain disorder, a healthy control group is taken. The stride intervals of the foot areanalyzed and compared for identification of disease, using machine learning based techniques.

II. LITERATURE SURVEY

Pushpa Rani et al.,[12] presents, a fast and effectual classification method called the Extreme Learning Machine(ELM) algorithm to classify the abnormal gait patterns. The t-test ranking algorithm is used for the classification of gait patterns. The study shows that when the number of categories for the classification task is large, ELM attains better and stable classification accuracy with less training time comparative to SVM. This system shows the result of ELM with PCA gives accuracy of 97.98% and T-Test gives the maximum accuracy of 99.21%.

Huang et al.[6] presents the kernelized version of ELM (KELM), which shows no randomness in providing connection weights between input and hidden layers. The Comparative result shows that KELM can achieve better performance with easier implementation and faster training speed in different classification tasks compared to SVM.

Hui-LingChen et al.,[8] proposed an efficient hybrid method, mRMR-KELM for early diagnosis of Parkinson's disease (PD). ELM and KELM used best possible parameters to train the ideal models for PD diagnosis with the help of mRMR filter for feature selection. It is noted that mRMR-KELM achieves the very optimistic classification accuracy of 96.47% via 10-fold cross-validation (CV) analysis can distinguish well enough between patients with PD and healthy persons.

Pushpa Rani [5] proposed a method Hybrid Extreme Learning Machine (HELM) which chooses the input weights and biases for the hidden nodes were constructed with the use of Analytical Network Process (ANP) algorithm. Experimental results prove that the Hybrid Extreme Learning Machine (HELM) technique for human gait pattern classification results in higher accuracy compared to the existing ELM and SVM techniques.

Yunfeng Wu et al. [13] presents a statistical study on the gait stride interval in ALS. The two statistical parameters such as the mean of the left foot step interval and the improved method of Kullback-Leibler divergence, from the probability density functions (PDF) of stride interval is estimated. This study got result of 82.8% accuracy by differentiating the gait patterns of the ALS patients and healthy control groups with the support of least-squares support vector machine (LS-SVM).

The studies show the possibility of gait being used in the classification and identification of neurological disorders, involving high computational cost, increased time with less accuracy rate. To overcome the above limitations, it is proposed to design and implement machine learning based gait analysis techniques targeting improved performance with minimum cost.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

III. IDENTIFICATION OF GAIT ABNORMALITY

The gait analysis method is proposed to identify the neurological disorders consists of three phases namely feature extraction phase, classification phase and identification phase. The feature extraction phase deals with preprocessing of gait video, silhouette feature extraction and ranking of gait features. The classification and identification phase deals with the training and testing of collected ranked gait features and finally identify the type of brain disorders.The block diagram of proposed system for Analysis of Gait abnormality is shown in Figure 1.

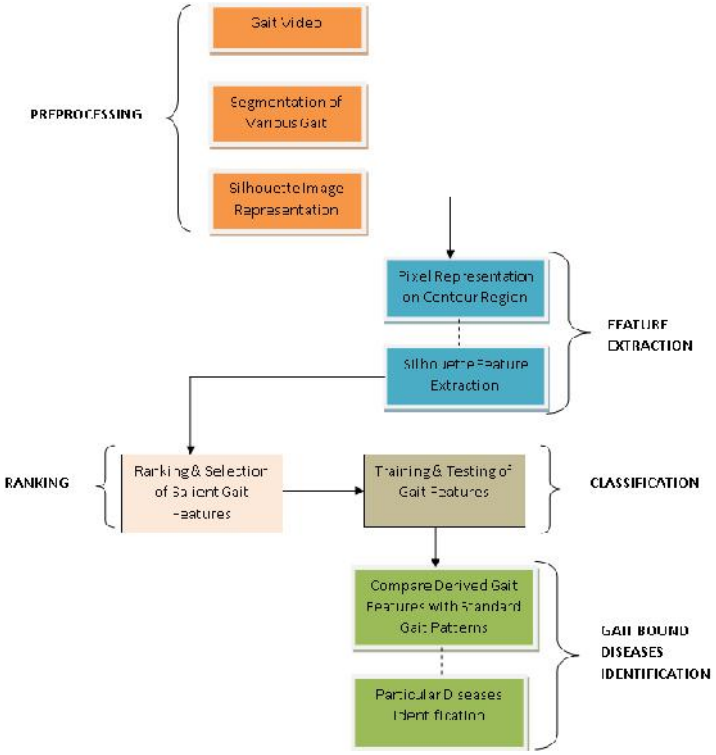


Fig1. Block Diagram of proposed system for Analysis of Gait abnormality

A. Feature Extraction Phase:

First step of the proposed method comprises of preprocessing of gait video, silhouette features extraction and ranking of gait features.

1) Preprocessing:

Initially gait video is taken and the adaptive gait silhouette extraction technique [4] is used to obtain the walking figure from the gait video. The median based background elimination method is applied and then some morphological operations are used to correct the noises and extracting the particular regions of silhouette image. The silhouette images are normalized to original size to do the feature extraction simpler and in slighter time.

2) **Silhouette Feature Extraction:**

The outermost Contour technique is applied here for the improved extraction of silhouette feature which collects the pixel from the contour region [9]. The right end pixel and the left end pixel on the outline region in every row of a silhouette image is measured as Outermost Contour. The centroid of the Outermost Contour is considered and the number of pixel on the outermost contour is noted. The desired silhouette features are detected and extracted by calculating the distance between each outermost contour pixel and the centroid.

3) **Ranking of Gait Features:**

Then ranking of gait features is performed by mRMRalgorithm to select and rank the essential features and avoid selecting redundant features for improved performance.

Minimum Redundancy Maximum Relevance (mRMR):

The mRMR is a feature selection filter technique implemented to select features which are associated to the target class (maximum relevance) and non-redundant subset features (minimum redundancy) [3]. The correlations between selected features and target class is discovered by using the collected common information. This mRMR technique is implemented to choose features based on maximizing the joint dependency of top ranking unique features on the target class and avoids choosing redundant features. The optimization standard of mRMR is as follows:

$$\max_{x_j \in X - S_{k-1}} \left[I(x_j, c) - \frac{1}{k-1} \sum_{x_j \in S_{k-1}} I(x_j, x_i) \right] \qquad (1)$$

where X is the complete set of gait features, c is the target class feature, $x_j$is the jth feature, S_{k-1} is the set of top k-1 features designated in the previous iterations, I is the mutual information, I (x_j, c) and I (x_i , x_j) show common information between distinct features $x_j$with class c and common information between features x_j and x_j, correspondingly.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

B. Classification Phase:

The classification of neurological disorders is done with the training and testing of collected ranked gait features. The Enhanced Extreme Learning Machine (EELM) technique is implemented for the classification purposes which combines the concept of ELM and SRC techniques. This technique avoids random input and provides choosing suitable inputs and hidden biases for gaining better result.

The learning speed of the Extreme Learning Machine(ELM) is extremely fast while it cannot handle noise well, whereas Sparse Representation based Classification(SRC) method illustrates eminent robustness to noise while it suffers high computational cost. In order to overcome their drawbacks and include their particular advantages, this work proposes an approach of combining ELM and SRC technique for classification of normal and abnormal gait patterns with minimum time.

C. Identification Phase:

The identification phase deals with particular recognition of the type of neurological diseases based on the gait patterns. For the accurate identification of brain disorders, intelligence-based gait analysis technique is implemented. For comparison of the gait pattern of patients suffering from neurological disorder, a healthy control group is taken. The stride intervals of the foot are compared for identification of particular brain disorder. The particular identification of the type of brain diseases based on the gait patterns is done by Sparse Representation based Classification(SRC) technique.

IV. RESULT AND DISCUSSION

The proposed work deals with the classification and identification of neurological disorder using gait analysis techniques. The data used in this study are gained from the CGA normative gait database. The Figure 2 shows the stride times of the healthy persons and persons with brain disorders [7]. It shows that the stride times associated with brain disorders are more flexible and have maximum deviation.

The table I illustrates the comparison of classifiers based on classification accuracy of movement disorders. This illustrates that the proposed EELM algorithm gives improved accuracy rate than the other methods. This study shows that the EELM has a higher accuracy rate of 99.3%, while that of other classifiers. The Figure 3 shows the graphical representation of classification accuracy of brain disorders for each classifier.

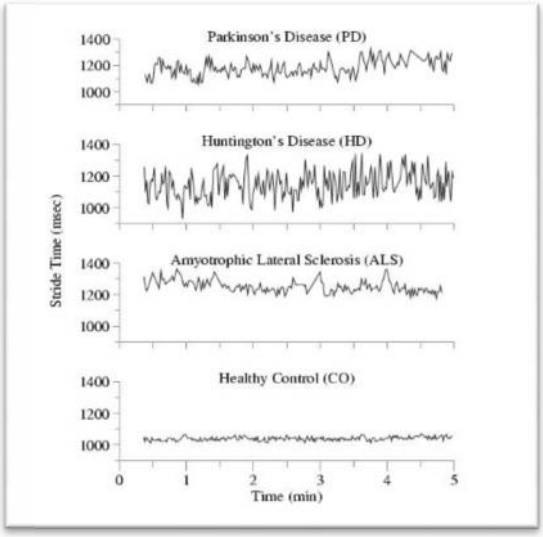


Fig. 2. Stride Time for each type of diseases

TABLE I:

COMPARISION OF CLASSIFIERS BASED ON CLASSIFICATION ACCURACY

CLASSIFIERS	ACCURACY (%)
SVM	94.5
ELM	95.8
HELM	99.2
EELM	99.3

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

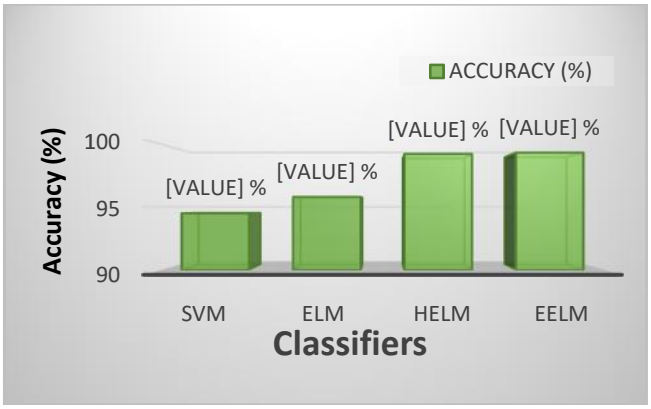


Fig 3. Histogram of classification accuracy for each classifier.

The table II shows the comparison of classifiers based on the time of classification of the gait bound disease. The result shows that the EELM algorithm have minimum time for classification of brain disorders but other classifiers show more time for classification. These analysis reveals that the EELM classified the disease at 0.36 sec, while that of other classifiers. The Figure 4 gives the histogram representation of time of classification of brain disorders by each classifier.

TABLE II:
COMPARISION OF CLASSIFIERS BASED ON CLASSIFICATION TIME

CLASSIFIERS	TIME (sec)
SVM	0.87
ELM	0.69
HELM	0.38
EELM	0.36

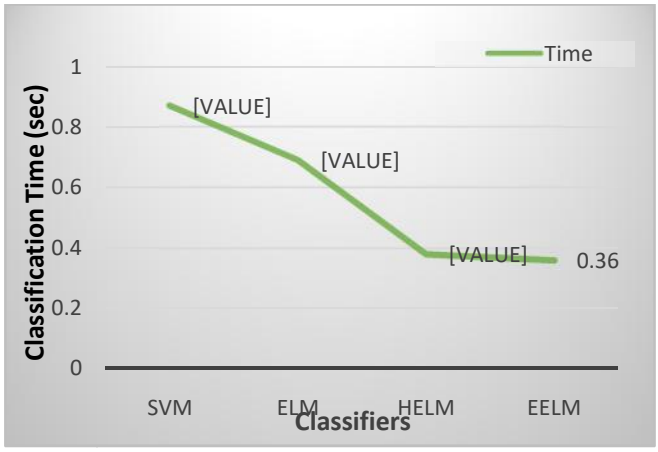


Fig 4. Histogram of classification time for each classifier.

V. CONCLUSION

An automatic diagnosis method for the classification and identification of brain disorders using gait analysis technique is proposed. This will result in the early diagnosis and treatment process for the neurological disorder. The proposed method deals with feature extraction, classification and identification phases for the detection of gait abnormality. The Enhanced Extreme Learning Machine (EELM) technique is implemented by the combination of Extreme Learning Machine (ELM) and Sparse Representation based Classification (SRC) for the training and testing of gait features and for the identification of brain diseases in particular. The EELM technique achieves higher classification accuracy with minimum time and less computational cost than the other algorithms for improved identification of gait diseases. In future some advanced machine learning algorithms can be implemented to enhance the classification accuracy added with reduced training time.

REFERENCES

1. Sandra Amor, Fabiola Puentes, David Baker, Paul van der Valk, "Inflammation in neurodegenerative diseases", Immunology, 2010 Feb; 129(2): 154–169.
2. T. Carletti, D. Fanelli, A. Guarino, A new route to non invasive diagnosis in neurodegenerative diseases?, Neuroscience Letters 394 (2006) 252–255

3. Peng, H., F. Long, and C. Ding, Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005. **27**(8): p. 1226-1238.
4. M.Pushpa Rani and G.Arumugam, "An Efficient Gait Recognition System for Human Identification using Modified ICA ",International Journal of Computer Science and Information Technology ,Vol 2, No I, pp55-67,2010.
5. M.Pushpa Rani, "Abnormal GAIT classification using hybrid ELM," *2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*, Toronto, ON, 2014, pp. 1-8.
6. Huang, G.B., et al., Extreme Learning Machine for Regression and Multiclass Classification IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics, 2012. **42**(2): p. 513-529.
7. The National Institutes of Health-sponsored Research Resource for complex physiological signals https://www.physionet.org/physiobank/database/gaitndd/
8. Hui-LingChen et al.,"An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson′s disease",Elsevier,NeurocomputingVolume 184, 5 April 2016, Pages 131-144
9. Lili Liu et al.,"Gait Recognition Based on Outermost Contour", International Journal of Computational Intelligence Systems, Vol. 4, No. 5 (September, 2011)
10. Rueterbories J, Spaich EG, Larsen B, Andersen OK. Methods for gait event de- tection and analysis in ambulatory systems. Med Eng Phys 2010;32:545–52. doi: 10.1016/j.medengphy.2010.03.007 .
11. Yang M, Zheng H, Wang H, McClean S, Hall J, Harris N. A machine learning approach to assessing gait patterns for complex regional pain Syndrome. Med Eng Phys 2012;34:740-746. doi: 10.1016/j.medengphy.2011.09.018.
12. M. Pushpa Rani, G.Arumugam, "Children Abnormal GAIT Classification Using Extreme Learning Machine", Global Journal of Computer Science and Technology,Vol. 10 Issue 13 (Ver. 1.0) October 2010, p. 66.
13. Yunfeng Wu and Sin Chun Ng," A PDF-Based Classification of Gait Cadence Patterns in Patients with Amyotrophic Lateral Sclerosis" 32nd Annual International Conference of the IEEE EMBS ,Buenos Aires, Argentina, August 31 - September 4, 2010
14. M Pushparani and D Sasikala, "A Survey of Gait Recognition Approaches Using PCA and ICA",Global Journal of Computer Science and Technology, May 2012
15. Athisakthi, Dr.M.Pushparani, A, "Detection of Movement Disorders Using Multi SVM", Global Journal Of Computer Science And Technology, Feb 2013.

Proposed Architecture for Sentiment Analysis in MongoDB Database

[1]V.Lakshmi Praba, [2]M.Bhuvaneswari
[1] Assistant Professor, Rani Anna Govt. College For Women, Tirunelveli, India
[2] Research Scholar, Reg.No: 17224012162051, Research Department of Computer Science, Rani Anna Government College for Women, Manonmaniam Sundaranar University, Abishekapatti, Tirunelvel - 627 012,Tamil Nadu, India.

ABSTRACT *Text Mining and Natural Language Processing form the core concepts of the emerging area - Sentiment Analysis. The increase in the use of Internet and Web have made its users rely on online for all their essential services like healthcare, products, fashion, education etc. They also express their opinions, feelings or attitudes towards a particular concept, service or brand new products as reviews, ratings, YouTube comments, debates, social network sharing etc. Sentiment Analysis is used to find the polarity of the opinions of the users in those sources. The choice of suitable approach of analysis and algorithms for sentiment analysis, classification and polarity detection makes its application more effective. This paper discusses on the available Sentiment Analysis approaches, algorithms and research areas emerging in this field. The dataset for sentiment analysis is normally stored in traditional databases but this paper introduces an emerging concept of NoSQL database MongoDB which handles real time data like user reviews and comments in an efficient and effective manner . It also proposes a hybrid architectural model of sentiment analysis to extract the advantages of its component approaches.*

Keywords: *sentiment analysis, feature selection, sentiment classification, lexicon approach, MongoDB*

I. INTRODUCTION

Sentiment Analysis also known as Opinion mining or Emotion AI refers to the use of natural language processing, text analysis, computational linguistics and biometrics to systematically identify, extract, quantify and study afftective states and subjective information from social media, blogs, ratings and reviews. It is an emerging field of research in text mining.The increase in the use of web and internet drives the need for Sentiment Analysis because users most often express their opinions or feelings as textual data. It aims to determine the attitude of a speaker, writer or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction or event. The attitude may be a judgement, evaluation, affective state or intended emotional communication.

Sentiment Analysis is a classification task that detects the sentiment of a text based on its polarity as Positive or Negative or Neutral, Good or Bad, Subjective or Objective. Based on the level of text selected for analysis Sentiment Analysis is mainly classified as – Document level, Sentence Level or Aspect Level. Document level classifies the whole document about a single topic as expressing positive or negative or neutral sentiment. Sentence Level analysis is performed on each sentence of a document to find whether it is objective or subjective. Aspect level analysis identifies the individual entities and their aspects in a document and then finds the polarity of each aspect of the entities. A single entity may have different polarities for different aspects. For example a user review may express a positive opinion on one aspect of a hotel but a negative opinion of another aspect of the same hotel. So aspect level analysis is used in applications where a detailed study of the opinions are essential.

Sentiment Analysis is largely performed on text but recent focus is on multimedia content with audio, video and images. The main sources of data for sentiment analysis are from the Internet like product reviews, ratings, tweets, debates, stock market information, news articles, youtube comments, amazon, social networking sites and some benchmark datasets. Sentiment Analysis is used in Business Intelligence, Bias identification of new sources, identifying appropriate advertisement content, politics, law, sociology, psychology, beauty, fashion, sports, health, traveller's experiences with hotels, tourist spots, airlines[1] etc. Sentiment Analysis faces certain challenges in analyzing dynamic events due to fast paced data, domain dependent classification techniques, use of contents like hashtags, emoticons, links, sarcasm etc., location oriented analysis, trustworthiness of user or robot reviews etc. [2]

Research Areas on Sentiment Analysis are mainly focused on Feature Selection, Sentiment Classification, Cross-Domain analysis, Emotion Detection, Resource Development etc. [3][4]. In this paper we discuss on the first three techniques of sentiment analysis mentioned above. This paper is organized as follows, Section 2 describes about Feature Selection, section 3 discusses about Sentiment Classification, section 4 explains Cross-Domain Analysis,section 5 discusses on MongoDb database, section 6 reviews the existing literature on sentiment analysis, section 7 presents the proposed architecture and section 8 presents the conclusion.

II. FEATURE SELECTION

A. *Preprocessing*

The preliminary step of Feature Selection is preprocessing of input text. The preprocessing methods used are[5][6]:

- Handling Negations like won't, can't etc.
- Deletion of URL links.
- Removing repeated letters in words Ex. Cooool to cool.
- Expanding acronyms.
- Removing stop words like the, is, at etc.
- Removing numbers.

- Upper case to lower case conversion.

B. *Features*

The features extracted from an input text can be categorized as – Sentiment Features, Punctuation Features, Syntactic Features, Semantic Features, Unigrams, Pattern-Related Features, Top words etc.[7]

a) Sentiment Features: Features that decide the sentiment polarity of the text like positive or negative words, emoticons, hashtags etc.
b) Punctuation Features: Features related to the use of punctuation mark like comma, fullstop, semicolon, colon, exclamation marks, parenthesis, apostrophe, double quotes etc.
c) Syntactic Features: Features like particles, interjections, pronouns, nouns, adjectives, verbs, adverbs etc.
d) Semantic Features: Features that reveal the meaning of words and the logic behind it like highly sentimental words, uncertainty words, opinion words or expressions etc.
e) Unigrams: Words that are used to represent the similar meanings, words that are given in a broader or specific sense etc.
f) Stemming Features: Parts of Speech words that have the same root word.
g) Top Words: Words that appear frequently in the training set.

C. *Feature Selection and Weighting Methods*

The features selected using feature selection methods are weighted using Feature Presence(FP) or Feature Frequency(FF) methods. FP method calculates feature value by considering only their presence or absence. FF method counts the number of features in the input text.

1) *Mutual Information (MI)* : Mutual Information method selects features that are informative of their classes and are not unevenly distributed among the sentiment classes. It concentrates on terms that appear less frequently in the input data. It is given by eqn (1). as

$$MI(f,c) = \sum_{c \epsilon C} \sum_{f} P(f,c) \log \frac{P(f,c)}{P(f)P(c)} \quad (1)$$

where P(f,c) indicates the joint probability distribution function, P(f) and P(c) denotes the marginal probability distributions of f and c, and C is the classes: POSITIVE and NEGATIVE.

2) *Information gain(IG)*:Information gain calculates whether a feature is relevant for sentiment analysis by analyzing its presence or absence in the text . It is calculated using eqn(2) as

$$IG(f,c) = -\sum_{c,c} P(c) \log P(c) + \sum_{f,f} P(f) \sum_{c,c} P(c|f) \log P(c|f) \quad (2)$$

P(c|f) is the joint probability where class C and feature f is occur together. P(c) denotes the marginal probability.

3) *Chi-square (χ^2)*: Chi-square is a measure of the deviation of the observed count of a frequency from the expected counts[3]. It is measured using eqn(3) as

$$\chi^2(f,c) = \frac{N(WZ-YX)^2}{(W+Y)(X+Z)(W+X)(Y+Z)} \quad (3)$$

W, X, Y, Z denotes the frequencies, indicates the presence or absence of feature in the sample. W is the count of samples in which feature f and c occurred together. f is the feature and c is the class. The following table shows what each symbol W, X, Y, Z indicates

TABLE: I
2×2 CONTINGENCY TABLE OF FEATURE (F) AND CLASS(C)

	c	ĉ
f	W	X
$\bar{f}$	Y	Z

4) *TF-idf(Term Frequency-Inverse Document Frequency)* :TF-idf is a weighting scheme that measures a word relevance in the dataset based on the frequency of appearance of word in the sample. It is calculated using eqn(4) as

$$TF\text{-}idf_i = t_{i,j} \times \log \left(\frac{N}{df_i} \right) \quad (4)$$

$TF\text{-}idf_i$ is the weight of a term i,$t_{i,j}$ is the frequency of term i in sample j. N is the total number of samples in the corpus. df_i is the number of samples containing term i.

III. SENTIMENT CLASSIFICATION

Sentiment Classification techniques can be broadly classified into Machine Learning Approach, Lexicon Based Approach and Hybrid Approach.

A. *Machine Learning Approach*

This approach makes use of syntactic and/or linguistic features for classification. It consists of a training set and a test set of data. Based on the training set classification the test data is classified. This approach is classified as:Supervised Learning, Unsupervised Learning and Semi-supervised Learning. A hybrid of the above learning methods can also be performed.

1) *Supervised Learning* : The sample data is classified into known class labels. The commonly used classifiers are Decision Tree Classifiers, Linear Classifiers, Rule Based Classifiers and Probabilistic Classifiers. Decision Tree Classifiers builds by continuously splitting the data based on a certain parameter to form a tree structure. The tree has two components - decision nodes and leaves. The decision nodes denote data split and decision leaves are the decisions or final outcomes. An example of Decision Tree Algorithm is ID3 – Iterative Dichotomiser. Linear Classifiers classifies based on the value of a linear combination of the characteristics .It is the fastest and most commonly used type of classifier. The two popular algorithms in this are Support Vector Machine(SVM) and Neural Network(NN) Algorithms. SVM determines linear separators in the search space which best separates the different classes. Text data is ideally suited for SVM classification because of its sparse nature and they tend to correlate to one another and get organized into linearly separable categories. It is best suited for classifying reviews according to their quality. NN takes real vectors as inputs. Words in input text should be represented in the form of real vectors which is a function of the word frequencies in the document. This is well suited for identifying non-linear patterns where there is no one-to-one relationship between input and output. It is a model of artificial neurons built across two or three layers. A set of weights are associated with each neuron to compute a function of its input. Rule Based Classifier models the data space with a set of rules. Based on the criteria to generate rules the training set constructs all the rules. The two criteria that govern the rule construction process are support and confidence. Support is the number of instances of the training set which are relevant to the rule. Confidence is the conditional probability of the rule that is satisfied. Probabilistic Classifier outputs a probability distribution of input over a set of classes. Popular algorithms in this classifier are Naïve Bayes Classification, Bayesian Network, Maximum Entropy classifer. Naïve Bayes Classifier computes posterior probability of a class based on the word distribution in the document. It makes use of Bayes theorem to predict the probability of a class and Naïve's assumption that all features are independent. Bayesian Network model is a directed acyclic graph whose nodes are random variables and edges represent conditional dependencies. This model assumes that all the features are fully related. This is less frequently used because of its computation complexity. Maximum Entropy Classifier is based on Principle of Maximum Entropy which selects the model with maximum entropy value from all the models that fits the current state of knowledge of the training data.

2) *Unsupervised Learning*:This approach of machine learning also has training set and test set but classifies data into unknown class labels. This method reveals the structure or relationship between different inputs. Clustering is the most common method of unsupervised learning which groups the input data into different clusters and puts the new input into the appropriate cluster. This method is more challenging because same data may be placed in different clusters depending upon the clustering process. Algorithms that can be used are Expectation Maximization, k- Means etc.

3) *Semi-supervised Learning*: In this approach some data is classified with known labels and a part of data is classified with unknown labels. A majority of the real life applications are representatives of this type of learning. Some of methods used are Generative models, Discriminative models like Smoothness, Graph/Manifold regularization method etc. Both Supervised and Unsupervised algorithms can be used.

B. *Lexicon Based Approach*

This approach is based on the idea that the polarity of a sentence is calculated from the polarity of the words in the text. This may be Dictionary Based Approach or Corpus Based Approach.

1) *Dictionary Based Approach* : The word of a text are matched with a lexicon. The lexica is constructed with a few seed words with positive and negative sentiments. The construction continues by searching synonyms, antonyms, n-grams from online dictionaries. The process continues until no more words are found. Now the lexica is built completely. This method fails to give domain or context specific meanings.

2) *Corpus Based Approach* : This approach finds context specific opinion words. It makes use of seed list of opinion words to find more words from a large corpus. It also considers connective words AND, OR, BUT, HOWEVER etc to determine two adjectives are of same or different orientations. Statistical or semantic approach are the two concepts of this approach. Statistical approach[8] makes use of statistical techniques to find the polarity of words by determining the frequency of occurrence of words, co-occurrence of words etc. Semantic approach[9] depends on different principles for computing similarity between words and assign similar sentiment values to semantically close words. Both semantic and statistical methods can be used in combination as a hybrid method.

IV. CROSS-DOMAIN ANALYSIS

Cross-domain analysis is the process of adapting a classifier which is trained using class labels of one domain (source domain) to classify the test data of another domain(target domain). For example trained data from a book review may be used to test the data extracted from a movie review. The trained source domain data can be used alone or in combination with some trained target domain data to handle test data of the target domain. For example trained data from one product review can be used to analyse untrained data of another product review. Cross-domain analysis poses many technical challenges because sentiment analysis is highly domain-specific. Moreover words or phrases that frequently in one domain might not appear or less frequently appear in another domain. The interpretation of a word in

one domain might give a different meaning in another domain. A classifier whose performance is excellent in one domain might be very poor when applied to another domain. The polarity of words might differ from source domain to target domain. For example the word 'Cheap' has positive polarity in the case of a hotel or product but it represents negative polarity in movie reviews where the user may comment the story line of the movie shows the cheap mentality of the writer. There are a variety of techniques used for Cross-Domain Analysis- Spectral Feature Alignment(SFA) Algorithm, Structured Correspondence Learning(SCL) Algorithm, Joint Sentiment Topic Model, Deep Learning, Topic Modeling, Thesaurus Based Approaches, Case Based Reasoning Approach, Feature Based Approach, Graph Based Approach, Distance Based Modeling, Domain Similarity and Complexity Approach, Knowledge-Enhanced Meta- Classifier Model etc.[4].

V. MONGODB

Traditional relational databases were initially used for handling data. But big data deals with an enormous amount of data with key features as velocity, veracity, volume and variety. Data may be of different formats and often semi structured or unstructured. These factors challenges the traditional databases and they failed to provide adequate solutions . This lead to the concept of NoSQL database which is a non-relational databases that eliminates the concepts of the normal tables, rows, columns etc. NoSQL databases is designed to handle huge amount of data and variety of data with proper time and cost constraints. Some of the usecase where the NoSQL databases used are social network graphs, search and retrieval etc. MongoDB, a NoSQL database, is a cross-platform, document oriented, open source database. It stores data in collections made of individual documents. Documents in this database are in schemaless binary representation of JSON format called BSON. It best suits to big data analytics because it proves effective in handling unstructured data and extracts full cloud computing and storage benefit by providing automatic sharding facility to store large amount of data in cloud. It also enables location based data analytics and operations, get real time data reporting and analytics, **Capitalize on sensor data and connected devices, Powering content management systems (CMS), Push out new versions of mobile apps fast, Personalize data to tailor user's experiences** . It also supports the regular database features like adhoc queries, indexing, aggregation, scaling, replication, load balancing, programming languages support especially Java etc. MongoDB has official drivers for major programming languages and development environments. MongoDB is well apt for real time analysis from user reviews, comments, browsing history, youtube comments, social networking site information etc. To summarise MongoDB is able to meet the data challenges of today which cannot be accomplished well with relational databases.

VI. REVIEW OF LITERATURE

A. *RESOURCE BUILDING*

Lorenzo Gatti et al.[10] derived SentiWords lexicon based on the prior polarity computation. The lexica was trained using Warr data set. The performance of the lexica was tested using classification and regression in SemEval and TreeBank data sets. The lexica experimentally proved of having high precision and high coverage than manually annotated sentiment lexicas with a precision accuracy of 0.602. Supervised Learning algorithm SVM is the process of building the lexica.

B. *Sentiment Preprocessing*

Wanxiang Che et al. [11] proposed a framework Sent_Comp as a preprocessing tool to compress the sentences used for Aspect-Based Sentiment Analysis. This tool classifies lexicons in each sentence based on nine rules for Chinese language and decides 'delete' or 'reserve' of lexicons and thus compresses sentence making it shorter and easier to parse. Corpora database is used for evaluating the tool and it proved to be domain independent and effective for Aspect Based Analysis with a good performance indicator values for precision, recall, F-score etc.

C. *Sentiment Analysis*

Xiaojiang Lei et al.[12] proposed a Sentiment Based Rating Prediction Method(RPS) for user reviews on products based on three factors user sentiment similarity, interpersonal sentimental influence and item's reputation similarity. Corpus Based Statistical based LDA algorithm. The dataset used is Yelp. This method was experimentally proved to improve recommendation performances and proves that these three factors make great contribution to rating prediction.

Yoonjung Choi et al.[13] investigated a knowledge based coarse-grained +/- effect Word Sense Disambiguation(WSD) approach to recognize whether word instances in a corpus are used with + effect, -effect or Null senses. This is done by constructing a coarse grained sense inventory by grouping senses and developing selectional preferences for sense group. This disambiguates the sense of a word in different contexts. The LDA algorithm is used which is a Corpus Based Statistical Approach. The performance was tested in Senseval-3 dataset and experimentally proved to disambiguate words with an accuracy of 0.83.

Wei Zhao et al.[14] proposed a novel deep learning framework for product review sentiment classification using weak supervision signals. This paper focussed on developing two deep learning frameworks- WDE-CNN and WDE-LSTM using Supervised Learning approach. The data for study as Amazon user reviews. WDE-LSTM proved to be capable of modelling long term dependencies in sentences and WDE-CNN was more efficient.

Rui Xia et al.[15] devised a distantly supervised learning framework for continuous sentimental analysis of social media sites. This method uses emoticons to label data and extracts four different types of knowledges. The algorithms used are Naive Bayes, Logistic Regression and PMI-SO in Twitter and the Chinese blog Weibo. The results showed the method was feasible and efficient for distant learning.

Mauro Dragoni et al.[16] presented a novel neural word embedding approach to exploit linguistic overlap between domains and build models for extracting polarity inference for documents of multiple domains. This paper summarises the whole review as a single distributed vector and identifies domain as a parallel task using a Supervised Learning approach. The approach is experimentally evaluated using Dranziera dataset and has shown an efficient precision, recall and f1 indicator.

Wei Wang et al.[17] studied the impact of the text content at various levels like title, blurb, first 100 words, detailed descriptions of crowdfunding projects using Lexical Based Approach and Supervised Learning algorithms like Association Rule mining, Conditional Random Field(CRF) and Support Vector Machine(SVM).The result showed that positive sentiment in the blurb and detailed description plays a vital role in promoting the crowdfunding campaigns.

Ngot Bui et al.[18] analysed the reasons of sentiment polarity change in user discussion threads in a Cancer Survivor Network. A Probabilistic Kripke Structure representation was used to represent and reason about prime causes of sentimental change. Support Vector Machine and Logistic Regression Classifiers of Supervised Learning Approach were used. The choice of classification threshold and classifier used decides the validity of this classification method.

Zhen Hai et al. [19] proposed a Supervised Joint Aspect and Sentiment Modeling Approach to model reviews and ratings, identify aspect level sentiments and predict overall sentiment of reviews. The dataset for study was taken as user reviews from Amazon CD and Game Reviews and Hotel Reviews from Trip Advisor. The model predicted sentiments at higher accuracy value.

Lanshan Zhan et al. [20] proposed a Dynamic Sentiment- Topic(DST) model which detects dynamic topics and also analyses the shift of public sentiment towards a specific topic by depiction using Probability Graph Model. The dataset taken for study was Chinese blog Weibo and algorithm Expectation Maximization was used. DST performance better in terms of perplexity which is the measure of the model's prediction ability.

Yonggan Li et al. [21] devised multiple algorithms to perform emotion sentence identification and classification, emotion tendency classification and emotion expression extraction of Weibo Sentiment Analysis. It proposed an unsupervised topic sentiment mixture algorithm for emotion detection Weibo Blog. The emotion factor extraction accuracy was on an average 70%.

Si Shi et al.[22] proposed a hierarchical framework for realtime tracking of Chinese Blog using complex event processing. It converts the microblog text to emotional microblog, classifies it and summarises mood using Online Batch Window Technique periodically. Naive Bayes and SVM algorithms were used. The average latency per event was stable at 2000ns which proves the effectivenss of the framework proposed.

Farman Ali et al. [23] developed a ontology based and SVM based information extraction and recommendation system for social robots to recommend for disabled user queries by considering the items polarity.

Mondher Bouazizi et al. [7] proposed a novel approach of classification of Twitter data into multiple sentiment classes and developed a tool SENTA for performing the classification. OpenNLP was used for feature extraction and Random Forest Classifier as used for classification. The performance indicators like accuracy, precision, recall, F-Measure showed improved results in the classification process.

Rita Georgina Guimaraes et al.[24] suggested age group as the most relevant parameter to sentimentally analyse user writings in Social Networks. A model based on Deep Learning was proposed for this. The proposed model was compared with the other popular algorithms like Multilayer Perceptron, Decision Trees, Random Forest, SVM and DCNN was able to classify better with a F-Measure value of 0.940.

Mondher Bouazizi et al. [25] proposed a pattern- based approach to classify sarcastic and non-sarcastic text of Twitter data. Feature Extraction was done using OpenNLP and Classification using SVM. The experimental results showed good performance indicator values.

D. *Sentiment Classification*

Zhang Yangsen et al. [26] studied a Recurrent Neural Network Sentiment Classification for Chinese MicroBlogs which helps to learn the deep structure of a sentence in blog reviews.

Yanghui Rao [27] proposed a Multi-labeled sentiment topic model- CSTM for adaptive social emotion classification by distinguishing context independent background theme and contextual theme of the input data.

Fangzhao Wu et al.[29] proposed a multi-domain sentiment classification approach which collaboratively decomposes the sentiment classifier into general sentiment knowledge and domain specific sentiment knowledge. It exploits sentiment relatedness between different domains to handle insufficiently labelled data. It extracts general sentiment knowledge using global model and specific sentiment using domain specific model. It proposed an Accelerated algorithm based on FISTA and a paralled algorithm to train dataset. Amazon dataset was taken for study. The parallel algorithm was effective in training the dataset. This models classification as a convex optimization problem. The time complexity was found to be less.

E. *Cross Domain Analysis*

Danushka Bollegala et al.[28] constructs a model for cross-domain sentiment classification. It constructs a cross domain classifier based on three objective functions that capture distributional properties of pivots of both domains , label constraints of source domain and geometric properties of unlabeled documents in source and target domains.

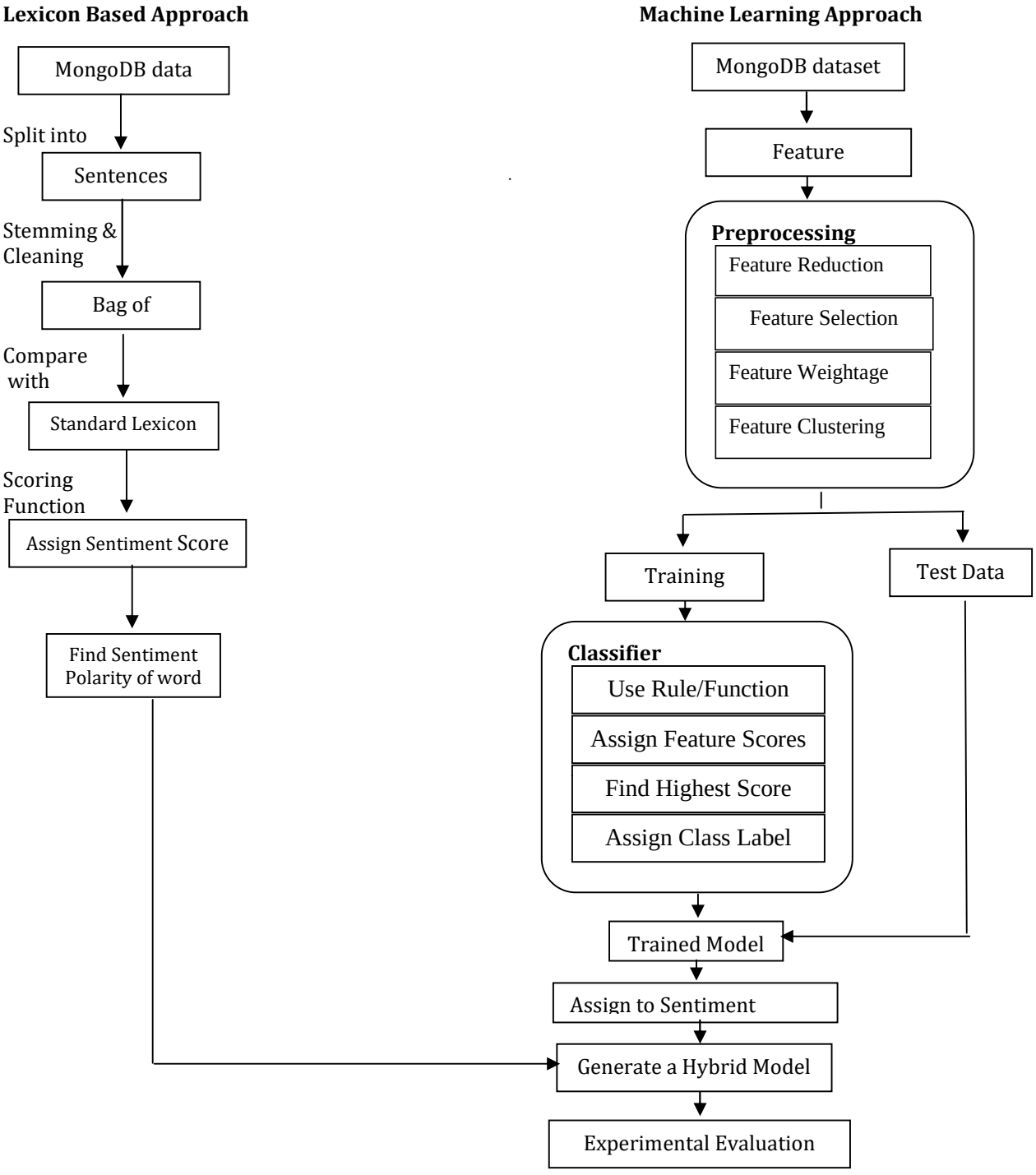
VII. PROPOSED ARCHITECTURE

Based on the analysis performed on the above papers a novel Hybrid model for sentiment analysis is proposed which combines the benefits of two popularly used approaches – Lexicon Based Approach and Supervised Machine

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Learning Approach. A lexicon based analysis is performed by splitting up the input data into sentences which is further tokenized and cleaned to form a bag of words. These words are then compared with any standard lexicons like Sentinet to assign sentiment score. Based on the score assigned the polarity of the word is determined as negative, positive or neutral. In machine learning approach the various types of features are extracted from the input data. The features undergo a preprocessing phase of feature reduction, feature selection, assigning weight to features according to its significance in the input and feature clustering to group related features. The input data is then divided into training data and test data. The training data undergoes a classification process. Depending on the classifier chosen using a rule or function a score is assigned to features. The highest score assigned decides the sentiment class to which the feature belongs. The trained model is used to assign class to the test data. The results of the two approaches are combined to develop the hybrid model that reveals the overall sentiment of the document and also highlights the polarity of specific words in different aspects. The dataset of the analysis will be stored in the recently developing MongoDB database. This database stores data as collections and documents which are non-schema based. This is well suited for storing unstructured user comments and reviews. This is a NoSQL database so the data retrieval process will be easier and efficient than the regular databases which uses a strict SQL format for retrieval. The model devised will be implemented and experimentally evaluated for its performance in the near future.

ARCHITECTURE DIAGRAM FOR THE PROPOSED MODEL



VIII. CONCLUSION

The detailed study of literature performed in this paper shows the significance of the various approaches in various contexts and their performance levels in different experimental environments. This review has also introduced to the recently emerging NoSQL database concept with reference to MongoDB database. This database is efficient in real time data analytics of user reviews and comments which are unstructured. The data for the implementation is to be stored in cross-platform, document oriented, open source MongoDB database. This review has enabled to propose a hybrid approach of sentimental analysis which utilizes the positive aspects of both Lexicon Based and Machine Learning Approaches of Sentiment Analysis. From the study, the proposed work uses the most successfully proven OpenNLP for feature extraction and SVM classifier or Naïve Bayes classifier for Machine Learning. Various lexicon based approaches will be comparatively studied to identify the best method and standard lexicon to achieve higher performance. What features to be selected and what weightage should be given to it will be analysed and identified in further study. The illustrated framework is under implementation applying different features and weights for the considered algorithms for achieving higher performance and accuracy.

REFERENCES

1. Ana Valdivia, M.Victoria Luzon and Francisco Herrera, " *Sentiment Analysis in TripAdvisor*", IEEE Intelligent Systems, 2017.
2. Monireh Ebrahimi, Amir HosseinYazdavar, and AmitSheth, " *Challenges of Sentiment Analysis for Dynamic Events*", IEEE Intelligent Systems, 2017.
3. WalaaMedhat, Ahmed Hassan, HodaKorashy, "*Sentiment analysis algorithms and applications: A survey*" ,Ain Shams Engineering Journal, 2014, pp 1093-1113.
4. Tareq Al-Moslimi, Nazlia Omar, Salwani Abdullah and Mohammed Albared, " *Approaches to Cross Domain Sentiment Analysis: A Systematic Literature Review*", IEEE Access Vol 5 2017, pp 16173-16192.
5. Zhao Jianqiang and GuiXiaolin, " *Comparison Research on Text Preprocessing Methods on Twitter Sentiment Analysis*", IEEE Access Vol 5 2017,pp 2870-2879.
6. Sujata Rani and Parteek Kumar, " *A Sentiment Analysis System to Improve Teaching and Learning*" , IEEE Computer Society, 2017.
7. Mondher Bouazizi and TomoakiOhtsuki, " *A Pattern-Based Approach for Multiclass Sentiment Analysis in Twitter*" , IEEE Access, Vol 5, March 2017.
8. Eric Cambria, Bjorn Schuller, Bing Liu, Haixun Wang, Catherine Havasi, "*Statistical Approaches to Concept-Level Sentiment Analysis*",IEEE Intelligent Systems, 2013.
9. Eric Cambria, Bjorn Schuller, Bing Liu, Haixun Wang, Catherine Havasi, " *Knowledge-Based Approaches to Concept-Level Sentiment Analysis*", IEEE Intelligent Systems , 2013.
10. Lorenzo Gatti, Marco Guerini, and Marco Turchi, " *SentiWords:Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis*", IEEE Transactions on Affective Computing, Vol 7, Issue 4, Oct-Dec 2016, pp 409-421.
11. WanxiangChe, Yanyan Zhao, HongleiGuo, Zhong Su, and Ting Liu, " *Sentence Compression for Aspect-Based Sentiment Analysis*", IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol 23, Issue 12, Dec 2015, pp 2111-2124.
12. Xiaojiang Lei, XuemingQian, " *Rating Prediction based on Social Sentiment from Textual Reviews*", IEEE Transactions on Multimedia, Vol 18, Issue 9, Sep 2016, pp 1910-1921.
13. Yoonjung Choi, JanyceWiebe and RadaMihalcea, " *Coarse-grained +/- Effect Word Sense Disambiguation for Implicit Sentiment Analysis*", IEEE Transactions on Affective Computing, Vol 8, Issue 4, Oct-Dec 2017, pp 471-479.
14. Wei Zhao, Ziyu Guan, Long Chen, Xiaofei He, " *Weakly-supervised Deep Embedding for Product Review Sentiment Analysis*", IEEE Transactions on Knowledge and Data Engineering, Vol 30, Issue 1, Jan 2018,pp 185-197.
15. Rui Xia, Jie Jiang and Huihui he, " *Distantly Supervised Lifelong learning for Large-Scale Social Media Sentiment Analysis*", IEEE Transactions on Affective Computing, Vol 8, Issue 4, Oct-Dec 2017, pp 480-491.
16. Mauro Dragoni, GiulioPetrucci, " *A Neural Word Embeddings Approach for Multi-Domain Sentiment Analysis*", IEEE Transactions on Affective Computing, Vol 8, Issue 4, Oct – Dec 2017, pp 457-470.
17. Wei Wang, Kevin Zhu, Hongwei Wang, Yen-Chun Jim Wu, "*The Impact of Sentiment Orientations on Successful Crowdfunding Campaigns through Text Analytics*", IET Software, Vol 11, Issue 5,Oct 2017,pp 229-238.
18. Ngot Bui, John Yen, and VasantHonavar, " *Temporal Causality Analysis of Sentiment Change in a Cancer Survivor Network*", IEEE Transactions on Computational Social Systems, Vol 3, Issue 2, June 2016,pp 75-87.
19. Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng, and Chunyan Miao, "*Analyzing Sentiments in One GO: A Supervised Joint Topic Modeling Approach*", IEEE Transactions on Knowledge and Data Engineering, Vol 29, Issue 6, June 2017, pp 1172-1185.
20. Lanshan Zhang, Xi Ding, Ye Tian, Xiangyang Gong, Wendong Wang, " *A Semi-supervised Topic Model Incorporating Sentiment and Dynamic Characteristic*", China Communications, Vol 13, Issue 12, Dec 2016, pp 162-175.
21. Yonggan Li, Xueguang Zhou, Yan Sun, Huanguo Zhang, " *Design and Implementation of Weibo Sentiment Analysis Based on LDA and Dependency Parsing*", China Communications, Vol 13, Issue 11, Nov 2016.
22. Si Shi, Dawei Jin, and GohTiong-Thye, "*Real-Time Public Mood Tracking of Chinese Microblog Streams with Complex Event Processing*" in IEEE Access, Vol 5, March 2017, pp 421-431.
23. Farman Ali, DaehanKwak, Pervez Khan, Shaker Hassan A. Elsappagh, S.M. Riazul Islam, Daeyoung Park, and Kyung-Sup Kwak, " *Merged Ontology and SVM-Based Information Extraction and Recommendation System for Social Robots*", IEEE Access, Vol 5, July 2017.
24. Rita Georgina Guimaraes, Renata L. Rosa, Denise De Gaetano, Demostenes Z. RodriGuez, and GracaBressan, " *Age Groups Classification in Social Network Using Deep Learning*" , IEEE Access, Vol 5, June 2017.
25. MondherBouazizi and TomoakiOhtsuki, " *A Pattern-Based Approach for Sarcasm Detection on Twitter*", IEEE Access, Vol 4, Sep 2016.

26. Zhang Yangsen, Jiang Yuru and Tong Yixuan, " *Study of Sentiment Classification for Chinese Microblog Based on Recurrent Neural Network*", Chinese Journal of Electronics, Vol 25, No 4, July 2016.
27. Yanghui Rao, " *Contextual Sentiment Topic Model for Adaptive Social Emotion Classification*", IEEE Intelligent Systems, Vol 31, Issue 1, Jan-Feb 2016.
28. Danushka Bollegala, Tingting Mu, John Y. Goulermas, " *Cross-domain Sentiment Classification using Sentiment Sensitive Embeddings*", IEEE Transactions on Knowledge and Data Engineering, Vol 28, Issue 2, Feb 2016,pp 398-410.
29. Fangzhao Wu, Zhigang Yuan and Yongfeng Huang, " *Collaboratively Training Sentiment Classifiers for Multiple Domains*", IEEE Transactions on Knowledge and Data Engineering, Vol 29, Issue 7, July 2017,pp 1370-1383.

A Recommendation System using Parallel Computing Techniques

G. Arumugam[1], M.Vithya[2] , S.Suguna[3]
[1]Professor and Head of the Department, Computer Science Department/Madurai Kamaraj University, Madurai.
[2] Lecturer, Sri Meenakshi Govt. Arts College for Women (A), Madurai-2.
[3]Assistant Professor, Sri Meenakshi Govt. Arts College for Women (A), Madurai-2.

ABSTRACT *Now a day's use of internet has been increased tremendously. Every day internet users generate 2.5 quintillion bytes of data from various sources, and thus leads to Big data analytics. Web usage mining is the type of web mining activity that involves discovering user access pattern from web log data. Web usage mining has three phases such as Data Preprocessing, Data Discovery and Data Evaluation. In this paper we have mainly focused on Data preprocessing. Data preprocessing is an important phase of Web usage mining phase, which is required to handle unstructured, heterogeneous and unwanted (noisy) nature of log data. Preprocessing consists of four phases namely, Data Extraction, Data Cleaning, User identification, Session Identification and Path completion. Identification of user session circumference and to extract travel path set in path completion processes are most important in the web mining for predictive prefetchning of user next request based on their navigational behaviour and recommended overall user interested pages. This paper presents techniques to identify the user sessions and also includes the work to extract entire travel path set for every users through session times. Learning graph is also constructed based on user maximal access sequence for predictive prefetching without searching the server page. The result shows the overall user interest pages and which pages are accessed by more than user, those prediction used to web administrator for improving their performance as well as economic. The analysis with Vizhamurasu News site server logs shows that the proposed approach provides better results in terms of time complexity and accuracy for hit rate prediction. The existing preprocessing and prediction algorithms are efficient but that are not scalable because when we increasing size of log file and also take much more computation time compared to proposed parallel computing techniques.*

Keywords: *Web Mining, Preprocessing, Session Identification, Path Completion, Learning Graph and Recommendation*

Introduction

Today internet is playing important role in our daily life, because millions of data generated in various ways. Those amounts of data on WWW are huge therefore it is very critical to store and manage, that leads to big data. It also produces problem in data accessing [1]. Web site is a group of web pages. Web pages may be text, images audio and videos. Whenever user accesses any website, log files are generated. Log file have information about each user access. Data stored in web log files in various formats like NCSAs Common log file format, W3C Extended log file or IIs log file formats. These log files are different types such as Error logs, Referrer logs and Access logs. Log files are located in different places like web server, proxy server and client browser. Analysis of these log files helps to extract knowledge about navigation behaviors of web user, increasing performance of web site, by improving web site design. Web mining is the applications of data mining techniques to discover data from web documents and services. The internet business model, which provides efficient way to accessibility of customer with less resources with minimum expenses [2]. So Web usage mining is to take hold of analyzes behavior patterns of web user. Web users are extracted by access patterns. The above processes are done by web logs, those loges are noisy and uncertain, so log files are needed to be cleaned. Preprocessing is an important and complex of the web architecture because which takes 80% of works. It includes the process of Data Cleaning, User identification, Session Identification and Path completion construction of user transactions. Data Cleaning is the process of removing irrelevant and duplication records. User Identification is the task of associating pages with same IP address, User agent and Operating system. User Identification is defined as a sequence of actions made by one user for a single navigation. A user made one or more sessions during a period of time. Path completion task is to identify missing pages and used to fill missing reference pages in a session. We analyzed various methods of user session Identification and Path completion process. We construct a learning graph based on user session sequences, which helps to predicting the user access patterns.

The organization of the paper is as follows: In Section III, architecture for preprocessing process is proposed. In section IV, description of methodologies and an algorithm is proposed for different phases of preprocessing and learning graph construction. Section V shows the experimental results of proposed algorithm for Viahz Murasu web server logs. The section VI deals with comparative analysis of existing and proposed methods. Finally, concluded the paper in section VII.

Related Works

In general, user session identification processes are analysed by the most common methods such as timeout, maximal forward reference and reference length. Shanta H Biradar proposed [3] Maximal forward reference method based on reference length, time window and content pages for analyzing users sessions and travel path transactions are also constructed. In [4] proposed an algorithm for every user travel path using referrer information. Nirali Honest, Anti Patel, Ban kim and Patel [5] proposed path completion technique for University website through navigational behaviour and discovering patterns of user activities. In [6] proposed a new algorithm for User Session Identification (USIDALG) for the activities related to the User, and Session Identification based on Access history list. They constructed learning graph to predictive perfecting user next request without searching the whole web server. Mingming Zhou [7] describes

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

the preprocessing techniques on e-learning logs for educators who are investigating student online reading behaviour using sequential pattern analysis. Priyanka Verma et al. Murat Ali Bayir et al. [8] analyzed mobile operators in Turkey server logs for prediction of user session. They measure the success of next page prediction through comparison of C-SRA, S-SRA, Navigation oriented and Time oriented methods. The experiments show that C-SRA is better in prediction among other three methods. Neelima et al. [9] presents algorithm for data cleaning, user identification and session identification and analyzed the performance in terms of web logs. Vikram Singh Chauhan et al. [10] discussed about the existing methods for pre-processing and applied navigation and time based techniques for session identification through parallel computing. They proved least time for completion compared with other methods. Mingming Zhou [11] describes the preprocessing techniques on e-learning logs for educators who are investigating student online reading behaviour using sequential pattern analysis. Priyanka Verma et al. [12] analyses the problems in existing techniques of cleaning and user identification. They also proposed possibility of improving performance of cleaning and user identification with experiments. The experiment shows that the quality is improved for user identification with the help of reduced unwanted log size. In paper [13] online reading behaviour are predicted based on events along with time dimension such as opening a window, closing window and clicking hyperlink. In paper [14] improved K-means clustering algorithm is used for identifying internet user behavior based on IP address, user-agent and time taken. Sessions are clustered according to their co-occurrence parameters in paper [15]. Recommender system is one among the essential applications of Web usage mining. The main objectives of Recommendation system are to present a structure with the feature of automatic recommendation on the basis of user's navigation patterns on the web. Paper [16] proposed product based Recommendation system for both registered and unregistered user, based on burst time visited. Paper [17] evaluate different websites in order to provide good recommendation by availability of information using personalized resource genitor extended algorithm. Paper [18] proposed hybrid page ranking model based on user historic accessed page. Paper [19] proposed weighted K-means clustering for MovieLens data set. Paper [20] focuses on providing real time dynamic recommendation to all registered and unregistered visitors of websites.

Proposed System

The above survey indicates the following issues in the existing Preprocessing techniques:

- The existing methods are only sufficient to static pages for user sessions identification.
- User sessions are made by unvisited pages also.
- More Computation time for large data set.
- More memory space needed.
- To generate recommendation pages.

In this work, system is proposed to solve the above issues. The Proposed system is based on Viahz Murasu server side logs. The server site structure is shown in figure 1.

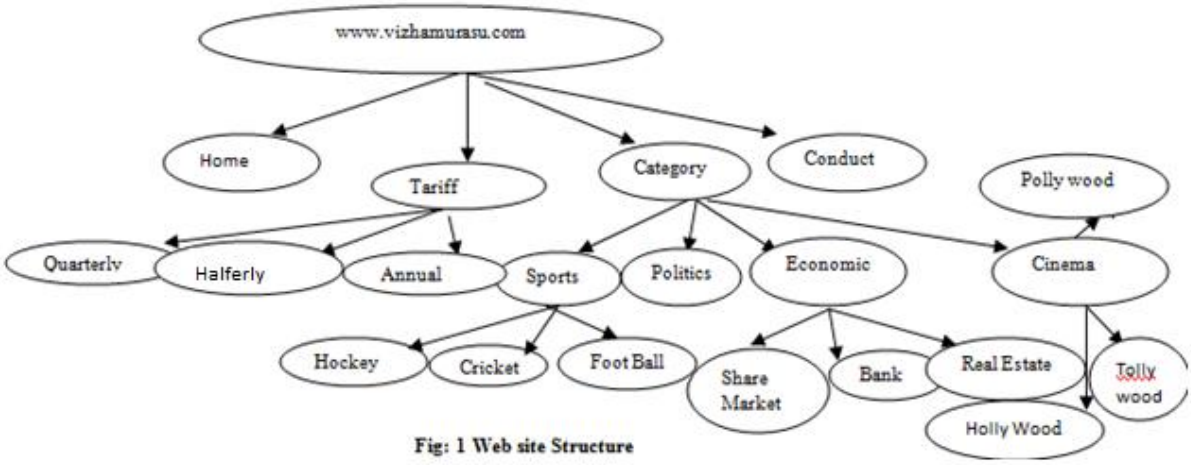


Fig: 1 Web site Structure

In general, log file is automatically created and maintained by web server whenever user access the web site with request for documents (HTML, PHP), images, text and other objects. This contains information about the user click stream on the website. Web logs are represented in any one of the available formats. In this work, extended common log format is used to represent the web logs. There are various attributes in web server raw log such as IP address, date and time of request, authentication details, Referrer URL, URL requested, browsing agent, status code, number of bytes transferred, access method, operating system and so on. Table 1 shows the sample web server log in extended common log format.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

117.222.157.35,Aug 9 17:04:50 017,http://vithya.vizhamurasu.com/index.php,id=25&prevpage=/index.php,Mozilla/5.0 (Windows NT 6.1; rv:54.0) Gecko/20100101 /log.php,CGI/1.1,43.255.154.124,GET,1502298290

Table 1: Sample log entry from Vizhamurasu web server

Data preprocessing is an important phase, it can be done by the processes such as Data Extraction, Data Cleaning, User Identification, Session Identification and Path Completion. In data extraction phase, log files are gathered from various web servers based on time duration for analysis. In data cleaning process only needed entries are extracted from raw server logs. User Identification phase is used to identify unique user through IP Address, browsing agent, operating system, and query string. Session identification annex set of request crated by unique user during his/her sessions. In path completion phase, missing user requests that are not logged into a log file (cached in client / proxy server side) are identified and complete navigation path for all these sessions are framed. The proposed architecture for preprocessing the server side web logs is depicted in fig. 1.

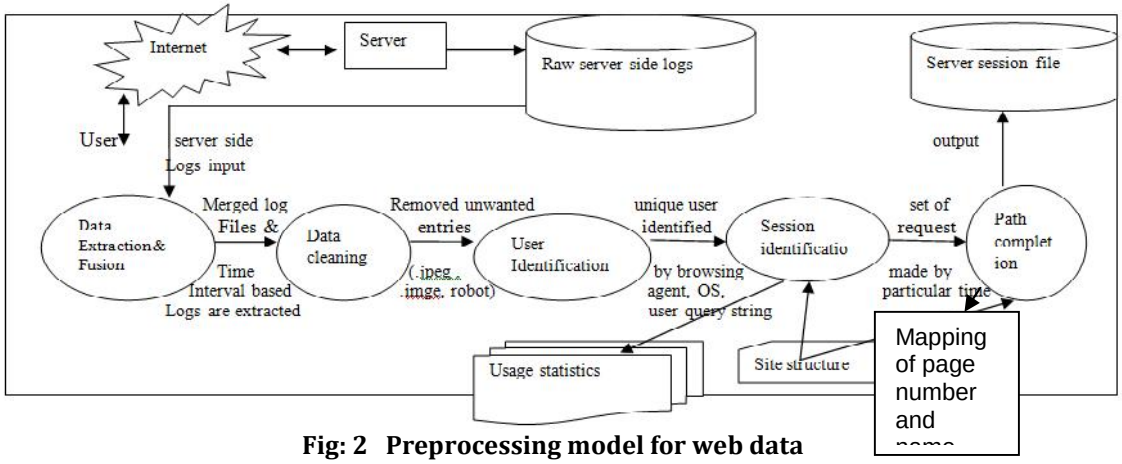


Fig: 2 Preprocessing model for web data

Proposed Preprocessing Methodologies

The raw vizhamurasu news website server logs are experimented in this stage. The web usage mining data are usually taken from three parts: server level, client level, and proxy level. Among them, the server level is the most convenient to predict user browsing behaviour in web usage mining. In general, the data preprocessing task involves data extraction, data cleaning, user identification, session identification and path completion.

Phase: I Data Extraction and Fusion Algorithm:

Data Extraction includes the merging of log files because the data size is too big, access data stored in different server those logs are collected from different sources. In this work, we have collected access log data in a variety of log files named as access _log1, access_log2…..access log N from different Web servers. After that data are extracted according to specified time interval. In this work, six months logs from 16-05-2017to 16-12-2017 are gathered.

Input: merged log file, start $_{time}$, end $_{time}$
Output: reduced sorted log file and within time interval
Begin
Current time= Date and time
for every entry in log file **do**
Read web log W= {w_1, w_2…., w_n}; of n instances
If (Given current $_{time}$ >= start $_{time}$) and (Given current $_{time}$ <= end $_{time}$) then
Sort file according to its date and time
Stored in new table
End if
End

Phase 2: Data Cleaning Algorithm

Data cleaning is the second step in our proposed methodology to remove all unwanted data. This method is used in data mining analysis more efficiently. The cleaned data include removal of unsuccessful requests, request with irrelevant HTTP methods like 404, 300, etc., removal of multimedia records such as video, graphics , image etc thus file extension is .js, .mp3, .jpeg etc. Robot or spider and uninterested session from log file are removed.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

```
Input: Stored new table
Output: Refine Log file
Begin
Mapper class
  Read fields (status codes) do
    For every in log file _table
              If Method ϵ [GET or POST] and then get all fields
              Else
                 File_ext ϵ[ jpeg,png,jpg,tif,mp3,css,js,swf,ico,cgi]  and
                 Robot_ext ϵ [robot.txt] then
                Remove this entry from log file
             End if
End
```

Phase 3: User Identification Algorithm

Cleaned log files are further processed for user identification process. The aim of this process is to recognizing every user's access characteristics. Different users are identified by different IP addresses. User Identification based on IP address provides poor result because different users use same IP address, and same IP address can be used by different Browsers. So some more entries such as user query string, Url, operating system, number of time accessed particular url is also considered for user identification.

```
Input: Refine log table
Output: identification of user
Begin
Mapper class
    Read fields in dataset
      For each entry in log table do
       If current IP, OS,  Browsing Agent, user Query string are not in Listof IP
        then add the current IP in ListofIP   then mark whole record as a new user and assign new user ID
  Else assign the old_userID
  End if
End
```

Phase 4: Session Identification Algorithm

A sequence of pages seen by a user during one visit is called session. Session is called as grouping the different actions of a single user in the web log files. When a new user starts page browsing, a new session is created. In user session identification methods, the most of the researchers are used Timeout, Reference Length and Maximal Forward Reference methods. Many researchers assume 30 minutes time duration between two pages requests. If predefines time exceeded then new session is started mechanically. In Reference Length method the amount of time a user spends on a page is based on whether the page is an auxiliary or content page. In Maximal Forward Reference, each session is the set of page sequences from the first page to final page before a backward reference is made.

```
Input: user identified table
Output: identification of user sessions
 Begin
Mapper class
   Read fields in data set
     For each entry in log table do
       If current session time >30 , refer url, Date , Time  and  query sting Next request page are not in Listof sessionID assign new sessionID for that log
entry then mark whole record as a new session and assign new sessionID
          Increment sessionID
        Else
           Assign the old sessionID
       End if
Reducer class
 Count the no of session made by single user
Count no of user are made same sessions
 End
```

Phase 5: Path completion Algorithm

In Path completion Phase, missing user requests that are not logged into a log file due to proxy / client side caching or use of back button to retrieve a page etc, are analyzed and complete user navigation path is framed for each and every session**.**

We have analyzed and complete user navigation path is framed for each and every session**.** The aim of path completion task is to provide better result of data preprocessing and improve mined pattern's quality. In this work missing pages are included in the user access path, duplicate pages are removed in successive access session and name of the pages are mapped with page number. Consider a path sequence P={p1,p2....pn}, here n is the last page in a single session travel set.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Input: User Session identified table
Output: The complete travel set session file
Begin
Mapper Class
Read fields in data set
For each entry in log table do
Read user session USID, USID= {U1, U2,,,,, Un) where n is the total number of sessions.
Take the first Url from the user session of accessed pages pi= {p1, p2,,,,,, .pn}
Maximum number of traversed page in a single session.
Calculate the length of page
If current path== link details, store the username, page number, page name
Find the previous and next page number and page name of links if the path
Then
Replace the actual name of the path
Else
Don't add in the path
End if
Repeat the above steps for all sessions.
End
To make a complete user navigation path is framed for each and every individual user session.

Experimental Results

We have collected one lakh access logs from Vizhamurasu Web server. In the first step, log files are merged into a single log file. We have used Regular expression (Rex) for log field extraction. All implementations are done on UBUNTU 12.04, 64 bit operating system with 8GB Ram and Intel core 13 processor. The table 2 depicts sample raw web server logs.

106.66.169.221,Sep 3:45:59 2017,http://vithya.vizhamurasu.com/category.php,id=6&prevpage=/category.php,Mozilla/5.0 (Linux; Android 5.0.2; vivo V1 Build/LRX22G) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/59.0.3071.125 Mobile Safari/537.36,/log.php,CGI/1.1,43.255.154.124,GET,1505447159,id=6&prevpage=/category.php,http://vithya.vizhamurasu.com/category.php,,/home/prabhupri/public_html/vithya/log.php,webmaster@vithya.vizhamurasu.com,80,,/log.php,,/category.php,6117.222.1573,Sep 15 3:45:59 2017,http://vithya.vizhamurasu.com/category.php,id=7&prevpage=/category.php,Mozilla/5.0 (Linux; Android 5.0.2; vivo V1 Build/LRX22G) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/59.0.3071.125 Mobile Safari/537.36,/log.php,CGI/1.1,43.255.154.124,GET,1505447159,id=7&prevpage=/category.php,http://vithya.vizhamurasu.com/category.php,,/home/prabhupri/public_html/vithya/log.php,webmaster@vithya.vizhamurasu.com,80,,/log.php,,/category.php,745.248.5.157,Oct 25 8:49:55 2017,http://vithya.vizhamurasu.com/,id=0&prevpage=,Mozilla/5.0 (X11; Ubuntu; Linux i686;rv:39.0) Gecko/20100101 Firefox/39.0,/log.php,CGI/1.1,43.255.154.124,GET,1508921395,id=0&prevpage=,http://vithya.vizhamurasu.com/,,/home/prabhupri/public_html/vithya/log.php,webmaster@vithya.vizhamurasu.com,80,,/log.php,,,0 103.207.141.220,Oct 25 11:09:46 2017,http://vithya.vizhamurasu.com/,id=51&prevpage=/index.php,Mozilla/5.0 (X11; Ubuntu; Linux i686; rv:39.0) Gecko/20100101 Firefox/39.0,/log.php,CGI/1.1,43.255.154.124,GET,1508929786,id=51&prevpage=/index.php,http://vithya.vizhamurasu.com/,,/home/prabhupri/public_html/vithya/log.php,webmaster@vithya.vizhamurasu.com,80,,/log.php,,/index.php,51117.222.154.151,Dec 17 13:02:29 2017,http://vithya.vizhamurasu.com/about.php,id=0&prevpage=,Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/58.0.2988.0 Safari/537.36,/log.php,CGI/1.1,43.255.154.124,GET,1513515749,id=0&prevpage=,http://vithya.vizhamurasu.com/about.php,,/home/prabhupri/public_html/vithya/log.php,webmaster@vithya.vizhamurasu.com,80,,/log.php,157.50.9.52,Dec 17 13:49:43 2017,http://vithya.vizhamurasu.com/contact.php,id=1524&prevpage=/contact.php,Mozilla/5.0 (X11; Ubuntu; Linux i686; rv:39.0) Gecko/20100101 Firefox/39.0,/log.php,CGI/1.1,43.255.154.124,GET,1513518583,id=1524&prevpage=/contact.php,http://vithya.vizhamurasu.com/contact.php,,/home/prabhupri/public_html/vithya/log.php,webmaster@vithya.vizhamurasu.com,80,,/log.php,,/contact.php,1524

Table: 2 Raw sample logs from vizhamurasu web server

These logs contain unwanted details like multimedia and robotic files. After the completion of data cleaning process, the cleaned web server log file is prepared to load into Hbase server. The cleaned logs are shown in fig. 3

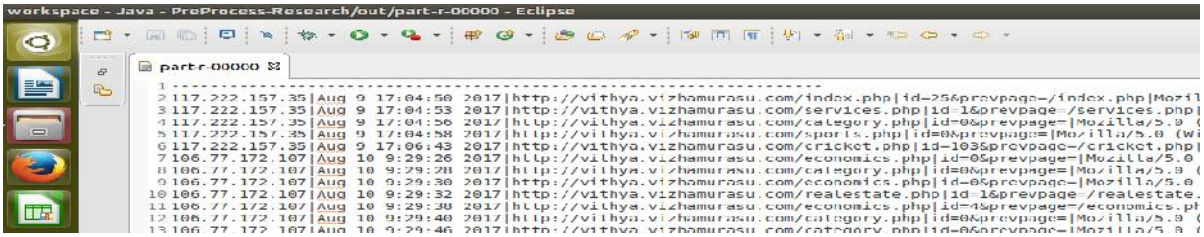


Fig: 3 Cleaned log file

Then individual user is identified based on the IP Address, Browsing Agent, OS, Date, time and User query string. The results are depicted in fig. 4

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

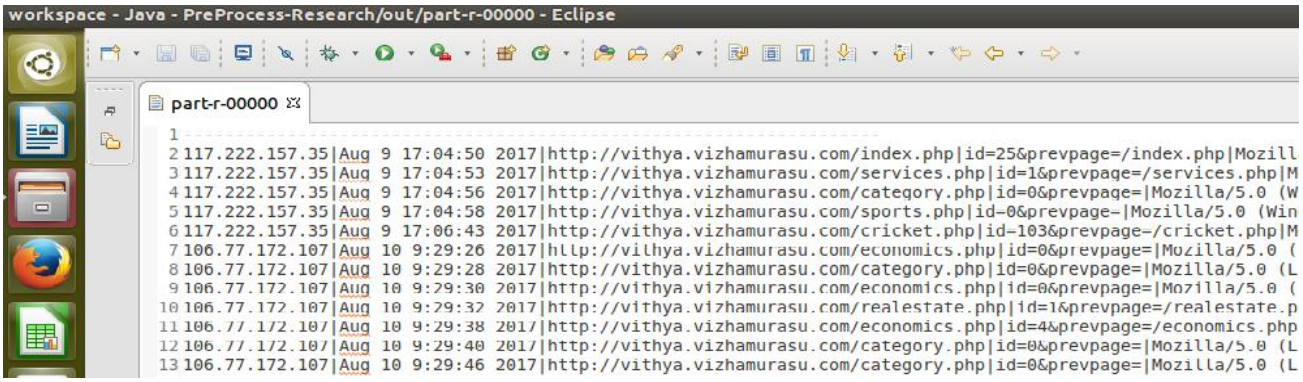
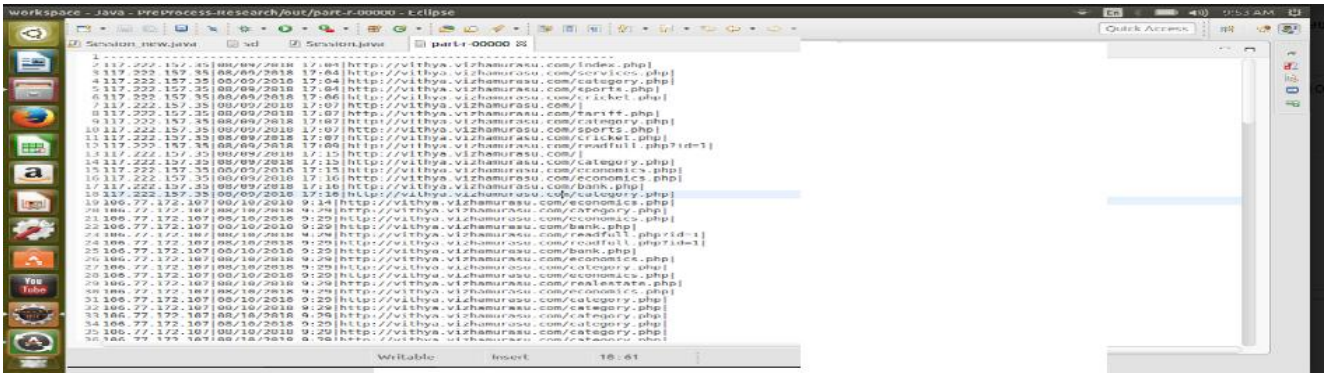


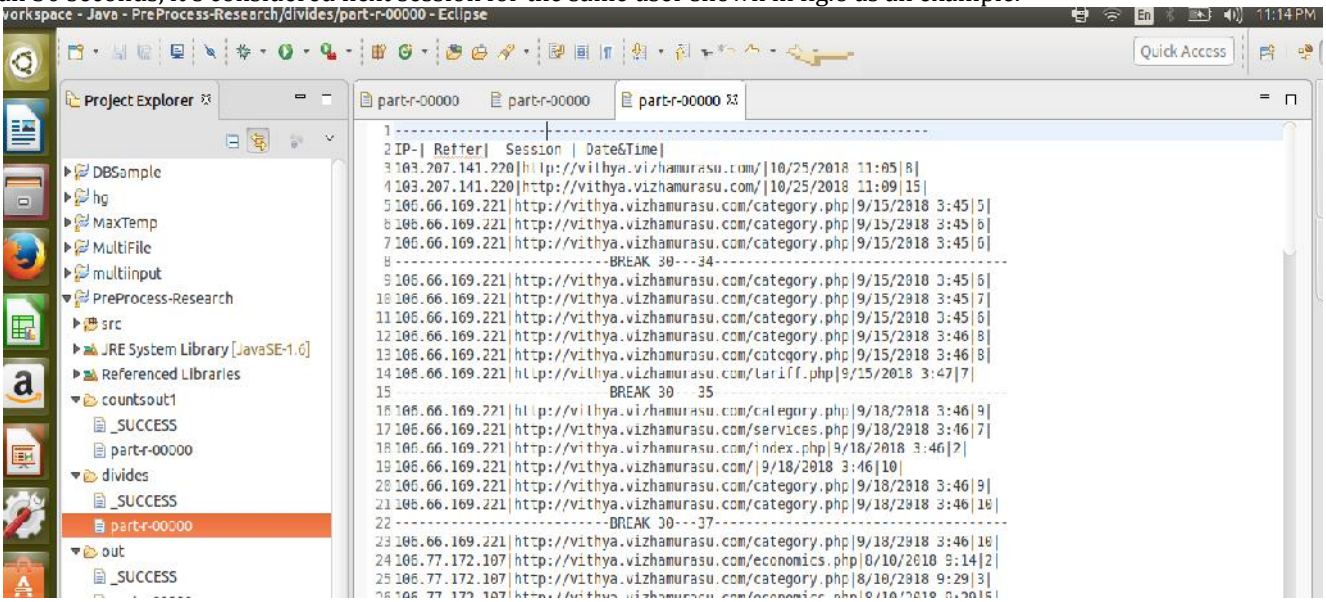
Fig : 4 User Identification

The each user sessions are identified based on date, query sting, url, content and the time spent on each page. The fig. 5 shows the number of sessions for each user.



Fig; 5 Session Identification

The user sessions are identified based on their spending time in particular page, date and time. If a user can spent more than 30 seconds, it's considered next session for the same user shown in fig.6 as an example.



Fig; 6 User Session for all IP Address

The table 3 shows the result path of one user session before path completion.for access

IP Address	Date	Time	Access Content/URL/ Query String	Referrer page number
106.77.172.107	08/10/2018	9:14	http://vithya.vizhamurasu.com/economics.php	4
	08/10/2018	9:44	http://vithya.vizhamurasu.com/category.php|	2
	23/09/2017	10:20	http://vithya.vizhamurasu.com/bank.php|	6
	23/09/2017	10:27	http://vithya.vizhamurasu.com/realestate.php	8

Table: 3 Access path of one user session

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Table 4 shows the result of the complete access path of corresponding user session. Our proposed path completion algorithm, appends the missing path using referrer number available in the web access log to each user session to generate complete path.

	User's access path 106.77.172.107
Page sequence	4-2-6-8
Combination	2-3-4, 1-2,1-2-3-4-6, 4-6-8
Path completion	2-3-4-6-8

Table: 4 Complete Access Path of User Sessions

In fig.7 shows the result of user complete access for every user sessions.

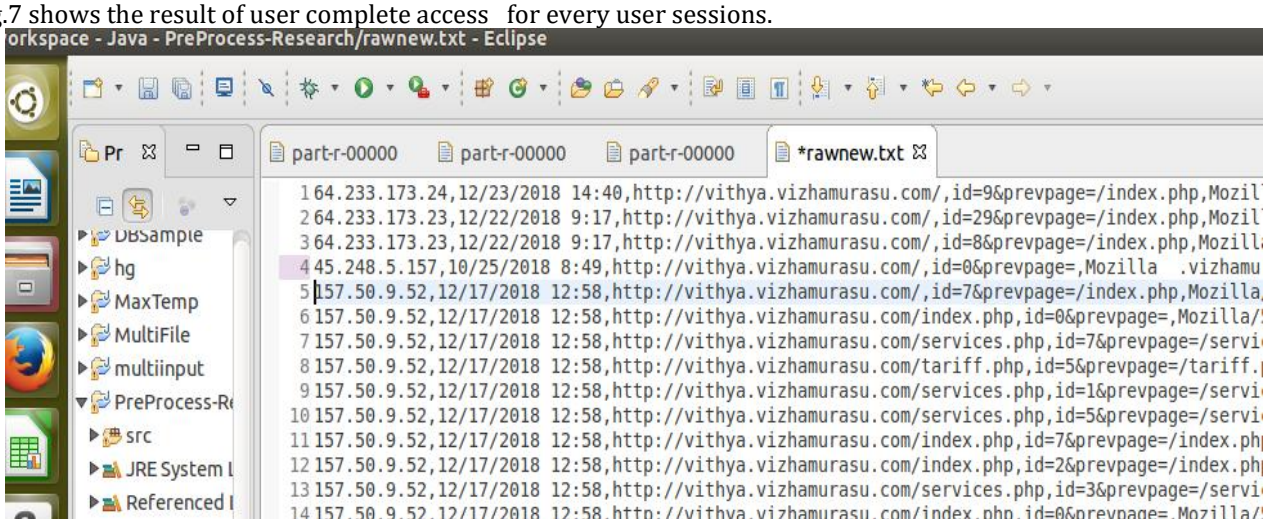


Fig: 7 Complete Access Path of User Sessions

Recommendation System

The aim of transaction identification is to identify user access frequency for predictive prefetching of user next request. Here two types of transactions are considered, i.e, travel path and content path .The figure 8 shows the number of times the particular path is accessed by a user. The learning graph is constructed from these accessed sequences and shown in figure 9 and figure 10. Figure 9 shows the learning graph for **IP** address 117.222.157.35 and the overall recommended learning graph is shown in figure 10. This learning graph is used for predicting the user next request, while the user is with the current page.

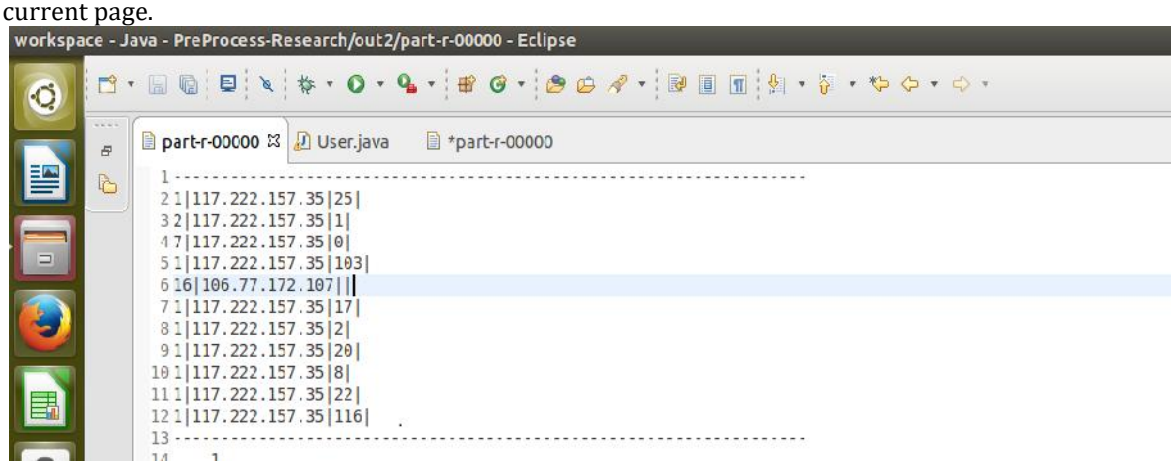


Fig: 8 User session sequences for IP Address 117.22.157.35

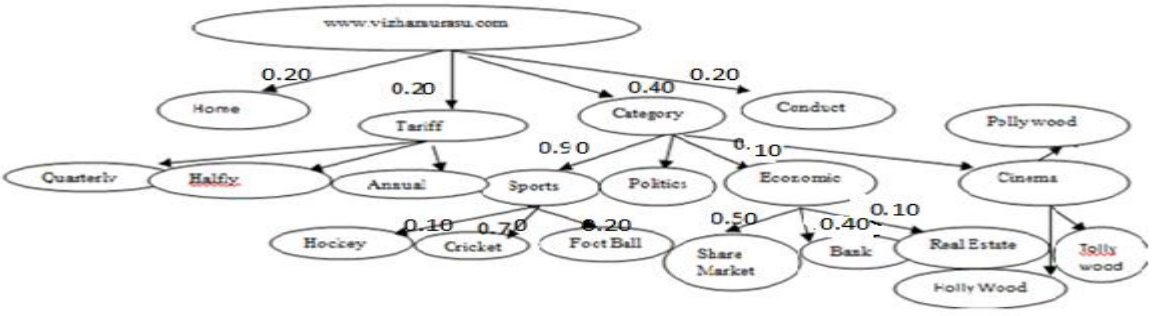


Fig: 9 learning for IP address 117.222.157.35

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

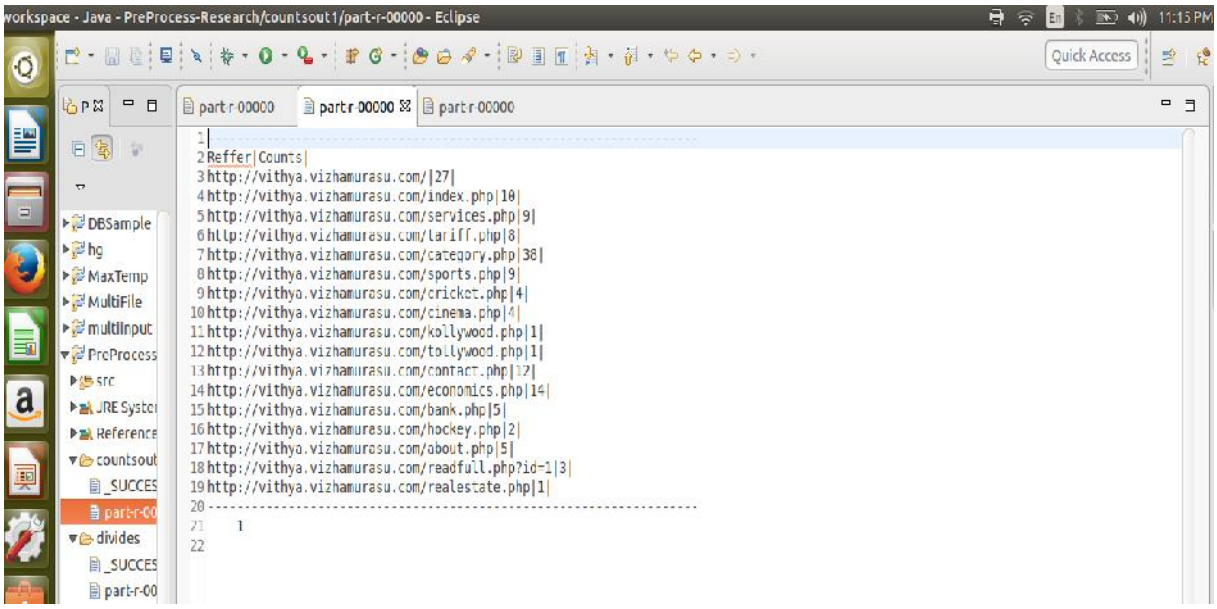


Fig: 10 learning graph for overall recommended Access pages

Comparative study

For assessing the performance of proposed method, two metrics are considered. One is time and another is hit rate accuracy. The proposed methods shows better performance because of the parallel computing techniques used and Hbase for storing unimaginable data set over MYSQL by existing methods. The analysis of comparison of overall execution time taken by existing Preprocessing algorithms and our proposed algorithm is depicted in figure 11.

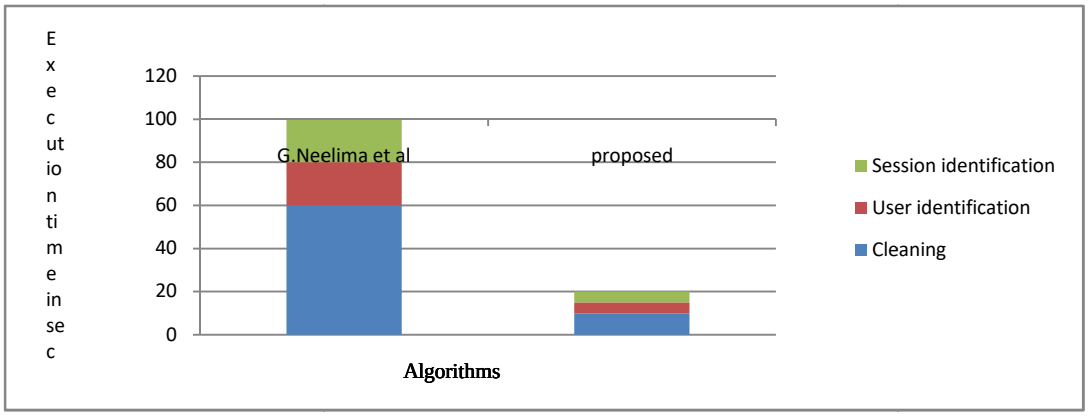


Fig:11 Comparison of execution time for overall Preprocessing phases

The figure 12 depicts the analysis of prediction accuracy comparison between our proposed algorithm and existing methods.

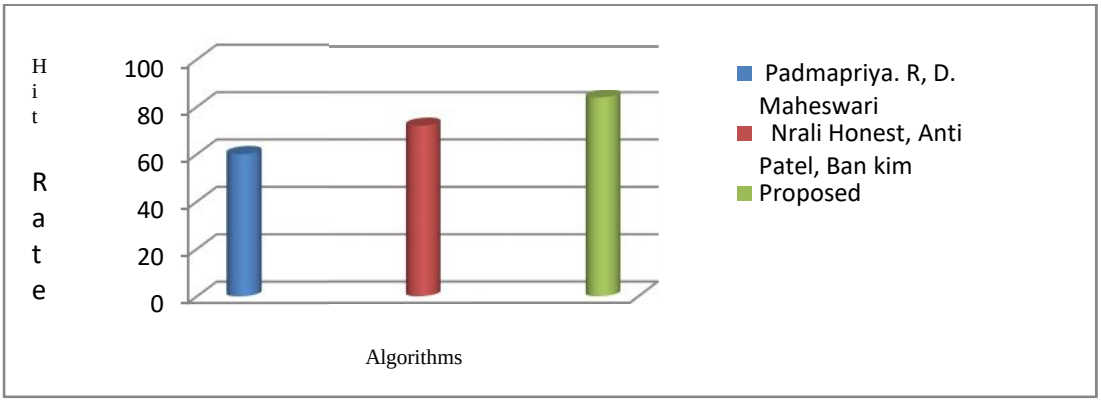


Fig: 12 Comparison of Prediction Accuracy level between Existing and proposed algorithms

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

In figure 13 depicts the performance metrics analysis of prediction accuracy comparison between our Proposed and existing algorithms.

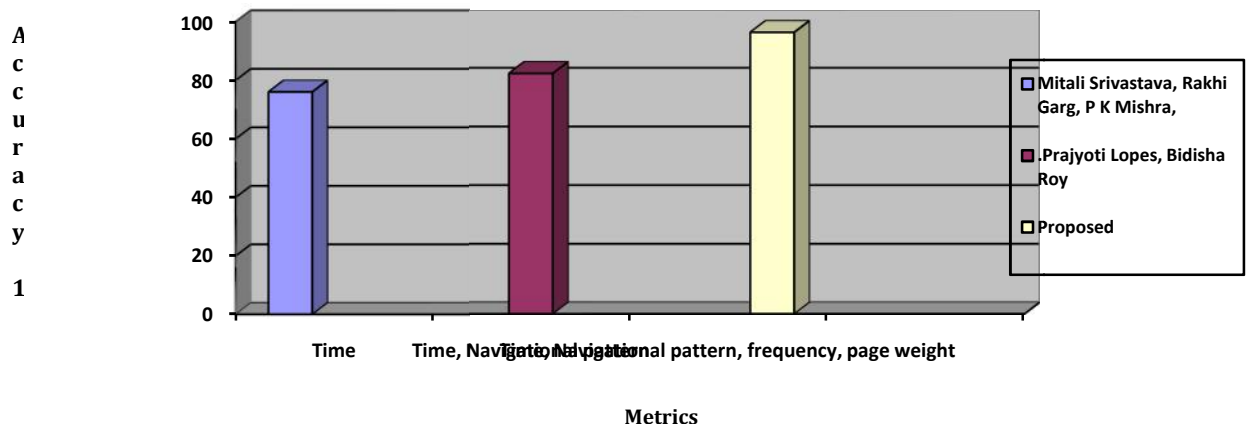


Fig: 13 Performance metric level comparisons between proposed and existing method

Conclusion

Web site is considered the most important tool for advertisement in web areas. One can identify user navigation behavior by the log records which is stored whenever user access the website. Log file includes heterogeneous, unstructured and large amount of data. It is necessary to perform Preprocessing before applying mining algorithm. The main goal of the Preprocessing is to cleaning and extraction of only needed data. We have proposed algorithm for Data Extraction, Data Cleaning, User identification and Session identification and path completion. Construction of learning graph is used to predict the user most interested pages. The existing methods deal with the static pages for user sessions identification and unvisited pages also predicted. In existing methods used only time metrics for prediction of user interesting pages it's not predict with high accuracy. In proposed method used more accuracy metrics such as time, navigation pattern, page weight and frequency of access. So high accuracy is achieved compared with existing methods. We proved that existing algorithms are not scalable because when we increasing size of log file and also take much more computation time compared to proposed parallel computing techniques and also get good accuracy level of prediction to user next requested pages.

References:

1. Brijesh Bakariya, Krishna K, Mohbey G.S. Thakur , "**An Inclusive Survey on Data Preprocess Methods Used in Web Usage Mining",** Advances in Intelligent Systems and Computing 202, Springer 2013.
2. RamKumari Tyagi, Anand jawdekar, "**An Advanced Recommendation System for E-Commerce Usrs"**, Symposium on Colossal Data Analysis and Networking(CDAN) 2016 IEEE.
3. Shanta H Biradar, "**An Efficient Path Completion Techniques for Web Log Mining**", International Journal of Computer Science Trends and Technology(IJCST), Vol: 4, Issue 3, May-June 2016.
4. Padmapriya. R, D. Maheswari , " **A Novel Technique for Path Completion in Web Usage Mining**", International Journal of Advanced Research, Ideas and Innovations in Technology, Vol: 3, Issue 2, 2017.
5. Nirali Honest, Anti Patel, Ban kim, " **Astudy of Path Completion Techniques in Web Usage Mining**", 2015 IEEE International Conference on Computational Intelligence & Communication Technology.
6. G. Arumugam, S. Suguna, "**Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs**",IEEE Explorer,Pages: 1-6, ISBN:978-2-9532-4431-1, June 2009.
7. Mingming Zhou, "**Data Preprocessing of Student e-Learning Logs**", Information Science and Applications 2016 Springer.
8. Murat Ali Bayir, Ismali Hakki Toroslu, "**Fining all Maximal Paths in Web User Sessions**", WWW'16 Comanion, April 11-15, 2016 ACM.
9. G.Neelima, Dr. Sireesha Rodda ," **Predicting user behavior through sessions using the Web log mining**", International Conference on Advances in Human Machine Interaction, March 3-5-2016 IEEE.
10. Vikram Singh Chauhan,, B.L Pal, " **Hybird Data Preprocessing: User Sessions Identification through Hadoop**", International of Computer Trends and Technology(IJCTT), Vol:28, No.4, October 2015.
11. Priyanga Verma, Dr.Nishtha Kesswani, "**Web Usage mining framework for data Cleaning and IP Address Identifaction"**.
12. S.Jagan Dr.S.P.Rajagopalan, "**A Survey on Web Personalization of Web Usage Mining"**, International Research Journal of Engineering and Technology, Vol.2, No.1, March 2015.
13. D.Dixit et al., "**Mining Access Patterns using Classification",** International Journal of Engineering Science and Technology, vol.2, 2010.
14. Shruthi Ramdas, Rithesh Pakkala P., Akila Thejaswi R., "**Determination and Classification of Interesting Visitors of Websites using Web Logs"**, International Journal of Computer Science and Mobile Computing, Vol.5, No.1, January 2016, pg:01-09.
15. Ravi Khatri, Daya Gupta, "**An Efficient periodic web content Recommendation based on Web usage mining"** 2015 IEEE 2nd International Conference on Recent Trends in Information System.

16. No'aman M.Abo A1-Yazeed, Ahmed M.Gadallah, Hesham A.Hefny, "**A Hybird Recommmendation Model For Web Navigation",** 2015 IEEE Seventh International Conference on Intelligent Computing and Information System (ICICIS'15).
17. Jyotsna Chandra, Annappa B, "An **improved Web Page Recommendation System using Partition and WebUsage Mining**", IPAC'
18. Prajyoti Lopes, Bidisha Roy, "**Dynamic Recommendation System using Web Usage Mining for E-commerce Users",** International Conference on Advanced Computing Technologies and Applications 2015, pp.60-69.
19. Xindong Wu AT, "**Data Mining with Big data**", IEEE Transactions on knowledge and data Engineering,vol: 26,no.1,January 2014.
20. S.SiddharthAdhikari ,DeveshSaraf, Mahesh Revanwar and Nikhil Ankam, "**Analysis of Log Data and Statistics Report Generation Using Hadoop"**,International Journal of Innovative Researchin Computer and Communication Engineering, vol.2, Issue. 4, April 2014.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Semantic Web based Recommender System in E-learning System

[1]B.Gomathi, [2]M.Thangaraj, [3]S. Suguna
[1]Research Scholar, Madurai Kamaraj University
[2]Professor, Department of Computer Science, Madurai Kamaraj University
[3]Assistant Professor, Sri Meenakshi Govt. Arts College for Women (A), Madurai-2, kt.suguna@gmail.com

ABSTRACT *E-learning is one of the distance learning methods via World Wide Web (WWW). The learner and instructor interact via learning content through the web. There are many of search techniques available today but retrieving meaningful information is difficult. However to retrieve accuracy and meaningful information intelligently, semantic web technologies are playing a major role. Also, semantically implemented e-learning system with Recommendations will effectively suggest and guide the learner for choosing the course as their need. After the preprocessing and filtering methods, list of recommendations are generated by using recommendation engine. In this work, a survey and different approaches of recommendation system for semantic web-based model in E-learning system is discussed. Collaborative Recommender System (CRS) for E-learning with the heuristic information retrieval technique is introduced to choose the right course for the right time. Based on different metrics CR System works with different indexing methods and clustering process to semantically group and retrieve the related materials of the course effectively. An experimental evaluation showed that the recommendation accuracy of the proposed system(CRS) was much higher than that of the other (non-CF) method in all four evaluation criteria (recall, precision, and half-life utility).*
Keywords *- Domain Ontology, E-learning, Recommendation, Semantic Web, OWL, RDF*

I. INTRODUCTION

Information Retrieval with semantic approach [31], [32] is the main issue nowadays. E-learning is a new-type education form characterized mainly of online interactive learning by taking the learners as the subject and utilization of the network technology, multimedia technology and other modern information technical means [37], [38]. E-learning provides abundant learning resources for students, who can receive the education as they need conveniently overcoming the limitations of time and space, presenting great flexibility and convenience. E-Learning recommendation system helps learners to make choices without sufficient personal experience of the alternatives [59].Collaborative Recommendation is suitable for learners in E-learning forum based on its performance metrics[65].

The basic architecture for e-learning process is shown in fig. 1. In this architecture, the academia interacts directly with the E-Learning main page. The Tutor manages the curriculum with E-Learning. Ontology construction is made for courses in E-learning. Adaptivity and assessments are updated through indexing. Knowledge base is in the repository and the ontology editor is used for the ontology construction for the elearning courses. Personalization is done for the elearning courses based on ontology creation. The main phases of elearning system is given below,

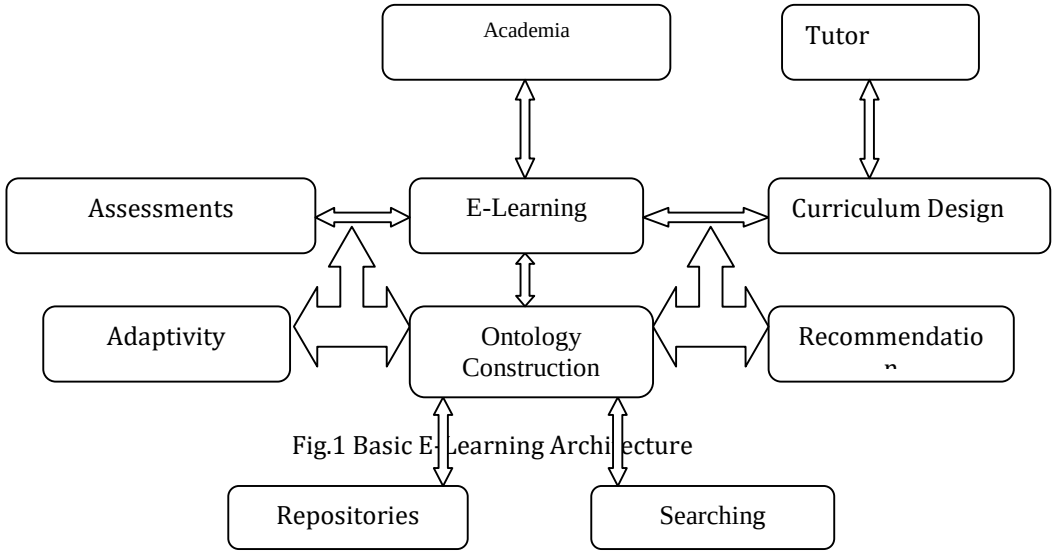


Fig.1 Basic E-Learning Architecture

II. SURVEY OF SEMANTIC E-LEARNING

Ontology Based Text Mining (OBTM) method is used to cluster the data according to their similarities. This method is efficient and effective for clustering the courses [1], [2]. Information Retrieval [3] with OBTM method not only used for text documents, it is also used to extract the posts on Social Networking Websites (SNW) [4]. In Social Network Analysis, OBTM method is used to analyze the unstructured data. In SNW, E-Learning [5] is the emerging technology nowadays. The importance of clustering process in E-Learning system is discussed in [6]. Text mining goal is to derive high quality information from text. The Weighted Text Graph based Text Mining Ontology make the effective Statistical Report in Text [7]. The preprocessing steps such as stop word removal, stemming, TF-IDF, clustering are discussed in [9], [20]. In order to save both space and time, the stop words are dropped at indexing time and then ignored at search time.

Stemming reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents [10], [21]. From the preprocessed text data the term-frequency was identified. But for the concept weight based calculation the k-Means clustering algorithms is used to comparing with other. Concept Based indexing technique with dynamic weight is used to effectively identify the document [11].

There are many ontology extraction tools available, for example TETDM [16] provide the Text Mining Skills for the Beginners. Text-To-Onto and On-To-Gen tools, which are fulfill the functional specifications. The Text-to-onto miner uses OWL as the container to hold the relationships [9]. Querying Text Based Documents in Ontology Repositories is an important task. The process of Clean, populate, enrich a knowledge base repository exploited to answer a complex queries [10].

The Ontology Based Information Extraction (OBIE) is the application of automatic text grading system to identify the correct and incorrect statements [11], [12]. The OBIE retrieved the data in relational databases using query process [22]. The variability of correct and incorrect statements can be reduced through the use of a hybrid configuration. Selection of research projects is an important research topic in research and development (R&D) project management [13]. Ontology Based Information Retrieval(OBIR) is capable of handling the numerous variations in same identity[17], and the keyword matching domain ontology are discussed in [18], [19]. The common Indexing includes Term frequency (TF) and Inverse document frequency (IDF). TF means repeated words are strongly related to content; IDF means uncommon term, which is more important in the documents. Indexing of the fetched web contents using RDF, OWL languages for effective information retrieval is the key issue of research nowadays. In deploying full text indexation technique for web contents, it is better to index the pages using ontology. Ontology based indexing will have small size [14].

In the OBTM method, ranking the results is also analyzed [23]. There are many ranking algorithms available. First, querying and keyword matching and the page-rank method is implemented [27]-[29]. After analyzing the basic concepts of OBTM the E-Learning with Semantic concepts are discussed. E-Learning architecture provides semantic-based services to students and tutors, in particular ways to browse and obtain information through web services.

Ontology is commonly encoded [8], [18] using Web Ontology Language (OWL) and Resource Description Format (RDF) with ontology Editor. There are more than 50 kinds of ontology editor, such as Protégé [88], Ontolingua, Apollo, Onto Studio, Swoop and Top Braid Composer. Among all Protégé is the free, open source ontology editor based on relationship, data properties, concepts and annotation property. Word Net [30] is a lexical database [31] for the common English dictionary. It is used to improve the quality of clustering. Word Net covers polysemy and synonymy for text representation. Word Net semantic similarity measured the word sense disambiguation.

Table 1, Table 2, and Table 3 shows the survey of few existing clustering algorithms, Indexing algorithms and Recommendation approaches.

TABLE 1
ANALYSIS OF CLUSTERING ALGORITHMS

Clustering Algorithms	Merits	Demerits
K-Means[8]	High cluster Speed	Non-hierarchical
Self Organized Map[9]	Large, complex data sets	Always requires nearby similar points
H-Agglomerative [23]	Easy to implement, quality of clusters	Difficult to identify the number of clusters
Particle Swarm Optimization[84]	Low computational cost	Too slow while clustering

TABLE 2
ANALYSIS OF INDEXING ALGORITHMS

Indexing Strategy	Description	Limitations
Triple store[38]	(S,P,O) stored in Triples form	Slow execution, requires many self joins
Vertical Partitioning[50]	Different table for all triples	Stored the data in different table
Property Table[16]	Data is stored in relational table	Unable to handle multi valued attributes

The paper is organized as follows. Section 3 deals with the proposed methodology for E-learning process. In section 4 performances of our proposed methodology is analyzed. Finally, concluded the paper in section 5.

II. PROPOSED METHODOLOGY

The main aim of this work is learning the courses systematically with the help of indexing and recommendation techniques. The resultant Collaborative Recommender System (CRS) is shown in fig. 2.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

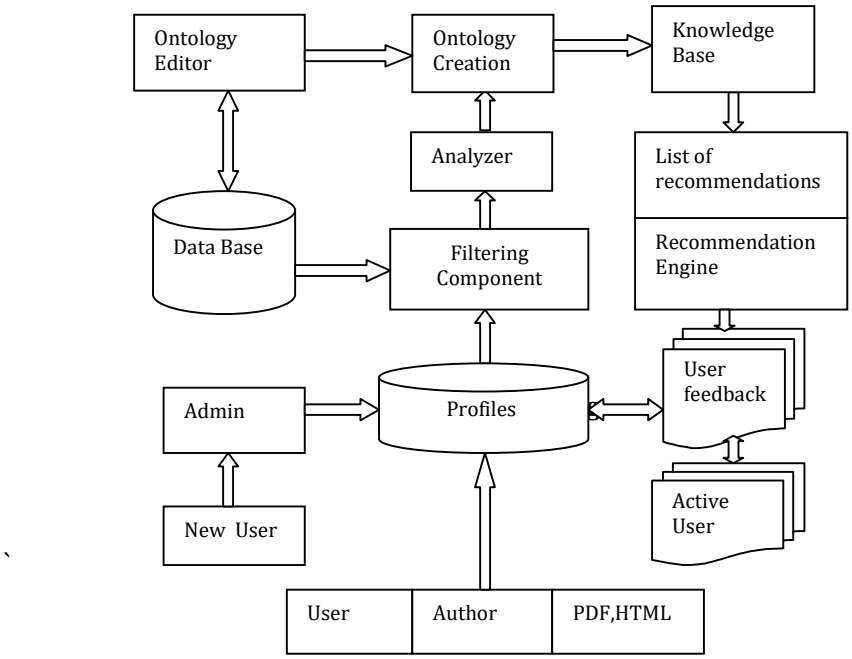


Fig.1 CRS System

The phases involved in our Semantic based educational system PCRS is:

- Preprocessing
- Clustering
- Indexing
- Recommendation
  - Collaborative Recommendation(CR)
  - Content Based Recommendation(CBR)
  - Demographic Recommendation(DR)
  - Utility Based Recommendation(UBR)
  - knowledge Based Recommendation(KBR)
- Ontology creation
  - Ontology Tools
  - Word net Ontology

A. Preprocessing

In this phase, the different document preprocessing methods are analyzed and applied for e-Learning process.

1) *Format conversion*: The documents are converted from the original to the desired one (like doc, img, pdf and so on).

2) *Stop word Removal*: In this phase, we remove the terms which are not useful for our content (i.e., is, by, under, that, etc.). Figure 3 shows the sample stop word removal process.

3) *Stemming*: We reduce the prefix and suffix terms of the word (i.e., running->run). Figure 4 shows the sample stemming process.

4) *White Space Removal*: We reduce the white spaces to save our memory storage.

5) *Part-of-Speech tagging (POST)*: The terms of the document compared with the dictionaries (i.e., names, verbs, adjectives, etc.,)

6) *User Intervention*: After all the preprocessing, the human intervention can be useful to improve the processing steps.

B. Clustering

After preprocessing process, the courses are clustered using K_Means Algorithm.

C. Indexing

Indexing parses and stores data to facilitate fast and accurate information retrieval. Tf - Idf can use a term-document matrix which describes the occurrences of terms in documents; it is a sparse matrix whose rows correspond to terms and whose columns correspond to documents. A typical example of the weighting of the elements of the matrix is tf-idf (term frequency–inverse document frequency): the weight of an element of the matrix is proportional to the number of times the terms appear in each document, where rare terms are up weighted to reflect their relative importance.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

D. Recommendation

This section mainly focused on different Recommendation methods such as Collaborative, Content based, Demographic, Utility based and knowledge based algorithms and finally proposed which one is best.

1) Collaborative Recommendation

The weight given to a Learning Object(LO) is determined by the correlation between that LO and the LO for whom to make a prediction. As a measure of correlation the Pearson correlation coefficient can be used. In this example a positive rating has the value 1 while a negative rating has the value –1, but in other cases a rating could also be a continuous number. The ratings of LO *X* and *Y* of the item *k* are written as X_k and Y_k, while $\overline{X}$ and $\overline{Y}$ are the mean values of their ratings. The correlation between *X* and *Y* is then given by:

$$r(X,Y) = \frac{\sum_k (X_k - \overline{X})(Y_k - \overline{Y})}{\sqrt{\sum_k (X_k - \overline{X})^2 \sum_k (Y_k - \overline{Y})^2}}$$

In this formula *k* is an element of all the items that both *X* and *Y* have rated.

2) *Content based Recommendation*

The content-based filtering systems are mostly used with the text documents. A standard approach for term parsing selects single words from documents. The vector space model and latent semantic indexing are two methods that use these terms to represent documents as vectors in a multi dimensional space.

3) *Demographic Recommendation*

A demographic recommender provides recommendations based on a demographic profile of the user. Recommended products can be produced for different demographic niches, by combining the ratings of users in those niches.

4) Utility based Recommendation

Utility-based recommender systems provide recommendations based on the computation of the utility of each item for the user. Some utility-elicitation methods have been developed on the basis of multi-attribute utility theory (MAUT) to represent a decision maker's complete preference

5*) Knowledge based Recommendation*

Knowledge-based recommender systems (knowledge based recommenders) [1][2] are a specific type of recommender system that are based on explicit knowledge about the item assortment, user preferences, and recommendation criteria (i.e., which item should be recommended in which context)

E. Ontology creation

Ontology is created with the use of Word Net English dictionary. Ontology is created with the concept like" is-a", "part-of", "sub-of" relations. It shows the relationship between the concepts that are holynym, hypernym, hyponym, synonym, etc. Ontology creation with different methods are focused in this paper[90].

III. EXPERIMENTAL RESULTS

The experimental results of preprocessing, clustering, indexing and ontology creation processes are shown in the following figures:

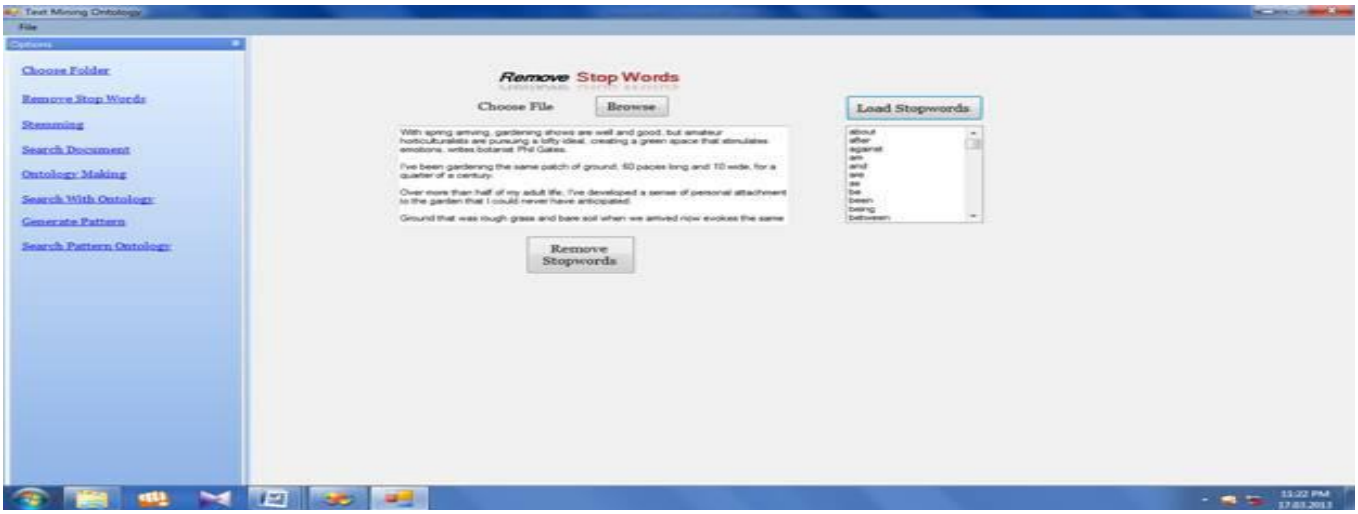


Fig. 1 Stopword Removal

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

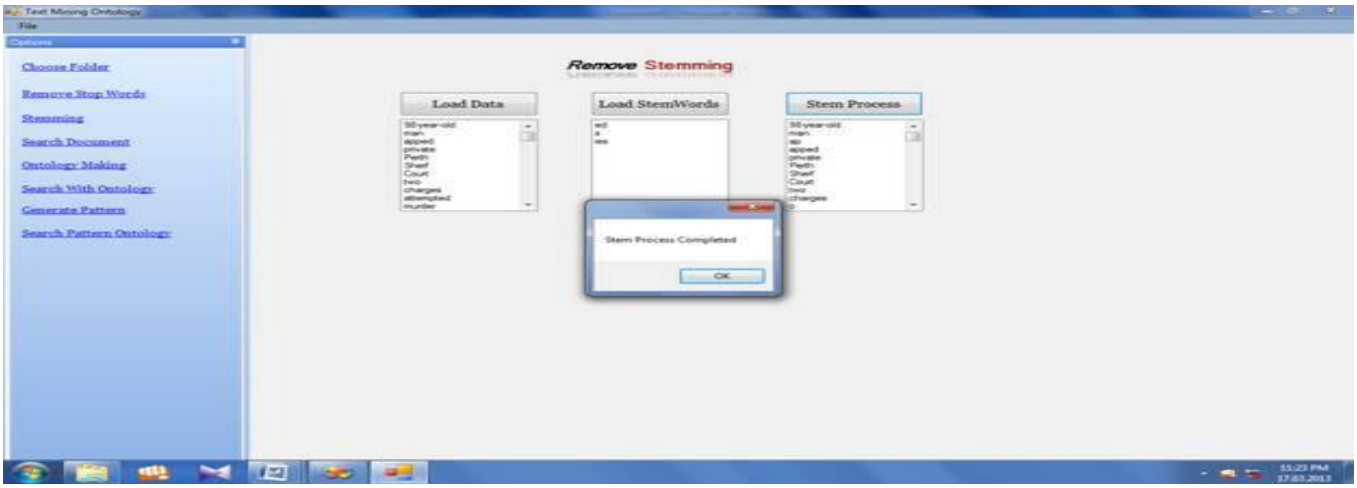


Fig. 2 Stemming Process

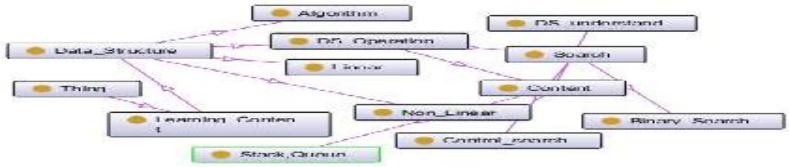


Fig. 3 Ontology Creation

IV. COMPARATIVE ANALYSIS

In this section, the performances of our system CR is compared with other Recommendations systems.

A. Dataset

For an experimental dataset, we extracted classes from RDF and OWL class library of J2SE (Java 2 Platform Standard Edition) Software Development Kit with the use of Protégé. We selected 150 commonly used learning object as programs P={p1, p2, ..., pm,} and 120 classes as library class files L={l1, l2, ..., ln}, which is a subset of P. The number of library class files n is smaller than that of programs m because we excluded 40 library class files (from L) that were not used by any program in P, Consequently, we made 150x120 size dataset from them.

Based on 150 commonly used learning objects the recommendation algorithms are classified under accuracy, instance used, user based, material based, link analysis with the rank 1 to 10.

Algorithms	CR	CBR	DR	UBR	KBR
Accuracy	1	2	2	4	5
Instance Based	2	4	6	4	7
User Based	2	1	4	6	3
Material Based	2	3	4	5	6
Link analysis	1	1	3	4	5

Table.1 Ranking for Recommendation Algorithms

B. Precision of each method

Based on rank the precision value of different recommendations is analyzed for learning objects.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

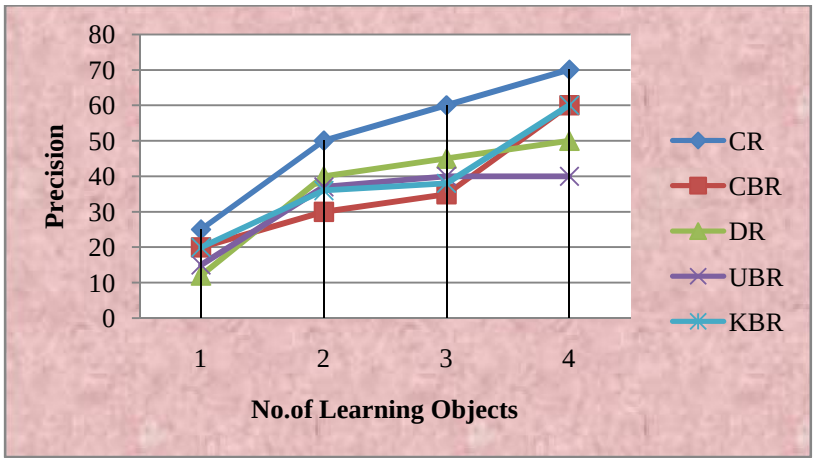


Fig. 1 Precision of each method

C. Half-life utility of each method

The half-life utility metric attempts to evaluate the utility of a recommendation list to a user. It is based on the assumption that the likelihood that a user examines a recommended object decays exponentially with the object's ranking. There is a 50% chance that the user will eventually examine it. This utility can be further normalized by the maximum utility (which is achieved when the user's all known ratings appear at the top of the recommendation list).

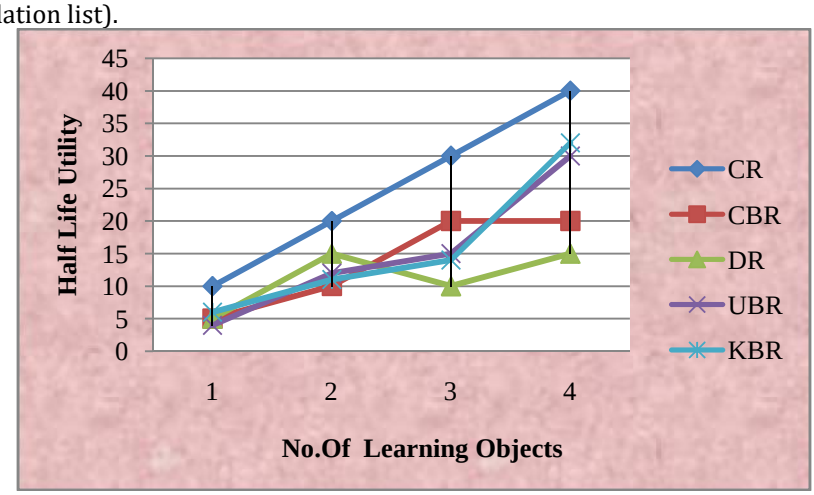


Fig.2 Half Life Utility

D. Recall of each method

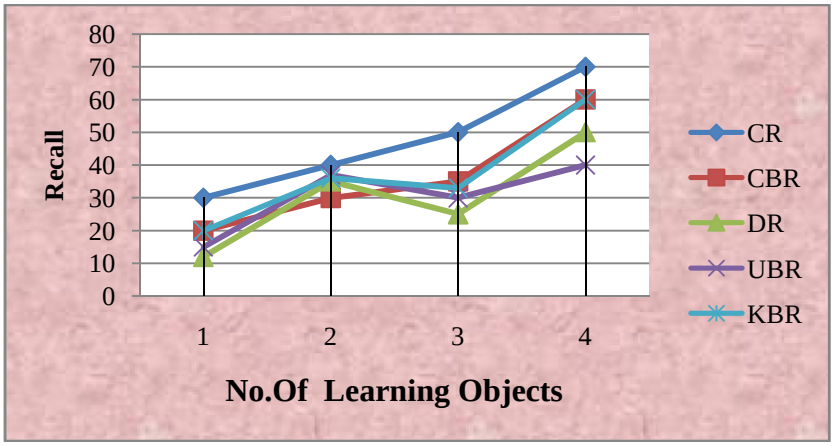


Fig.4 Recall of each method

E. The relation between recall and precision

Precision and recall are the most popular metrics based on this. For a target user i, precision and recall of recommendation, Pi(L) and Ri(L), are defined as

$$Pi(L) = di(L)/ L \text{ , } Ri(L) = di(L) /Di$$

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Where di(L) indicates the number of relevant objects (objects collected by i that are present in the probe set) in the top-L places of the recommendation list, and Di is the total number of i's relevant objects. Averaging the individual precision and recall over all users with at least one relevant object, we obtain the mean precision and recall, P(L) and R(L), respectively

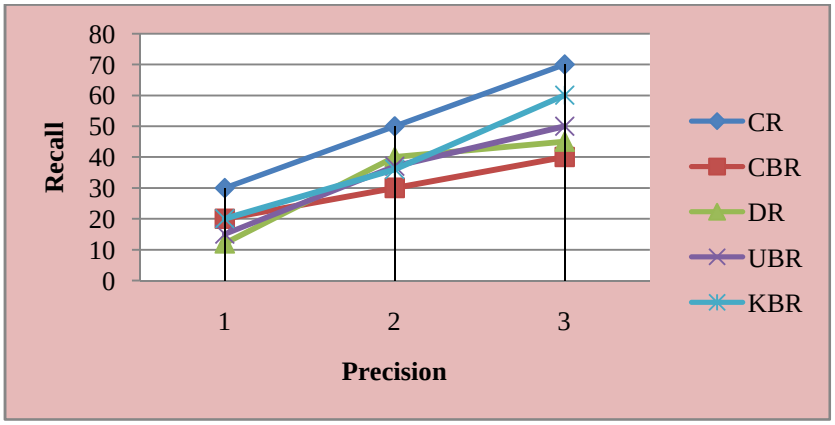


Fig.5 Comparison of Precision and Recall

V. CONCLUSIONS AND FUTURE WORK

Research works in the field of E-Learning is represented by a broad spectrum of applications, ranged from virtual classrooms to remote courses or distance learning. Web-based courses offer obvious advantages for learners by making access to educational resource very fast, just-in-time and relevance, at any time or place. In this paper, based on our previous work, we present the Semantic Web-Based model for our e-learning system. In addition we present an approach for developing a Semantic Web-based e-learning system, which focus on the RDF data model and OWL ontology language. Compared with other Recommendation systems, CRS gives the better performance based on the Precision and Recall. In future, Recommendation process in CRS is to be incorporated.

REFERENCES

1. Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu An, "Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection" IEEE Transactions On Systems, Man, And Cybernetics – Part A:Systems And Humans, Vol.42, No.3, May 2012.
2. Gong Guangming, Jiang Yanhui, Wang Wei, Zhou Shuangwen, "A Clustering Algorithm Based on the Text Feature Matrix of Domain-Ontology", Third International Conference on Intelligent System Design and Engineering Applications,2013.
3. Aarti Singh, Anu Sharma, "A Framework for Semantics and Agent Based Personalized Information Retrieval in Agriculture" 978-9-3805-4416-8,2015
4. K. M. Sam, C. R. Chatwin," Ontology-Based Text-Mining Model For Social Network Analysis" 978-1-4673-0110-7/12/$31.00 ©2012 IEEE
5. Osmar R.Zaiane,Jun Luo,"Towards Evaluating Learner's Behaviour in a Web-Based Distance Learning Environment",IEEE,ISBN:0-7695-1013-2,2001.
6. Hamid Mousavi, Deirdre Kerr, Markus Iseli, Carlo Zaniolo," Mining Semantic Structures from Syntactic Structures in Free Text Documents", IEEE International Conference on Semantic Computing, 2014.
7. Deepak Agnihotri, Kesari Verma, Priyanka Tripati, "Pattern and Cluster Mining on Text Data",2014 Fourth International Conference on Communication Systems and Network Technologies.
8. S.Suguna, B.Gomathi, "Comparison between Clustering Algorithms Based On Ontology Based Text Mining Techniques", International Journal of Advanced Research in Computer Science.
9. S.Logeswari, Dr.K.Premalatha, "Biomedical Document Clustering Using Ontology based Concept Weight" 2013 International Conference on Computer Communication and Informatics (*ICCCI* -2013), Jan. 04 –06, 2013, Coimbatore, INDIA
10. T.Preethi, Mrs.R.Lakshmi, "An Implementation of Clustering Project Proposals on Ontology based Text Mining Approach"
11. Vicky Min-How Lim, Tong-Ming Lim, Wong Siew Fan, "Text-to-Onto Miner: A Concept Driven and Interval Controlled Ontology Builder" 2013 10th International Conference on Information Technology: New Generations
12. Alda Canito, Paulo Maio and Nuno Silva,"An Approach for Populating and Enriching Ontology-based Repositories" 2013 24th International Workshop on Database and Expert Systems Applications
13. Jie Tao, Amit V. Deokar, Omar F. El-Gayar, "An Ontology-based Information Extraction (OBIE) Framework for Analyzing Initial Public Offering (IPO) Prospectus"
14. Fernando Gutierrez, Dejing Dou, Adam Martini, Stephen Fickas and Hui Zong, "Hybrid Ontology-based Information Extraction for Automated Text Grading" 2013 12th International Conference on Machine Learning and Applications
15. Devendra Singh Rathore Dr. R.C.Jain Babita Ujjainiya, " A Text Mining Method for Research Project Selection using KNN", 978-1-4673-6126-2/13/$31.00_c 2013 IEEE
16. Vandana Dhingra, Komal Kumar Bhatia, "SemIndex : Efficient Indexing mechanism for Ontologies",978-1-4799-5173-4/14/$31.00©2014 IEEE
17. G. Arumugam, S. Suguna ,"Optimal Algorithms for Generation of User Session Sequences Using Server Side Web User Logs",Published in IEEE Explorer - June 2009. Pages: 1-6, ISBN: 978-2-9532-4431-1 (Paris Conference).

18. Rina Nakagochi, Kayo kawamoto,Wataru sunayama, "Acquissition of Text-Mining Skills for Beginners Using TETDM",2013 IEEE 13th International Conference on Data Mining Workshops.
19. Yujie Yang, Yunpeng Cai, Wenshu Luo,Zhifeng Li, Zhenghui Ma, Xioolu Yu and Halibo Yu, "An Ontology- based Approach for Text Mining of Stroke Electronic Medical records",2013 IEEE International Conference on Bioinformatics and Biomedicine.
20. HAO-DI LI,QING-CAI CHEN,XIAO-LONG WANG,"A COMBINED MEASURE FOR TEXT SEMANTIC SIMILARITY", Proceedings of the 2013 International Conference on Machine Learning and Cybernetics,Tianjin,14-17 July,2013
21. Feng Hu,Yu-feng Zhang, "Text Mining Based on Domain Ontology",2010 International on E-Business and E-Government
22. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu," Effective Pattern Discovery for Text Mining" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.
23. Yao-Tang Yu,Chien-Chang Hsu,"A Structured Ontology Construction By Using Data Clustering And Pattern Tree Mining", Proceeings of the 2011 International Conference on Machine Learning and cybernetics.
24. Luis Tari, Jo¨ rg Hakenberg, Yi Chen, Tran Cao Son, Graciela Gonzalez, and Chitta Baral , "Incremental Information Extraction Using Relational Databases" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.
25. Chunchen Liu, Jianqiang Li," Semantic-based Composite Document Ranking", 2012 IEEE Sixth International Conference on Semantic Computing.
26. Qi Yu1,Cao Jiang2,Wang Jianxin2," Automatic Evaluation of Domain-Specific Ontology Based on Ontology Co-relation Network", 2013 Fifth Conference on Measuring Technology and Mechatronics Automation
27. Poonam Chahal, Manjeet Singh, Suresh Kumar," Ranking of Web Documents using Semantic Similarity" 2013 International Conference on Information Systems and Computer Networks .
28. Wenhu Tang1, Long Yan2, Zhen Yang2, Qinghua Henry Wu1,2," Improved document ranking in ontology-based document search engine using evidential reasoning", Published in IET Software Received on 1st February 2013,Revised on 10th April 2013,Accepted on 24th April 2013 doi: 10.1049/iet-sen.2013.0015
29. Jinwoo Kim, Dennis MeLeod, "A 3-Tuple Information Retrieval Query Interface with Ontology based Ranking", IEEE IRI 2012
30. Liu Xinhua, Zhang Xutang, Li zhongkai, "A Domain Ontology-based Information Retrieval Approach for Technique Preparation", Elsevier, International Conference on Solid State Devices and Materials Science,2012.
31. Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, "A Semantic Approach for Text Clustering using Wordnet and Lexical Chains", Expert Systems with Applications 42,pp.2264-2275,2015.
32. Tugba Ozacar, "A tool for producing structured interoperable data from product features on the web", Elsevier, Information Systems,Vol.56,pp:36-54,2016.
33. Miriam Fernandez, Ziqi Zhang, Vanessa Lopez, Victoria Uren, Enrico Motta, "Ontology Augmentation: Combining Semantic web and Text Resources" ACM 978-1-4503-0396-5,2010
34. Andrea Zielinski, Jurgen Bock, "A Case Study on the Use of Semantic Web Technologies for Learner Guidance", ACM 978-1-4503-3473-0, 2015.
35. W.Yan, C.Zanni-Merk, D.Cavallucci, P.collet, "An ontology-based approach for inventive problem solving", Elsevier, Engineering Applications of Artificial Intelligence,Vol.27,pp:175-190,2014
36. Li Yaxiong, Zhang Jianqiang, Dan Hu, "Text Clustering Based on Domain Ontology and Latent Semantic Analysis", IEEE, International Conference on Asian Language Processing,2010.
37. David Monk, "Using Data Mining for e-Learning Decision Making", The Electronic Journal of e-Learning ,Volume 3 Issue 1,pp 41-54,ISSN:1479-4403,2005.
38. P.Sheba Alice ,A.M.Abirami, A.Askarunisha, "A Semantic Based Approach to Organize eLearning through Efficient Information Retrieval for Interview Preparation",ICRTIT,IEEE,ISBN:978-1-4673-1601-9,2012.
39. Santi Caballe, David Britch, Leonard Barcalli, Fatos Xhafa, "A Methodological to Provide Effective Web-based Training by using Collaborative Learning and Social Networks", Eighth International Conference on Complex, Intelligent and Software Intensive Systems,IEEE,ISBN:978-1-4799-4325,2014.
40. Reema Sikka, Amita Dhankhar, Chaavi Rana, "A Survey Paper on E-Learning Recommender System", International Journal of Computer Applications, Volume 47-No.9,June 2012.
41. Daniela Resende Silva, Marina Teresa Pires Vieira, "Using Data Warehouse and Data Resources for ongoing Assessment of Distance Learning" , IEEE,0-473-08801-0,2002
42. Leticia dos santos Machado, Karin Becker, "Distance Education: a Web Usage Mining Case Study for the Evaluation of Learning Sites" Proceedings of the 3 rd IEEE International Conference on Advanced Learning Technologies,0-7695-1967-9,2003.
43. Divna krpan, Slavomir Stankov, "Educational Data Mining for Grouping Students in E-learning System", Proceedings of the ITI 34th Int. Conf. on Information Technology Interfaces, doi:10.2498 June 25-28,2012
44. Sana HAMDI, Alda LOPES GANCARSKI, Amel BOUZEGHOUB, Sadok BEN YAHIA, "Enriching Ontologies from Folksonomies for eLearning: DBpedia case", 12th IEEE International Conference on Advanced Learning Technologies, DOI 10.1109,2012.
45. Isabel Guitart, Joaquim More, Jordi Duran, Jordi Conesa, David Baneres, David Ganan, "A Semi-automatic system to detect relevant learning content for each subject",International conference on Intelligent Networking and Collaborative system", IEEE ,iSBN:978-1-4673-7695-2,2015.
46. Charles Tijus, Francois Jouen, Sebastien Poitrenaud, Anna Scius-Bertland, Pierre Collet, Michelle Molina, Paul Bourgine, "Know-How modeling for e-Learning", IEEE RIVF International Conference on Computing & Communication Technologies – Research, Innovation, and for the Future, 978-1-4799-1350-3,2013.
47. Zaffer Ahmed Shaikh, Shakeel Ahmed Khoja, "Towards Guided Personal Learning Environments", IEEE 14th International Conference on Advanced Learning Technologies, 978-1-4799-4038, 2014.
48. Osmar R.Zaiane, Jun Luo, "Towards Evaluating Learners' Behaviour in a Web-Based Distance Learning Environment" 0-7965-1013-2,IEEE 2001.
49. Chakkrit Snae, Michael Brueckner, "Ontology-Driven E-Learning System Based on Roles and Activities for Thai Learning Environment", Interdisciplinary Journal of Knowledge and Learning Objects, Volume 3,2007
50. G.Suresh Kumar, G.Zayaraz, "Concept relation extraction using Naïve Bayes classifier for ontology-based question answering systems", Journal of King Saud University – Computer and Information Sciences, pp:13-24,1319-1578,2015.

51. Quanyu Wang, Xingen Yu, Guilong Li, Guobin Lv, " Ontology-Based Ecological System Model for e-Learning", International Journal of Information and Technology, Vol 2, No.6,DOI:10.7763,2012
52. Mohamed Koutheair Khribi, Mohamed Jemni, Olfa Nasraoui, "Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval", Educational Technology & Society,12(4),pp:30-42,ISSN 1436-4522,2009
53. Margo Hanna, "Data mining in the e-learning domain", Campus-Wide Information Systems, Vol 21,NO.1 pp.29-34,DOI:10.1108/10650740410512301,ISSN 1065-0741,2004
54. TORSTEN LEIDIG, "L³ – Towards an Open Learning Environment", ACM Journal of Educational Resources in Computing, Vol.1, No.1, Article #5,11 pages,2001
55. Changjie Tang, Rynson W.H.lau, Qing Li, "Personalized Courseware Construction Based on Web Data Mining",0-7695-0577-5,IEEE,2000
56. A Sai sabitha, Deepti Mehrotra, "A Push Strategy for delivering of Learning Objects using meta data based association analysis(FP-Tree)",IEEE International Conference on Computer Communication and Informatics,Jan 04-06,Coimbatore,India,978-1-4673-2907-1,2013
57. Martina Holenko Diab, Natasa Hoic-Bozic, "Recommender System for Web 2.0 supported eLearning", IEEE Global Engineering Education Conference,Turkey,978-1-4799-3190-3,2014.
58. Vincenza Carchiolo, Alessandro Longheu, Michele Malgeri and Giuseppe Mangioni, "Courses Personalization in an E-learning Environment", Proceedings of the 3 rd International Conference on Advanced Learning Technologies,0-7695-1967-9,2003
59. Huiyi Tan,Junfei Guo, "E-learning Recommendation system", IEEE International Conference on Computer Science and Software Engineering,978-0-7695-3336-0,2008
60. Sana HAMDI, Alda LOPES elf ,"Enriching Ontologies from Folksonomies for eLearning:DBpedia case",12 th International Conference on Advanced Learning Technologies,IEEE,ISBN:978-0-7695-4702-2,IEEE Computer Society,2012.
61. Sarma Cakula, Maija Sedlenieece, "Development of a Personalized e-Learning model using methods of Ontology", ICTE in Regional Development, Dec, ,pp:113-120,1877-0509, 2013
62. B.Saleena, S.K.Srivatsa, "Using concept similarity in cross ontology for adaptive e-Learning systems", Elsevier, Journal of King Saud University – Computer and Information Sciences, 1319-1578,pp:1-12,2015.
63. Marija Blagojevic, Zivadin Micic, "A web-based intelligent report e-learning system using data mining techniques", Elsevier, Computers and Electrical Engineering, vol.39, pp:465-474,2013.
64. Javier Enrique Rojas Moreno, "Adaptation of Learning Strategies in Learning Objects for using Learning Styles", Elsevier, International Conference on Future Computer Supported Education, IERI Procedia, Vol.2, pp:808-814,2012.
65. Eugenijus Kurilovas, Inga Zilinskiene, Valentina Dagiene, "Recommending suitable learning paths according to learners' preferences:Experimental research results", Elsevier, Computers in Human Behavior,vol.51,pp:945-951,0747-5632,2014.
66. Anbuselvan Sangodiah and Lim Ean Heng ," Integration of Data quality Component in an Ontology Based Knowledge Management Approach for E-learning System", IEEE International Conference on Computer and Information Science,pp.105-108,978-1-4673-1938-6,2012.
67. Sungjin Cho, Jeon-Young Kang, Ansar-UI-Haque Yasar, Luk Knapen, Tom Bellemans, Davy Janssens, Geert Wets, Chul-Sue Hwang, "An Activity-based Carpooling Micro simulation using Ontology", Elsevier, The 4th International Conference on Ambient Systems, Networks and Technologies,vol.19,pp:48-55,2013.
68. Arundhati Walia, Neeraj Singhal, A.k.Sharma,"A Novel E-learning Approach to add more Cognition to Semantic Web", IEEE International Conference Intelligence & Communication Technology, 2015.
69. Sharifullah Khan, Muhammad Safyan, "Semantic matching in hierarchical ontologies", Elsevier, Journal of King Saud University – Computer and Information Sciences,Vol.26,pp:247-257,2014.
70. Fabio Augusto Procopio de Paiva, Jose Alfredo Ferreira Costa, Claudio Rodrigues Muniz Silva, "A Hierarchical Architecture for Ontology Based Recommender Systems", BRICS Congress on Computational Intelligence & 11th Brazillian Congress on Computational Intelligence,2013.
71. Frank Goossen, Wouter IJntema, Flavius Frasincar, "News Personalization using the CF-IDF Semantic Recommender", ACM 978-1-4503-0148-0,2011.
72. YAO TANG YU, CHIEN-CHANG HSU, "A STRUCTURED ONTOLOGY CONSTRUCTION BY USING DATA CLUSTERING AND PATTERN TREE MINING" IEEE Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 2011.
73. S.C. Punitha, K. Mugunthadevi, and M. Punithavalli , "Impact of Ontology based Approach on Document Clustering" International Journal of Computer Applications (0975 – 8887) Volume 22– No.2, May 2011.
74. Feng Hu, Yu-Feng, "Text Mining Based on Domain Ontology", IEEE International Conference on E-Business and E-Government, 2010.
75. Smita Bachal, Prof.S.M.Sangave, "Survey on An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", International Journal of Advanced Research in advanced Engineering,Vol.1,Issue 2,April 2014.
76. S.Subbaiah, " Extracting Knowledge using Probabilistic Classifier for Text Mining", Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, February 2013.
77. Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, Guangquan Zhang, "Recommender system application developments : A survey", Elsevier, Decision Support Systems,Vol.74,2015.
78. Nikolas Galanis, Enric Mayol, Marc Alier, Francisco Jose Garcia-Penaivo, "Supporting, evaluating and validating informal learning. A social approach", Elsevier, Computers in Human Behavior,Vol.55,pp:596-603,2016
79. Katerina Kostolanyova, Stepanka Nedbalova, "Individualization of foreign language teaching through adaptive eLearning",Springer International Publishing Switzerland,pp.159-170,2016
80. Markus Krause,"A Behavioral Biometrics based Authentication Method for MOOC,s that is Robust against Imitation Attempts",ACM,2014

81. Johann M.Marquez-Baarja, Guillaume Jourjon, Alexander Mikroyannidis, Christos Tranoris, John Domingue, Luiz A.DaSilva, " FORGE: Enhancing eLearning and Research in ICT through remote experimentation", IEEE Global Engineering Education Conference,2014
82. Marta A rguedas, FatosXhafa, Thanasis Daradoumis, "An Ontology about emotion awareness and affective feedback in elearning" ,IEEE International Conference on Intelligent Networking and Collaborative Systems,2015
83. Angel Fidalgo-blanco, Maria Luisa Sein-Echaluce, Francisco Jose Garcia-Penalvo, Miguel Angel Conde-Gonzalez, "Learning Content management systems for the definition of adaptive learning environments", IEEE, 2014
84. Sara Alaee, Fattaneh Taghiyareh, "A Semantic Ontology-based Document Organizer to Cluster eLearning Documents", IEEE Second International Conference on Web Research,2016
85. Sohail Sarwar, Zia UI Qayyum, Muhammad Safyan and Rana Faisal Munir , "Ontology Based Adaptive,Semantic E-Learning Framework", Springer Science Business Media Singapore,2016
86. Jose Angel Olguin, Francisco Javier Carrilo Garcia,Ma. De la Luz Carrilo Gonzalez,Antonia Mireles Medina, Julio Garcia Cortes,"Expert system to engage CHAEA Learning Styles, ACRA Learning Strategies and Learning Objects into an E-Learning Platform for Higher Education Students", Springer International Publishing AG 2017
87. Karsten O.Lundqvist, Guy Pursey and Shirley Williams," Design and Implementation of Conversational Agents for Harvesting Feedback in eLearning Systems", Springer-Varlag Berlin Heidelberg 2013
88. The Protégé Ontology Editor and Knowledge Acquisition System. [Online]. Available: http://protege .stanford.edu/
89. Fabio Augusto Procopio de paiva, "A Hierarchical Architecture for ontology based Recommender systems", Brics Congress on Computational Intelligence & 11 th Brazillian Congress on Computational Intelligence,978-1-4799-3194-1,2013.
90. Suguna,Sundaravdivel,Gomathi," A Novel Semantic Approach in E-learning Information Retrieval System", 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th & 18th March 2016, Coimbatore, TN, India.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Quality Determination of Indian Pulse Seed using Imaging Techniques

[1]SalomeHemaChitra H, [2]Thangaraj M, [3]Suguna S
[1]Dept of Computer Science, Sri Meenakshi Govt. Arts College for Women (A),
Madurai – 625 002, Tamil Nadu
[2]Dept of Computer Science, Madurai Kamaraj University, Madurai,
Madurai – 625 021, Tamil Nadu
[3]Dept of Computer Science, Sri Meenakshi Govt. Arts College for Women (A),
Madurai – 625 002, Tamil Nadu
salomechitra_2@yahoo.com
thangarajmku@yahoo.com
kt.suguna@gmail.com

ABSTRACT *Seed is the most basic entity of agriculture, which governs the quality and yield of its production. Without good seeds the investment on fertilizers, water, pesticides, and other inputs will not be worth. It is necessary to improve the quality of seed for ensuring the high efficiency, quality and productivity of agriculture production. The four basic parameters for the seed quality is Physical qualities of the seed within the specific seed variety, Physiological qualities which refers to the aspects of performance of the seed, Genetic quality which relates to specific genetic characteristics of seed variety and Seed Health which refers to the presence of diseases and pests within a seed variety. The traditional method for quality determination of the seeds in industries is human-intensive and time-consuming while analyzing the quality parameters like Vigorness, Germination level, foreign particles, trueness of the seed, etc. In existing, the color trueness of the seed is evaluated by Histogram Color Pixel Intensity, Vigorness of the seed is estimated using textural pattern analysis and binarization method is used to determine the Germination level of the seed. Thus in this paper, Seed Color Purity Analysis, Vigorness Analysis, Germination Emergence Analysis are considered for determining the quality aspects of the pulse seeds. The above stated quality aspects are analyzed through four processing components such as Image Acquisition, Pre-processing, Quality Analysis and Quality Results. Our proposed method outperforms in 94% than existing method in terms of performance accuracy ratio, recognition rate and reliability rate.*
Keywords: *Indian pulse seed, image processing techniques, quality analysis, vigorness, thermal image, germination emergence.*

Introduction

The seed is the first determinant of the future plant development and it is the master key to success with the good crop productivity with increase in the economical profits. Benefits from the breeding can only be transferred to the farmer if good quality seed is released, and farmers expected outcome is achieved through only by quality trueness of seed varieties. Presently, the seed quality determined by involving the seed into various stress tests and immerses the seed into various chemical solutions that could distort the originality of seed. To overcome these disputes, our seed quality determination method involving an image analysis techniques to avoid distortion of a pulse seed. The variety testing of pulse seed may be aimed to identify variety, to discriminate different varieties, to check the genetic purity or to provide a characterization of the variety. Germination capability and seed vigor are the key objective to achieve good cropping. To measure the seed quality assessment, many factors are considered in the germination test, for instance, seed quality additives, species purity, physical purity, pests and infection, seed germination and seed vigor. Though, it is extremely complex to recognize which kind of seeds should be superior used in seed germinations. Thus, this study is applied the image processing and computer vision method instead of using only human visualization. The traditional way to behavior the germination test, vigor test and purity test is based on human capability, it takes times and high labor to conduct the quality test in the seed quality control process. Thus this context introduces a frame for determining the quality aspects of the Pulse seed in more accurate and efficient manner. The image analysis techniques are included to reduce the manual process for the analysis of quality pulse seed for good cropping. This system employed with various form of image set like true color Image, Thermal image and Gray-scale images for the better quality determination that would guarantee for the high-quality harvesting. The morphological features are extracted and evaluated to analyze the aspects of the quality would satisfy an assured quality of the pulse. The remaining concepts of this context are described in the following chapters. In section 2, the related work is discussed. The section 3 declared the proposed methodology. In section 4, the results and discussion is illustrated.

RelatedWork

The paper [1], provides the quality of rice grains based on its size and use the different varieties of rice grains for testing. The scheme is developed using set of images and are categorized using decision-tree based classification technique. The consequences are established to be expectant. In paper [2], proposed a technique that distinct with the assist of automated image processing mechanism on MATLAB. In these paper three parameters; types of wheat, Foreign Particle and Admixture of Agmark values are distorted over to computerized structure for advanced quality examination of wheat. The paper [3], discussed and suggested another method in rice grading for Malaysia's type of rice using image processing method based on several features like is length, color and shape. In paper [4], analysis is performed on basmati rice granules to evaluate the performance using image processing and Support vector machine (SVM) is

implemented based on the features extracted from rice granules for classification grades of granules. The paper [5] deals with the quality assessment of rice grains based on its size. Based on the size the grains are graded as (grade 1, grade 2 and grade3). Here they considered different varieties of rice grains for testing like Basmati, sona masuri, boiled rice, egg rice etc and classified using decision tree based classification.

The paper [6] represents an approach to determine the quality of rice seed by identified the impurities in the rice seed with texture analysis and linear binary pattern. The paper [7] presented evaluation of moisture content of cereals by texture analysis in image processing. The paper [8] represented an approach for quality determination by evaluating germination emergence by color and morphological measurement. In paper [9], proposed method for quality assessment of Indian Basmati rice grains using Top-hat Transformation which attained high quantity of accuracy in prospering the effects of the Non-uniform Illumination than Computer revelation Inspection. Paper [10] proposes a model that uses color and geometrical features as characteristics for categorization. The rating of rice illustration is completed according to the dimension of the grain kernel and presence of impurities. The paper [11], developed a system for automated seed vigor assessment. This system contains a flatbed scanner which is used to capture the images of seedlings; this scanner is interfaced with computer. The images obtained were processed by computer to calculate the vigor index based on sample mean of various statistics acquired from morphological features of the image seedlings. The system was tested for lettuce seedlings grown in dark for three days. In paper [12] proposed an image processing computer application to automatically assess the vigor of three-day-old soybean seedlings. The soybean seedlings were segmented from the background and converted into various digital formats. These representations were used to segment the seedlings into normal and abnormal categories. The normal seedlings are further processed to perform length measurement. The paper [13] developed a system for computer-aided image analysis of digital images to evaluate seedling growth as a measure of seed vigor. The paper [14] proposed image processing techniques for grading of rice samples based on their sizes. Image thus acquired is then converted to binary image to which they apply morphological functions and by finding the possessions of the associated mechanism in the image the things features were removed. Based on the substance features, stem graphs were designed and the grain seeds which have smaller values than a threshold were redundant. At last they compute the percentage of full time-span grains in the sample image to status the quality. In paper [15] proposed image processing techniques for identifying two varieties of rice based on their shape and size. Image of a sample grains spread on the black or butter paper were arrested using a digital camera, the edge recognition function were achieved to compute the Geometric parameters. Based on these constraints they classified rice seeds into three parts namely normal, long and small rice seeds and exhibited the count of normal, long and small rice seeds on screen.

Proposed Methodology

The proposed method considered ten varieties of Indian Pulse Seed namely COBLK6, RG7, COGR8, CO5, CORG6 and COCP7. Each of these seeds are pre-processed and deployed for quality analysis. In the manual process an experts inspect the individual pulse seed, based on the features like major axis length, minor axis length and its area and then grade the pulse seed. The same features are used in automated method for grading of the pulse. In addition to this, the features of the varieties of the seed is extracted from the image set and included for the quality determination analysis. Traditionally, the quality of pulse seed is analyzing using manual process which leads to much time-consuming and more labor-intensive, but the computer aided vision system is established to reduce the time consumption in determining the quality aspects of the pulse seed.

Farmer's expected outcome is achieved through only by quality trueness of seed varieties. Presently, the seed quality determined by involving the seed into various stress tests and immerses the seed into various chemical solutions that could distort the originality of seed. To overcome these disputes, the seed varietal identification involving image analysis techniques is proposed to avoid distortion of a pulse seed. The variety testing of pulse seed may be aimed to identify variety, to discriminate different varieties, to check the genetic transparency or to afford a classification of the variety. Germination capability and seed vigor are the key objective to achieve good cropping.

Thus in this paper, a novel frame for determining the quality aspects of the pulse seed is recognized which includes:

- Seed Color Purity Analysis
- Vigorness Analysis
- Germination Emergence Analysis

Fig 1 shows the framework for Quality Determination procedure for Indian Pulse Seed,. The steps involved here are Image Acquisition, Pre-processing, Quality analysis and Quality results.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

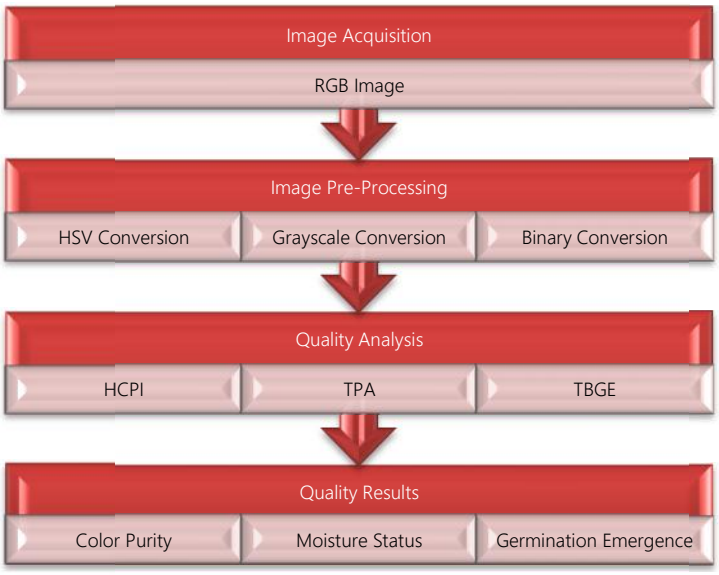


Fig 1. Framework for Quality Determination of Indian Pulse Seed

Seed Color Purity Analysis

The seed color purity analysis of the pulse seed is analyzed to classify and grade the seed by its quality using its color. The manual process of color purity would lead to decrease in test accuracy, expensive, human errors and time-consuming, but this system of quality determination would provide accurate and error-less purity results of the pulse seed. Fig 2 shows components of seed color purity analysis.

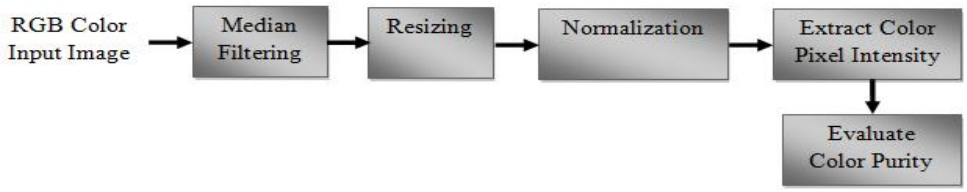


Fig.2.Process Flow of Seed Color Purity Analysis

Image Acquisition

In this module, the images of the pulse seed was capture using high-quality camera and stored as the image set. This unit consists of digital CCD camera and illumination unit consisting of two lamps adjusted at angle with respect to a seed position for correct field view of camera and a base covered with black cloth for placing seed samples. All images are 300dpi resolution and 240 x 218 pixels in PNG format. The resulted images show relatively bright with dark background. Fig.3 shows a sample data set for healthy and unhealthy seeds for further processing.

Healthy pulse seed sample set	Unhealthy pulse seed sample set
	

Fig.3.Indian Pulse Sample Dataset of healthy and unhealthy seeds

Pre-processing

The image pre-processing step is used to enhance the representation and to improve the quality of the image for the quality analysis. The input image for this color purity analysis would be RGB color image and it is pre-processed to obtain the color pixel intensity for the analysis of color purity. The *Median Filter* is pertained to improve the quality of the pulse seed image. It is often desirable to perform noise reduction on the image. Such noise reduction is the step used to enhance the image for the later processing. Then the images are rescaled to 128 x128 pixels using *Bi-cubic Interpolation* technique. At last the resized image is processed to remove the effect of color variations using *Normalization.* The normalization that changes the range of pixel intensity values which normalizes the poor contrast due to glare. Fig.4 shows the result set for our proposed seed color purity method.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

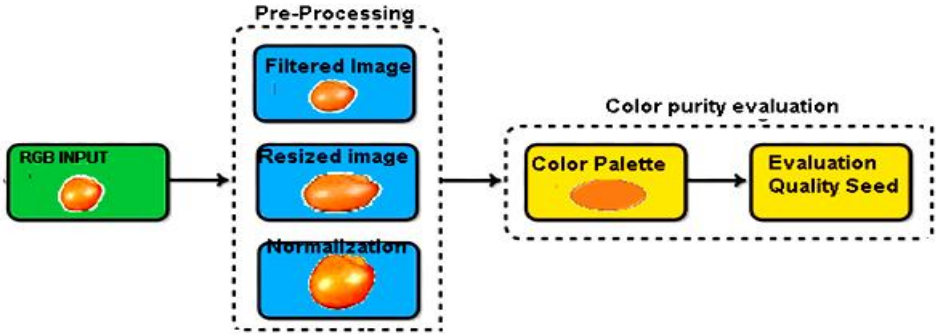


Fig.4. Image Result for Pre-processing in color purity determination

Quality Analysis

The applications of computer image processing in the fields of pulse seed clarity test efficiently advance the technical level of the seed purity recognition. In this processing component, the quality of the seed is evaluated based on the color pixel intensity for various seed varieties. Color pixel intensity for RGB is extracted from the sample healthy and unhealthy sample seed images. Then the color pixel intensity is classified as Healthy Seed-color purity and Unhealthy Seed-color purity and framed seed color palette for various seed. It decides that the extracted color intensity of the preprocessed test seed images is matched with healthy seed color palette classified color intensity would define the seed as good for cultivation and the case doesn't satisfy the color intensity would reject the pulse seed.

Quality Results

The evaluated color pixel intensity for sample seed varities is classified as a Color Palette and shown in Fig.5. This representation of color palette (a) and (b) depicts the color shade for the Healthy Seed and Defeated Seeds. This classified color shade is compared with the color intensity value of the test sample pulse seed image. This module obtains accurate color purity of the healthy pulse seed when compared with the existing technique and decline the unhealthy seeds that don't satisfy the range of the strong color-shade.

(a) (b)

Fig. 5 Color palette representation of pulse seed: (a) Healthy Seed (b) Dameged Seed

Vigorness Analysis

Vigor analysis is a significant module of seed testing because it's more sensitive test than germination, and because loss of vigor may be noted much earlier than loss of germination and also essential for seed production companies and commercial growers to evaluate and develop production and post-harvest techniques, to make inventory management and sales decisions and to justify premium prices to have confidence in the performance of their crops . The components of the vigorness analysis as shown in Fig.6.

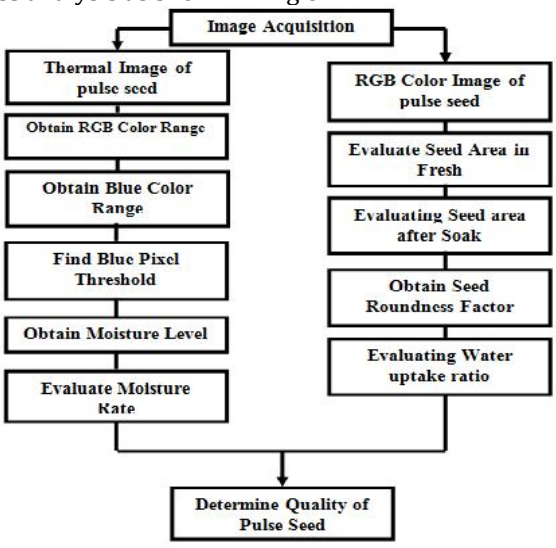


Fig 6. Process Flow of Vigorness Analysis of Pulse Seed

Moisture content is the most essential aspect manipulating objective and automatic possessions of pulse crop seeds. Seeds may lose or attract water during and after grinding. In this work, the vigorness test is performed using RGB color image and thermal image. The roundness factor is employed in RGB image to analyze the vigorness ratio and the blue pixel intensity value is evaluated from the thermal image, then the blue pixel intensity is processed to find out the water uptake or moisture content ratio from the pulse seed.

Image Acquisition

Moisture pulse seed Image acquisition of this unit is processed like previous acquisition procedure as stated in section 3.1.1 for various time intervals like 0hr, 6hrs, 12hrs, 18hrs and 24hrs. Thermal images of seed samples were taken with the IR camera (SC7600; FLIR Systems) that had a resolution of 320x240 pixels. Thermal image system can record the temperature. Thermal images have hot and cold spots which are used to predict the moisture content level of the pulse seed image easily. A sample set used for seed roundness factor analysis is same as Fig.3 in section 3.1.1. And sample set for thermal image processing is shown in Fig.8.

Pre-Processing

The input image might be either RGB color Image or Thermal Image. The pre-processing is employed after the image acquisition for further process of vigorness test. In this component, pre-processing is the process used to enhance the images which are increasing the chances for success of other processes. The pre-processing includes two different processes for analyzing the vigorness of the seed using thermal image and RGB color image. The soaked Pulse seed RGB image is pre-processed by smoothening and contrast enhancement. The thermal image is pre-processed to obtain the RGB color ranges from the RGB channels. From this color range, the blue color threshold value is evaluated. This technique engaged to detect the blue color pixel values from the images contains RGB color pixels, because the blue color portion of the thermal images of the pulse seed depicts the moisture content. The blue color pixel area of the pulse seed image is detected using color threshold value. The threshold value that selects the blue pixel section from the image by thresholding the blue pixel values and automatically situating other colors R, G by zero would be eliminated.

Quality Analysis

Moisture content in the seed is required to be adequate for the seedling growth. If less than the range it doesn't germinate and also if it is superior to the moisture range then it may petrify. After the pre-processing of thermal image of the pulse seed, evaluated the blue color mean intensity value. This mean intensity value would extract the blue pixel range of the image and depicts the water uptake level of the seed. Then the moisture content level is evaluated to find out the quality of seed that is further used for germination process and for the cultivation. The moisture content ratio is evaluated by the equ.1:

$$\text{Moisture Ratio (M)} = \frac{\text{Total Image size (S)}}{\text{Blue Pixel Ratio(R)}} \times 100 \quad \text{----- (1)}$$

The estimated moisture ratio is verified with moisture ratio range to qualify the seed and evaluate the water uptake of the seed. If the moisture ratio is between 7% to 9 % then the seed is determined as quality seed for healthy cropping elsewhere, the seed is declined to be low-quality that is not good for cultivation and doesn't met the level of water uptake.

In the second phase, the pre-processed RGB color pulse seed image is contributed as the input image. Seed can be represented by its swelling and growth, which can be measured by means of ROI property in image analysis techniques. The morphology shape of the seed is elongated in corresponding dimensions. Swelling process could be measured the increase in size (area, perimeter, roundness, spericity, geometrical diameter) and shape. The seed growth is observed and analyzed the quality of the seed by its area. For evaluating the vigor assessment the *roundness factor* is employed. The water uptake ratio or moisture content ratio determines the germination aspect of the pulse seed.

Quality Results

The Fig.7 depicts the analysis of seed roundness factor over time interval. This segmented image would illustrate that the roundness area and the diameter of the seed would increase for the good vigor assessment. This would come to a decision with good quality seeds and decline the bad seeds.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

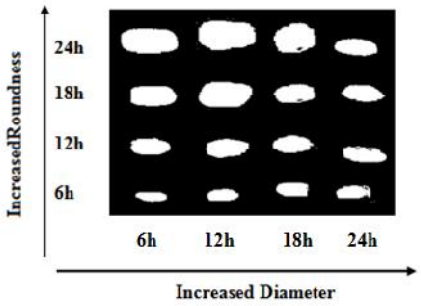
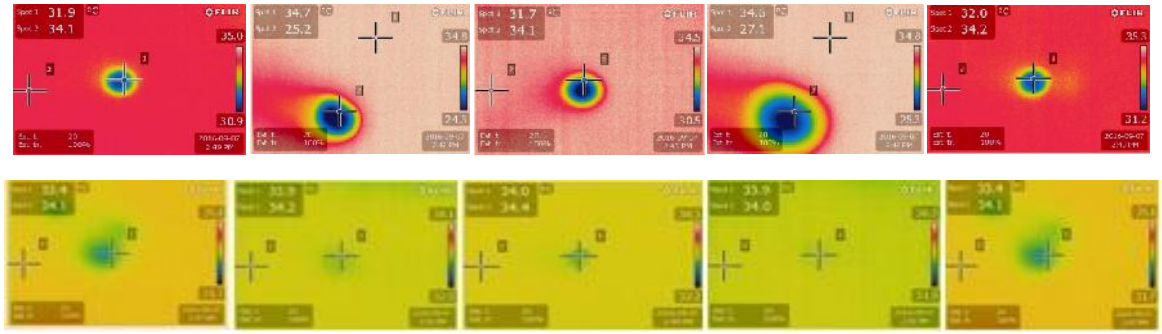


Fig.7.Pulse seed vignorness by roundness factor determination

In Fig. 8, the sample thermal image set for vigorness analysis is shown. Fig 8, (a) (b) (c) (d) depicts the sample seed image that defines the moisture content in term of blue color pixel over 0h ,6h , respectively



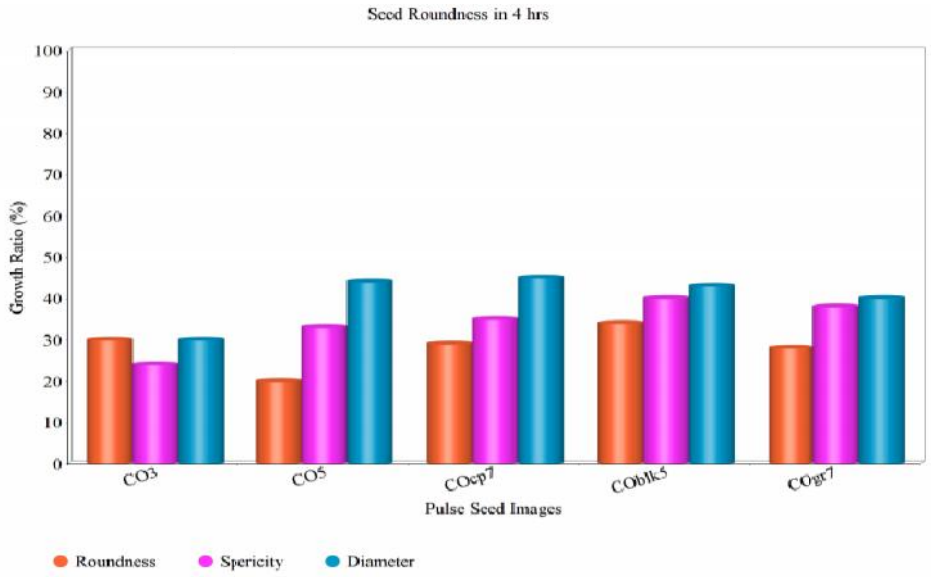
(a) (b) (c) (d) (e)

Fig. 8. Sample of Thermal image for wet and dry pulse seed (a) (b) (c) (d) (e) Moisture seed water uptake by 6h, 12h, 18h and 24h.

The Fig 9 shows the detection of moisture region by thresholding the blue color pixel for wet seed.

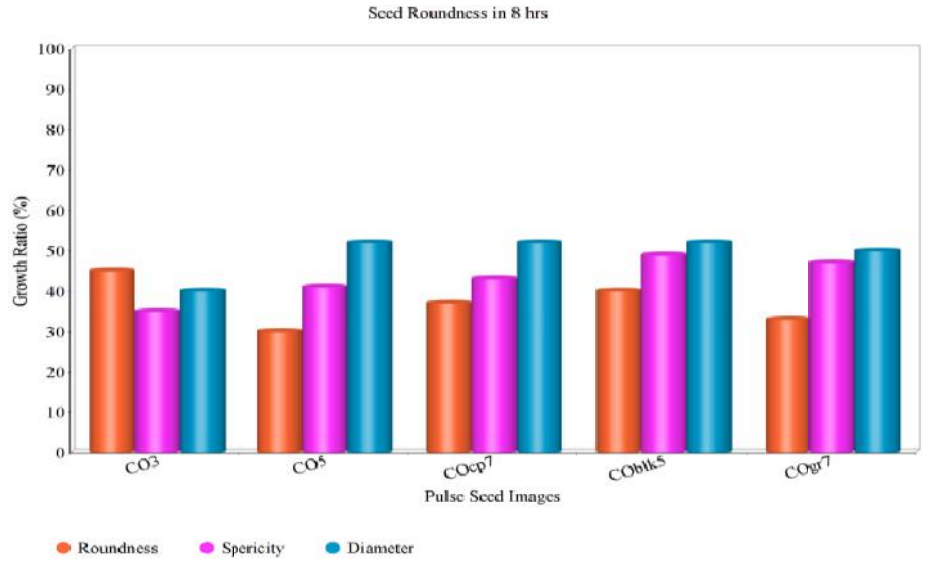
(a) (b) (c) (d) (e)

Fig. 9 . (a)(b)(c)(d) detection of moisture region by blue pixel thresholding

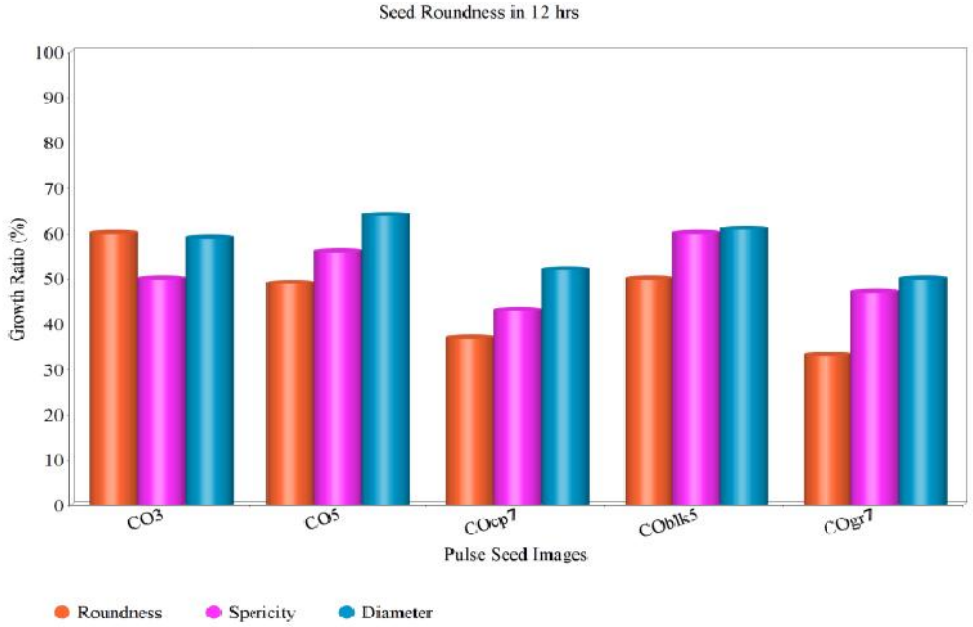


(a)

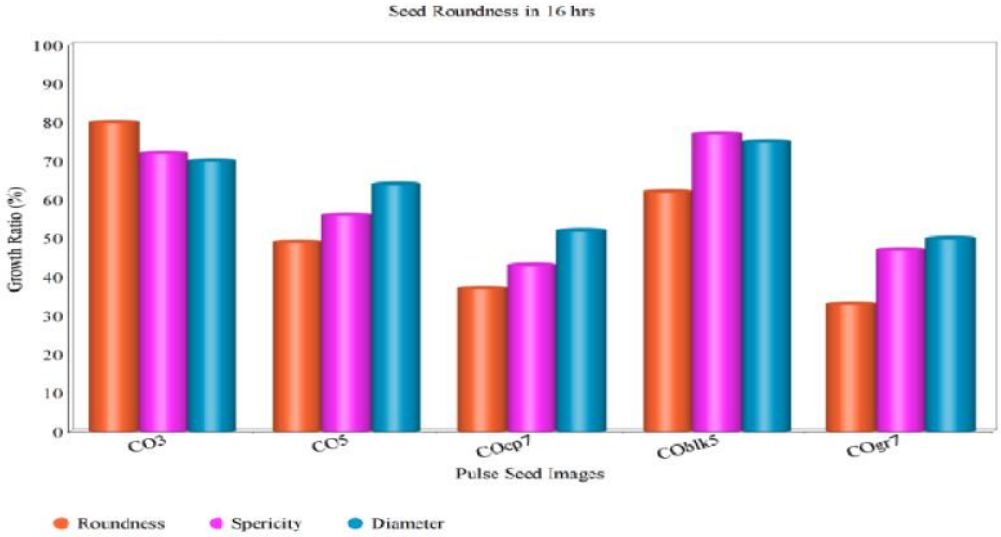
e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.



(b)



(c)



(d)

Fig.10. (a)(b)(c)(d) Seed Roundness Increase in Area observed in Time for every 4 hours

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

The increase in roundness area for the set of sample in time is observed for every 6 hrs during water treatment as shown in Fig.10. The increase in roundness area depicts that the seed would uptake sufficient water in certain time for its further vigorness.

Germination Emergence Analysis

Germination is distinct as "the appearance and improvement of seedling to a phase where the features of its necessary structures specify whether or not it is proficient to expand promote into a plant under constructive conditions in the soil". Germination is usually approved out in germination cabinet under prohibited surroundings.

Seed vigor has been determined by both germination rate and seedling increase time. Germination rate events the rapidity of germination which is generally signified by time to 50% germination, while seedling growth rate is evaluated on a real time basis by measuring rate of elongation of the radical per unit time. A vigor test cannot replace a germination test but rather supplements it with more information about seed quality. Naturally, the germination test will approximately forever have the superior result because the assessment constraints are more pardoning than those used in the vigor test; so the germination analysis notifies producers how their seed will execute in optimal conditions.

Image Acquisition

In this module, the images of the pulse seed was capture using high-quality camera and stored as the image set is depicted in the section 3.1.1. Fig.11 shows a component for evaluation of germination emergence.

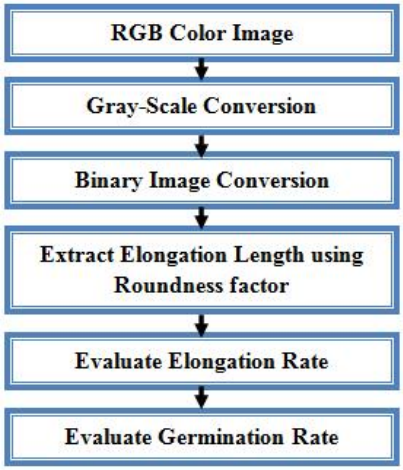


Fig 11. Process Flow of Germination Emergence

Pre-processing

The RGB color image of the pulse seed is input for this process component. The RGB image is pre-processed by Gray-scale Conversion for the noise removal of the original image. The gray-scale conversion is undergone for the process of enhance the image to situate the germination emergence rate. Then the binary conversion is handled with this image. The binary conversion of the image leads to identify the complex edges from the image. Typically the colors used for the binary image are black and white. The color used for the objects in the image is white (foreground) and the rest of the image will be filled by the color black (background). This would assist in find out the elongation length and analyze the quality of germination of the pulse seed using imaging techniques.

Quality Analysis

A digital image of a pulse seed can be regarded as a two-dimensional object which can be measured in size, shape and color density during the progress stage of germination by computer image examination skill. Seed transforms its organic structure passing from a quiescent stage to a proliferating one, and any morphological dissimilarity can be connected with the equivalent distinction of seed geometry and color space mechanism. The germination emergence test that includes the identification of seeds such as Normal Germinants, Abnormal Germinants and Ungerminated seeds. The cumulative number of seeds which have developed into seedlings of normal and healthy appearance with all essential structures of a seedling is referred to as Normal Germinants. The cumulative number of seeds which have germinated during test period but in which seedlings show abnormal or unhealthy appearance is known as Abnormal Germinants. Ungerminated Seed which have not germinated by the end of test period which contains dead seeds, hard seeds and empty seeds.

After the pre-processing completion the binary image is employed with imaging technique of roundness factor to extract the elongation portion from the image. The roundness factor eliminates the seed part from the image and extracts the elongation radicle from the image. Finally the elongation radicle area is evaluated to find the elongation rate for the pulse seed using this elongated image. The elongation length of the seed is evaluated for the elongation rate which congregates the sufficient range of the elongation rate. The radicle length of the seed must be between

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

20% and 80% approximately. If the length leads to less than 20% of its rate then the seed doesn't germinated and the seed is declined for the cultivation.

Quality Results

The germination rate for the pulse seed area is illustrated in this section. The elongation rate for the subclasses of the germination seed is defined in the fig 12. This Fig 12 illustrates the Gray-scale and binary image for the germinated seed image. Here the roundness is detected for the germinated seed to evaluate elongation rate. Then the elongation length of the seed is extracted by the roundness factor to evaluate the elongation rate. Vigorness index is also important factor for quality determination. Vigor index is obtained by geometrical parameter like seed length and germination rate in %. Germination rate and vigor is obtained by the following formula (2) and (3)

Germination rate % = (Number of seedling seeds/Number of total seeds) X 100 _____ (2)

Vigor index (VI) = [seedling length (cm) × germination percentage] _____(3)

(a) (b) (c) (d)

Fig.12 Germination emergence determination

(a) Gray-scale Image (b) binary Image (c) Roundness detection (d) Elongation length

Performance Evaluation

The performance of our proposed quality analysis technique on the India Pulse seed would be analyzed in this section. The seed area of set of samples is observed for the vigorness assessment is evaluated in time using the parameters roundness, sphericity and geometrical diameter which might increase or stable in certain time period and listed in the Table 1.

Pulse Seed Sample	6h			12h			24h		
	Roundness	Spericity	Geometrical Diameter	Roundness	Spericity	Geometrical Diameter	Roundness	Spericity	Geometrical Diameter
vigna mungo	1.16	0.5028	9.5799	2.16	0.7841	10.8802	2.38	0.8461	11.2621
vigna mungo	1.12	0.5289	9.6608	2.12	0.7244	10.3311	3.33	0.8767	10.5126
vigna Radiata	1.06	0.5138	9.3303	2.18	0.7021	11.5727	2.25	0.8024	11.6167
vigna radiata	1.07	0.5011	7.4562	2.57	0.7652	10.9335	2.96	0.8203	11.1587
vigna unguiculata	1.32	0.5023	8.3233	2.34	0.7283	10.2377	2.72	0.8814	10.6326
vigna unguiculata	1.38	0.5038	9.2462	2.05	0.7894	10.6696	2.34	0.8484	10.8348
cajanus cajan	1.27	0.5126	9.2586	2.92	0.7167	10.6281	3.12	0.8211	10.8105
cajanus cajan	1.19	0.5114	9.8819	2.67	0.8967	10.8799	3.01	0.8085	11.1458
vigna mungo	1.33	0.5226	9.9124	2.92	0.7424	10.4579	3.34	0.8235	10.8919
vigna mungo	1.35	0.5342	9.3216	2.05	0.7425	10.1527	2.82	0.8891	10.3592

Table.1 Vigorness assessment geometrical parameter

The Table 2 shows the RGB color range for the thermal image of the pulse seed. This table illustrates that the color region in the thermal image would be within the range of this table to determine the vigorness. From this Table, the blue color threshold value is obtained to evaluate the moisture ratio.

Color	ColorRange
Red	0-124
Green	3-218
Blue	0-135

Table 2. RGB Color pixel range from thermal Image

The Table 3 illustrates the moisture level in % for the quality seed and unquality seed. The high-quality seed must acquire 7% to 9% of moisture content, but in low-quality seed the moisture content might be less than 7% will be declined.

Pulse Seed quality	Moisture level in %
Viable Seed	7%-9%
Non-Viable Seed	<7%

Table.3. Moisture level ratio for quality determination of pulse seed

The table 4 depicts that RGB pixel range in thermal image for the dry seed image samples to determine the water uptake level. The table shows the RGB pixel range for the thermal images for the dry seed and soaked seed is evaluated. The RGB pixel range of the soaked seed would be evaluated for the specific time interval. This would depict the moisture content uptake of the seed samples. The blue pixel range for dry seed lays between 16 and 22 shows the moisture ratio of unqualified seed and also blue pixel range lies between 118 and 124 shows the moisture ratio of qualified seed .

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Condition	Image	Red Pixel			Green pixel			Blue pixel			Moisture	Seed
Dry Seed	1	198			186			18			<4	NV
	2	205			177			22			<5	NV
	3	187			187			17			<3	NV
	4	180			187			16			<3	NV
	5	184			174			17			<3	NV
Pulse Seed Condition	Image	Red pixel			Green pixel Range			Blue pixel Range For			Moisture Ratio	Seed
		6h	12	24h	6h	12h	24h	6h	12h	24h		
Soaked Seed	1	20	19	196	154	186	177	80	112	124	>9	V
	2	20	18	188	157	177	156	102	108	120	>7	V
	3	20	18	179	152	155	160	90	115	123	>8	V
	4	21	21	198	157	140	135	65	102	122	>7	V
	5	21	21	207	157	154	135	87	107	118	>8	V

V-Viable Seed, NV – Non Viable Seed

Table.4. RGB pixel range for dry and wet seed over time interval

This table 5defines the final germination rate of the seed image. The variety of the sample images of the pulse seed is analyzed and the germination growth rate is calculated in % .In this table table5 seed image in lot D is the most vigorness and seed image in lot E is the least vigorness as they have the high and low vigor index respectively.

Seed Lot	*Germination Rate% Germination Rate*	*Seedling Length Seedlength*	*Vigor Index*	*Vigor status Vigor*
A	***79***	***85***	***6715***	***High-vigor***
B	***48***	***56***	***2688***	***Low- vigor***
C	***86***	***74***	***6364***	***High- vigor***
D	***88***	***91***	***8008***	***High- vigor***
E	***55***	***45***	***2475***	***Low- vigor***

Table.5.Germination emergence evaluation

V COMPARITIVE ANALYSIS

In this section, our proposed IPSQD-2(Indian Pulse Seed Quality Determination-2) method is compared against existing method IPSQD-1(Indian Pulse Seed Quality Determination-1) in terms of Performance Accuracy ratio, Regocnition Rate and Reliability Rate for quality analysis of Indian Pulse Seed.

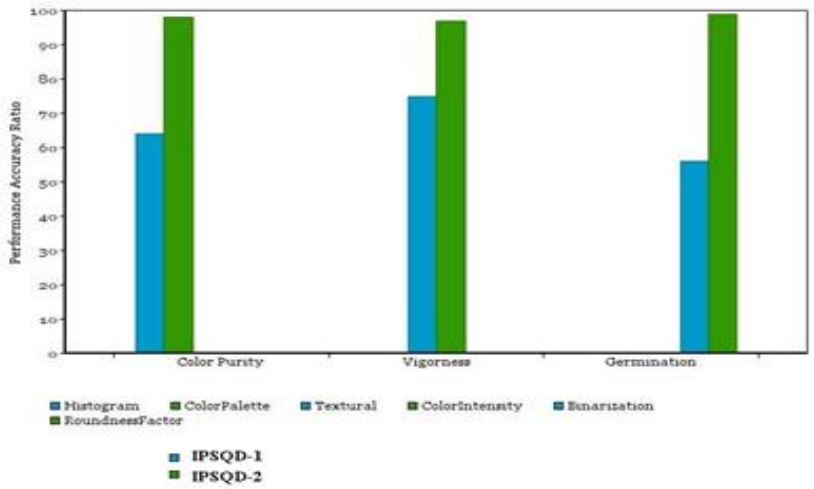


Fig.13. Performance accuracy ratio for Quality seed analysis

The performance accuracy ratio is measured in the Fig 13. The quality parameters (Color Purity, Vigorness and Germination) of pulse seed is evaluated and compared with existing IPSQD-1 methodologies to measure the Performance Accuracy Ratio. The quality parameters of the proposed IPSQD-2 methods provide accurate results with less-time consumption when compared with the existing methods. The HCT methods have some lack in detecting the

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

quality seeds and unquality seeds in accurate manner. Thus this system, evaluates the quality parameters with high-quality approach to classify the defeated and healthy seeds.

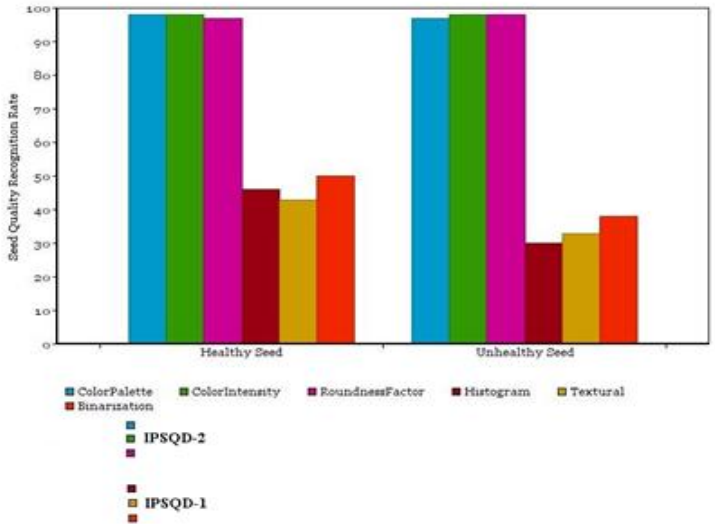


Fig.14. Recognition rate for quality analysis methods

In Fig 14 the recognition rate for the quality of the seed is shown. The recognition rate is based on the time and accurate ratio. The proposed methods would provide high-recognition rate for both healthy and defeated seeds from the image set when compared with the existing methods.

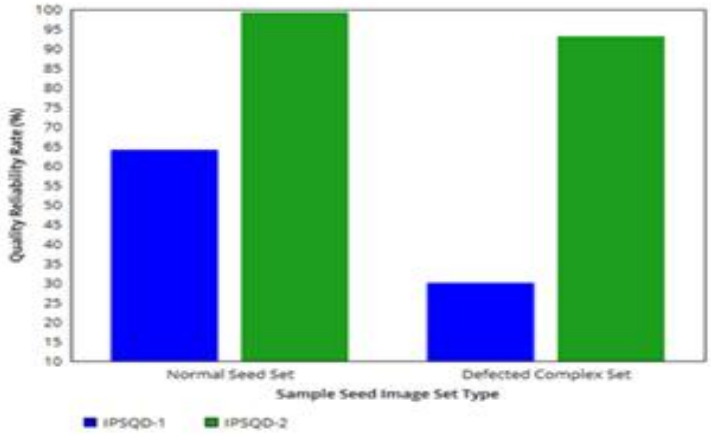


Fig.15 Reliability rate for quality analysis methods

The reliability ratio for the quality aspects for the different image set type is shown in the Fig 15. The reliability ratio is measured based on the outcome of the quality aspects for both normal image set and defeated complex set. The existing methods would lack in providing healthier outcome for the complex set. Thus these proposed methods would provide enhanced and reliable results based on both normal and complex image set.

Figure 16 shows the quality recognition rate for various quality aspects such as color purity, vigorness and germination emergence with the sample pulse seed varieties that gives sensibly high detection rate.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

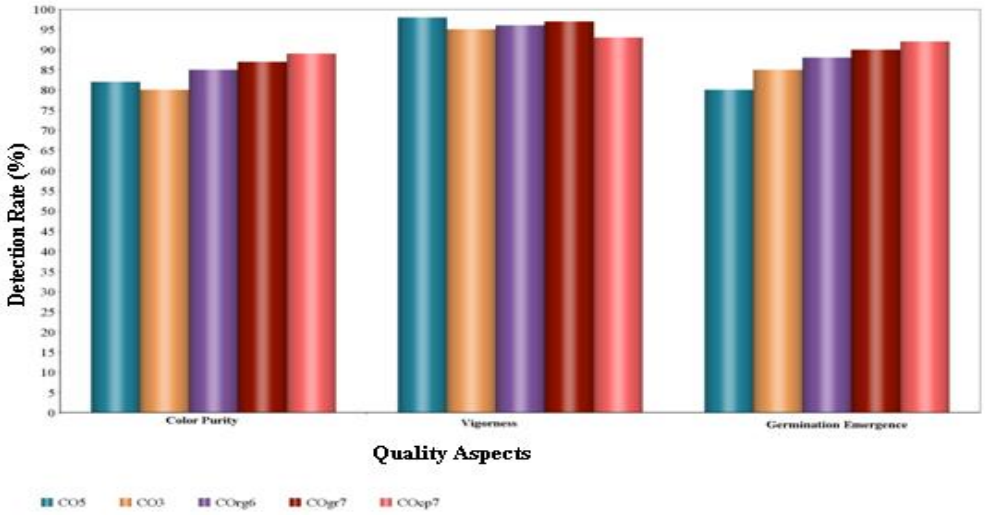


Fig.16. Detection Rates of the Quality Aspects

In this phase results for Indian pulse seed quality determination is discussed. The results discussed for the quality aspects color purity, germination rate and vigorness. The results indicate that the color purity analysis of proposed system can increase the determination of healthy and unhealthy seeds by more than 90%. The error rate for determination is decreased to 5% from the manual color and other existing techniques for color prediction.

In the seed germination the precision rate various from varieties of seeds are 89%, 78%, 90%, 95% and 88% for CO3, CO5, COcp7, COrg6 and COgr7 respectively. The average access time was 3.41 seconds per Indian pulse seed image. The moisture characteristics of Indian pulse seed is observed through our proposed Method. The seed moisture content for moisture pulse varieties varies from 5.2% to 7%. In the second case moisture content for dry seed is varied from 2% and 3%. The observation of moisture ratio and seedling rate is increased from 70% to more than 90% through our algorithm.

The overall result is observed that the quality ratio, reliability ratio increased from 80% to more than 93% from existing and manual identification and recognition system of Indian Pulse seeds.

Conclusion

Image analysis, with its capability to imitate individual intellect in behavior visual information, is an essential technology that will find many requests in contemporary varietal recognition and seed qualifications. This paper describes the method of different machine vision techniques for pulse seed quality evaluation. It illustrates how the imaging technology is useful in monitoring seed imbibitions, germination activities and determination of seed size, shape constraints. Recently, the furthermost efforts have paying attention on constructing non-destructive methods with ability of computer hardware of image processing and its combination with prohibited ecological situation systems. The main advantage of this proposed method is faster, cheaper, and easier to be performed and less labor-intensive and also offered good accuracy ratio, reliability rate about 94%. In addition, it is also operator-independent, for physical visual examination and organization were eradicated. So it is an ideal alternate for the conventional manual quality determination method for the evaluation of pulse seed in technical research and propagation programs.

References

[1]. Prof. V.B.Raskar, Swapnil Thorat, Praviin Mane, Eeshwar Tak, “Rice Grading Quality Analysis for Agmark Standards”, 2016
[2]. Amit Bhande, Dr.S.V.Rode, “Quality Identification of Wheat by Using Image Processing”, 2016
[3]. Wan Putri N. W. M. Tahir, Norhaida Hussin, “Rice Grading Using Image Processing”, 2015
[4]. Sunil S. Telang, Prof. Sandip Buradkar, “Review Paper on Analysis and Grading of Food, Grains Using Image Processing and SVM”, 2015
[5]. Vidya Patil, V. S. Malemath, “Quality Analysis and Grading of Rice Grain Images”, 2015
[6]. Ece Olcay Güneş ; Sercan Aygün ; Mürvet Kırcı ; Amir Kalateh ; Yüksel Çakır,” Determination of the varieties and characteristics of wheat seeds grown in Turkey using image processing techniques”, IEEE, Aug.2014
[7]. B. Lurstwut and C. Pornpanomchai,” Application of Image Processing and Computer Vision on Rice Seed Germination Analysis”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 9 (2016) pp 6800-6807.
[8]. Sontisuk Teerachaichayut and Wipawee Yokswad, Anupun Terdwongworakul,” Application of Image Analysis for Determination of Mangosteen Density”, Journal of Advanced Agricultural Technologies Vol. 2, No. 2, December 2015
[9]. Sheetal Mahajan, Sukhvir Kaur, “Quality Analysis of Indian Basmati Rice Grains using Top-Hat Transformation”, 2014
[10]. Megha R. Siddagangappa , Asso.Prof. A. H. Kulkarni , “Classification and Quality Analysis of Food Grains”, 2014

[11]. Y. Sako, M. B. McDonald, K. Fujimura, A. F. Evans, and M. A. Bennett 'A System for Automated Seed Vigor Assessment', 2001.

[12]. A.L. Hoffmaster, K. Fujimura, M.B. McDonald, M.A. Bennett, 'An Automated System for Vigor Testing Three-Day-Old Soybean Seedlings', 2003.

[13]. K. Oakley, S. T. Kester, R.L. Geneve, 'Computer-aided Digital Image Analysis of Seedling Size and Growth Rate for Assessing Seed Vigor in Impatiens', Seed Sci & Technology, 32, 907-915, 2004

[14]. Jagdeep Singh Aulakh, Dr. V.K. Banga: grading of rice grains by image processing. International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 4, pp 1-4, June – 2012 ISSN: 2278-0181.

[15]. Chetna V. Maheshwari,[3] q-curve approach for quality analysis of indian oryza sativa ssp indica(rice), International Journal of Advanced Technology in Engineering and Science, Volume No.01, Issue No. 03, March 2013.

Role of Feedback Analytics Recommender in organising Workshop

[1]Meenatchi V.T, [2]Thangaraj M, [3]Gnanambal S, [4]Gayathri V
[1]Department of Computer Application & Information Technology, Thiagarajar College, Madurai.
[2]Department of Computer Science, Madurai Kamaraj University, Madurai.
[3]Department of Computer Science, Raja Doraisingam College, Sivaganga.
[4]Department of Computer Application, NIT, Tiruchi

ABSTRACT *Collecting Feedback from the participants is now a days common. Extensive analysis from the feedback, always pave way to effective analytical report. Through the report, decisions are made precisely. The proposed work implements a feedback suggester framework, by which some recommendations are suggested for future improvements in conducting workshop.*
Keywords: *Feedback, Analytics, Decision, Suggester, Workshop, Recommendations*

Introduction

Feedback Analytics is an upcoming concept. It is becoming vital since, people always prefer improvement in their work done in their workplace and learning environment. Feedback provides the ability to capture the satisfaction of the workshop participants or the users and in turn facilitate and improvise the existing system. Normally, the goal of feedback analytics is to find the assessment of the content, collect more suggestions and ideas for future improvement. This Feedback Analytics can be carried out effectively through Waikato Environment for Knowledge Analysis (WEKA), an open source data mining tool.

We have undergone our study through five main parts, in which part 1 introduces the subject matter, while part 2 discusses the various issues in feedback analytics and part 3 deals with the proposed framework. Part 4 presents the Experimental study and Results and finally Part 5 records the conclusion.

Related Work

In [1] the better work place through the analysis of strengths and weaknesses of the employees were focused and novel approaches to competency analytics based on event-driven, continuous peer feedback were devised. The work in [2], assess the feedback in student learning development. It discusses about drawback of the traditional technology which is used in universities, Gradebook which does not allow the tracking of student engagement with their feedback and then digital foot printing data was applied to gather detailed information on students' access and engagement.

In [3] lab sessions characterization has been carried out which integrated both qualitative and quantitative parameters. A supporting tool was proposed which used the students' logs, learning analytics and visualization techniques for monitoring and awareness mechanisms that leverages the detected problems and thus improve the learning and assessment processes.

The characteristics of learning are shared by the combination of Learning Analytics (LA), Formative Assessment (FA) and immediacy are facilitated through the mobility of learners. This paper [4] discusses the analytical application called Quiz My Class Understanding (QMCU) investigated the significance of the combination between LA and FA techniques. The conducted case study assures that the utilization of QMCU student's centered mobile dashboard increases the student's engagement.

In [5], presents a real example of application of learning analytics framework which supports a real case of practical learning in a 3D virtual environment by analyzing the in depth the problems that arises there from. As a result, the automatization of a learning process inside a virtual world with the subject of quality assurance in pharmacy laboratories were the outcome of the system.

This paper [6] mainly deals with feedback analysis, sentiment analysis and word-count. The paper enforces the feedback as an important criteria, it proposes that the feedback, not only when it highlights weaknesses but also for strengths. It also suggests that analysis of feedback when done in wrong way, then the result of analysis will also be wrong. The feedback analysis system was done using Map-Reduce framework for processing large data set.

In [7]**,** an automated analytics system monitors the students attending online lectures from a remote location and provides feedback to the teacher. The study conducted three different experiments with sessions from online courses and observations concluded that the automated system efficiently differentiates the attention level of students and helped the teachers to improve their style of instruction.

In [8], Feedback Evaluation is necessary for any institute to maintain and monitor the academic quality of the system. An automatic evaluation system based on sentiment analysis, was proposed, in which feedback is collected in the form of running text and sentiment analysis is performed with supervised and semi supervised machine learning techniques.

In [9], Feedback analysis is a valuable tool for analyzing multistage feedback amplifiers. Students, however, often find this topic confusing and difficult to learn. Reasons for this difficulty and confusion include applying feedback analysis to inappropriate circuits and failure to explicitly state and check the basic assumptions of feedback analysis. Another source of confusion is failure to keep the source and the load separate from the basic amplifier making the

feedback model less physical. In this paper, these pitfalls are discussed and a straightforward methodology for analyzing feedback amplifiers is presented. Several examples of feedback analysis are presented.

Proposed Work

The proposed framework is coined as Feedback suggester framework and is shown in the fig 1. The framework contain four main phases namely,

a) Input phase
b) Classification phase
c) Knowledge extraction phase
d) Report generator phase

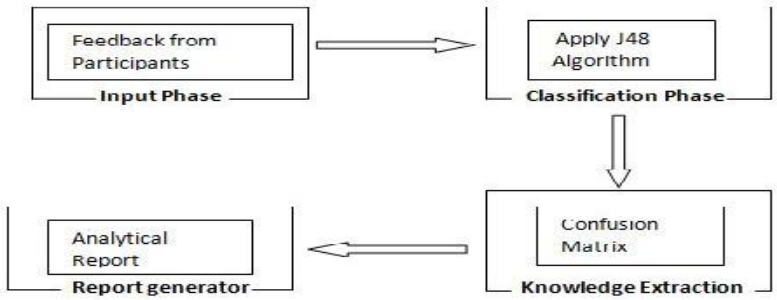


Fig 1: Feedback suggester framework

The details of the phases are,

a) Input phase

The study was carried out after a state level workshop in a reputed educational institution. The workshop was conducted in 2 sessions. Session I was technical lecture from an eminent person. Session II was Hands-on session in a data mining tool. Both the sessions were very much interactive. The participants were given a feedback form. Feedback was collected after two sessions were completed.

The feedback attributes taken into account were, i) Commendable ii) Presentable iii) Interesting and Informative iv)Class. These attributes are used to assess the information related to Content Delivery, Languageof communication and Hands – on Applicability of the sessions.

b) Classification phase

This phase applies J48 classification algorithm in the feedback data for classifying the participants based on their comments. J48 is a decision tree based algorithm which is a WEKA implementation of C4.5 algorithm[10]. The participant's comments were aggregated as Outstanding, Excellent, Good and Satisfactory. Confusion matrix is generated as the outcome of this phase and is shown in fig 2 and fig 4.

c) Knowledge extraction phase

From the classification results derived, this phase mainly identifies, which class label has scored well. The results are visualized as bar chart as shown in fig 3 and fig 5.

d) Report generator phase

From the Visualization of graphs, we can come to conclude that how many participants mainly focus on technical lecture and how many are more interested in Hands-on. Reports are generated as Data table as shown in Table II and Table III.

Pseudo code for the proposed framework is depicted in Table I,

Step1: Input collected from P_n with a_i

Step2: Apply J48 algorithm ∀ ai of Pn

Step3: for each ai, classification analysis made through C_m

Step 4: K_n derived from C_m

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

The variables P_n denote the n number of participants feedback, a_i denote the feedback attributes ranging from 1 to 4, C_m represents the Confusion matrix and K_n derived through C_m .

Experimental Results

The feedback suggester framework has been implemented with WEKA 3.7.5. The set of experimental study explored the feedback form entries, collected from the workshop participants with several attributes on performance.

Session I: Technical Lecture

The outcome of the analysis for this session is got in the form of confusion matrix and the interpretation is given below.

Table II: Data table for Technical Lecture

Number of participants	Comments
20	Outstanding (O)
2	Excellent (E)
1	Good (G)
4	Satisfactory (S)

```
=== Confusion Matrix ===

 a  b  c  d   <-- classified as
20  0  0  0 |  a = O
 2  0  0  0 |  b = E
 0  0  4  0 |  c = S
 0  0  1  0 |  d = G
```

Fig 2 Confusion Matrix for Session I: Technical Lecture.

Confusion matrix is a powerful tool used by the classifier for analyzing the instances of different classes. For m classes, confusion matrix is a table of size at least m×m. The correctly classified instances are represented in diagonals.

This table is a comparison of positive instances versus negative instances. True positive rate is determined by the number of instances that are correctly classified whereas false positives are represented by incorrectly classified instances. These measures are used for analyzing the ability of the classifier.

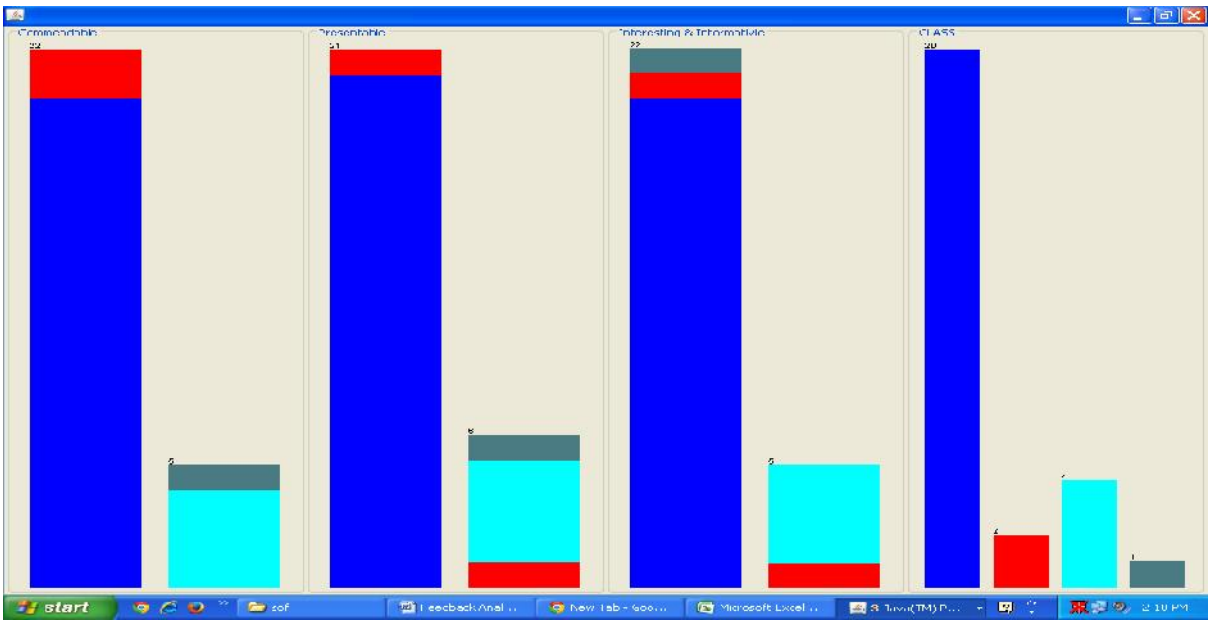


Fig 3: Interpretation of Technical Lecture

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

The bar chart shows the data distribution of instances which is related to the technical lecture session. The colors in the chart are represented as blue for outstanding, red for excellent, green for good and gray for satisfactory. The chart concludes that the feedback analysis results more in outstanding category.

Session II: Hands-on

The outcome of the analysis for this session is got in the form of confusion matrix and the interpretation is given below.

Table III: Data table for Hands-on

Number of participants	**Comments**
16	Outstanding
5	Excellent
4	Good
2	Satisfactory

```
=== Confusion Matrix ===

 a  b  c  d   <-- classified as
16  0  0  0 |  a = O
 0  4  0  0 |  b = G
 0  2  0  0 |  c = S
 5  0  0  0 |  d = E
```

Fig 4: Confusion Matrix for Session II: Hands-On

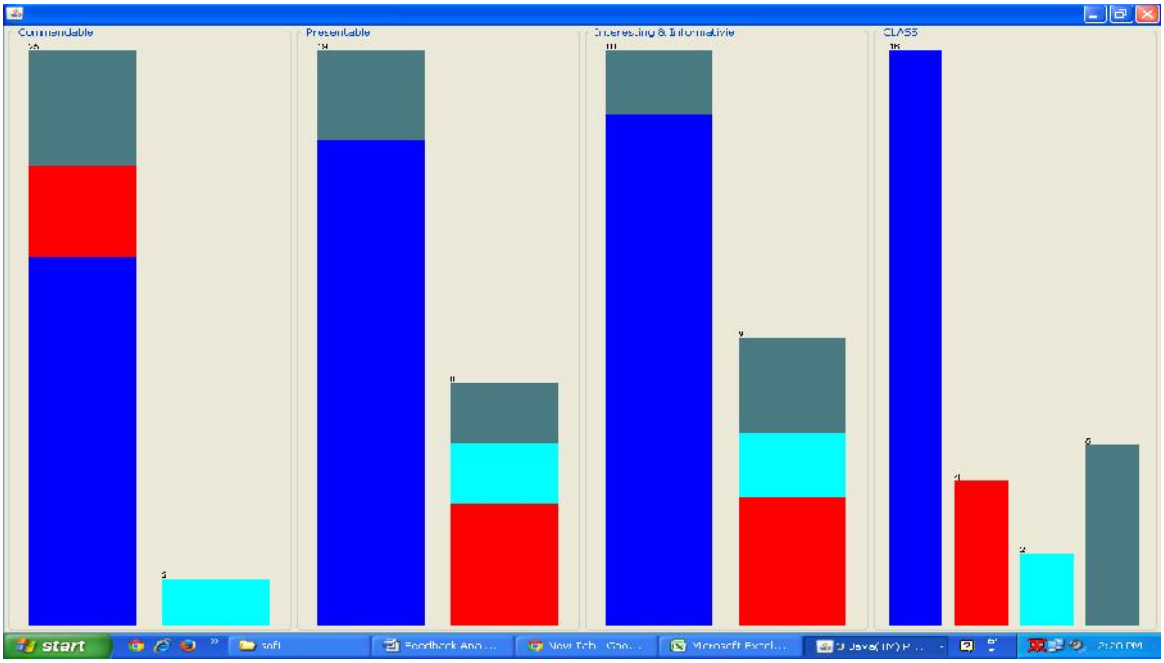


Fig 5: Interpretation of Hands-on

Conclusion

Feedback analytics is very much needed for future planning of organizing workshop. Thus the proposed framework suggests the organizers to have improvements in prepared plans. This system is also utilized by the

Tutors/Trainers to improve their skills related to content delivery and language adaptation. In future the developed framework will be extended to analyze the behavior/perception of workshop participants.

REFERENCES

[1] Eddie Walsh, Mirjam Neelen and Evangelos Kapros, "Competency Analytics in the Workplace through Continuous Peer Feedback", IEEE 17th International Conference on Advanced Learning Technologies (ICALT), 2017, doi: 10.1109/ICALT.2017.150

[2] Mireilla Bikanga Ada and Mark Stansfield, "The Potential of Learning Analytics in Understanding Students' Engagement with Their Assessment Feedback", IEEE 17th International Conference on Advanced Learning Technologies (ICALT), 2017 ,doi: 10.1109/ICALT.2017.40

[3] Israel Gutiérrez Rojas; Raquel M. Crespo García, "Towards Efficient Provision of Feedback Supported by Learning Analytics", IEEE 12th International Conference on , doi: 10.1109/ICALT.2012.171

[4] Naif Radi Aljohani, Hugh Davis,"Learning Analytics and Formative Assessment to Provide ImmediateDetailed Fee dback Using a Student Centered Mobile Dashboard",Seventh International Conference on Next Generation Mobile Apps, Services and Technologies,2013. Czech Republic. doi**:** 10.1109/NGMAST.2013.54

[5] Juan Cruz-Benito, Roberto Therón, Francisco J. García-Peñalvo, Cristina Maderuelo, Jonás Samuel Pérez-Blanco, Hinojal Zazo and Ana Martín-Suárez, "Monitoring and feedback of learning processes in virtual worlds throughanalytics archite ctures: A real case", 9 th Iberian Conference on Information Systems and Technologies (CISTI), Spain, 18-21 June 2014. doi**:** 10.1109/CISTI.2014.6877097

[6] Kusum Yadav, Manjusha Pandey and Siddharth Swarup Rautaray," A proposed framework for feedback analysis system using big data tools", 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I- SMAC), India. 10-11 Feb 2017. doi: 10.1109/I- SMAC.2017.8058238

[7] Divya Dinesh, Athi Narayanan S and Kamal Bijlani, "Student analytics for productive teaching/learning", 2016 International Conference on Information Science (ICIS), Kochi, India. 12-13 Aug. 2016.

[8] Alok Kumar and Renu Jain," Sentiment analysis and Feedback Evaluation", IEEE 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), Amritsar, India. 11 January 2016. Doi: 10.1109/MITE.2015.7375359.

[9] D. R. Hertling and J. A. Connelly," Effective teaching of feedback analysis",Mixed- Signal Design, SSMSD '99. Southwest Symposium on Tucson, AZ, USA, USA. 11-13 April 1999. Doi: 10.1109/SSMSD.1999.768596

[10] M.Thangaraj and S.Gnanambal,"Rule based Decision Support System for Aiding Vitamin D Deficiency Management' on *Indian Journal of Science and Technology*, Vol 7(1), 48–52, January 2014. ISSN: 0974-6846.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Web Services and Tools for Real Time Analytics

[1] Aruna Devi .P, [2] Chamundeeswari .M
[1]Department of Computer Science (SF), V.V.Vanniaperumal College for Women, Virudhunagar
[2]Department of Computer Science (aided), V.V.Vanniaperumal College for Women, Virudhunagar

ABSTRACT *The most valuable currency in idea economy is the huge amount of data that is made easily available around us. The Success of the organization could be achieved if they use these huge data for predictive planning that brings big decision into their market which helps them to reduce the risk. This lead to the evolution of Big data analytics. Big data refers not only to these enormous amounts of data but also data with high velocity and variety, which is difficult to handle using traditional tools and techniques. The possible solution to handle the large volume and velocity of data is to migrate to cloud for analytics. Cloud computing is the service that is provided over internet from the data centres that is on the cloud to the consumers on demand basis. One of the well known and popular cloud service providers is the Amazon Web Services. Big data analytics on cloud helps in determining valuable decisions by considering the data patterns and the relationship between them. It plays a successful role in performing meaningful real-time analysis on huge volume of data. This paper aims to analyze tools that can be applied to big data on cloud which helps the organisation to afford minimum cost on compute and storage, as well as the success rate obtained due to bringing this valuable decision fast into the market.*
Keywords – *Content Delivery Network, Domain Name System, Streaming Data, Glacier, Web Services.*

I. Introduction

Big Data and Analytics is a combined hardware/software architecture that congregates vast quantities of distinct data for speedy analysis. With the increase in storage capabilities and methods of data collection, large amounts of data have become effortlessly available. Every second, more and more data is being produced and needs to be stored and analyzed in order to extract value. The analysis produces meaningful information. The result is that you can make earlier and better-informed business decisions. The benefit is competitive advantage and (hopefully) increased profitability for required operation [1]. There are structural failures throughout history due to the traditional approach. So we are in need of an infrastructure that helps us in fast processing and storage. This need could be fulfilled with the help of cloud computing. Cloud Computing is virtualized compute power and storage delivered via platform-agnostic infrastructures of abstracted hardware and software accessed over the Internet. These shared, on-demand IT resources, are created and disposed of efficiently, are dynamically scalable through a variety of programmatic interfaces and are billed variably based on measurable usage [2]. The main power of cloud computing lies in the way data is stored, how it is transmitted and accessed. A virtualized platform with management capabilities like availability, automated load balancing and fault tolerance reduces infrastructure cost and maintenance cost [6]. There are lots of web services available today for data analytics. IBM, Microsoft, Amazon, Google and adobe. All provide their own services for analytics.

This paper focuses on AWS that provides on demand cloud computing capabilities to organization and customers on pay as you use basis. This technology is available to customers through internet to have full fledged virtual cluster of computers for their requirement. The customers can login and configure their virtual system just like their own physical systems. Based on the needs of the customers the customer can use as many number of services from AWS. Services provided by them are managed services where the user need not worry about deploying and managing the infrastructure, it also provides industry standard security to each user system. It operates globally among many geographic regions.

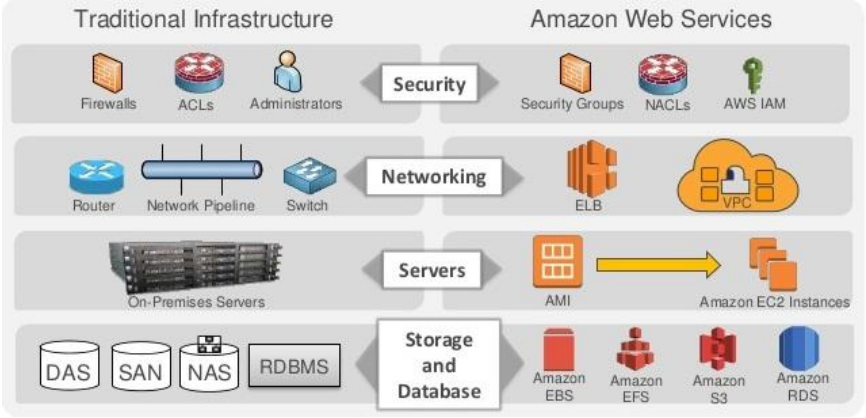


Fig. 1. Traditional Vs AWS Infrastructure and Services

We are in need of right tool for big data analytics which helps the organization to stay on top in the market. We need some proper infrastructure with the hardware in place and this could be available with the emergence of AWS Cloud Services which provides a lot of managed service. Fig. 1 shows the comparisons of the traditional and AWS

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

infrastructure which helps to handle enormous amount of data with high performance, processing and storage power. It helps the organization to save upfront hardware cost with the cost associated on demand basis. It also helps the organization by providing managed service where lot of care is taken by the provider itself. Let us look at the architecture that will be needed for the big data analytics on cloud.

In this paper we focus on the factors that motivate us to move the analytics on cloud. We also explore various tools and methods that the Amazon Web Services provide for the profit of the organization. The infrastructure of the cloud provides the organization with scalable, flexible and high resource utilization with low capital investment. The cloud also provides them pay as you use which is much favourable for the organization. The rest of the paper is organized as follows: Section II describes the infrastructure needed for big data, Section III explores the characteristics of big data and the tools that support their usage, Section IV summarizes the steps in big data analytics, Section V says the tools for real time big data analytics, Section VI describes the benefits of the analytics on cloud, Section VII says the case study of organization that uses cloud services for analytics, finally Section VIII concludes the paper

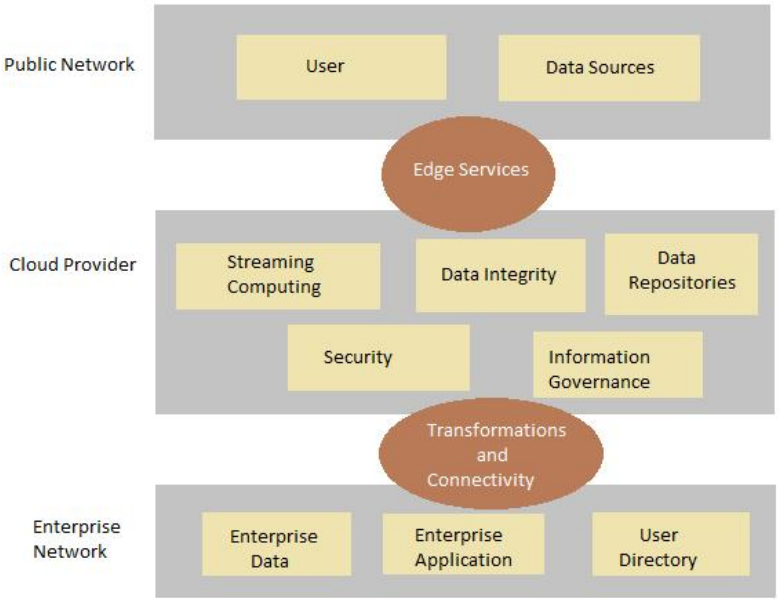


Fig. 2. Cloud Architecture for Big Data Analytics

The public network is injecting large volume and velocity of data that is needed for analytics to provide the knowledge behind it. The analysis of data for decision making is done on cloud. Edge service is the way by which the data flows safely to the processing system. It is done with the help of Domain Name System, Content Delivery Network, Firewall and Load Balancer. The cloud provider is responsible for providing managed services for streaming, computing, storing and information governance. Transformation and Connectivity enables the enterprise with the secure connection.

AWS provide the organisation with the highly scalable, secure Big Data applications that is fast and do not require hardware to procure and no infrastructure to maintain. It provides the necessary services that we need for data collection, storage, processing, analyzing, and visualizing big data on the cloud.

II. Importance of Infrastructure to Big Data

With the technology evolution and the augmented multitudes of data flowing in and out of organizations daily, there has become a need for faster and more effective ways of analyzing such data. There are three major reasons for the companies to invest in Big Data, and those reasons correlate directly for the need of proper infrastructure [5].

1. **Access**: The visibility of customers and their behaviours can create better interactions and emerge a need to develop new products and services.
2. **Speed**: There is a need to process huge data, as fast as possible, which have as many insights in real time applications.
3. **Availability**: It is necessary to have 99.9 percent uptime (or better) so we can quickly make decisions or correct errors when needed without any inconvenience to the customers.

This leads us to make decision on how to handle all of our data with proper infrastructure. Once if the architecture is right with its hardware it is easy to handle the large volumes of data which moves us forward with our Big Data strategy. Today's infrastructure has come remarkably far from the past, and we should look for a solution that's designed to handle high-performance loads.

There arise a need for new tools and methods for an organization with terabytes of data, it has also presented an issue of managing this large volume of data. Analyzing these large volumes of data often becomes a difficult task as well.

The solution in today's fast revolving world is to use the cloud service. It was the emergence of cloud computing which made it easier to provide the best of technology in the most cost-effective packages. Cloud computing not only reduced costs, but also made a wide array of applications available to the smaller organisations for their big data analytics.

III. Characteristics of Big Data & Tools on Cloud

The main characteristic that makes data "big" is the sheer **volume.** It makes no sense to focus on minimum storage units because the total amount of information is growing exponentially every year. In 2010, Thomson Reuters estimated that the world was "awash with over 800 exabytes of data and growing." For that same year, EMC, a hardware company that makes data storage devices, thought it was closer to 900 exabytes and would grow by 50 percent every year. No one really knows how much new data is being generated, but the amount of information being collected is huge. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set [3].There are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time [4].This large enormous volume of data could be easily managed in Cloud with the help of managed services of Amazon S3 Bucket, Amazon EBS, Streaming large volume of real time data is done with the help of Kinesis Firehose, Kinesis Stream. Amazon Kinesis [8] is a cloud-based service for real-time data processing over large, distributed data streams. It streams data in real time with the ability to process thousands of data streams on a per-second basis. The service, designed for real-time apps, allows developers to pull any amount of data, from any number of sources, scaling up or down as needed.

Object Lifecycle Management helps to maintain the large volume of data in an easy way. Lifecycle configuration enables you to specify the lifecycle management of objects in a bucket. The configuration is a set of one or more rules, where each rule defines an action for Amazon S3 to apply to a group of objects. These actions can be classified as follows [6]:

- **Transition actions** – In which objects transition is made to another storage class. For example, you may choose an object transition as STANDARD_IA (IA, for infrequent access) storage class 30 days after creation, or archive objects to the GLACIER storage class one year after creation. This helps to keep the long term historical archived data to be stored on the glacier.
- **Expiration actions** – In which you specify when the objects expire. Then Amazon S3 deletes the expired objects on your behalf.

Variety is one of the most interesting developments in technology as more and more data is digitized. Traditional data types are structured data which include data on a bank statement like date, amount, and time. These data fit neatly in a relational database. Structured data is augmented by unstructured data, the data available on Twitter feeds, audio files, MRI images, web pages, web logs which can be captured and stored but doesn't have a meta model (a set of rules to frame a concept or idea — it defines a class of information and how to express it) that neatly defines it. Unstructured data is a fundamental concept in big data. The best way to understand unstructured data is by comparing it to structured data. Think of structured data as data that is well defined with the set of rules. For example, money will always be numbers and have at least two decimal points, names are expressed as text, and dates follow a specific pattern. On the other hand, in unstructured data, there are no rules. A picture, a voice recording and a tweet all can be different but express ideas and thoughts based on human understanding. One of the goals of big data is to use technology to take this unstructured data and make sense of it [5].

Velocity is the frequency of incoming data that needs to be processed. Think about the data on SMS messages, Face book status updates, or credit card swipes that are being sent on a particular telecom carrier every minute of every day. These media's are good examples that have a good appreciation of velocity. The data is massively and rapidly-expanding, but it's also noisy, messy, constantly-changing, hundreds of formats and virtually worthless without analysis and visualisation. Streaming application like Amazon Web Services, Amazon EMR, Elastic search Service, Kinesis & Lambda are examples of applications that can handle these velocity of data.

Variability refers to constantly changing data. This is particularly the case when gathering data relies on language processing. Words don't have static definitions, and their meaning can vary widely in context which is the challenge for the big data analytics. Companies have to develop sophisticated programmes which can understand the context and decode the precise meaning of words from it.

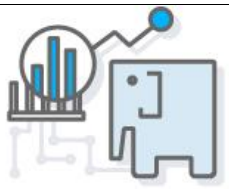
Verasity refers to the widespread agreement about the potential value of Big Data; the data is virtually worthless if it's not accurate. This is particularly true in programmes that involve automated decision-making, or feeding the data into an unsupervised machine learning algorithm. The results of such programmes are only as good as the data they're working with. Sean Owen, Senior Director of Data Science at CloudEra, expanded upon this: 'Let's say that, in theory, you have customer behaviour data and want to predict purchase intent. In practice what you have are log files in four formats from six systems, some incomplete, with noise and errors. These have to be copied, translated and unified. Big Data is the messy, noisy nature of it, and the amount of work that goes in to producing an accurate dataset before analysis can even begin.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Visualization Once it's been processed, you need a way of presenting the data in a manner that's readable and accessible this is where visualization comes in. Visualization can contain dozens of variables and parameter to present this information that makes the observation clear. It is one of the challenges of Big Data.

Let us now have a quick look of the services that the Amazon provide to build a powerful services to process, analyze, visualize data easily and cost effectively. These services help to concentrate on the V's of Big data Analytics with much care by providing much managed services.

Table I. Analytics Services Provided By AWS

S.No	Tool Name	Description	LOGO
I.	**Storage for Big Data**		
1.	Amazon S3	Highly Reliable, Secure, And Scalable Object Storage For All Your Data	
2.	Amazon Aurora	It is a relational Database engine that combines the speed and reliability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases.	
II.	**Computing for Big Data**		
1.	Amazon EMR	Fully managed Hadoop framework in minutes. Scale your Hadoop cluster dynamically and pay only for what you use.	
2.	Amazon Athena	Easily analyze data in Amazon S3 using SQL. We can start analyzing data immediately with no need load your data into Athena, it works directly with data stored in S3.	
3.	Elastic Search Service	Makes it easy to deploy, operate, and scale Elasticsearch.	
4.	Kinesis	Easiest way to work with streaming data.	
III.	**Datawarehousing**		
1.	Amazon Redshift	Configure and deploy a data warehouse within minutes. It handles all the work needed to manage, monitor and scale the data store.	

IV. Steps in Big Data Analytics

- **Collect**. Collecting the raw data like transactions, logs, mobile devices and more which is the first challenge that many organizations face when dealing with big data. A good big data platform makes these step easier, allowing developers to ingest a wide variety of data that can vary from structured to unstructured that occurs at any speed and vary from real-time to batch.
- **Store**. Any big data platform needs a secure, scalable, and durable repository to store data prior or even after processing tasks. Depending on your specific requirements, you may also need temporary stores for data in-transit.
- **Process & Analyze**. This is the step where data is transformed from its raw state into a consumable format which usually by means of sorting, aggregating, joining and even performing more advanced functions and algorithms. The resulting data sets are then stored for further processing or made available for consumption via business intelligence and data visualization tools.
- **Consume & Visualize**. Big data is all about getting high value, actionable insights from our data assets. Ideally, data is made available to stakeholders through self-service business intelligence and agile data visualization

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

tools that allow for fast and easy exploration of datasets. Depending on the type of analytics, end-users may also consume the resulting data in the form of statistical predictions that in the case of predictive analytics or recommended actions in the case of prescriptive analytics [5].

Fig. 3. Steps in Big Data Analytics

V. Real Time Big Data Analytics Tools on Cloud

Amazon Kinesis Firehose

Firehose is fully managed service for delivering real time streaming data to destination buckets. Easily load massive volumes of streaming data into AWS. Firehose buffers incoming streaming data for certain size and certain period of time before delivering to destination. Enable near real-time big data analytics with existing BI tools and dashboards [9].

Amazon Kinesis Streams

Build your own custom applications that process or analyze streaming data. Continuously capture and store terabytes of data per hour. We can use Kinesis Stream for rapid and continuous data intake and aggregation. The type of data may be log data, application logs, social media, market data feeds, and web click stream data. The response time for the data intake and processing is fast in real time. The quicker response is made possible because of the power of parallel processing. Data which is intake from the data stream ensures durability and elasticity. The data can be immediately consumed on data arrival [9].

Amazon Kinesis Analytics

Easily analyze streaming data with standard SQL without having any processing framework. Kinesis Analytics takes care of everything required to run your queries continuously and scales automatically to match our volume and throughput rate of incoming data. The kinesis analytics is capable of ingesting the data automatically and recognize the data format and suggest the suitable schema then execute the stream data with SQL queries and sends the processed results to Amazon S3or Amazon Redshift or custom destination [8].

Process Flow for big data analytics

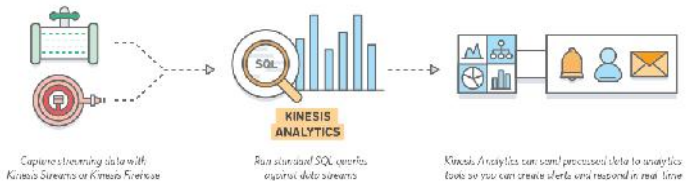


Fig 4. Capturing, Analysing & Visualising Real Time Big Data

VI. Benefits of Big Data Analytics on Cloud

- Powerful Real-time Processing – Enable to analyze and respond in real time with low processing latency.
- Fully managed – Do not require us to provision or manage any infrastructure.
- Automatic Elasticity – Scales up and down the infrastructure as and when needed.
- Easy to use – provides interactive tool to process structured and unstructured data.
- Standard SQL – No need for any complex processing frameworks.
- Pay only for what you use [6].

VII. Organization using Big Data on Cloud

Thomson Reuters uses AWS as the platform for its Product Insight analytics solution, which can process more than 4,000 events per second and scales automatically to accommodate increases in traffic during breaking news. Product Insight captures and processes data using Amazon Kinesis and AWS Lambda, stores it using Amazon S3, and secures it using AWS Key Management Service [10].

Toyota Tsusho uses AWS to quickly scale their data processing of traffic data from over 50 thousand vehicles, while reducing costs up to 35 percent. Toyota Tsusho uses AWS products such as Amazon EC2, Amazon Kinesis, and DynamoDB to process large amounts of data in a scalable and reliable way .

CrowdChat aggregates conversations on the Internet and social media networks and then unifies them according to hash tags so that users can easily find topics of interest. The company uses AWS to run its web application and big data workloads as well as its development and testing environments. By using AWS, CrowdChat can scale its platform to handle traffic spikes during large events and store more than 250 million documents [11].

VIII. Conclusion

In today's fast decision making process we believe that there is significant importance on Big Data Analytics. It could be most perfect to implement the system on cloud so that it provides more insight and benefits for the organization. If the architecture with good compute system is properly exploited and applied on the cloud, it gives the organization free from the burden of purchasing and maintaining the infrastructure. It also helps them with the proper channel of focussing their analytics with the managed services where the hardware, security and scaling of computing is done with the great care by the service provider. The pay for what you use is also a most recognized criterion in which all the organization is satisfied and also come forward to use the cloud services. Finally the Big Data Analytics on Cloud play a major role in fast decision making by using their technical computing environment.

References

[1] *Barry schoenborn , Big Data Analytics Infrastructure, A Weiley Brand, IBM Edition.*
[2] *Suruchee V.Nandgaonkar, Prof. A. B. Raut, A Comprehensive study on cloud computing, IJCSMC, Vol. 3, Issue. 4, April 2014, pg.733 – 738*
[3] *Nada Elgendy& Ahmed elragal , Big Data Analytics A Literature Review Paper, Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings (pp.214-227)*
[4] *Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial In sights, pp. 26–28(2012)*
[5] *https://aws.amazon.com/big-data/what-is-big-data/*
[6] *https://aws.amazon.com/what-is-cloud-computing/*
[7] *T.Swathi, K.Srikanth, S. Raghunath Reddy, Virtualization in cloud computing, IJCSMC, Vol. 3, Issue. 5, May 2014, pg.540 – 546*
[8] *https://aws.amazon.com/kinesis*
[9] *https://docs.aws.amazon.com/streams/latest/dev/introduction.html*
[10] *https://aws.amazon.com/solutions/case-studies/thomson-reuters/*
[11] *https://aws.amazon.com/solutions/case-studies/big-data/*

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Weather Analysis and Prediction: A Survey with Visual Analytic Perspective

[1]G.Arumugam, [2]Suguna Sangaiah, [3]G.Sudha
[1]Senior Professor and Head Department of Computer Science, Madurai Kamaraj University, Madurai
[2]Assistant Professor of Computer Science, Sri Meenakshi Government Arts College for Women(A), Madurai
[3] Assistant Professor of Computer Science,M.V.Muthiah Government Arts College for Women, Dindigul

ABSTRACT *To forecast weather, we need to analyse large set of data in order to save lives, reducing risk, improving quality of life, enhance the profitability and humanity. As nature of data and data source become more complex, voluminous, the ability to explore, analyse and to provide robust decision making become critical but it is essential in many situations. Visual analytics help us in such situations; Visual analytics is a powerful, interactive, automated technique to enhance the decision making ability, which is powered by data model and data visualizations. This survey presents the various research work carried out by research community on data visualization, data models developed and visual analytic techniques along with its merits and limitations.*
Keywords*:Visual Analytics, Data Visualization, SVM, Regression model, Forecasting algorithms, Machine Learning.*

Introduction

Weather is everything that affects our lives on daily basis and has power to make a permanent change to human life, society, economy and environment. Weather Forecasting or Prediction may be long-term prediction or short-term prediction. Anyway, it need historical data to analyse the present state with empirical one. Based on the similarity, we can come to the conclusion of weather forecast or prediction results. The model which transform the current weather data in to form in which comparison with empirical data decides the accuracy in weather prediction. Construction of such model is not an easy job; we cannot say single model works well for all sort of data.

A Picture worth thousand words; Representing data in visual form leads to easy exploration, discovery, and enable new findings, insights. Thus, make analytic process simple and fast. Data visualizations are broadly classified in to the following.

Table I :Classification Of Visualizations

Typeof visualization	***Data Requisite***	***Suitable Applications***
Scientific Visualization	Structured data	Medial, seismic applications
Information Visualization	No inherent structure	Social media, financial sector related applications
Visual Analytics	High dimensional, multi variate data, spatio temporal data	Geographical applications. Applications handles complex, multivariate, high velocity data.

Visualizations support following types of charts

Table II :Types Of Visualizations

Name of the chart	***Functionality***
Line chart	Shows comparison or changes of values of several variables
Bar chart i. Water fall chart or progressive chart	Shows comparison of quantities which are clustered shows variation in data in par with previous data
Scatter plot i. Bubble chart	Shows joint variation of two attributes. Suitable for statistical analysis like correlation, dependency and regression. Suitable when values differ from several order of magnitude.
Pie chart	Used to compare a part in a whole. Suggested on limited components used for comparing and precise accuracy is not aimed.
Box plots	Suitable for visualizing structured data and converted into percentile format. Extreme values are represented as whiskers. Box plots are also incorporate statistical measurements such as minimum, maximum, lower quartile, upper quartile and mode. Suitable for financial services and applications.
Word cloud	Applicable in visualizing structured and semi structured data. Word clouds are formed with help of taxonomy and ontology. Association between clouds can be constructed by identifying degree of association between various word clouds.
Network diagram	Suitable for deriving counter intelligence and law in-forcement to map clandestine of convert espionage ring.
Correlation matrix	Suitable for visualize high velocity data; identifies coupling between attributes.
Histogram	Provides visual distribution of data and shows how data will change on filtering on particular measurement.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Decision trees	Group of values having strong relationship are transformed as bin; relationship between bins is represented as branch which refines various elements influencing analytic process.
Sankey diagram	Tracks the path analysis to identify dynamics of how transaction moves through the system. Visualization consist of series link nodes attributing it frequency and isolated flow concludes the actions which are failed to deliver.

Visual Analytics

Visual analytics combines automated analysis techniques supported by interactive visualizations and data model for effective understanding, reasoning, decision making on large, complex and voluminous data set.

Visual analytics acquires knowledge and analytical intelligence by two supportive as well as iterative processes.

i. Data model
ii. Data visualization

As we have discussed on data visualizations in detail, visualizations strengthen the decision making ability. Visual analytics develops data model, which cope up with data visualization in interactive manner. This interactivity aids visual analytics not to deviate from analytic domain of concerned application. Data model can be implemented by using data mining, neural networks, machine learning, and statistical models. Machine learning models are popularly used to gain new derived knowledge from empirical data. The strength of visual analytic is most promising in weather analysis and forecasting applications.

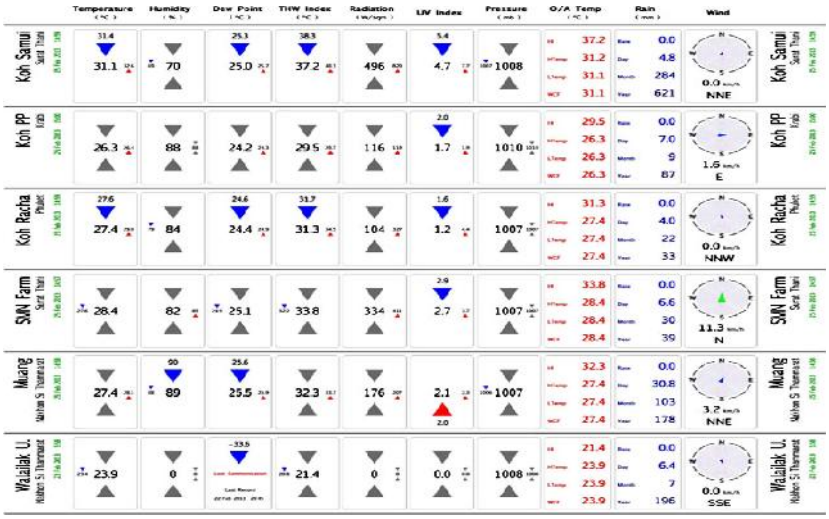
Survey of Related Works And Comparative Analysis

KrisanadejJaroensutasinee et al [1] have developed integrated online weather data analysis and visualization tool called 'Ecoinfows'. The weather data were collected from 18 automatic weather data servers and 8 loggers of Thailand. This application also provides near real time weather portal. Thus, the authors have collected data from multiple sources representing various weather attributes such as temperature, humidity, temperature-humidity-wind index, radiation, UV index and pressure and wind direction with velocity. Weather data archive was explored, cleaned and converted in order to provide monthly data and yearly data. Ecoinfows have produced eight different visualizations focusing on various weather analysis.

Table III :Visualizations Used And Functionality

Visualization	***Insight derived***
Timeseries visualization	Temperature variation in a city for a specific period of day-interval.
Histogram	Relative temperature frequency of city for a given period of days.
Overlay histogram	Compare temperature variations in two cities on same period
3D moving histogram	Average humidity recorded in 12-month for given a year
Scatter plot	Cluster analysis
Regression Scatter plot	Correlation of temperature with relative humidity on a given month Regression between two city temperatures in a specific period. Data distribution analysis given as normal distribution and binomial distribution

Ecoinfows have come up with overall temperature and rainfall summery frame, a visual representation comprised of all 8 parameters recorded at various cities. The output of this entire work is shown in summarised visual representation given as:



----------a---------- ------b------ -- c ---

Fig1. Visualizations of weather data : a weather parameters, b overall temperature and rainfall in a summary frame, and (c) wind rose visualization

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

We have included the research work of Yu Zhang et al.[2] in our survey since it uses Bigdata and neural network concepts which also suitable in visual analytics. The authors have collected historical weather data from year 2000 and used it as ground truth (statistical model). They used big data to store historical weather data captured at every 10 minutes interval by satellite. It is used to compare current scenario in order to analyse and forecast. Jet streams (curve shape formed in satellite image where warm and cold air met at high altitude) causing Coriolis effect – a atmospheric circulation. The low pressured scenario is known as through; high pressured scenario is known as ridge. Cloud movement are tracked as optical flow. Its estimation is computed by lucas – kannade algorithm. Vortex core are the identified since it is the significant signature of storm. Vortex core are associated with geographical map. Ensemble learning algorithm is used here to make predictions by combining results of multiple individual decision trees obtained from random subset of training data.

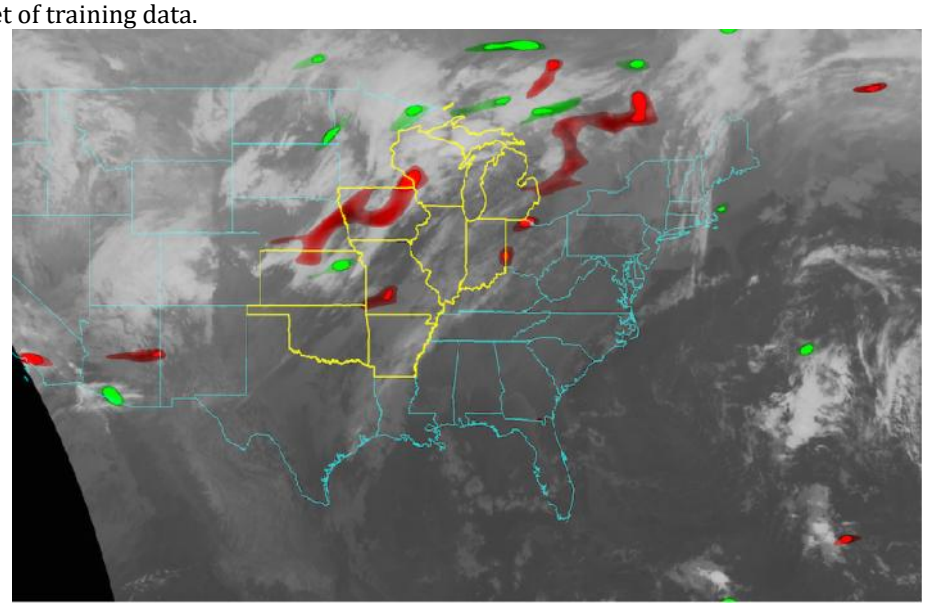


Fig 2 .Vortex cores and their expansions

Our survey includes this work of Saroj Kr. Biswas et al [3] since neural network and machine learning algorithms are used which are most promising data models in visual analytics. The authors implemented this work in MATHLAB. They developed NARX NN (Non linear Auto Regressive with eXogenous Neural Network) which work with artificial intelligence learning and problem solving approach works on knowledge gained at past, known as Case Based Recording (CBR). NARX NN predicts next value in time series by tracking time series value at past which influences current one. CBR act as repository of historical experiences comprised of four cyclical process:

1) Retrieval : retrieves similar cases; match the retrieved case with current one in terms of Euclidean distance; compare past by dissimilarity score. select the past one having least similarity score.
2) Reuse : retrieve more cases which match with current one.
3) Revise : integrates useful information about new case

NARX NN model was trained by eight possible seasons such as winter, spring, autumn, winter + spring, spring + summer, summer + autumn, autumn + winter. The number of hidden layer and number of neurons' in each hidden layer are not fixed. Lavenberg – Marquardt back propagation algorithm was used as training algorithm and the learning rate was 0.01; 1000 epochs are required in training the network. Mean Squared Error was used as stopping the training process. The model have concluded that NARX NN architecture have given optimal input-output delay. They suggested the inclusion of more number of weather attributes, incorporating historical data.

Mahamad B Pathan [4] has contributed the statistical data model applicable to climate science. Authors have made a building model of climate data using statistics in order to expose insights in climate data. He used numerical example containing three variables describing climate data – minimum temperature, maximum temperature and rain fall. He fetched 100 year data from Indian Meteorological Department and transformed it as mean minimum temperature, maximum mean temperature and mean rain fall. He exercised single and multiple regression models. He also calculated correlation coefficient and variance; analysed the result in ANNOVA. The author has concluded that correlation between rainfall and temperature range has shown stronger impact than minimum and maximum temperature values. He also concluded that multiple regression model of rainfall on maximum temperature and minimum temperature have produced better result. He also suggested that the inclusion of climate influencing factor such as green house gases may give precise and better analysis.

Md. Tanvir AlamAnik et al [5]have constructed the dataset from Bangladesh Agriculture Research Council which has provide year wise – monthly average rainfall at 32 different rainfall stations. The author have chosen data visualization to plot rain fall data. And deliver to Bangladesh farmers, who are the backbone the country thru ICT. Farmers need only short-term forecasting to save their crops and the author did so. The yearly data was split into three ranges based on monsoon.

i. Pre-monsoon – March to May ii. Monsoon – May to October iii. Post monsoon – October to February

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

and implemented using R tool. Moran's I and Geary's C methods are used to calculate the correlation and used the following visualizations to derive the insights.

Table IV :Visualizations Used And Insight Derived

Visualization	***Insight***
Bubble chart	Average rainfall in a year station Vs average rainfall
Histogram	Average rainfall VS rainfall frequency
3D scatter plot	Log of monsoon rainfall at a station
Line chart	Individual station rainfall – month Vs rainfall measure

Author has suggested the inclusion of google map as future work. This work handles spatial data with short term forecasting.

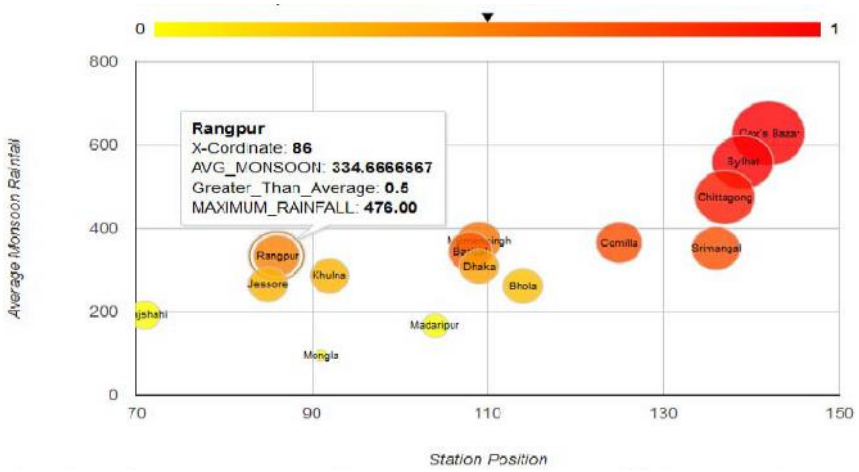


Fig 3 . Bubble chart based representation of Monsoon season's rainfall data. X and Y-axis show station position and average Monsoon rainfall unit respectively

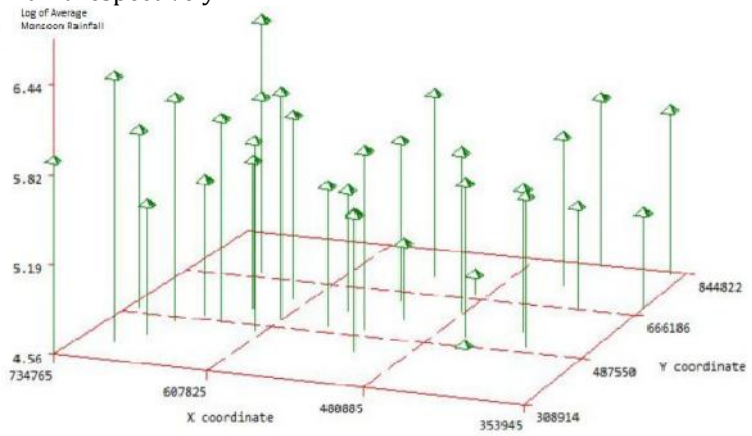


Fig 4 .3D visualization of average monsoon rainfall. X and Y-axis present annual rainfall and its frequency respectively

A.Diehl et al [6] have developed Visual Interactive Dash board (VIDa) to provide interactive visualization interface and provided multiple views such as timeline with geo referential maps, integrated web map view, forecast operational tool, curve pattern selector, spatial filters and linked meteogram. This work was passes thru four major stages. They are

1) GDA – Global Data Assimilation system, used to provide observational data for further phases.
2) WT1 – model simulation convert the observation into every 12-hours numerical weather data.
3) WT2 – predictions post–processing–simulation outputs are post processed to derive new atmospheric information from the results.
4) WT3 – visualization pipeline – generates the post processed output into 2D plots and hosted in site.

Minimaps are generated to show spatial information; whereas temporal information is done as curve pattern analysis. The authors have concluded that integrating the work with empirical data to handle ensemble, uncertainties and model errors. It is also suggested to include more, relevant parameters in minimaps to give robust, more accurate forecast results.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

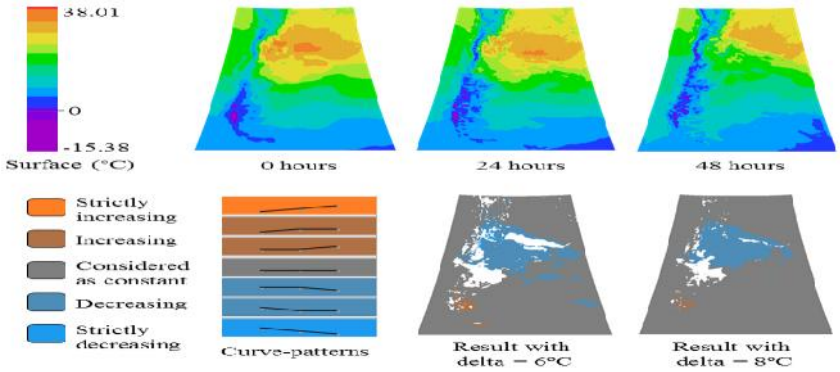


Fig 5. Trend analysis: the largest temperature drops, just after the passage of the cold front, are indicated in cyan tones. Orange tones in the south of the map indicate increasing temperature. Results are presented using two different delta values: 6_ and 8_ Celsius.

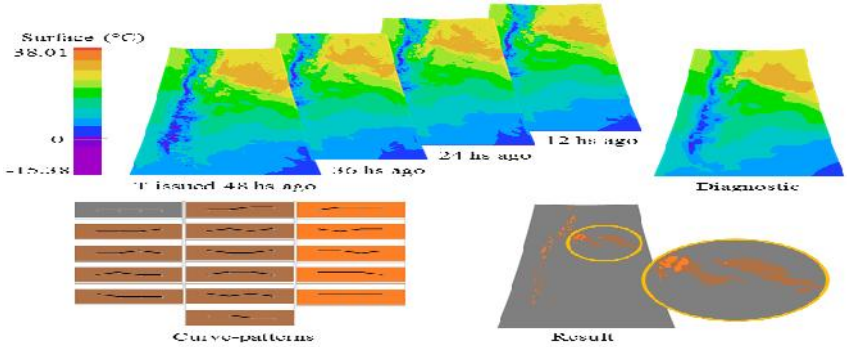


Fig 6. Forecast verification using multiple runs. It shows the passage of a cold front. This is observed as errors in the forecasted temperature shown in oranges tones.

P.Samuel Quinan et al [7] have developed a open tool WeaVER, for visualizing weather forecasting. The objective of the work was to categorize weather related problems and isolating the data relating to weather forecasting. The author have discussed about the following weather related problems

1) Forecasting type –
   a. Deterministic forecasting
   b. Ensemble forecasting – forecasting done by two or more simulation results coupled on same time frames, different parameters, boundaries and initial values.
2) Locating and relating specific feature – features such as 20% humidity, low pressure (not in numerals).

The result of the work is plotted as following visualizations:

1) Interactive spaghetti plots – colour coded one to differentiate and enable direct comparison among distributions of multiple isocontour features with in single plot.
2) Interactive contour box plots – summarizes the distribution of isocontour based features across ensemble.

The authors have concluded that formal valuation of interactive contour box plots and spaghetti plots need multiple isocontour based features. They also suggest new research on comprising non – isocontour based features are also to be developed.

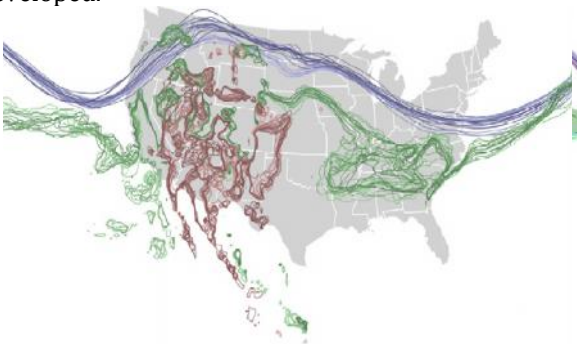


Fig 7. interactive spaghetti plots

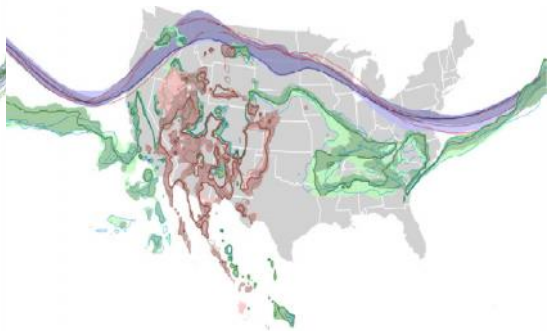


Fig 8. interactive contour boxplots

Wei-Ta chu et al [8] have contributed researchers by developing image2weather dataset containing more than 180,000 images to promote similar research in weather analysis. This work construct data set describing the weather features intelligently obtained from a photo image. The photos or images are acquired from various resources like google map, pictures posted at flicker and weatherground.com portal. These images are processed and classified into any one of five weather types (sunny, cloudy, snowy, rainy, and foggy). They used convolutional neural network to

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

extract the feature and SVM for classification. Every image is attributed with related properties such as weather type, temperature and pressure. It is suggested weather estimation model development as future work.

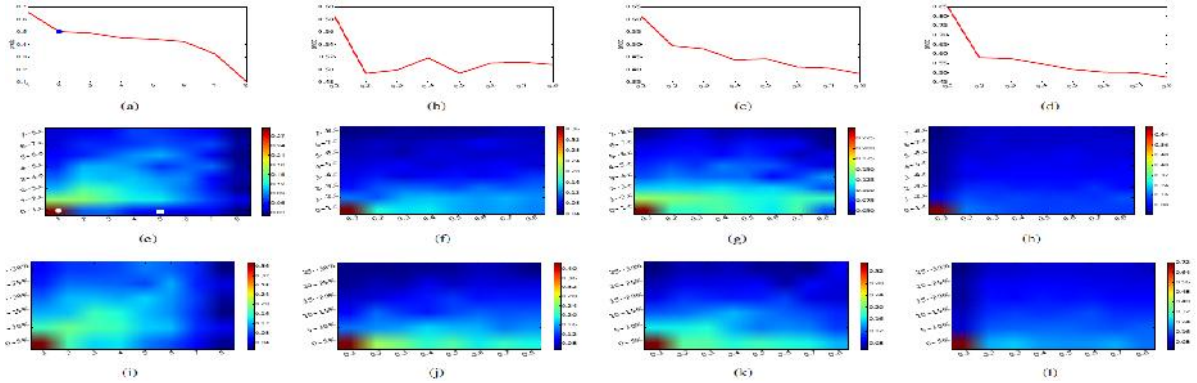


Fig. 9. Relationships between various properties and weather properties. (a)–(d) The prob. of sharing the same weather type vs. time/color/texture/intensity distances. (e)–(h) The prob. of temperature distance within a range vs. time/color/texture/intensity differences between photo pairs. (i)–(l) The prob. of humidity distance within a range vs. time/color/texture/intensity differences between photo pairs.

T.R.V.Anatharajan et al [9] have developed a tool which takes inputs – maximum temperature, minimum temperature and rainfall regarding certain sample period of days. Author claims that analysis of sample data and prediction based on linear regression model achieves more than 90% accuracy, based on sample data. They declare that branch of AI, Machine learning has been proved as robust method, in prediction and analysis of given dataset.

Since weather prediction cannot be a supervised one; hypothesis functions aiming at predict a hypothesis close to the output. Then cost function is used to compute minimum distance between hypothesis curve and output curve. Then differential equation, gradient decent is used to minimize cost function after repeated iteration. The entire work was developed in MATHLAB. The authors have concluded that same work could be done using Neural Network and classification models; but it cannot result in nearest value probability of how the day is going to be. They also stated that 1/7 of their prediction is accurate.

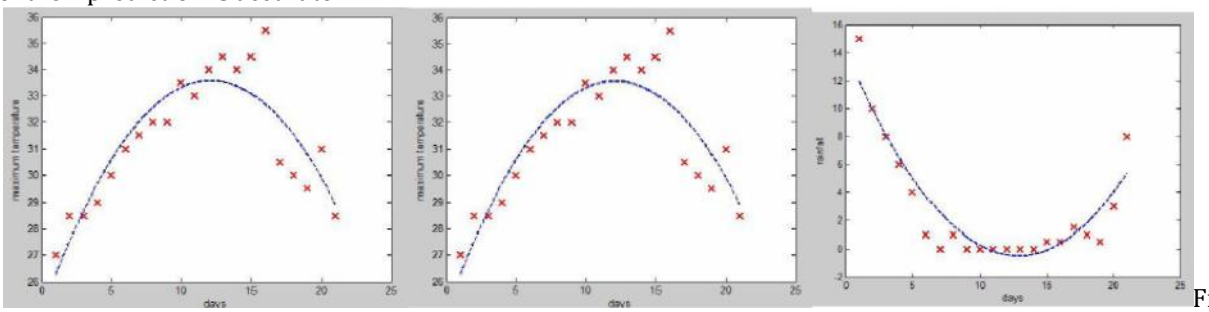


Fig 10.
Maximum temperature Vs Days Minimum temperature Vs Days Rainfall vs. days

The objective of Mr.C.P.Shabariram et al [10] work is to predict maximum rainfall, upcoming rainfall maximum rain storm location based on hourly and overall storm parameters. This work handled spatio temporal data and the storm parameters are categorized such as local storm, hourly storm, overall storm, overall storm(local and hourly), location based storm, event related storm and measuring the tornado waves. This work was implemented in Bigdata Hadoop and Map Reduce framework. SVM classifier of Data Mining works based on spatio temporal characteristics and also extracts the features for classification in MapReduce. Map process took care of partitioning spatio temporal data; Reduce function handled overlapping of spatial data. This work has used National Weather Service Dataset.

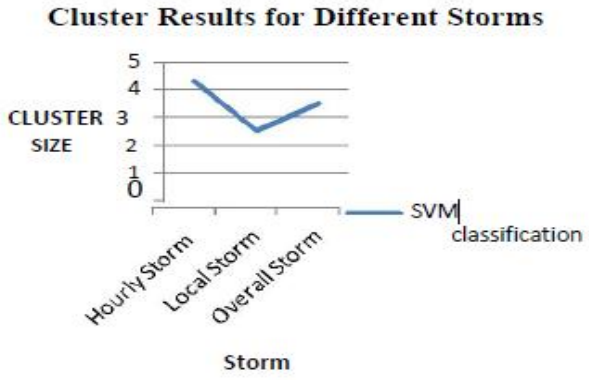


Fig 11. Expected Results of different Storms

Liu et al [11] have discussed about interactive visualisations combined with machine learning which really helps in solving real world problems using Artificial Intelligence. They have presented comprehensive analysis and classified relevant work into three categories like – Understanding, Diagnosis and Refinement.

(i) Understanding : Liu et al broadly classified the visualization-approach into two categories

(a) Point based technique – It reveals relationship between neural network components such as learned representations. Point based techniques facilitates confirmation of hypothesis on neural networks and identification of previously unknown relationship between neural network components. But point based techniques have failed to reveal topological information such as, role of different neurons at different layers and interaction between them.

(b) Network Based Technique – This technique represents neural network as DAG (Directed Acyclic Graph) and encode information from network by size, colour and glyph.

The above discussed techniques are capable of displaying visualizations on limited number of neurons. The complexity raises when number of neurons in neural network increases or number of neuron in each layer increases and thus interaction between them are increases in exponential. In such cases Convolutional Neural Network (CNN) helps us; meanwhile, to represent the CNN as visualization, the visualization is designed to have a representative layer; in which a representative from each layer participates. Biclustering algorithm is used to bundle the edges and reduce visual cluster.

(ii) Diagnosis: Current techniques utilises prediction score distributions of model to evaluate error severiority and study how score distributions correlate with misclassification and selected features. This paper has stated two visualizations namely

(a) Confusion wheel – to show samples which are misclassified in to radical view.

(b) Square tool – to show prediction score at multiple level.

CNN visualization allows machine learning experts to debug training process that fails to converge or does not achieve an acceptable performance.

(iii) Refinement – After gaining understanding of how machine learning models behave and why they do not achieve desirable performance, Visual Analytic system provides interaction techniques or capabilities for improving supervised or unsupervised models.

This paper has proposed visual notation called Neighbour Joining Trees to select the training data to train the model to achieve desirable training model which reduces training time. UPTOPIAN is a Visual Analytic system for refining topic model results and it uses scalable algorithm to learn correlated topic models. The authors have concluded that even though machine learning is widely used to solve real time applications; it fails to explain their decisions and actions to the users. As an initial remedy, Probabilistic program induction algorithm was developed and build on simple stochastic programs to represent the concepts.

Cong Feng et al [12] have proposed new method in time series characteristic analysis approach to visualize and quantify wind time series diversity. They emphasis that ANN, ARIMA and SVM have shown conflicting results on various dataset. To analyse time series characteristics, authors have devised six characteristic indices (CI), each representing different wind features and they are

(i) Strength of trend CI1 – Long increase / decrease in time series
(ii) Strength of seasonality CI2 – wave like fluctuations of constant length
(iii) Skewness coefficient CI3 – Univariate distribution computed by 3rd moment of random variable (Pearson's moment)
(iv) Kurtosis coefficient CI4 – Kurtosis distribution computed by 4th moment of random variable (Pearson's moment)
(v) Nonlinearity CI5 – to measure nonlinear structure in time series
(vi) Spectral Entropy CI6 – to describe uncertainty, complexity and chaotic in time series.

These six CIs creates high dimensional data as extracted from stream of time series. To reduce the highdimensionality, principal component analysis (PCA) is used. Normalization on CI values are mandatory before exercising the PCA process.

Next, the result is shown as two types of visualizations.

(i) Scatter plot – to characterise the diversity
(ii) Convex polytope – By applying 2D quick hull algorithm and general dimensional beneath beyond algorithm to plot polytope surface area. The diversity volume is formed by Delaunary triangular Algorithm.

Authors have tested their method with five different datasets namely GEFCom2012, GEFCom14, SURFRAD, WIND toolkit and CompNWP which delivers different wind features collected by ground station sensors.

A.Diehl et al [13] have proposed a novel approach to break down automatic process using the experience and knowledge of users, may be expert and creates Visual workflow to aid probabilistic weather forecasting (PWF). In PWF, numerical weather prediction system is used to estimate occurrence probability of particular weather phenomenon. Hamill and Whittaker have introduced Reforecast Analog Regression (RAR) method which also be suitable for PWF which have following advantages

(i) Relationship between systematic model and scale atmospheric circulation patterns can be captured which in turn benefits better estimation.

(ii) Weather analysis are not only projected as probabilities but also quantified as sample mean, standard deviation, maximum and minimum values in order to detect extreme weather events.

Visual Analytic work flow consist of three tasks.

(i) Parameterization loop – This task is used to set spatio and temporal parameters for RAR. The result was shown as visualization named ubermap, which allow the user to explore different thresholds, accumulation ranges.

(ii) Probabilistic forecast loop – This task aids user to visualize uncertainty at different levels of detail. This loop produces the visualizations on probabilistic forecast for given metrological variable, analyse the associated numerical forecast results and summarises information regarding observations and uncertainty.

(iii) Analog loop – This task has analog viewer, a tool which allow user to access information about forecast analog and their associated observations as well as statistical aggregations and summary.

The authors have tested their work using GFS(Global Forecast System) and CMORPH of NOAA dataset.

Himanshi Jain and Raksha Jain [14] have emphasised the need of adaptation of BigData in weather analysis and forecasting activity. They have discussed on the following applications.

(i) Industry –
   (a) Agriculture and food industry – Weather forecasting helps farmers to prepare for droughts, over watering and soil erosion. Rainfall prediction is mandatory for food protection management and scarcity issues.
   (b) Tourism industry- As tourism has major impact on income of the country, weather forecasting aids tourists to plan their tour according to the climate of the country to be visited.
   (c) Construction industry- Construction can be affected by wind, temperature, moisture, humidity and dew. Flood forecasting helps in construction sustainable and build strong buildings.
   (d) Sports Industry – Preparation of tournament, play grounds are effectively planned with the aid of accurate short term forecasting.
   (e) Transportation – Airways, shipping, railways, roadways and even unmanned vehicles are benefitted by weather forecasting information in order to plan diversions, risk evaluation and mission planning.

(ii) Disaster Management – Good weather predictions saves lives, reduce damages, reservoir management, flood preventive measurement and save us from major economic losses.

(iii) Energy – Short term weather forecasting like wind speed, forecast, cloud coverage, temperature prediction is necessary for wind and solar energy power stations for its effective running.

Challenges in weather forecasting systems are

(i) Retrieving uninterrupted volume of empirical weather data.
(ii) Coupling Technology and infrastructure to support simulation and analytical work flow.
(iii) Adaptation of suitable forecasting model.
(iv) Integrating variety of data from different sources with no-compromise.
(v) Very costly to Implant new sources (like launching new satellite, planting new ground stations with sensors) to capture weather related data.

Sudha. G and Suguna Sangaiah [15] have discussed about the application of visual analytic techniques and various visualizations applicable in healthcare domain. This paper has illustrated the way in which visualizations and visual analytic techniques used in accessing the disease acute level, by providing the similar cases observed in the past; based on the collective information, the insights are derived which used to tailor the treatment schedule and plan could be optimally scheduled.

Conclusion

This survey paper has summarized the past work done on visualization, visual analytic and related approaches applicable in Weather Forecasting Domain. Wide ranges of visualizations are available and data models ranging from statistical model to Neural Networks are joined hands in visual analytic technique to provide accurate and robust forecasting. Each model has its own advantages and limitations thus the interactive visualization also has an impact since it gets tightly coupled with the result of the model. Our future work is focused on implementing and deploying various functional models like datamining, Neural network, Machine Learning and Statistical Models in the Visual Analytic architecture and to generate interactive visualizations. Our future work is to develop ensemble models through Visual Analytics architecture and to analyse the performance while adapting various functional models and to aid the domain experts to highlight, provide the new insights concern to the underlying problem dealt with.

References

[1] KrisanadejJaroensutasinee, WittayaPheera and Mullica Jaroensutasinee, Online Weather Data Analysis and Visualization tools for Applications in eco informics, Springer-Verlang Heidelberg 2013.

[2] Yu Zhang, Stephen Wistar, ose A. Piedra-Fernandez, Jia Li, Michal A. Steinberg and James Z.Wang, Locating visual storm signatures from satellite images, IEEE, 2014, IEEE International conference on BigData.

[3] Saroj Kr. Biswas, Nidul Sinha, Biswajit Purkayastha and LeincyMarbaniang, Weather prediction by recurrent neural network, International Journal of Intelligent Engineering Informatics, Vol 2, Nos2/3, 2014, InderScience Enterprises Ltd.

[4] Mahamad B Pathan, Modeling Data from Climate Science using Introductory Statistics, Journal on Climatol Weather Forecasting, 2015, http://dx.doi.org/10.4172/2332-2594.1000129.

[5] Md. Tanvir AlamAnik, Krishna Pada Dhali, S M Ferdous Hossain and FatemaTujJohara, Spatial Data Visualization Methodologies in ICT4D research, 18th International Conference on Computer and Informaion Technology (ICCIT), IEEE, 2015.

[6] A.Diehl, L.Pelorosso, C.Delrieuax, C.Saulo, J. Ruiz, M.E. Groller and S.Bruckner, Visual Analysis of Spatio-temporal data: Applications in weather forecasting, Euro Graphics on Visualization (Euro VIS) Volume 34(2015), No.3.

[7] P.Samuel Quinan and Miriah Meyer, Visually comparing Weather features in forecasts, IEEE Transactions on visualizations and computer graphics, Volume 22, No.1, Jan 2016.

[8] Wei-Ta chu, Xiang-You Zheng and Ding-Shiuan Ding, Image2Weather: A larger scale image dataset for weather property estimation, 2016, IEEE second international conference on Multimedia and BigData. DOI.10.1109/BigMM.2016.9

[9] T.R.V.Anatharajan, G.AbishekHariharan,K.K.Vignjith and R.Jijendirankushmita, Weather monitoring using Artificial Intelligence, 2016 International Conference on Computational Intelligence and Networks, DOI 10.1109/CINE.2016.26.

[10] Mr.C.P.Shabariram, Dr.K.E.Kannammal and Mr.T.Manojprabakar,Rainfall Analysis and rain storm prediction using map reduce framework, 2016 IEEE, International conference on computer communication and Informatics (ICCI-2016).

[11] Shixia Liu, Xiting Wang, Menchen Liu, Jun Zhu,Towards better analysis of machine learning models: A visual analytics perspective, Elsevier B.V. Visual Informatics, 2017 DOI: http://dx.doi.org/10.1016/j.visinf.2017.01.006

[12] Cong Feng, Erol Kevin Chartan, Bri-Mathias Hodge, Jie Zhang, Characterizing Time Series Data Diversity for Wind Forecasting, ACM 2017 ISBN 978-1-4503-5549-0/17/12..https://doi.org/10.1145/3148055.3148065

[13] A.Diehl, L.Pelorosso, C.Delrieux, K.Matkovic, J.Ruiz, M.E. Groller, S.Bruckner, Albero: A Visual Analytics approach for Probabilistic Weather Forecasting , Computer Graphics Forum, 2017, DOI: 10.1111/cgf.13279 Volume 36(2017) Number 7.

[14] Ms.Himanshi Jain and Ms.Raksha Jain, Big Data in Weather forecasting: Applications and Challenges, IEEE 2017, International Conference on Big Data Analytics and Computational Intelligence (ICBDACI), 978-1-5090-6399-4/17/

[15] Sudha G. and Suguna Sangaiah, A Survey on Contribution of Visual Analytics in Health Care Domain (April 19, 2018). 2018 IADS International Conference on Computing, Communications & Data Engineering (CCODE). ELSEVIER ,SSRN: https://ssrn.com/abstract=3165309 or http://dx.doi.org/10.2139/ssrn.3165309.

Analysis Of Tree And Rule Based Classifier Algorithms For Laptop Utilization Dataset

[1]V.Lakshmi Praba & [2]K.Vettriselvi
[1]Assistant Professor [2]Research Scholar,
[1, 2]Dept. of Computer Science, Rani Anna Government College for Women, Tirunelveli, India
[1]E-mail: vlakshmipraba@rediffmail.com [2]E-mail: vettriselvi25@gmail.com

ABSTRACT *Laptops are used by the students for various purposes like - education, coaching, browsing of information, communication, playing games and watching movies. This work is carried out to study and examine higher education student's perceptions related to the usage of laptop by analyzing its utilization characteristics using classification algorithms. To carry out this task, 'Laptop Utilization by students' has been taken as dataset for analysis. The data is collected from 100 postgraduate women students of Rani Anna Government College, Tirunelveli by supplying questionnaires. Classification using - Tree based and Rule based classification algorithms were supplied with this data set. Tree based algorithms applied are J48, Random Tree and Rep Tree. Rule based algorithms applied are, Decision table, OneR and ZeroR. Weka tool is used for carrying out this analysis. The most suitable algorithm in classifying the dataset is identified.*
Keywords: *Data mining, Classification algorithms, Laptop Utilization.*

I. Introduction

The World Wide Web (WWW) overwhelms us with information; meanwhile, every choice we make is recorded. As the volume of data increases, inexorably, the proportion of it that people understand decreases alarmingly. Lying hidden in all this data is information. Laptops are used by everyone in different fields. Nowadays the main users of the laptops are college students. They used it in a variety of environments such as programming, preparing the assignments, project work, and for entertainment. This research uses laptop utilization as data set. The information is collected from 100 students who are using laptops. From this data set, it is possible to identify the usage of laptops among students. The main objective of the research is to determine the most suitable classification algorithm using the data set of Laptop utilization by the students.

This process is done by using algorithms in data mining applying Weka tool. The dataset helps to analyze, for what purpose the students are using laptops more - whether for education or entertainment or internet usage. Seven classes are identified to separate the students' category as how they spend the valuable time in laptops.

Some standards and terms for Accuracy Measure:

The accuracy (AC) is the proportion of the total number of predictions that were correct.

- ***True positive*** (TP) is the proportion of positive cases that were correctly identified.
- ***False positive*** (FP) is the proportion of negative cases that were incorrectly classified as positive.
- ***Precision and recall***: Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. Recall is the true positive rate for the class [2] [3].
- ***F-measure*** (also known as F1 or F-score) is a measure of test's accuracy. It considers both the Precision and the Recall of the test to compute the score. It can be interpreted as a weighted average of the Precision and the Recall, where 1 is its best value and 0 its worst. The F-measure only produces a high result when Precision and Recall are both balanced, thus this is very significant.
- ***ROC Area:*** A Receiver Operating Characteristics (ROC) curve is a technique for visualizing, organizing and selecting classifiers based on their performance.

II. Literature Survey

Jiawei Han and Micheline Kamber [1] in their book they have elaborated about the data mining concepts and techniques.

Hong Hu, Jiuyong Li, Ashley Plank [2] in their paper presented five classification methods, namely C4.5, BaggingC4.5, AdaBoostingC4.5, LibSVMs and Random Forest for comparison using seven Microarray data sets, with or without gene selection and discretization.

Sushilkumar Kalmegh [3] in his paper analyses the performance evaluation of REP Tree, Simple Cart and Random Tree classification algorithm. According to his findings, the efficiency and accuracy of Random Tree is better than REP Tree, and Simple Cart.

S. Syed Shajahaan, S. Shanthi, V. ManoChitra [4] presents the application of decision trees for predicting the presence of breast cancer. The performance of conventional supervised learning algorithms viz. Random tree, ID3, CART, C4.5 and Naive Bayes were also analysed in this work.

Bernhard Pfahringer [5] in his paper introduced a novel general regression method that combines model trees with random forests. According to his finding, the training and optimization of Random Model Trees scales better than Gaussian Processes Regression for larger datasets.

Mining Social Networking Data for Classification Using REP Tree presented by Dr. B.Srinivasan, P.Mekala [6], focuses on demonstrating the workflow of social media data sense-making for educational purposes by integrating both qualitative analysis and large-scale data mining techniques.

Payal P.Dhakate, Suvarna Patil, K.Rajeswari, Deepa Abin [7] in their paper presents about the basics of data mining, preprocessing and different classification techniques using diabetes dataset. WEKA tool was used in this work

Pavel Berkhin [8] in his paper presents a review on different types of clustering algorithms in data mining.
Jyotismita Goswami [9] in his paper presents a tutorial overview of the main clustering methods used in Data Mining. Concepts of few classification algorithms were also presented in this work.

Weka Machine Learning Project [10] illustrates the Weka Tool with examples and working functionality of various data mining algorithms.

M.Sivagami, V.Lakshmi Praba [11] in their paper, presents Tree Based Classification algorithms and some Hierarchical Clustering Algorithms.

III. Tools And Algorithms Applied

A. WEKA Tool

WEKA is open source software issued under the GNU General Public License [4]. WEKA tool is used to implement the algorithms for Classification. The algorithms are applied directly to a dataset.

B. Classification Techniques

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining has many analytical tools for analyzing data from many different dimensions, categorize it, and summarize the identified relationships.

Data Classification refers to categorization, the process in which the data elements are differentiated based on certain criteria. An algorithm that implements classification is known as a classifier. The term "classifier" also refers to the mathematical function that is implemented by a classification algorithm which helps to map input data to a particular category. Classification is a data mining algorithm that creates a step-by-step guide for how to determine the output of a decision based on the input, and to move to the next node and the next until one reach a leaf that tells the predicted output [5].

C.Tree Based Algorithms

J48 Classifiers

J48 Algorithm can predict both 'nominal' and 'numeric' attribute values. This algorithm uses 'most relevant attribute' from the dataset to determine the prediction values. Hence it is better to have all the attributes rather the only relevant attributes. Using all the data set for J48 Algorithm, the prediction efficiency increases.

Random Forest

Random trees have been introduced by Leo Breiman and Adele Cutler. The algorithm can deal with both classification and regression problems. A random tree is a collection (ensemble) of tree predictors that is called forest. In a random forest, each node is split using the best among the subset of predicators randomly chosen at that node. The random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of "votes". [6] [7] [8].

REP Tree

Reduced Error Pruning (REP) Tree is fast decision tree learning. It builds the decision tree based on the information gain as the splitting criterion, and prunes it using reduced error pruning. Rep Tree uses the regression tree logic and creates multiple trees in different iterations. After creating multiple trees it selects the best one from all generated trees which is considered as the representative. In pruning the tree the measure used is the mean square error on the predictions made by the tree.

D. Rule Based Algorithms

Decision Table is an accurate method for numeric prediction from decision trees. It is an ordered set of *"If-Then"* rules that makes it more compact and understandable. Decision table algorithms are found in the Weka classifiers under Rules. It summarizes the dataset with a "decision table" which contains the same number of attributes as the original

dataset. A new data item is allocated a category by searching the line in the decision table that is equivalent to the values contained in the non-class of the data item.

OneR

This algorithm creates a single rule for each attribute of training data and then picks up the rule with the least error rate. To create a rule for an attribute, the most frequent class for each attribute value is determined. The most frequent class is the class that appears most often for that attribute value. A rule is a set of attribute values bound to their majority class. The algorithm is based on ranking all the attributes based on the error rate.

ZeroR

ZeroR classifier determines the majority category for a class. It is useful for determining a baseline performance that can be considered as a benchmark for other classification methods.

IV. Methodology

The work carried out is to identify the efficient classification algorithm among Tree classifiers and Rule classifiers. J48, Random Tree and Rep Tree classification methods are considered under Tree classifiers. Decision Table , OneR and ZeroR classification algorithms are considered under Rule Classifiers for classifying the considered dataset using WEKA tool.

A. Data Collection

A direct survey is used to collect the data for this study. This has been distributed to the students randomly. A total of 100 sets have been distributed. Initially the questionnaire was distributed and essential directions and instructions were given to them. Though any time restriction was not given for filling in the details, the students took around 20 to 30 minutes for filling the data.

B. Dataset Description

In this study, all data is considered as instances and features in the dataset are known as attributes. The data set consists of 25 attributes and 100 records/instances that are used Laptop Utilization by higher education students. The data set detail is as given in Table1.

Table 1: Dataset for Laptop Utilization

Dataset Name	Number of Attributes	Number of Records/Instances
Laptop Utilization	**25**	**100**

V. EXPERIMENTAL ANALYSIS AND RESULTS

The Tree and Rules classification algorithm that were discussed in the previous section were applied for the considered dataset and the obtained results were analyzed. The performance of all the three algorithms under Tree Classifiers is tabulated in Table 2 for the considered performance metrics. The various parameters that are considered for analysis are, Correctly Classified Instances, Incorrectly Classified Instances, Kappa Statistic, Mean absolute error and Root mean squared error.

Table 2: Comparison of the Different Tree Classifiers

TREES	J48	RANDOM TREE	REP TREE
Correctly Classified Instances	92	99	83
Incorrectly Classified Instances	8	1	17
Kappa statistic	0.8904	0.9867	0.7742
Mean absolute error	0.0317	0.0274	0.0701
Root mean squared error	0.1259	0.089	0.1872

From the observation, it is evident that, when compared with REP Tree, the other two are better and the Random Tree classifier produces the best result for correctly classifying the instances which leads to the better performance in other considered parameters too.

The performance of all the three algorithms under Rules Classifiers are tabulated in Table 3 for the considered performance metrics. The various parameters that are considered for Functions classifiers are taken for this method too.

Table 3: Comparison of the Different Rules Classifiers

Parameters	Decision Tree	OneR	ZeroR
Correctly Classified Instances	76	60	43
Incorrectly Classified Instances	24	40	57
Kappa Statistic	0.6476	0.3725	0
Mean absolute error	0.1656	0.1143	0.2164
Root mean squared error	0.2549	0.3381	0.3275

From the observation, it is evident that, when compared with ZeroR, the other two are better and the Decision Tree classifier produces the best result for correctly classifying the instances for the considered dataset with 100 instances, which leads to the better performance in other considered parameters too.

Accuracy Measures	Tree classifiers			Rule Classifiers		
	J48	Random Forest	Rep Tree	DT	OneR	ZeroR
TP Rate	0.92	0.99	0.83	0.76	0.6	0.43
FP Rate	0.047	0.001	0.04	0.161	0.248	0.43
Precision	0.926	0.991	0.842	0.688	0.367	0.185
Recall	0.92	0.99	0.83	0.76	0.6	0.43
F-Measure	0.916	0.99	0.827	0.709	0.455	0.259
ROC Area	0.991	0.994	0.971	0.92	0.676	0.5

Table 5 depicts a clear picture of the overall result of all the considered classification algorithms. The categories considered are basically Education, Internet and Entertainment. Various combinations of these basic categories are also included. The actual data and the result obtained using each algorithm is shown in the table.

Table 5: Results for Classification algorithms

Category	Actual Data	J48	RANDOM TREE	REP TREE	Decision Tree	OneR	ZeroR
Education	43	40	43	36	33	26	18
Internet	18	17	18	15	14	11	8
Entertainment	6	6	6	5	5	4	3
Edu+Int	7	6	7	6	5	4	3
Int+Ent	6	6	6	5	5	4	3
Ent+Edu	9	8	9	7	7	5	4
Edu+Int+Ent	11	10	11	9	8	7	5

To conclude, from the considered classification algorithms under Rule based and Tree based classifiers, it has been observed that Random Tree gives the best result. From the collected dataset, it is found that 43% of the students are utilizing the laptop for education purpose.

Conclusion

In this paper, it has been discussed about Tree and Rules based classifiers for Laptop Utilization dataset. The Tree based classifiers considered are, J48, Random Forest and REP tree. Rule based algorithms taken for consideration are Decision table, OneR and ZeroR. The aim is to determine the most suitable algorithm that can correctly predict the classes of different attributes using in Weka tool. Collected data is about how the students are using the laptops, based on the purpose of usage; purposes are categorized in data set.
Based on the considered performance metrics, the experimental results obtained are tabulated for each category of algorithms. It is observed that, among the six classification algorithms considered, Random Forest produces best results under the considered category. In future, other related algorithms can also be tried in similar line.

Acknowledgement: This research work is supported by UGC (MRP) grant

REFERENCES

1. Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques", Third Edition (The Morgan Kaufmann Series in Data Management Systems, 2006
2. Milan Kumari, Sunila Godara, "Comparative Study of Data Mining Classification Methods in cardiovascular Disease Prediction", IJCST, Vol. 2, Issue 2, 2011, pp. 304-308.
3. Hong Hu, Jiuyong Li, Ashley Plank, "A Comparative Study of Classification Methods for Microarray Data Analysis", published in CRPIT, Vol.61, 2006.
4. Sushilkumar Kalmegh, Analysis of WEKA Data Mining Algorithm REP Tree, Simple Cart and Random Tree for Classification of Indian News, International Journal of Innovative Science, Engineering & Technology(IJISET), Vol. 2 Issue 2, February 2015, ISSN 2348 – 7968.
5. S. Syed Shajahaan, S. Shanthi, V. ManoChitra, "Application of Data Mining Techniques to Model Breast Cancer Data", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 11, November 2013, pp- 362-369.
6. Bernhard Pfahringer, "Random model trees: an effective and scalable regression method" University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/~bernhard
7. Dr. B. Srinivasan, P.Mekala, "Mining Social Networking Data for Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014 pp-155-160.
8. Payal P.Dhakate, Suvarna Patil, K. Rajeswari, Deepa Abin, "Preprocessing and Classification in WEKA Using Different Classifier", Int. Journal of Engineering Research and Applications, Vol. 4, Issue 8 (Version 5), August 2014, pp-91-93.
9. Pavel Berkhin, Survey of Clustering Data Mining techniques, Accrue Software, Inc, 2000.
10. Jyotismita Goswami, A Comparative Study on Clustering and Classification Algorithms, International Journal of Scientific engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, ISSN: 2395-3470, 2015.
11. Weka Machine Learning Project, http://www.cs.waikato.ac.nz/~ml/index.html
12. M.Sivagami, V.Lakshmi Praba, "A Study and Analysis of Laptop Utilization Using Tree Based Classification and Hierarchical Clustering Algorithms", International Journal of Trend in Research and Development, Volume 3(5), ISSN: 2394-9333 www.ijtrd.com **IJTRD | Sep-Oct 2016**

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Rani Anna Government College for Women

(Re-accredited with "A" grade by NAAC)
Tirunelveli-627008
Research Department of Computer Science
UGC Funded Project
on

WOMEN EMPOWERMENT THROUGH EFFECTIVE UTILIZATION OF LAPTOP

QUESTIONNAIRE

Personal Details

Instructions: For each of the statements below, please tick (✓) in only one box that best describes yourself or you and your opinion.

Gender	☐ Male ☐ Female
Age	
Residential Area	☐ Rural ☐ Urban
Country	
College Name Course Name	
Year	

GENERAL	Yes	No
1. Do you own a laptop?		
2. If yes, is it a freely distributed laptop? 3. According to your opinion, is laptop a must for education purpose?		
4. Usage of laptop helped you to gain more knowledge.		
5. Is laptop a distraction for your studies?		
6. Has laptop helped you to score more marks in Theory Exams.		

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

7. Has laptop helped you to score more marks in practical Exams.		
8. Do you use Internet?		

8a, If yes, How do you get internet service?

☐ From College ☐ Net Connection at home

☐ Browsing Center ☐ Any other source

8b, Daily how much time do you spend on working with laptop?

☐ Below 1 Hour ☐ 1-2 Hours

☐ 2-3 Hours ☐ More than 3 Hours

8c, How much money do you spend monthly for internet?

☐ Less than ₹50 ☐ Less than ₹100

☐ Less than ₹150 ☐ More than ₹200

8d, Which Browser do you use surfing the internet?

☐ Google Chrome ☐ Opera

☐ Mozilla Firefox ☐ Internet Explorer

GROUP I (Education Purpose)	Strongly agree	Agree	Disagree	Strongly disagree
1. I use my laptop for academic purposes (ex: notes taking).				
2. I take better notes when I have access to a laptop.				
3. I am more concentrated and focused when I can view the lecture notes on my laptop.				
4. The laptop replaces my hard copy text book.				
5. I use my laptop for project typing work?				
6. I use my laptop for documentation.				
7. I use my laptop for programming purpose.				
8. I use the laptop for Excel sheets only.				

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Group II (Internet Purpose)	**Strongly agree**	**Agree**	**Disagree**	**Strongly disagree**
1. I use the laptop for checking my e-mail.				
2. I use the laptop for reading news. 3. I use the laptop for applying exams.				
4. I use the laptop for applying jobs				
5. I use the laptop for online purchase. 6. I use the laptop for net-banking.				
7. I use the laptop for finding information on internet.				
8. I use the laptop for downloading purpose (music, movie, game). 9. I use the laptop for visit a social networks (ex: twitter).				
Group III (Entertainment)	**Strongly agree**	**Agree**	**Disagree**	**Strongly disagree**
1. I use the laptop for chatting on online.				
2. I use the laptop for watching movies.				
3. I use the laptop for listening to music.				
4. I use the laptop for watching videos.				
5. I use the laptop for viewing images.				
6. I use the laptop for playing games.				
7. I use the laptop for designing purpose (Paint, Photoshop, etc.).				

SIGNATURE

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Data Integrity Techniques of Private Verification in Outsourcing Storage

[1]Senthil Kumari P, [2] Nadira Banu Kamal A. R.
[1] Ph.D Scholar, Alagappa University, Karaikudi
[2] Principal, Mohamed Sathak Hamid College of Arts and Science for Women, Ramanathapuram

ABSTRACT *Cloud computing has become predominant entity in today's world. Cloud computing acts as the fifth utility for providing computing after the other four utilities. Even though Cloud computing has a lot of advantages, any Cloud computing research work addresses three critical problems namely security, integrity and privacy. The privacy problem can be preserved by encryption followed by the search process. The challenging scenario in the cloud computing field is the data contribution from several data owners and search process by several users. This paper proposes Key Derivation Policy (KDP) followed by two integrity verification proofs: Proof of Data Possession (PDP) and Proof of Retrievability (PoR). Secret key derivation policy is common to both the integrity proofs. PDP is verified by Message Authentication Code (MAC). PoR is verified by MD5. Second integrity verification technique also includes two client authentication techniques: One Time Password (OTP) and dynamic missing number puzzle. The major advantage of the proposed approach is that the integrity verification is done by private auditing, that is the data consumer itself. Need for Third Party Auditing (TPA) which is trusted partially is eliminated.*
Keywords:*Proof of Data Possession (PDP), cloud computing, data integrity, One Time Password (OTP) and dynamic missing number puzzle.*

Introduction

Cloud computing provides several services among which "Storage as a service" is an important service. The desirable properties of Cloud computing such as scalability, elasticity, fault-tolerance and pay-per-use make it as an unavoidable commercial trend. Service Oriented Architecture, autonomic and utility computing, widespread adoption of hardware virtualization, low cost computers and storage devices and availability of high capacity networks have made the enormous development in cloud computing. The confidentiality of the patient's Personal Health Records should be preserved. Current data security approaches focus only on data security in which cryptographic solutions are followed by the random key generation processes. But, the prevailing security technique suffers from minimum data integrity. Loss of key in the conventional cryptographic techniques crash the original data provided by the data owner. Fig.1 shows the system model of the integrity verification technique.

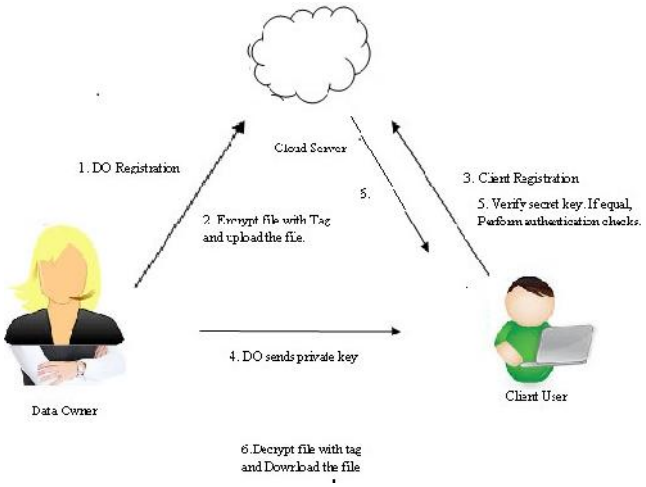


Fig. 1. System Model of Integrity Verification Technique

Downloading the entire file for its integrity verification is not an advisable solution due to the I/O and transmission cost expensiveness across the network. Access rights are given to some users and forbidden for other users. This process is called as "access control". Secure, scalable access control in cloud computing is achieved by "Attribute Based Encryption (ABE)". "One-to-many" encryption service is provided. One encrypted file can be decrypted by multiple users whose attributes conform to the access policy. The most suitable technology for data access control is Cipher text Policy ABE, because the data owner is given more direct control on access policies and the policy checking is performed "inside the cryptography". CP-ABE becomes expensive because of attribute revocation problem. Re-keying and re-encryption operations are performed by the data owner to avoid the revoked user from accessing the future data.

A Key-Policy based ABE is based on the association of attributes and decryption keys of the user. In the KP-ABE scheme, a cipher text relates to the set of attributes. The decryption key of the user is associated with a monotonic tree access structure. The user can decrypt the cipher text, only when the user attribute related with the cipher text satisfies the tree access structure. Hierarchical Attribute Set Based Encryption (HASBE) technique increases the efficiency of user revocation in multiple value assignment environments. The key escrow problem induced in HASBE technique is

considered by Multi Authority Attribute Based Encryption (MA-ABE). The shared files between the data owner and user have hierarchical structure. A number of hierarchy subgroups located at different access levels share a group of files. The storage cost of cipher text and time cost of encryption could be saved, if the files in the same hierarchical structure could be encrypted by an integrated access structure. Three security control mechanisms such as authentication, encryption and data verification technique are incorporated into a single, stand - alone system in this paper. The main drawback of the ABE technique is the increase in the computational cost for key generation and encryption. To overcome this problem, this paper proposes a secret key derivation process and MAC and MD5 verification to ensure data integrity and authentication techniques to verify data security in the cloud services.

The novel contributions of the proposed efficient key derivation policy and integrity verification techniques are listed as

- Effective users' addition and removal are done by multi-attributes-based secret key generation process.
- Capability to solve the simultaneous multi-users/ keys handling problem.
- Time complexity is reduced by the hash-based mapping and the attributes decomposition-based secret key generation and the secure data transfer level is improved.
- Integrity is verified by MAC and MD5.

The rest of the paper is structured as follows: Section II includes the existing works related to the ABE encryption techniques, integrity verification techniques, one time password and software puzzle authentication techniques for the cloud computing applications. Section III describes the detailed description of the proposed integrity verification techniques with key derivation process and authentication techniques. Section IV illustrates the results of the proposed technique and section V presents the conclusion and future work of this paper.

Related Works

This section describes the conventional encyrption techniques for the cloud computing applications. A scalable, flexible and fine-grained access control of the outsourced data, Hierararchical Attribute-Set-Based Encryption (HASBE) is proposed by Wan et al. [1]. The access control of the outsourced data in the cloud computing was efficient and flexible in the proposed scheme. A hierarchical encryption scheme combining the identity based encryption and cipher text policybased encryption systems was proposed by Wang et al. to achieve fine-grained access control [2]. Proxy and lazy re-encryption techniques are used efficiently to revoke the access rights of the users in the proposed scheme. A revocable Identity Based Encryption (IBE) was proposed by Li et al. to deploy a hybrid private key for each user [3]. The key generation complexity was reduced in the proposed scheme, while improving the efficiency and security.

The cipher text is encrypted with a tree access policy chosen by a data owner, while the corresponding decryption key is created with respect to a set of attributes in a CP-ABE scheme [4]. The set of attributes associated with a decryption key should satisfy the tree access policy associated with a given cipher text. Great flexibility in access control was provided in Attribute Set Based Encryption (ASBE) which enforces dynamic constraints on combining attributes to satisfy a policy [5]. The user revocation problem was solved efficiently in ASBE by assigning multiple values to the same attribute, which is not possible in CP-ABE. The leakage of data was eliminated by using a Two Round Searchable Encryption (TRSE) [6]. Top-k multi-keyword retrieval was supported by this approach.

The following paragraphs describe the conventional integrity verification techniques. An efficient and secure publlic verification of data integrity scheme was proposed by Yuan Zhang et al. to protect against external adversaries and malicious auditors [7]. A random masking technique was adopted to protect against external adversaries in the proposed scheme and required users to audit auditors' behaviors to prevent malicious auditors from fabricating verification results. A theoretical framework for the design of PORs was proposed by Bowers. K. D et al [8]. PoR is a compact proof by a file system (prover) to a client (verifier) that a target file F was intact, so that a client can fully recover it.

Different variants of PoR problems such as knowledge-soundness vs. information-soundness, bounded-use vs. unbounded-use are developed by Dodis et al., and gave nearly optimal PoR schemes for each of these variants [9]. Their contributions either generalized the prior PoR constructions or gave the known PoR schemes with the required properties. An efficient Key Derivation Policy (KDP) was proposed by Senthil Kumari et al. for enhancing data security and integrity in the cloud [10]. In the proposed method, local key was generated from the data attributes, i.e., file attributes. Then private key and secret key are generated. MAC verification process was used to validate the data integrity. A survey of techniques and tools used for cloud data integrity verification was presented by Princelly Jesu. A et al. [11]. Depending on the type of data and its size, the method selection was performed.

An application based on Hadoop and MapReduce framework was implemented by Rajat Saxena et al. [12]. Variant of the Paillier homomorphic cryptography system with homomorphic tag and combinatorial batch codes were the building blocks of their technique. A novel integrity auditing scheme for cloud data sharing services characterized by multi-user modification, public auditing and practical computational / communication auditing performance was proposed by Vedire Ajayani et al. [13]. User impersonation attack can be resisted in their scheme, which was not considered in existing techniques that supported multi-user modification. In the case of the untrusted server, a scheme which made use of Merkle Hash Tree (MHT) and AES algorithm to maintain data integrity was proposed by Poonam M. Pardeshi et al. [14]. The proposed scheme used Third Party Auditor (TPA) which acted on behalf of client for data integrity checking and sent an alert to notify the status of the stored data.

The following paragraphs describe the existing one time password technique for the cloud computing applications. A novel OTP algorithm was proposed by Hoyul Choi et al. and compared it with the existing algorithms

[15]. The proposed scheme was secure against MITM (Man-in-the-Middle) attack and MITPC/Phone (Man-in-the-PC/Phone) attack by using a CAPTCHA image. An adversary could know a valid OTP value and be authenticated with this secret information in the presence of those attacks. A framework model to authenticate cloud users in secured way using One Time Password (OTP) was proposed by Boopathy et al. [16].

A secure and accessible multimodal authentication method was proposed by Kristin S. et al. that used a one-time-password client installed on a mobile phone which allowed usage by people whose functional impairments adversely affected their ability to use existing solutions [17]. The method used a one-time-password (OTP) client installed on a mobile phone that replaced dedicated OTP generators. A "One Time Password" user authentication module along with Advanced Encryption Standard (AES) was proposed by Ramesh K et al. for encrypting the owners personal Health Record before uploading it onto the semi trusted cloud server [18].

The following paragraphs describe the existing software puzzle techniques for the cloud computing applications. A software puzzle scheme was proposed by Deepu et al. to deal with Denial of Service (DoS) attacks of certain types [19]. The algorithm was such that an attacker was unable to solve the puzzle in time. Qiao Yan et al. discussed the new trends and characteristics of DDoS attacks in cloud computing, and provided a comprehensive survey of defense mechanisms against DDoS attacks using Software Defined Networking (SDN) [20]. Their work can help to understand how to make full use of SDN's advantages to defeat DDoS attacks in cloud computing environments. A method to prevent DOS/ Distributed DOS attackers from inflating their challenge solving capabilities was presented by Pankaj Kumar. S et al. [21]. The proposed scheme used CPU generated puzzle which was efficient against the GPU (Graphics Processing Unit) generated puzzle.

Rupali Anil Suravase et al. conducted an exhaustive survey of the techniques used for defending the internet from DOS attacks [22]. They proposed a software puzzle scheme that randomly generated only after a client request was received at the server side that gave better performance as compared with the previous techniques. Shui Yu et al. proposed a dynamic resource allocation strategy to counter DDoS attacks against individual cloud customers [23]. When a DDoS attack occured, they employed the idle resources of the cloud to clone sufficient intrusion prevention servers for the victim. Yongdong Wu et al. introduced a new client puzzle referred to as software puzzle [24]. Unlike the existing client puzzle schemes, which published their puzzle algorithms in advance, a puzzle algorithm in the present software puzzle scheme was randomly generated only after a client request was received at the server side.

By using the key, the authenticity of data will be protected and only the one who has the key can check the data authenticity and integrity. In order to overcome the data security and integrity problems, this paper proposes a secret key generation process and two integrity verification techniques to ensure the data security and integrity in the cloud.

SECRET KEY GENERATION PROCESS AND INTEGRITY TECHNIQUES

This section describes the secret key generation process which is common to both the integrity techniques. This paper mainly focuses on the key security for the outsourced data in the cloud servers along with the integrity techniques. The secret key generation algorithm provides secure access control mechanism and data access policies.

3.1. Secret Key Generation Process

The extraction of the data and user attributes are performed initially. Then, any two file attributes are randomly selected. The AND operation is performed on the selected attributes. The resultant value of the AND operation is the local key. The exclusive OR (XOR) operation is performed with the local key and user attribute and a private key is generated. Then, the Hashing operation is performed to convert the private key into a secret key. When the users need to retrieve data, their request is transferred to the data owner. The data owner sends the secret key directly to the user. Using this secret key, the user can decrypt the cipher text obtained from the cloud, to get the original plain text.

3.2. ABE Process

The ABE process defines an access control policy, so that the user can access the data, if the data attributes and user attributes satisfy the defined policy. The attribute-set-based encryption prevents the combination of the attributes across the multiple sets. The user access control policy is defined on the basis of corresponding attributes of the users. Initially, the users have to send the request to the key authority, to access a service. The key authority performs computation using the user attributes and private key issues. Then, the key authority issues the public parameter to the service provider, to encrypt the service response. The key security for the message transmission is computed to ensure the security and integrity.

3.3. Secret Key Generation Algorithm

The key generation algorithm depends upon the two issues such as a secret key and the attributes of the user. Let file name (fname) and file extension (fext) be the two data attributes. Local key L_K is generated by the intersection of fname and fext. User name is used as the third attribute (uname).

The private key P_K is generated by using the Ex-or operation of the L_K and un. The secret key S_K is generated by hashing the private key P_K. The cost function is performed using the secret key S_K and the selected file. Finally, the secret key is used for encryption. Following algorithm illustrates the secret key generation process.

Secret Key Generation Algorithm

Input: File Extension Fext, File name Fn, User name un
Output: K_E = KeyGen (Hash (P_K))
Step 1: Start
Step 2: Generate L_k = LocalKeyGen (Fn **intersect** Fext**)**
Step 3: Generate P_K = PrivateKeyGen (un **Exor** L_k)
Step 4: K_E = KeyGen (Hash (P_K))
Step 5: CF = Encrypt (K_E, F)
Step 6: Data Owner STORES CF into cloud
Step 7: Client REQUESTS for File Download
Step 8: DF=Decrypt (K_E, F)
Step 9: Stop

3.4. Integrity Verification Techniques

The secret key generation process is common to both the integrity verification techniques. First, integrity is verified by MAC. Second, integrity is verified by MD5. This sub-section describes both the integrity techniques.

3.4.1. MAC Verification Technique

MAC verifies the data integrity by using a secret key shared between the data owner and user. MAC standard defines the cryptographic checksum, which is obtained from passing the data through a message authentication algorithm along with the user attributes. The MAC utilizes a session key to detect both accidental and intentional data modifications. The outsourced data file F consists of a finite ordered set of blocks S_1, S_2….S_m. A direct way to ensure data integrity is pre-computation of the MACs for the entire data file. The data owner pre-computes MACs of the file using a set of secret keys and stores them locally, before data outsourcing. For each time during the auditing process, the data owner reveals a secret key to the cloud server and requests for a fresh keyed MAC for verification. MAC verification process enables high data integrity, since it covers all data blocks.

The algorithm to implement MAC verification is as follows:

MAC Verification

Input: Key(K_E), message (M)
Output: Hash Concatenation
Step 1: Start.
Step 2: Check the size of keys greater than the block size.
Step 3: Calculate the hash function, otherwise add the zero pads to the hash function.
Step 4: Calculate the underlying hash function for key sizes within the block and XOR function.
Step 5: Calculate the concatenated hash output.
Step 6: Stop.

MAC utilizes a session key and message to detect both concatenated data modifications in hash function. The data owner pre-computes MACs of the file using a set of secret keys and stores them locally, before data outsourcing. For each time during the auditing process, the data owner reveals a secret key to the cloud server and requests for a fresh keyed MAC for verification. MAC verification process enables high data integrity, since it covers all data blocks.

For a large file, downloading and calculating the MAC of the file is an overwhelming process and takes a lot of time. Also, it is not practical since it consumes more bandwidth. Therefore, there is a need for using a lighter technique, which is calculating the hashing value. The advantage of MD5 integrity verification is that it is simple and implementable.

3.4.2. MD5 Verification Technique

The simplest form of proof of retrievability takes the hash of block using a keyed hash function. Data Owner uses MD5 Cryptographic hash function of the private key as the secret key. The data owner encrypts the file with the secret key and sends the file to a remote server. When the data owner needs to check his data retrievability, he sends his key and asks the server to send the hash values by using his key in order to compare them with the hash values that data owner has. This method is used by the data owner for private auditing.

In the second method, Data Owner sends the secret key to Data User (client). When the client needs to check his data retrievability, he sends the secret key and compare it with the hash value stored in the data base. If they are the same, the client is authenticated to download the file which he wants from the Cloud Service Provider (CSP) which gives Storage as a Service. This method is used by the data user for private auditing for integrity verification. Data User is given with further authentication checks in order to verify the security. Client is performing One Time Password authentication technique. Overall MD5 integrity verification technique is illustrated in Figure 2.

3.4.2.1. One Time Password Authentication Technique

Registration process should be done by both the Data Owner and Data User separately in the Cloud Service Provider's Cloud Repository. All the details of the data owner and Client including their mobile number and email address are collected in the cloud data repository. Separate login and control panel user id and password are given for

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

both the data owner and client respectively. Data owner's user name of the control panel is taken into consideration for or operation with the local key and private key is generated. MD5 hash of the private key is considered as the secret key. It is given to the authorized client by the data owner. In the Key integrity checking, if the secret key values are the same, then the client is further checked by One Time Password CAPTCHA code which is sent to the client's mobile number. If the valid OTP Code is entered by the client, then he is further checked by the dynamic missing number puzzle authentication technique before he is given valid permission to download the file which is uploaded by the data owner.

3.4.2.2. Dynamic Missing Number Puzzle Authentication Technique

Dynamically random numbers are generated for a 3 x 3 matrix. This original table is sent to the client's email address. If the correct CAPTCHA code is entered by the client, he is given with the dummy table with three number positions as empty. The client has to verify the original table which is submitted by the cloud server to client's email address. Missing numbers should be entered by the client correctly. If the numbers are entered correctly, then the client can download the file. The steps are summarized below.

1. Initialize 3x3 Random Array by using CPU Host Techniques.
2. Start and generate Random number of initialized Array Values
3. Compare The Generated Value with the existing one. If already existed, call regenerate a new random number.
4. Then Identity the Current Missing Table Identity and Generate Dummy and Original.
5. Call SMTP Protocol and send Mail to valid customer ID and Keep the original puzzle table at Data Base.

RESULTS AND DISCUSSION

This section presents the results obtained when we execute our research work.

4.1. MAC VERIFICATION

The proposed research work is implemented using Java Swing as a front end tool, IDE – Net Beans 7.1, Back end – Wamp Server in Core 2 duo system with hard disk capacity 500 GB. This section presents the comparative analysis of the performance parameters such as computational time, computational overhead, average time to derive the keys and average time to generate the keys of the proposed KDP with the existing techniques CP-ABE, EPPDR, pseudo random key generation and subset cover.

4.1.1. Encryption Time

The time required to complete the encryption process is termed as computational time. When the number of attributes involved in the process increases, it increases the encryption time. The time for encryption increases to the maximum value in traditional CP-ABE methods. The proposed KDP provides the minimum time required for the encryption process for different number of attributes which is shown in Figure 2.

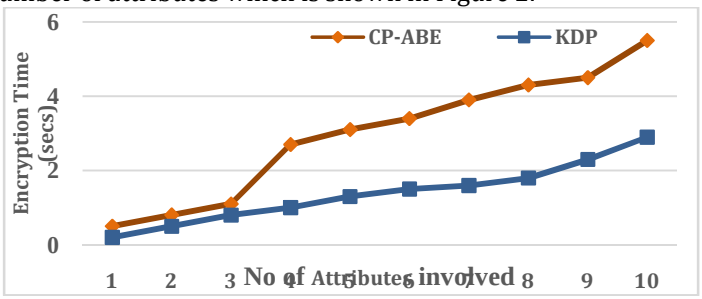


Figure 2. Encryption Time Vs. No. of Attributes

4.1.2. Computational Overhead

The measure of the capability of the network to withstand the emulation attackers is called the computational overhead. When the number of attackers increases, the overhead is limited to achieve the authentication. The proposed KDP algorithm provides the minimum overhead compared to existing EPPDR approach which is shown in Figure 3.

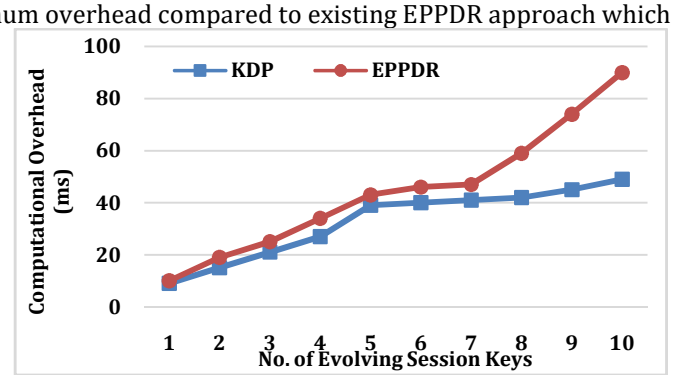


Figure 3. Computational Overhead Vs. No. of Evolving Session Keys

4.1.3. Average Lifetime to Derive Keys

The life time is the important parameter in the design of the network. The speed of the packet transmission depends upon the life time to derive the keys of the data transmission when the network is in high traffic. The interval for key update increases, then the average lifetime to derive the keys is computed using the existing pseudo random key generation algorithm and the proposed KDP algorithm. The simulation results confirm the effective increase in the

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

lifetime. The interval for updating process is more than the average lifetime for derivation of keys. The proposed KDP method provides the minimum average lifetime compared to the pseudo random key generation algorithm which is shown in Figure 4.

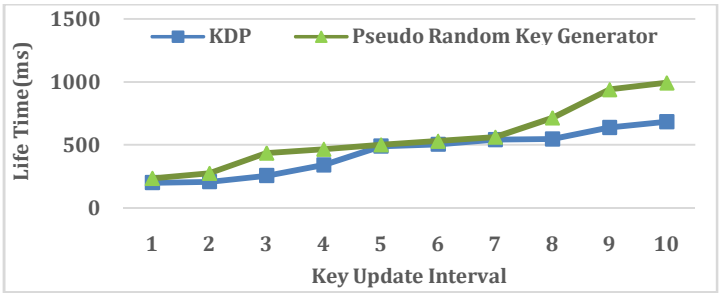


Fig.4. Average lifetime Vs. Key Update interval

4.1.4. Average Lifetime to Generate Keys

The key is the important parameter in the design of the network. The time to generate the keys depends upon the key update interval. The interval for key update increases, then the average lifetime to generate keys is computed using the subset cover and the proposed KDP algorithm. The simulation results confirm the effective increase in the average lifetime. The interval for updating process is more than the average lifetime for derivation of keys. But, using proposed KDP algorithm provides the minimum average lifetime compared to subset cover which is shown in Figure 5.

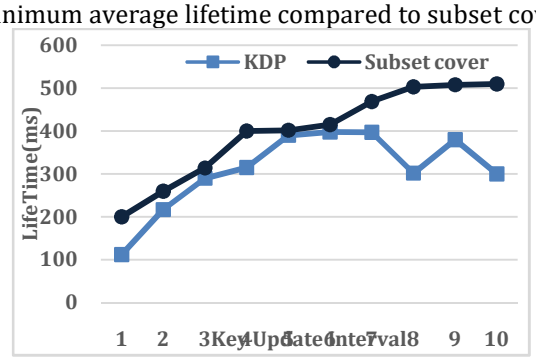


Fig. 5. Average lifetime Vs. Key Update Interval

4.2. MD5 VERIFICATION

The proposed research work is implemented using Microsoft Visual C# as a programming tool, IDE – ASP .Net, Back end – Microsoft SQL Server 2005 in Core 2 duo system with hard disk capacity 500 GB.

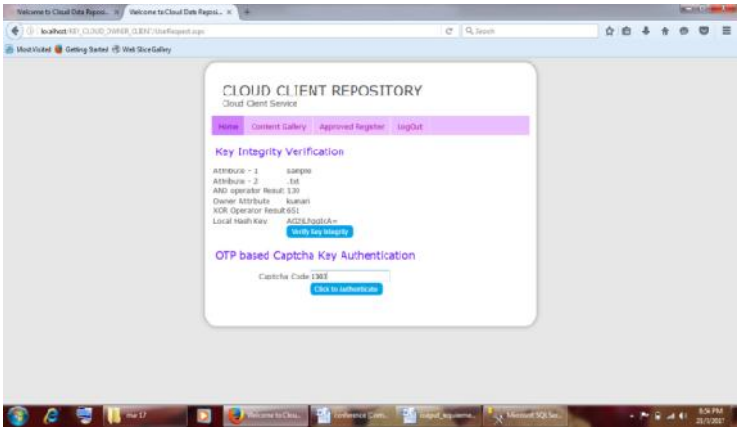


Fig. 6. OTP Authentication

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

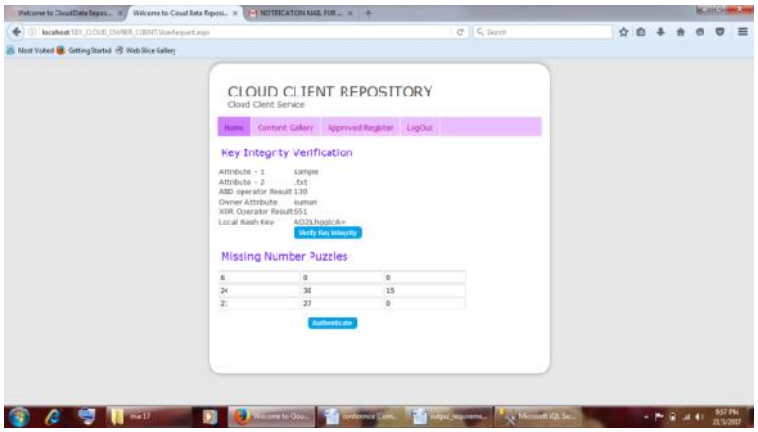


Fig. 7. Dynamic Missing Number Puzzle

If the data owner with data attributes file name "sample" and file extension ".txt" with user name "kumari" uploads the file "sample.txt", then we get the AND-XOR values. If the client wants to download the file sample.txt, then he gets the values which are sent by the data owner. If the client clicks Verify Key Integrity, then he gets OTP code to be entered in his mobile when the local hash key values are the same. It is depicted in Fig. 6. If the correct OTP code is entered, then the next authentication technique dynamic missing number puzzle is given as in Fig. 7. If the missing numbers are entered by the client from his email address which has original values, then the client can download the file "sample.txt".

If the client click "Authenticate", then he gets the next page which asks him whether he wants to open or save the downloaded file. If he selects "save" option, then encrypted file will be decrypted and the client can read the plain text which is originally uploaded by the data owner.

CONCLUSION AND FUTURE WORK

In this paper, secret key generation process and integrity techniques are proposed to ensure high data security and integrity in the cloud for secure data transmission. However, there are many disadvantages such that the data owner needs to store many keys in order to use one each time. Also, the number of checking is limited by the number of keys since the remote server could store all keys and the hash values and use them when it is asked to prove having that file. Processing overhead depends on the data consumer either the data owner or the data user. Hence, the future work shall be extended to remove all the drawbacks.

REFERENCES

1. Z. Wan, J. e. Liu, and R. H. Deng, "HASBE: a hierarchical attribute-based solution for flexible and scalable access control in cloud computing, "*IEEE Transactions on Information Forensics and Security",* Vol. 7, pp. 743-754, 2012.
2. G. Wang, Q. Liu, J. Wu, and M. Guo, "Hierarchical attribute-based encryption and scalable user revocation for sharing data in cloud servers," computers & security, vol. 30, pp. 320-331, 2011.
3. J. Li, X. Chen, C. Jia, and W. Lou, "Identity-based encryption with outsourced revocation in cloud computing", IEEE Transactions on Computers, 64(2), pp. 425-37, 2013.
4. J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy Attribute Based Encryption," in *Proc. IEEE Symp. Security and Privacy*, Oakland, CA, 2007.
5. R. Bobba, H. Khurana, and M. Prabhakaran, "Attribute-sets: A practically motivated enhancement to attribute-based encryption," in *Proc. ESORICS*, Saint Malo, France, 2009.
6. Jiadi Yu, Peng Lu, Yanmin Zhu, Guangtao Xue and Minglu Li,"Toward Secure MultiKeyword Top-k Retrieval over Encrypted Cloud Data", IEEE Transactions on Dependable and Secure Computing, Vol. 10, No. 4, July/August 2013, pp. 239-250.
7. Yuan Zhang, Chunxiang Xu, Hongwei Li and Xiaohui Liang," Cryptographic Public Verification of Data Integrity for Cloud Storage Systems", IEEE Cloud Computing published by the IEEE computer society, September/October 2016, pp.44-52.
8. Bowers K. D., A. Juels, and A. Oprea, "Proofs of retrievability: theory and implementation ", in *CCSW '09: Proceedings of the 2009 ACM workshop on Cloud computing security*, (New York, NY, USA), ACM, 2009, pp. 43–54.
9. Dodis Y., S. Vadhan, and D. Wichs, "Proofs of retrievability via hardness amplification," in *TCC '09: Proceedings of the 6th Theory of Cryptography Conference on Theory of Cryptography*, (Berlin, Heidelberg), Springer-Verlag, Mar - 2009, pp. 109–127.
10. Senthil Kumari. P and Nadira Banu Kamal.Dr. A. R, "Key Derivation Policy for Data Security and Data Integrity in Cloud Computing", Journal of Automatic Control and Computer Science, Vol. 50(3), ed : Springer,DOI:10.3103/S0146411616030032, July 2016, pp. 165–178.
11. Princelly Jesu. A and Ramesh Kumar. S, "A Survey of Techniques and Tools Used For Cloud Data Integrity Verification", *International Journal of Innovative Research in Computer and Communication Engineering* , Vol. 4(2), Feb 2016, pp.1923-1928.

12. Rajat Saxena and Somnath Dey, "Cloud Audit: A Data Integrity Verification Approach for Cloud Computing", Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016), Procedia Computer Science 89 (2016), pp. 142 – 151.
13. Vedire Ajayani, K. Tulasi and Dr P. Sunitha, " Public Integrity Auditing for Shared Dynamic Cloud Data with Group User Revocation", *International Journal of Advanced Technology and Innovative Research,* Vol. 8(16), Oct-2016, pp. 3146-3152 .
14. Poonam M. Pardeshi and Deepali R. Borade, "Improving Data Integrity for Data Storage Security in Cloud Computing", *IJCSNS International Journal of Computer Science and Network Security*, Vol. 15(6), June 2015, pp.75-82.
15. Hoyul Choi, Hyunsoo Kwon and Junbeom Hur, "A Secure OTP Algorithm Using a Smartphone Application", IEEE Seventh International Conference on Ubiquitous and Future Networks ICUFN Aug - 2015, pp. 476-481.
16. Boopathy D and M. Sundaresan, " Framework Model and Algorithm of Request based One Time Passkey (ROTP) Mechanism to Authenticate Cloud Users in Secured Way", *Third International Conference on Computing for Sustainable Global Development (INDIACom ,*Mar - 2016, pp.3898 - 3903.
17. Kristin S. Fuglerud and Oystein Dale, "Secure and Inclusive Authentication with a Talking Mobile One-Time-Password Client", IEEE Security & Privacy, co-published by IEEE Computer and Reliability Societies, March/April 2011, pp.27-34.
18. Ramesh K and Ramesh S, "Implementing One Time Password Based Security Mechanism for Securing Personal Health Records in Cloud", *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT),* pp. 968-972.
19. Deepu. S. D, Dr. Ramakrishna. M.V, " Software Puzzle Counterstrike for Denial of Service Attack ", *International Journal of Innovative Research in Computer and Communication Engineering",* Vol. 4(4), April 2016, pp. 8061-8065.
20. Qiao Yan, F. Richard Yu, Qingxiang Gong and Jianqiang Li, " Software-Defined Networking (SDN) and Distributed Denial of Service (DDoS) Attacks in Cloud Computing Environments: A Survey, Some Research Issues, and Challenges", IEEE Communications Surveys & Tutorials, Vol. 18(1), First Quarter 2016, pp. 602-622.
21. Pankaj Kumar, S. S. Ahluwalia and Tharun Kumar S. V. , " Using Software Puzzle for Reducing DDos / Dos Cost on SSL / TLS ", *International Journal of Computer Applications,*Vol. 151 (4), Oct-2016, pp. 23 - 27.
22. Rupali Anil Suravase. Miss and Mr. Pramod Patil, " Survey on: Software Puzzle for Offsetting DoS Attack", *International Journal on Recent and Innovation Trends in Computing and Communication,* Vol. 4(5), May - 2016, pp. 413 - 415.
23. Shui Yu, Yonghong Tian, Song Guo and Dapeng Oliver Wu, "Can We Beat DDoS Attacks in Clouds?", IEEE Transactions on Parallel and Distributed Systems, Vol. 25(9), Sept - 2014, pp. 2245 - 2254.
24. Yongdong Wu, Zhigang Zhao, Feng Bao, and Robert H. Deng, "Software Puzzle: A Countermeasure to Resource-Inflated Denial-of Service Attacks", IEEE Transactions on Information Forensics and Security, Vol. 10(1), Jan 2015, pp.168-177.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Computational tools used for Macromolecular DNA Nanotechnologywith Molecular Docking

[1] Kiruba Nesamalar E, [2] Chandran C.P

[#1]Research Scholar, Ph.D - Category B, Research & Development Centre, Bharathiyar University, Coimbatore

*[*2] Associate Professorof Computer Science, AyyaNadarJanaki Ammal College (Autonomous), Sivakasi*

*[*2] Corresponding Author*

[1]kirubanesamalar@gmail.com

[2]drcpchandran@gmail.com

ABSTRACT *DNA nanotechnology is a branch of nanotechnology concerned with the design, study and application of synthetic structures based on DNA. DNA nanotechnology takes advantage of the physical and chemical properties of DNA rather than the genetic information it carries.It is notoriously difficult to observe, let alone control, the position and orientation of molecules due to their small size and the constant thermal fluctuations that they experience in solution. Molecular self-assembly with DNA enables building custom-shaped nanometer-scale objects with molecular weights up to the megadalton regime. It provides a viable route for placing molecules and constraining their fluctuations in user-defined ways, thereby opening up completely new avenues for scientific and technological exploration. Here, we review progress that has been made in recent years toward the state of an enabled DNA nanotechnology. DNA nanostructures must be rationally designed so that the individual nucleic acid strands will assemble into the desired structures. This process usually has two steps Motif Design: To design the secondary structrure, there are different approaches: Tile-based structures , Folding structures, Dynamic assembly, Sequence design. This step has similar goals to protein design. The sequence is designed to favor the desired target structure and to disfavor other structures. In this step is important the symmetric minimitzation inside the sequence*

***Keywords**: Mathematics; Teaching Method; Performance; Students' variables; Teachers' variables.*

I. INTRODUCTION

A. Macromolecular DNA

The term macromolecule turned into coined by way of Nobel laureate Hermann Staudinger in the 1920s, although his first relevant e-book in this subject handiest mentions high molecular compounds(in excess of one,000 atoms). At that point the phrase polymer, as brought by Berzelius in 1833, had a different meaning from that of today: it genuinely become another form of isomerism for instance withbenzene and acetylene and had little to do with length.[1]

DNA stands for deoxyribonucleic acid. DNA is gift within the cells of each dwelling element. It carries the chemical instructions and genetic facts to assist organisms expand and function.DNA is simplest nanometers across, but if you could get to the bottom of all the strands from just one mobile and line them up cease to give up, you'd have a thread meters long! Most of our DNA—ninety nine.5%—is similar to every other person's, however a small quantity is particular. Simplest equal twins have exactly the same DNA as any other man or woman. The human genome (all of our genetic material) is contained in 46 lengthy, thin "threads" called chromosomes. Every person inherits 23 chromosomes from each of our dad and mom, for a complete of 46 chromosomes. Human DNA is fabricated from two lengthy strands that are twisted together, in a shape referred to as a double helix. It looks like an extended, spiral ladder.Fig.1 indicates the Double helix of DNA.

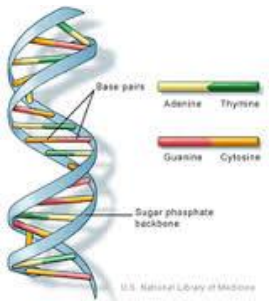


Fig.1. DNA Double helix

The sides of the ladder are fabricated from devices known as nucleotides. The rungs of the ladder are of fabricated from 4 different varieties of molecules, known as chemical base pairs:

- Adenine (A) and Thymine (T)
- Guanine (G) and Cytosine (C)

The base pairs constantly be a part of within the identical manner. A and T always be part of collectively, and G and C usually be a part of together. Humans have over three billion chemical base pairs in our DNA. The commands that

assist our our bodies develop and stay are carried within the collection of the bottom pairs, and are known as genes [2]. Our genes carries the chemical instructions to make the tens of hundreds of various proteins that our bodies need in an effort to develop and characteristic. To construct a brand new protein, the unique gene that is responsible unwinds from its chromosome. A single-stranded reproduction of its instructions is made. This reproduction is called RNA, short for ribonucleic acid. RNA leaves the nucleus of the cellular to discover ribosomes, which assemble proteins, and amino acids, that are the building blocks of proteins. RNA presents a template for the protein production, teaching the ribosomes to enroll in the amino acids in a selected order [3].

B.DNA Nanotechnology

In autumn 1980, the crystallographer Nadrian C. Seeman became in a campus pub, when he noticed Escher's depth, a woodcut which represents a lattice. This photo inspired him an idea, that changed into a new approach to achieve the shape by using X-Ray diffraction with out the complex crystallization manner [4]. This new technique consisted in a 3-d DNA lattice which may be used to orient difficult-to-crystallize molecules. Considering the fact that this first touch with the DNA as a cloth till in recent times, some researchers had been running with this biomolecule and that they carried out terrific results, as an instance DNA Origami or 2d lattice [5].

Nano is the medical term meaning one-billionth (1/1,000,000,000) It comes from a Greek phrase meaning "dwarf." A nanometer is one one-billionth of a meter. One inch equals 25.Four million nanometers. A sheet of paper is about a hundred,000nanometers thick. A human hair measures roughly 50,000 to a hundred,000nanometers across. Your fingernails develop one nanometer every 2nd. (other gadgets can also be divided by a billion [6]. A unmarried blink of an eye is ready one-billionth of a yr. An eyeblink is to a yr what a nanometer is to a yardstick.) Nanoscale refers to measurements of 1 – one hundred nanometers. A pandemic is ready 70 nm lengthy. A mobile membrane is about nine nm thick. Ten hydrogen atoms are approximately 1 nm. On the nanoscale, many commonplace materials showcase unusual residences, such as remarkably decrease resistance to strength, or quicker chemical reactions. Nanotechnology is the manipulation of cloth at the nanoscale to take gain of these houses. This often way running with person molecules. [7] Nanoscience, nanoengineering and different such phrases discuss with the ones sports applied to the nanoscale. "Nano," by itself, is frequently used as quick-hand to refer to all or any of these sports.

DNA nanotechnology uses DNA and different nucleic acids as structural materials, taking benefit of their ability to self-assemble as a way to create new, nanoscale systems. Packages of DNA nanotechnology encompass DNA computing, where DNA arrays perform computation as they bring together, and nanoarchitecture, wherein DNA is used as a template to collect different molecules [8].

In nanotechnology, self-assembly is predicated on chemical and bodily forces to guide molecules into arranging themselves into predictable, ordered structures. The final shape is determined through properties of the molecules which might be used, in addition to the controlled environment wherein they grow[9].

Self-meeting is a "backside up" manufacturing method, which means that it starts offevolved with small pieces and builds them up to a larger shape. This is in contrast to "top down" strategies, in which larger blocks of substances are pared all the way down to create a smaller structure. DNA is an critical material in self-assembly research for many reasons. Researchers now have a terrific knowledge of ways DNA bureaucracy; they are able to predict the shape that a given molecule of DNA will absorb answer; and they could without difficulty get DNA molecules of any series they need [10].

Paul Rothemund at the California Institute of technology is getting to know the ability of DNA to help build smaller, quicker computer systems and different examples of nanotechnology. Collectively with a group of different researchers, he is operating on a mission to use organic substances together with DNA to create nonbiological shapes and patterns proven in Fig.2 [11].

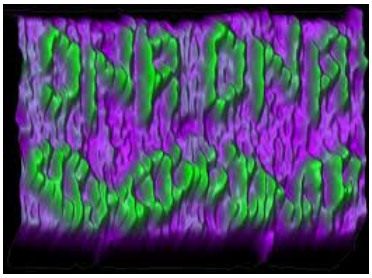


Fig.2. DNA Shapes

B. DNA Origami

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Rothemund designed and self-assembled nanoscale shapes and styles crafted from DNA. Examples include a smiley face, a map of North the united states, snowflakes, and a picture of a double helix DNA strand. To create these shapes and patterns, Rothemund used a process he calls "DNA origami." Is proven in Fig.3 an extended, single-stranded genome changed into used as a "scaffold." Many quick strands of synthetic DNA, or "staples," have been used to fold the lengthy strand right into a shape, along with a smiley face. Rothemund then used staple strands to beautify some of these shapes with patterns, including the map of North the us [12].

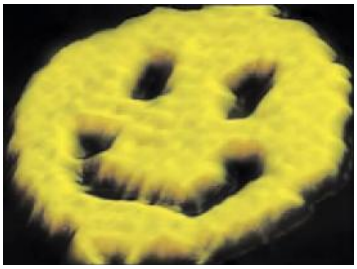


Fig.3 DNA Origami

To create the smiley face Rothemund first designed the geometrical shape, then discovered how the scaffold strand will be folded in stacks to fill inside the form. Then, the usage of a pc, he discovered the sequence of short staples important to fold the scaffold within the right places. Each staple has a left 1/2 that joins the long strand in a specific region and a right half of that joins it similarly down the strand. Subsequent, long, scaffold strands and quick, staple strands of DNA have been blended in a buffer answer. The answer become heated to almost boiling, then cooled. Because the combination cooled, the fast strands bonded to the lengthy strands, folding them up into tiny, 100 nm shapes. The end result became 50 billion DNA smiley faces floating round in unmarried drop of answer! About 72% of the faces were nicely formed. Fig.Four. Indicates the developing a one hundred nm smiley face with the system of DNA origami [13].

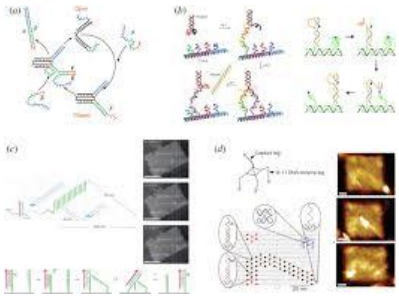


Fig.4. Creating a 100 nm smiley face with the process of DNA origami.

The patterned shapes Rothemund created with the DNA origami method have a spatial resolution of about 6 nm, and are composed of approximately two hundred character pixels. They are 10 instances greater complicated than formerly-created arbitrary patterns. DNA origami is confined by the duration of unmarried-stranded DNA that is available. On its personal, this method can't be used to create huge systems [14].

To extend things out of DNA, Rothemund is taking part with a scientist named Erik Winfree (additionally at Caltech) on a method referred to as algorithmic selfassembly of tiles. On this technique, rectangular tiles of DNA join together to make larger, -dimensional structures. DNA origami is used to start the technique. Rothemund predicts that during 5-10 years it will be possible to use a technique known as molecular programming to create circuits and computer memory. Further within the future, Rothemund envisions that molecular programming would possibly permit us to build technologies from the bottom up [15].

II. Databases

I) ISA-TAB Nano

To develop a specification that enables the import/export of data on nanomaterials and their characterizations to/from nanotechnology sources. Improvement of a trendy bendy sufficient to guide the numerous (100+) diverse characterization assays finished via the nanotechnology network – improvement of a preferred that could constitute the complicated systems of nanomaterials and their additives – identity of the minimal statistics required to reap crossmaterial comparison. Affords a fashionable tab-delimited format for describing data related to: Investigations, materials (Nanomaterials), studies (Specimens) Assays. Leverages and extends the research/look at/Assay (ISA-TAB) layout – general tab-delimited report format evolved via the eu Bioinformatics Institute (EBI) for representing a

selection of assays (e.G. MAGE-TAB) and era types. Supports ontology-based totally curation – Nanomaterials and ideas from the NanoParticle Ontology (NPO) as well as other ontologies. Desk I shows thenanoparticle associated ontology sources.

Table I List of nanoparticle-related ontology resources.

S.NO	Name	Description	URL
1.	Gene Ontology	The representation of gene and gene product attributes across species and databases	[16]
2.	Cancer Open Biomedical Resource	A tool for indexing cancer nanotechnology informatics knowledge	[17]
3.	Zebrafish Anatomy	A structured, controlled vocabulary of the anatomy and development of the zebrafish.	[18]
4.	Phenotype Quality ontology	Primarily defining composite phenotypes and phenotype annotation.	[19]
5.	BioPortal	BioPortal provides access to commonly used biomedical ontologies and tools for their analysis.	[20]

III. Molecular Modeling

The systems of nucleic acids and lots of proteins were determined with the aid of crystallography, nuclear magnetic resonance, electron microscopy, and lots of different strategies. Structural biology affords records at the static structures of biomolecules. However, in reality, biomolecules are fairly dynamic, and their motion is crucial to their characteristic. Extraordinary experimental strategies are to be had to help observe the dynamics of biomolecules. Computational power maintains to boom, and the development of latest theoretical methods gives wish of solving medical problems on the molecular degree.

All of the theoretical methods and computational strategies which are used to version the conduct of molecules are described as molecular modeling. Macromolecules may be studied only with molecular mechanics in view that different quantum strategies based totally at the Schrodinger equations which includes ab initio, semi-empirical, and density purposeful concept (DFT) strategies require good sized computational time. Molecular mechanics uses classical physics and is based on pressure subject.Desk II suggests the Molecular modelling and Visualization equipment.

Table II List of Molecular Modeling Visualization Tools

S.NO	Name	Operating System	URL
1.	Jmol	Windows/GNU Linux/Mac OS X	[21]
2.	RasMol	Windows/GNU Linux	[22]
3.	Raster3D	GNU Linux/Unix/Mac OS X	[23]
4.	PyMOL	Windows/GNU Linux	[24]
5.	Xeo	Windows/GNU Linux/Mac OS X	[25]

IV. Molecular Docking

Docking is an crucial computational tool inside the drug discovery technique and is used to especially expect protein-ligand interactions. The 2 simple functions of docking software are docking accuracy and scoring reliability. Docking accuracy suggests how comparable the prediction of ligand binding is to the ligand conformation that is determined experimentally, while scoring reliability ranks ligands based totally on their affinities. Docking accuracy and scoring reliability are used to assess the looking set of rules and the scoring capabilities, respectively, of docking software. The numerous searching algorithms utilized in docking software program vary with respect to randomness, velocity, and the location covered.

Maximum of the looking algorithms perform nicely when examined towards the acknowledged shape. In contrast, scoring functions are not often successful, and a number of issues had to be addressed to improve the docking functions. Many varieties of docking software program are presently available; AutoDock is a software that is quite accessed and freely available. At gift, as protein-nanoparticle complexes are hard to observe using experimental procedures, computational gear show promise. Structural models for carbon nanomaterials which include carbon nanotubes and fullerenes are to be had, and hence many protein-nanoparticle interactions may be studied computationally. Fullerenols, which are derivatives of fullerenes, are presently in trial for diagnostic and therapeutic uses, although inadequate data is to be had about the structural interactions and toxicity of fullerenols in biosystems.Desk III indicates the molecular docking tools.

Table III List of Molecular Docking Tools.

S.NO	Name	Operating System	URL
1.	Autodock 4	Windows/GNU Linux/Mac OS X	[26]
2.	Dock	Windows/GNU Linux/Mac OS X	[27]
3.	GOLD	Windows/GNU Linux	[28]
4.	FlexX	Windows/GNU Linux	[29]
5.	Surflex-dock	Windows/GNU Linux	[30]

V. Growth of Bioinformatics and Applications for Nanotechnology

Within the 1980s, it became obligatory to deposit all published DNA sequences in a valuable repository. At the equal time, a wellknown changed into followed by way of journals to make gene and protein sequences freely accessible to all and to compile all records inside the form of openly available databases. The open accessibility of DNA sequences in databases consisting of GenBank and the availability of protein systems in databases together with the Protein data bank have motivated many researchers to broaden powerful strategies, tools, and sources for huge-scale information analysis. Concurrently, boom in computational pace and memory garage capacity has brought about a new technology in the evaluation of biological records. Bioinformatics has emerged as a powerful field and nearly one thousand databases are presently publicly to be had at the side of a huge range of bioinformatics equipment. The boom and development of bioinformatics can offer treasured instructions for making use of the equal practice to gain nanoinformatics development.

VI. Specific Challenges and opportunities in DNA Nanotechnology

The databases and gear presented here highlight the growing sources that are to be had to customers. Their development is observed via the continuous expansion within the quantity of databases and various pc packages and the new database projects inclusive of ISA-TAB-Nano, caNanoLab, and Nanomaterial Registry will facilitate data sharing, records standards, and, relying on the boom of nanomaterials statistics, the development of methods and gear precise to the nanolevel. Furthermore, the growth of those fields basically calls for that the ISA-TAB-Nanowellknown be followed via journals and different businesses to ensure constant representation of nanotechnological information. It need to be noted that open accessibility and the liberty to use published DNA sequences in databases consisting of GenBank has recommended scientists to build effective strategies, tools, and resources which have substantially enriched the sphere of bioinformatics.

VII. INDIAN SCENARIO OF DNA NANOTECHNOLOGY

The ninth 5-12 months Plan (1998-2002) had stated for the first time that national centers and core agencies have been set up to sell studies in frontier regions of S&T which blanketed superconductivity, robotics, neurosciences and carbon and nano materials. Planning fee supported wide variety of such R&D programmes beneath fundamental research (GOI 1998).

In 2001-2002, the DST installation an professional institution on “Nanomaterials: science and devices”. The authorities identified the need to initiate a Nanomaterials technology and era undertaking (NSTM) within the tenth 5 year Plan (2002-07) after thinking of the trends in nanotechnology. A strategy paper changed into developed for helping on a protracted-time period foundation each simple studies and application oriented programmes in nanomaterials (DST 2001)

The 10th 5 year Plan (2002-2007) report diagnosed numerous regions for task mode programmes inclusive of era for bamboo products, pills and pharmaceutical studies, instrument development such as development of equipment and device, seismology, and also nano science and era (GOI 2002).

Nano technological know-how and era task (NSTM) changed into anticipated to offer preferred thrust to research and era development on this location (DST 2006). The eleventh five-12 months Plan (2007-2012) categorically mentioned projects to create excessive value and big effect on socio-monetary shipping involving nano cloth and nano devices in fitness and ailment. The generous 11th 5 yr Plan price range allocation of Rs. A thousand crore became earmarked for the Nano task when it turned into launched in 2007 (GOI, 2007).

VIII. GLOBAL TRENDS OF DNA NANOTECHNOLOGY

Global governments have launced many nanotechnologyspecific tasks/programmes to leverage the prospects of nanotechnology for social and financial gains. In 2005 itself, more than sixty two countries launched national nanotechnology-unique sports international over(Maclurcan 2005). In 2003, Taiwan launched its national technological know-how and generation Programme for Nanoscience and Nanotechnology. Taiwan has also evolved Nano Mark that is the world's first government-mounted machine for certifying nano-products. Globally, governments presently spend about US$ 10 billion in step with year on nanotechnology studies and improvement. Through the give up of 2011, the total government investment for nanotechnology research international turned into more than US$ sixty five billion, which is predicted to upward push to US$ a hundred billion with the aid of 2014. When figures for enterprise studies and various other varieties of non-public investment are taken into consideration, which had been notion to have passed government investment figures as a long way returned as 2004, it is expected that nearly a quarter of one trillion dollars will have been invested into nanotechnology via 2015 (Cientifica 2011). The nanotechnology marketplace prospects, as envisaged by means of various marketplace research our bodies, are forecasted to be very wonderful; Lux research, countrywide technology basis (NSF) and Cientifica projected these to be around one thousand billion dollar enterprise by way of 2015.

REFERENCES

[1] SergiMontané Bel, "Nanotechnology based DNA", UNB publications, Universal Autonoma de Barcelona, 2014.
[2] Calladine CR and Drew HR, "Understanding DNA", 2nd edn. San Diego, CA: Academic Press, 2007.
[3] National Center for Biotechnology Information (2001) http:// www.ncbi.nlm.gov:80/entrez/query.fcgi?db=Structure [type in ADNA, for example, for a list of A-DNA crystal structures].
[4] Ng H, Kopka ML and Dickerson RE, "The structure of a stable intermediate in the A-DNA helix transition", Proceedings of the National Academy of Sciences of the USA 97: 2035–2039, 2010.
[5] Sinden RR, "DNA Structure and Function", San Diego: Academic Press,2014.
[6] Makarucha A.J., Todorova N., Yarovsky I., "Nanomaterials in biological environment: A review of computer modelling studies", Eur. Biophys. J. 40:103–115, 2013.
[7] De la Iglesia D., Garcia-Remesal M., de la Calle G., Kulikowski C., Sanz F., Maojo V, "The impact of computer science in molecular medicine: Enabling high-throughput research". Curr. Top. Med. Chem.13:526–575,2013.
[8] Yang S.T., Liu Y., Wang Y.W., Cao A. Biosafety and bioapplication of nanomaterials by designing protein-nanoparticle interactions. Small. 9:1635–1653, 2013.
[9] Nanomaterial Registry. Available online:https://www.nanomaterialregistry.org/
[10] The ACTION-Grid White Paper on Nanoinformatics. Available online:http://www.action-grid.eu/documents/Final%20White%20Paper%20-%20Nano.pdf.
[11] Nie S., Xing Y., Kim G.J., Simons J.W, "Nanotechnology applications in cancer" . Annu. Rev. Biomed. Eng.9:257–288,2014.
[12] Love S.A., Maurer-Jones M.A., Thompson J.W., Lin Y.S., Haynes C.L. Assessing nanoparticle toxicity.Annu. Rev. Anal. Chem. 5:181–205,2012.
[13] Pinheiro, A. V, Han, D., Shih, W. M. & Yan, H. Challenges and opportunities for structural DNA nanotechnology. Nat. Nanotechnol. 6, 763–72 ,2011.
[14] Jiang, Q. et al. DNA origami as a carrier for circumvention of drug resistance. J. Am. Chem. Soc. 134, 13396–13403, 2012.
[15] Feldkamp, U. & Niemeyer, C. M. Rational design of DNA nanoarchitectures. Angew. Chemie - Int. Ed. 45, 1856–1876. 2006.
[16] The Gene Ontology. [accessed on 14 March 2014]. Available online: http://www.geneontology.org/
[17] caOBR—A Tool for Indexing Cancer Informatics Resources.Available online: http://www.bioontology.org/caOBR.
[18] Zebrafish Anatomy and Development Ontology. Available online:http://www.obofoundry.org/cgi-bin/detail.cgi?id=zebrafish_anatomy.
[19] primarily defining composite phenotypes and phenotype annotation.
[20] BioPortal provides access to commonly used biomedical ontologies and tools for their analysis.
[21] Jmol. Available online: http://jmol.sourceforge.net/
[22] RasMol. Available online: http://www.bernstein-plus-sons.com/software/rasmol/
[23] Raster3D. Available online:http://skuld.bmsc.washington.edu/raster3d/raster3d.html.
[24] PyMOL.Available online: http://www.pymol.org/
[25] Xeo. Available online: http://sourceforge.net/projects/xeo/
[26] AutoDock 4. Available online: http://autodock.scripps.edu/
[27] Dock.Available online: http://dock.compbio.ucsf.edu/
[28] GOLD.Available online:http://www.ccdc.cam.ac.uk/Solutions/GoldSuite/Pages/GOLD.aspx.
[29] FlexX. Available online: http://www.biosolveit.de/FlexX/
Surflex-Dock. Available online: http://www.tripos.com/index.php?family=modules,SimplePage,,,&page=surflex_dock&s=0

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

HUMAN AUTHENTICATION MATCHING WITH 3DSKULL AND GAIT

[1.]T. INDUMATHI , [2]M. PUSHPARANI
[1.]Assistant Professor, Dept. of Computer Application, Shri ShankarlalSundarbaiShasun Jain College for women, Chennai.
Indumathi1979@yahoo.co.in
[2.] Professor & Head, Dept. of Computer science, Mother Teresa Women's University,Kodaikanal
drpushpa.mtwu@gmail.com

ABSTRACT *The multimodal human identification is focused in this paper are captured as the images such as the face and walking style are captured using web camera by using the web camera . As per Human Identification Systemsthe image can then be considered for further analyses of gait and skull characteristics. The propose system to identify a skull by using a correlation measure between the 3D skull and 3D face in terms of morphology, and measure the correlation coefficient is a measure that determines the degree to which two variables movements are associated correlation using SIFT Canonical Correlation Coefficient Analysis (SCCA). The 3D skull data as the probe and 3D face geometric data as the gallery and the proposed system undergo the match of the skull with enrolled 3D faces by the correlation measure between the Probe and the Gallery. The GAIT Detector extracts from an image, a number of frames (attributed regions) in a way which is consistent with (some) variations of the illumination, viewpoint and other viewing conditions figure.This dataset contains samples of 16 different individuals that have been taken at 0, 45, 90 degrees of angles. Afterwards, using Uncorrelated Multilinear Discriminant Analysis (UMLDA) algorithm for the challenging problem of Gait Recognition and for feature extraction the silhouette images hasbeen from the gait samples. Finally, for training and testing purpose Neural Network for mat-lab tool is used. Based on hidden layer, selection of training algorithm and setting the different parameter for training they have created different model of neural network. we will then undergo the test for the combination of NN+SVM, Knearest Neighbour Classification. Where all these experiments are done here on CASIA gait database and input video.*

Keywords: *SIFT Canonical Correlation Coefficient Analysis(SCCA), Scale Invariant Feature Transform (SIFT),UMLDA,Neural Network, SVM, Knearest neighbour Classifier*

I.INTRODUCTION

Gait biometric :- Gait as a biometric method has some advantages such as being difficult to hide, steal, or fake. Furthermore, gait can be recognizable from distance. The most other biometrics can be captured only by physical contact or at a close distance from the recording probe.

The binary sequences are three-dimensional objects naturally represented as third-order tensors in a very high-dimensional tensor space, with the spatial row, column and the temporal modes for the three dimensions. To deal with these tensor objects directly, classical vector-based linear feature extraction algorithms such as the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) need to reshape (vectorise) the input into vectors in a very high-dimensional space, resulting in high computation and memory demand. Furthermore, the input reshaping breaks the structure and correlation in the original data and thus the redundancy and structure in the original data is not fully utilized. In this paper, a uncorrelated multilinear discriminant analysis (UMLDA) is proposed to extract uncorrelated discriminative features directly from tentorial data based on the Fisher's discrimination criterion. In the next section, basic notations and multilinear algebra are introduced and the tensor-to-vector projection (TVP) is formulated as a number of elementary multilinear projections (EMPs). The UMLDA is then derived in Section 3 section 4 and the method problem in gait recognition is analysed and a regularization procedure is introduced to tackle this problem.

Skull biometric: In skull identification, nearly all of the methods depend on accurate extraction and representation of the relationship between the skull and face. However, it is very difficult to extract this complex relationship. Because this work aims to identify human face is from skull.This paper proposes a skull identification method that matches a skull with enrolled 3D faces, in which the mapping between the skull and face is obtained using some analysis as below;

Firstly, we build a skull face database. Secondly, Here a statistic method is adopted to estimate outlook from subclass of skull-face database using Principle component analysis and Linear discriminant analysis LDA . In order to improve the accuracy of the result, we select the suitable organ (eyes, nose and mouth) for the statistic result based on anatomy principle from the database and achieve the organ and face integration to build the final outlook, a method to build a joint statistical 3D model of the skull and face is presented. This model is then used to reconstruct a face from available skull data. The idea is similar to, but uses a statistical shape model of both the skull and the face for the reconstruction task. The Third approach for enhancing the matching performance of AAM is to modify the fitting algorithm of AAM itself, by proposing a novel fitting algorithm or enhancing the existing fitting algorithms. In this perspective [14] have proposed a fast AAM using the SIFT canonical correlation coefficient analysis (SCCA), which has modelled the relation between differences of the image and the model parameter for improving the convergence speed of fitting algorithm. We propose to identify an unknown skull through using a correlation measure between the 3D skull and 3D face in terms of the morphology, and measure the correlation coefficient is a measure that determines the degree to which two variables' movements are associated. The range of values for the correlation coefficient is -1.0 to 1.0. If a calculated correlation is greater than 1.0 or less than -1.0, a mistake has been made. A correlation of -1.0 indicates a perfect negative correlation,

while a correlation of 1.0 indicates a perfect correlation. Using SIFT canonical correlation coefficient analysis (SCCA) [13]. We use the 3D skull data as the probe and 3D face geometric data as the gallery, and match the skull with enrolled 3D faces by the correlation measure between the probe and the gallery.

Finally, the suitable organs (eyes, nose and mouth) are selected from organ database and we achieve the integration between organ and outlook to estimate final outlook of the skeleton remains. we show that the region-based strategy is better than the holistic strategy in terms of the extraction of the relationship between the skull and face.

II. DATA ACQUISITION

A. SKULL AND DATABASE

There exist two Each entry or item of skull-face database consists of two parts: skull model and face model. Here, we briefly describe the common method to achieve model digitalization and introduce the modal of our database. An entry (i.e. a sample) in our database consists of a skull surface coupled with a skin surface. For facial reconstruction, only the skull surface is known. These surfaces are represented by 3D meshes (vertices and triangles). For each individual in our database, original meshes are reconstructed from CT data of the subject (Figure 1). The original CT slice images were processed by the model after filtering out the noise to extract the skull and face borders. The 3D skull and skin surfaces are reconstructed by a marching cubes algorithm [25], and they are represented as triangle meshes that include approximately 1 50,000 and 2 20,000 vertices, respectively. All of the heads are substantially complete.

B.GAIT

The motional individual silhouette must be detected before getting the gait feature. Back ground subtraction is the relatively simple and new approach to find silhouette from image.

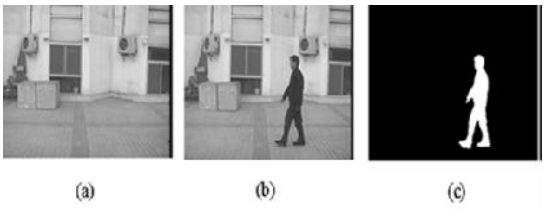


Fig. 1. Example of gait detection. (a) Background image; (b) Original image; and (c) Extracted silhouette

In our experiment the camera is assumed to be static and that body in the field of view is not occluded from each frame. The whole process of silhouette extraction is described as follows:

• To obtain an approximate background from the image sequence of a walking people, first mean image is computed by averaging the gray-level at each pixel over the entire image sequence in Fig.1 (b). Let Ik(x,y),k=1,2,…. N, represent sequence of N images. Back ground images b(x,y) can be computed by

$$b(x,y)=\text{median}(Ik(x,y)),$$

$$k=1,2,\ldots. N \qquad (1)$$

• Moving object is extracted by back ground subtraction.

Gait biometric :-

Gait as a biometric method has some advantages such as being difficult to hide, steal, or fake. Furthermore, gait can be recognizable from distance. The most other biometrics can be captured only by physical contact or at a close distance from the recording probe.

The binary sequences are three-dimensional objects naturally represented as third-order tensors in a very high-dimensional tensor space, with the spatial row, column and the temporal modes for the three dimensions. To deal with these tensor objects directly, classical vector-based linear feature extraction algorithms such as the Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) need to reshape (vectorise) the input into vectors in a very high-dimensional space, resulting in high computation and memory demand. Furthermore, the input reshaping breaks the structure and correlation in the original data and thus the redundancy and structure in the original data is not fully utilized. In this paper, a uncorrelated multilinear discriminant analysis (UMLDA) is proposed to extract uncorrelated discriminative features directly from tentorial data based on the Fisher's discrimination criterion. In the next section, basic notations and multilinear algebra are introduced and the tensor-to-vector projection (TVP) is formulated as a number of elementary multilinear projections (EMPs). The UMLDA is then derived in Section 3 section 4 and the method problem in gait recognition is analysed and a regularization procedure is introduced to tackle this problem.

Skull biometric:

In skull identification, nearly all of the methods depend on accurate extraction and representation of the relationship between the skull and face. However, it is very difficult to extract this complex relationship. Because this work aims to

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

identify human face is from skull.This paper proposes a skull identification method that matches a skull with enrolled 3D faces, in which the mapping between the skull and face is obtained using some analysis as below;

Firstly, we build a skull face database. Secondly, Here a statistic method is adopted to estimate outlook from subclass of skull-face database using Principle component analysis and Linear discriminant analysis LDA . In order to improve the accuracy of the result, we select the suitable organ (eyes, nose and mouth) for the statistic result based on anatomy principle from the database and achieve the organ and face integration to build the final outlook, a method to build a joint statistical 3D model of the skull and face is presented. This model is then used to reconstruct a face from available skull data. The idea is similar to, but uses a statistical shape model of both the skull and the face for the reconstruction task. The Third approach for enhancing the matching performance of AAM is to modify the fitting algorithm of AAM itself, by proposing a novel fitting algorithm or enhancing the existing fitting algorithms. In this perspective [14] have proposed a fast AAM using the SIFT canonical correlation coefficient analysis (SCCA), which has modelled the relation between differences of the image and the model parameter for improving the convergence speed of fitting algorithm. We propose to identify an unknown skull through using a correlation measure between the 3D skull and 3D face in terms of the morphology, and measure the correlation coefficient is a measure that determines the degree to which two variables' movements are associated. The range of values for the correlation coefficient is -1.0 to 1.0. If a calculated correlation is greater than 1.0 or less than -1.0, a mistake has been made. A correlation of -1.0 indicates a perfect negative correlation, while a correlation of 1.0 indicates a perfect correlation. Using SIFT canonical correlation coefficient analysis (SCCA) [13]. We use the 3D skull data as the probe and 3D face geometric data as the gallery, and match the skull with enrolled 3D faces by the correlation measure between the probe and the gallery.

Finally, the suitable organs (eyes, nose and mouth) are selected from organ database and we achieve the integration between organ and outlook to estimate final outlook of the skeleton remains. we show that the region-based strategy is better than the hol• Image processing operation likes Erosion, dilation are applied to improve the quality of extracted silhouette, and reduce noise.

III. HUMAN RECOGNITION GAIT AND SKULL

My research paper is multimodal human identifications are captured in a web cam captures images face and walking style. The image Fig4. can be considered and analyses gait and skull characteristics as per human identification systems. This paper proposes a skull identification method that matches a skull with enrolled faces, in which the mapping between the skull and face. This project, we propose to identify skull through an algorithm between the 3D skull and 3D face in terms of the morphology. We use the 3D skull data as the probe and 3D face geometric data as the gallery, and match the skull with enrolled 3D faces by the correlation measure between the probe and the gallery. Considering that the strength of the correlation between the skull and face in different craniofacial regions is not the same, we propose a region fusion strategy to measure the correlation between the skull and face more reliably and to boost the identification capability.

Gait recognition could also be used from a distance, making it well-suited in identifying an individual. We use the term gait recognition to signify the identification of an individual from a video Sequence of the subject walking. This does not mean that gait is limited to walking, it can also be applied to running or any means of movement on foot. We propose to use multi-view and multi-modal biometrics from a single walking image sequence. As multi modal cues, we adopt not only face and gait but also the actual height of a person, all of which are simultaneously captured by a single camera.

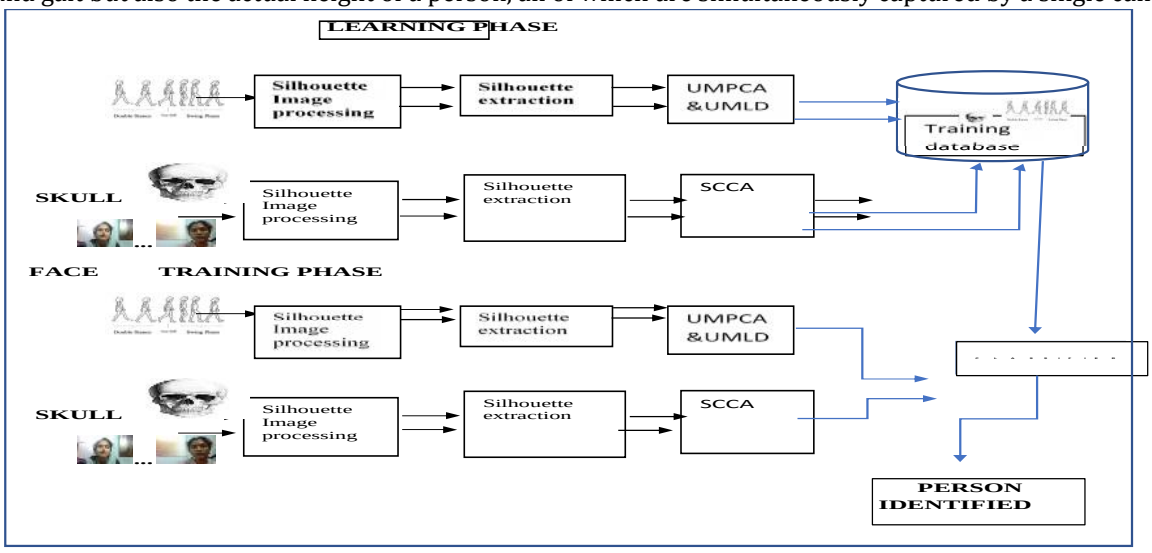


Fig. 2. Human recognition

The final result of the Gait data verified in the gait database using some algorithm by applying and checking for distance of person, walking speed, angle of the leg, pattern of walking is then processed in identifying the person.

IV. THE PROPOSED METHOD

In skull identification, nearly all of the methods depend on accurate extraction and representation of the relationship between the skull and face. However, it is very difficult to extract this complex relationship. Because this work aims to identify an skull by looking for its corresponding face skin from a 3D face gallery, we measure only the correlation between a skull and a face skin, and do not need an
accurate relationship.

A. Statistical Shape Model

our experiments we implemented Principal Component Analysis (PCA) and .Linear Discriminant Analysis (LDA) Given an s-dimensional vector representation of each face in a training set of M images, PCA tends to find a t-dimensional subspace whose basis vectors correspond to the maximum variance direction in the original image space. To identify an unknown image, that image is projected onto the face space as well to obtain its set of weights. By comparing a set of weights for the unknown face to sets of weights of known faces, the face can be identified. If the image elements are considered as random variables, the PCA basis vectors are defined as eigenvectors of the scatter matrix ST defined as:

ST= $\sum_{i-1}^{M}(x - \mu).(x - \mu)T$ (1)

M

where μ is the mean of all images in the training set and xi is the ith image with its columns concatenated in a vector.

LDA finds the vectors in the underlying space that best discriminate among classes. For all samples of all classes the between-class scatter matrix SB and the within-class scatter matrix SW are defined by:

SB = $\sum_{i-1}^{c} M . (x - \mu).(x - \mu)T$ (2)

SW=$\sum_{i-1}^{c}\sum x$ £X $(x - \mu i).(x - \mu i)T$ (3

B.SCCA in the Shape Parameter Spaces

As described above, each training skull or face skin can be projected into its shape parameter space.

The Scale Invariant Feature Transform in computer vision systems and pattern recognition,feature descriptors extracted from an image's gray values are usually used. Scale invariant feature transform (SIFT) is one of the best descriptors for feature matching .The SIFT algorithm transforms image data into scale-invariant coordinates relative to local feàtures .

In computer vision systems and pattern recognition,feature descriptors extracted from an image's gray values are usually used. Scale invariant feature transform (SIFT) is one of the best descriptors for feature matching .The SIFT algorithm transforms image data into scale-invariant coordinates relative to local features and is based on four major stages:

• Scale-space extrema detection: The image is first convolved with a series of Gaussian filters at different scales. Then, adjacent Gaussian images are subtracted to produce the difference-of-Gaussian images. Scale space extrema in the difference-of-Gaussians are regarded as the most stable scale-invariant features.

• Determination of keypoint location: Scale-space extrema are interpolated to obtain subpixel accuracy.Candidate keypoints with low contrast and those that are located along an edge but unstable to small amounts of noise are eliminated.

• Orientation assignment: This stage is the orientation assignment to each keypoint, based on local image gradient directions. This allows for the representation of each keypoint relative to this orientation, achieving invariance to image rotation. Peaks in orientation histogram are supposed to be dominant directions.

• Keypoint descriptor assignment: The previously describe d steps assigned the location, scale, and orientation of each keypoint. The motivation for the computation of a more complex descriptor is to obtain a highly distinctive keypoint and invariant as possible to variations. Each resultant SIFT descriptor is a 128-element feature vector. After the keypoint descriptor has been calculated, keypoints are matched by using the minimum distance method, where an exhaustive search between all keypoints in both images is performed.

In order to increase the stability of matching results,the ratio between the distance of the closest neighbor and the distance to the second closest neighbor is calculated to reject the matches. More details about SIF.Despite the outstanding characteristics of the SIFT, it has some problems with image registration. Even after the identification of matching candidates after removal of incorrect initial matches as described above, there are still many false matches due to feature points located in some similar structures, which lead to a further outlier removal.

That is, each skull has p feature variables, while each face skin has q feature variables. Let XN × p and YN ×q denote the training skull and face data matrices, respectively, and let N denote the size of the training samples. The aim of SCCA is to find two sets of basis vectors, wx £Rp and wy £Rq , that maximize the correlation coefficient between the components t1

= Xwx and u1 = Ywy, i.e.,

$$J=\max \frac{cov(t_1,)}{w_x w_y \sqrt{Var(t_1) \times Var(u_1)}}$$
$$=\max_{w_x, w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x \times w_y^T C_{yy} w_y}}$$

where the covariance matrices Cxy = XTY, Cxx = XTX and Cyy = YTY.

Let W = wxT, wyT T and Cyx = YTX, then Equation (4) can be solved by computing a generalized eigen-value decomposition problem, as follows:

$$AW = \lambda BW \quad (9)$$

where

$$A = \begin{pmatrix} 0 & Cxy \\ Cyx & 0 \end{pmatrix}, B = \begin{pmatrix} C_{xx} & 0 \\ 0 & C_{yy} \end{pmatrix}. \quad (10)$$

More details on the derivation and solution of SCCA can be found in [13].

Thus, we obtain two subspaces that consist of the basis vectors wx £Rp and wy £Rq , respectively, for the skull and face skin. Assume that Wx denotes the subspace projection matrix whose columns are the basis vectors wx, and Wy is the matrix that corresponds to the basis vectors wy. For a skull-and-face pair, let x denote the shape parameter vector of the skull, and let y denote the shape parameter vector of the face. Then, the feature vectors of the skull and face in the ECCCA subspace are

$$X_c = W_x T_x$$
$$Y_c = W_y T_y$$

We define the matching score between the skull and face as Follows

$$r(XcYc) = \frac{XcYc}{XcYc}$$

where _·, ·_ denotes the inner product operation.

C. Uncorrelated Multilinear Principal Component Analysis and Linear Discriminant Analysis

As a multilinear extension of PCA, UMPCA not only obtains features that maximize the variance captured, but also enforces a zero-correlated constraint, thus extracting uncorrelated features in a similar way to that of the classical PCA. Motivated by Fisher face algorithm, we use UMPCA algorithm for tensor image feature extraction and dimension reduction and then get low dimensionality images which are ready for applying LDA. Finally, we implement the nearest neighbour classifier to classify face images based on its computed LDA features.

C.1. Uncorrelated Multilinear Principal Component Analysis

In this section, UMPCA algorithm is introduced in detail based on the analysis introduced in [14]. The UMPCA objective function is first formulated. Then, the successive variance maximization approach and alternating projection method are adopted to derive uncorrelated features through TVP. The problem to be solved is formally stated as follows.

A set of M tensor object samples {$\mathcal{X}$1,$\mathcal{X}$2,⋯,$\mathcal{X}$M} are available for training and each tensor object $\mathcal{X}$m∈ℝI1×I2×⋯×IN assumes values in the tensor space ℝI1⊗ℝI2⋯⊗ℝIN, where In is the n-mode dimension of the tensor and ⊗ denotes the Kronecker product. The objective of the UMPCA is to find a TVP, which consists of P EMPs{up(n)∈ℝIn×1,n=1,⋯,N}p=1P, mapping from the original space ℝI1⊗ℝI2⋯⊗ℝIN into a vector subspace ℝP(withP<ΠInNn=1)

ym=$\mathcal{X}$m×n=1N{up(n)T,n=1,⋯,N}p=1P (13)

The P EMPs {up(n)T,n=1,⋯,N}p=1P are determined by maximizing the variance captured while producing features with zero correlation. Thus, the objective function for the Pth EMP is:

{up(n)T,n=1,⋯,N}p=1P = argmaxΣ(ym(p)−ȳp)2=STpyMm=1

Subject to up(n)Tup(n)=1 andgpTgq‖gp‖‖gq‖=δpq, p,q=1,⋯,P (14)

Where δpq is the Kronecker delta defined as δpq={1, if p=q0, otherwise (15)

To solve the UMPCA problem (14), we follow the successive variance maximization approach. The P EMPs

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

{up(n)T,n=1,⋯,N}p=1P are sequentially determined in P steps, with the pth step obtaining the pth EMP

Step 1: Determine the first EMP {u1(n)T,n=1,⋯,N}p=1P by maximizingST1y.

Step 2: Determine the second EMP {u2(n)T,n=1,⋯,N}p=1P by maximizing ST2y subject to the constraint thatg2Tg1=0.

Step 3: Determine the third EMP {u3(n)T,n=1,⋯,N}p=1P by maximizing ST3y subject to the constraint that g3Tg1=0 and g3Tg2=0.

Step 4: p(p=4,⋯,P): Determine the pth EMP {up(n)T,n=1,⋯,N}p=1P by maximizing STpy subject to the constraint that gpTgq=0 for q=1,⋯,p−1.

In order to solve for the pth EMP{up(n)T,n=1,⋯,N}, we need to determine N sets of parameters corresponding to N projection vectors, up(1), up(2),⋯,up(N), one in each mode.

C.2. Linear Discriminant Analysis

A classical linear discriminant analysis (LDA) is then applied to obtain an UMPCA+LDA approach for recognition, similar to the popular of PCA+LDA. Consider c classes existing with M samples. The within-class matrix is defined in the form

SW=Σ(ym−ȳcm)(ym−ȳcm)Tm, and ȳc=1NcΣymcm=cm (16)

The between-class scatter matrix is defined as

SB=ΣNc(ȳc−ȳ)Cc=1(ȳc−ȳ)T, and ȳ=1MΣymm (17)

The optimal projection matrix V is chosen as follows:

Vlda=argmax|VTSBV||VTSwV|=[v1v2⋯vM] (15)

Where the {vm,m=1,⋯,M} is the set of generalized eigenvectors of SB and SW corresponding to the m largest generalized eigenvalues{λm,m=1,⋯,M}: SBvm=λmSWvm. Thus, the discriminant feature vector zm is obtained as: zm=VldaTym, and a classifier can then be applied.

E. Neural Network (NN)

Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output as shown in the graphic below Fig.3.

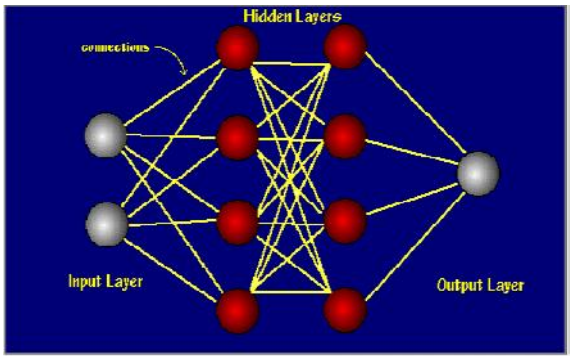


Fig 3. Neural Network

The neural network can be defined as an interconnection of neurons. Neural networks make use of the fact that the recall of information can be effected in two ways. The recall can be performed in the feed forward mode. The feed forward has no memory. Feed forward network's behaviour does not depend on what happened in the past but rather what happens now. The network responds only to its present input. The area of Neural Networks probably belongs to the borderline between the Artificial Intelligence and Approximation Algorithms. The NNs are used in (to name few) universal approximation (mapping input to the output), tools capable of learning from their environment, tools for finding non-evident dependencies between data and so on. The Neural Networking algorithms (at least some of them) are modelled after the brain (not necessarily - human brain) and how it processes the information.

VI. RESULTS AND DISCUSSION

The used dataset is the 3D skull-and-face skin pairs of the 250 subjects described in Section II. Five-fold cross-validation was used to evaluate the proposed method. We randomly chose 7 non-overlapping data groups from the dataset, and each group included the 3D skull-and-face skin pairs of 30 subjects. For each fold, 30skulls in one group were used as probes to test the proposed method, and the remaining 120 pairs of skull-and-face skins constitute the training set. In other words, we have 275 test skulls in all five folds. For each test skull, all of the 220 face skins constitute the gallery. The correct identification rate reported in the experiments is the average of the five folds. In this paper, we define that a test skull is identified at rank n, if the matching score of the correct match is of rank n. The identification rate at rank n is defined as the ratio of the cumulative count of the numbers of test skulls identified at rank n or less to the total number of test skulls. The correct identification rate is defined as the identification rate at rank

V.EVALUATION OF SKULL IDENTIFICATION USING LDA AND PCA:

Both Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are linear transformation techniques that are commonly used for dimensionality reduction. PCA can be described as an "unsupervised" algorithm, since it "ignores" class labels and its goal is to find the directions (the so-called principal components) that maximize the variance in a dataset. In contrast to PCA, LDA is "supervised" and computes the directions ("linear discriminants") that will represent the axes that maximize the separation between multiple classes.In Fig.8 compare the shape parameter.

Table 2 shape parameter LDA AND PCA

Region	Profile	forehead	Eye	nose	Mouth
Skull	74	67	30	33	65
Face	50	45	30	4 1	55

II. Evaluation of skull identification using Enhance canonical correlation coefficient analysis and SIFT

The used dataset is the 3D skull-and-face skin pairs of the 208 subjects described in Section. Five-fold cross-validation was used to evaluate the proposed method. For each fold, 40 skulls in one group were used as probes to test the proposed method, and the remaining 168 pairs of skull-and-face skins constitute the training set.

The identification rate at rank n is defined as the (c)ratio of the cumulative count of the numbers of test skulls identified at rank n or less to the total number of test skulls. The correct identification rate is defined as the identification rate at rank 1

For each fold, we built a enhance correlation analysis model in which all of the 176 eigenvectors are used to construct the statistical shape models. A total of 150 test skulls was correctly identified at rank 1. The correct identification rate attained 81.5%. It verifies that there indeed exists some relationship between the skull and face, and it also shows that the proposed skull identification method is effective.

In this part, the former four test groups are referred to as the probability training set, while the latter 40 test skulls are referred to as the evaluation set. We perform statistical analysis for the 25600 (4 × 40 × 160) matching scores that were derived from the probability training set. The mean and standard deviation are (0.3205, 0.0811) for the positive matching and (0.0009, 0.0806) for the negative matching. Figure 8a and 8b show the histograms of the matching scores for the positive and negative matching, respectively. From Figure8, we can see that both the positive and negative matching score approximately follow a Gaussian distributionFigure 8c shows the two standard Gaussian distributions. We can see that the scores of the positive and negative matching are separable. The conditional probability P(w1| E) in Equation (14) can be set as the average correct identification rate of the four folds that the probability training set belongs to, i.e., 80.6%. The prior probability P(E) can be set to 1 because all of the face data of the test skulls exists in the face gallery in this experiment. According to the Bayes decision rule, we evaluate the identification results of the evaluation set, i.e., the 40 test skulls in the remaining one fold. Five of the 6 false identification results and all of the 34 correct results are classified as positive. If we set P(E) as 0.5, i.e., we have no prior information about whether the face data of the test skull exists in the face gallery, then one correct identification result is classified as negative and 3 false results are classified as positive.

Table3 Enhance correlation coefficient analysis

Region	**Profile**	**Forehead**	**Eye**	**Nose**	**Mouth**
Skull	98	77	65	75	85
ace	98	75	80	81	85

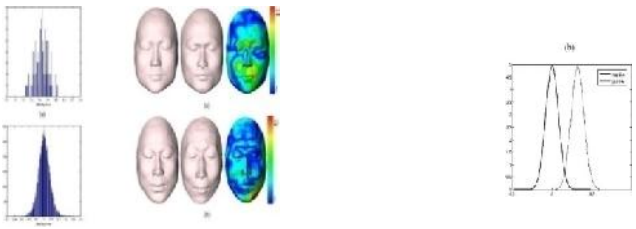
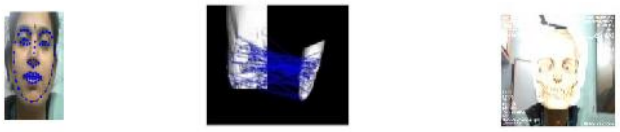


Fig. 4. Histograms of matching scores for the positive matching (a), negative matching (b), and class conditional probability density functions (c).



A b c

a.Active appearance model(AAM), b.Enhance Canonical Correlation Coefficient Analysis(ECCA), c. Scale Invariant Feature Transform (SIFT)

other case is that the global morph-logical similarity between the mismatched face and the true face is high, and it can

be misclassified as positive matching. Figure 9 shows two samples of the second case. We can see that the mismatched faces are visually similar to the true faces. Figure 8 also shows the comparison coloured from blue to red according to the geometric distance between the two faces. The average distances of these two samples are 2.7 mm and 2.56 mm, respectively. From the coloured comparisons, we can see that the distance between the mismatched faces and the true faces is very small in most of the regions. The global morphological similarity between two faces leads to similar shape model parameters, which contribute to the false matching due to a holistic model. For this reason, we propose the region-based method.

In this thesis, first step is extraction of foreground objects i.e. human and other moving objects from input video sequences or binary silhouette of a walking person is detected from each frame and human detection and tracking will be performed. After getting binary silhouettes of human beings model based approach is used to extract the gait features of a person. To compare the performance of silhouette-based ones, the results are gait image+UMPCA+UMLDA uncorrelated multilinear discriminant analysis (UMLDA) algorithm for the challenging problem of gait recognition.At last neural network is used for training and testing purpose. We have created different model of neural network based on hidden layer, selection of training algorithm and setting the different parameter for training. And then we will test for the combination of NN+SVM classification. Here all experiments are done on gait database and input video. Here all experiments are done on CASIA gait database In addition, MATLAB software with its neural network toolbox is used.As can be seen in table 1 the recognition accuracy achieved with 0 degrees' view point in significant lower for compared 90 degrees and 45-degree view points. With 45 degrees and 90-degree view point we obtained 88.45% and 89.45% recognition accuracy. This could be best view of gait information from 90-degree view points. However, when we three views, we found the recognition accuracy was 98.88% significantly higher than each of the viewpoints.

TABLE I: Recognition Accuracy

Numbers	Method	View Angle	Recoginition Accuracy
1	UMPCA+UMLDA	0° Left View	78%
2	UMPCA+UMLDA	0° Right View	85.23%
3	UMPCA+UMLDA	45° Left View	88.45%
4	UMPCA+UMLDA	45° Right View	86.56%
5	UMPCA+UMLDA	90° Left View	87.45%
6	UMPCA+UMLDA	90° Right View	89.45%
7	UMPCA+UMLDA	View total Recognition	97.88%

This shows of multiple views video footage with long range of video it is possible to perform large scale identification with high level accuracy, using simple subspace feature(UMPCA+UMLDA) and classifier NN+SVM. Such simple approaches can lead to real time and real world intelligent video surveillance system the begin of new generation of security system of public surveillance deployments.

The performance of proposed feature method was evaluated against different classifiers including a wide range of paradigms (neural network NN with, Knearest neighbour KNN, Support Vector Machines (SVM)).

Classifier Or Feature used	Gait	Skull	Result
NN	98.56%	98.55%	98.65%
KNN	89.78%	91.22%	92.21%
SVM	92.55%	95.92%	94.21%

Table II Accuracies of different classifier using our proposed Method.

We performed experiment on CASIA gait database. This database is divided into 3 degrees. As can be seen in table 1 the recognition accuracy achieved with 0 degrees' view point in significant lower for compared 90 degrees and 45-degree view points. With 45 degrees and 90-degree view point we obtained 85.35% and 88.45% recognition accuracy

VI. CONCLUSION AND FUTURE WORK

In skull identification, nearly all of the methods depend on accurate extraction and representation of the relationship between the skull and face. However, it is very difficult to extract this complex relationship. Because this work aims to identify human face is from skull. This paper proposes a skull identification method that matches a skull with enrolled faces, in which the mapping between the skull and face is obtained using enhance canonical correlation coefficient analysis with scale invariant feature transform (SIFT). In this work, we only use CT scan data to validate the proposed method. 3D face data acquisition by CT is infeasible for real applications because of the intrinsic radiation for the livings

and the cost of the system. Compared with CT scan, 3D face modelling from 2D images is a convenient and non-intrusive way. By this way, we can construct a 3D face database from a 2D database of persons in a nation. Therefore, future work will focus on how to live face change to live skull modal to our database applying the proposed technique to 3D face models reconstructed the field of public security. This paper presents a gait recognition approach using specific features, like centre of mass, step size length and cycle length. Here neural network is Feed forward back propagation network is being used to identify being used for recognizing people. Our results show that these features are more effective to identify people from distance. learning and performance function and result shows recognition rate of 97% with 4 layers and 30 neurons Recognition The work can be extended to develop new multimodal biometric system in which, gait can be combined with other biometrics system. So it can be use as one of the good reliable way of authentication.

VII.REFERENCES

[1] P. Claes, D. Vandermeulen, S. De Greef, G. Willems, J. G. Clement, and P. Suetens, "Bayesian estimation of optimal craniofacial reconstruc-tions," Forensic Sci. Int., vol. 201, nos. 1–3, pp. 146–152, 2010.

[2] M. Berar, F. M. Tilotta, J. A. Glaunès, and Y. Rozenholc, "Craniofacial reconstruction as a prediction problem using a latent root regression model," Forensic Sci. Int., vol. 210, nos. 1–3, pp. 228–236, 2011.

[3] J. Huang, M. Zhou, F. Duan, Q. Deng, Z. Wu, and Y. Tian, "The weighted landmark-based algorithm for skull identification," in Proc. 14th Int. Conf. Comput. Anal. Images Patterns (CAIP), LNCS 6855. 2011, pp. 42–48.

[4] Y. Hu, F. Duan, M. Zhou, Y. Sun, and B. Yin, "Craniofacial reconstruc-tion based on a hierarchical dense deformable model," EURASIP J. Adv. Signal Process., vol. 2012, p. 217, Oct. 2012.

[5] G. M. Gordon and M. Steyn, "An investigation into the accuracy and reliability of skull-photo superimposition in a South African sample," Forensic Sci. Int., vol. 216, nos. 1–3, pp. 198.e1–198.e6, 2012.

[6] [6] FuqingDuan, Yanchao Yang, Yan Li, Yun Tian, Ke Lu, Zhongke Wu, and Mingquan Zhou" Skull Identification via Correlation Measure Between Skull and Face Shape" IEEE transactions on information forensics 8and security, vol. 9, no. 8, august 2014

[7] M. Pushpa Rani, G.Arumugam "An Efficient Gait Recognition System For Human Identificatin Using Modified ICA" International Journal of Computer science& information Technology,Vol.2,No.1,February 2010

[8] M. Pushpa Rani, G.Aru D Sasikala "A Survey of Gait Recognition Approaches Using PCA and ICA" International Journal of Computer science& information technology.

[9] T.Indumathi, Dr.M. Pushpa Rani "automatic door opening using gait identification for movement as gesture "as organized By GSTF Journal of Engineering Technology (JET) Vol4 No1,Aug2016

[10] Dr. M. Pushpa Rani, T.Indumathi "human authentication by matching 3d skull with face image using enhance ccca with sift" as organized International Journal of Applied Engineering Research (IJAER).ISSN 0973-4562,Volume 10,Number 14,2015

[11] Dr. M. Pushpa Rani, T.Indumathi "monitoring and controlling locker system using skull identification " as organized International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)ISSN 2347-6710,(0nline) ISSN 2319-8753 Vol. 4, Issue 7, July 2015

[12] T.Indumathi ,Dr. M. Pushpa Rani ""multimodal human authentication by matching 3d skull and gait " as organized IEEE international conference on 2nd world congress on computing and communication technologies , ISSN 978-1-5090-5574-6,Feb 4

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Multi-kernel K-means Clustering based Kidney Stone Segmentation for Ultrasound Images

[1]Balamurugan, S. P. & [2]Arumugam, G.
[1]Assistant Professor/Programmer, Department of Computer and Information Science,
Faculty of Science, Annamalai University, Tamilnadu, India.
[2]Senior Professor & Head, Department of Computer Science, Madurai Kamaraj University, Tamilnadu, India.

ABSTRACT *The main objective of this paper is to design and develop an approach for kidney stone segmentation using clustering approach. At present, kidney stone segmentation is one of the vital procedures in surgical and treatment planning in ultrasound images. To date, in clinical practice kidney stone segmentation is done manually making it obsolete. Being time-consuming, tracing manual stone has become difficult and it solely depends on the operator. Therefore, in this work, we proposed a technique to segment the stone part separately from the abnormal image using multi-kernel k-means clustering algorithm. To achieve the concept, we comprised the proposed system into two modules such as (i) preprocessing and (ii) segmentation. At first, in the preprocessing stage we removed the noise present in the input image using Gaussian filter. After preprocessing, we accurately segmented the abnormal position using multi kernel k-means clustering algorithm. The experimentation results show that the proposed system achieved better results associated with the available methods.*
Keywords: *Ultrasound image, linear kernel, quadratic kernel, RBF kernel, k-means clustering, segmentation.*

I. Introduction

A common pervasive and a well known issue that affects the human urinary system is renal calculi, which is commonly known as kidney stones or urinary stones. Kidney stones are typically formed in the kidney and pass out of the body in urine stream. A small stone may go unnoticed without causing symptoms. If a stone grows to more than 5 millimeters it can cause blockage of the ureter resulting in severe pain in the lower back or abdomen. A stone may also result in blood in the urine, vomiting, or painful urination. About half of people will have another stone within ten years. Most stones are formed due to a combination of genetics and environmental factors. Causes for kidney stones include high calcium levels in urine, obesity, certain types of food, some medications, calcium supplements and not drinking enough fluids. Stones form in the kidney when minerals in urine are at high concentration. The diagnosis is usually based on symptoms, urine testing, and medical imaging. Of late, an extraordinary emphasis in the field of restorative imaging is laid on kidney stone, renal whole segmentation. The programmed segmentation has tremendous potential in clinical solution by relieving doctors from the burden of manual marking; although programmed segmentation is not yet a full blown clinical practice. Automatic stone-area segmentation poses serious challenges for the fact that renal stones are highly heterogeneous in terms of shape, color, texture, and position and they often twist other anatomical structures.

Ultrasound scanning can image organs like liver, gallbladder, spleen, pancreas, kidneys, bladder, uterus, ovaries, and baby in pregnant patients. The ultrasound image is instantly visible on a video screen and the radiologist will solidify those images which are essential for the conclusion. Telemedicine has come to serve the general public cutting time and cost of travel and offering best medicinal care. Tele-radiology which was internet has bolstered ultrasound scanning by serving the locales where radiologists are not available. In the absence of therapeutic specialists, the delay in getting reports still complicates the crisis. As a result, there is a need for a device to distinguish the abnormal from the normal so that untrained radiologist can offer remedial choices in treating the patients with care.

The main objective of this paper is kidney stone segmentation using multi-kernel k-means clustering for ultrasound images. In the first instance, we removed the noise present in the image as the noise image hinders the segmentation performance. The paper is organized as follows: Section 1 offers an introduction, Section 2 presents the review of related work and Section 3 explains the proposed methodology of the kidney stone segmentation method. Section 4 provides the results and discussion of the technique. Here the kidney stones are segmented from ultrasound images and the experimental results are noted. Finally, the conclusion of proposed method is given in section 5.

II. Review of Literature

In medical diagnosis, numerous researchers have projected many methods for kidney stone detection and segmentation. Among them some of the works are analyzed here; Segmentation of Calculi from Ultrasound Kidney Images was one of the segmnetation method explained by Tamilselvi and Thangaraj [4]. Here, they used a Region Indicator with Contour Segmentation Method for segmenting kidney stone. But in this method, the calculi detection accuracy was not satisfactory and it has produced high complexity in the calculi detection process. To overcome the difficulties, Anjit Rajan et al. [6] have explained detection of renal-stones, segmented the renal regions and calculated the area of the renal which is occupied by kidney stones. Three kinds of Ultrasound kidney images namely, Normal (NR), Medical Renal Diseases (MRD) and Cortical Cysts (CC) images are classified based on texture properties.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Ashish K. Rudra et al. [7] have accurately segmented kidney from very low contrast MRI data using graph cut and pixel connectivity. A connectivity term was introduced in the energy function of the standard graph cut via pixel labeling. The labeling process was formulated according to Dijkstra's shortest path algorithm. Moreover, Mariam Wagih Attia et al. [10] have explained the kidney stone image classification using PCA and neural network. Here, five types of classification is carried out such as Normal, Cyst, Stone, Tumor and Failure. This method offered correct classification with high accuracy; even though this method is difficult. Also Nikita Derle and Devidas Dighe et al. [11] have developed diagnosis of Kidney Disease Using DCE-MRI Images in 4D. Dynamic contrast enhancement magnetic resonance imaging (DCE-MRI) is an imaging proficiency, used for calibrating different parameters homologous to suffuse, capillary leakage, and convey rate in tissues of various organs and diseases detection. This model provided accurate diagnosis but high SNR is the limitation of this system.

Divya Krishna et al. [12] have created irregularity discovery for a kidney using FPGA-based IoT empowered convenient ultrasound imaging framework. The calculation was executed in two phases. In the first stage, a Look Up Table (LUT) based approach was utilized to separate normal and anomalous kidney images. In the second stage, after confirming the abnormality, Support Vector Machine (SVM) with Multi-Layer Perceptron (MLP) classifier trained with extracted features is used to further classify the presence of stone or cyst in the kidney. Here, the selected feature values of kidney images with cyst and stone are having very small difference and cannot find unique range of values to classify the image.

III. Proposed methodology for kidney stone segmentation system

The main objective of the proposed methodology is kidney stone segmentation using two stages. Figure 1 shows the details of the proposed system. The first stage is preprocessing. Here, the input images are processed for noise removal. The next stage of the proposed system is segmentation; which separates the stone part from the original image using multi-kernel k-means clustering algorithm. Finally, it extracts a stone portion from this ultrasound image.

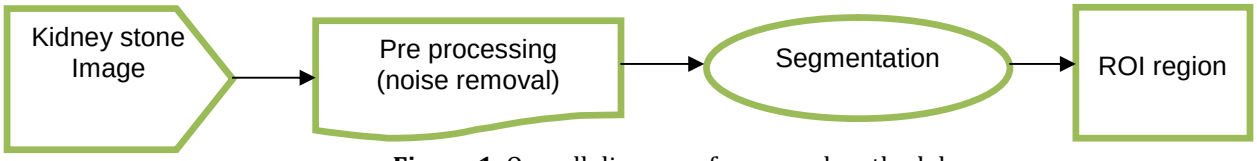


Figure 1. Overall diagram of proposed methodology

a. Preprocessing

Preprocessing is an important process of ultrasound images because the occurrence of noise components is more on Ultrasound image compared to other images like CT and MRI. Basically, the ultrasound images are mainly corrupted by a speckle noise. Therefore, noise removal is a critical process in medical ultrasound images. In this paper, for noise removal, we used Gaussian filter. The input image is passed through a Gaussian filter to reduce the noise as well as improve the image quality. The process involved in preprocessing is given in Figure 2.

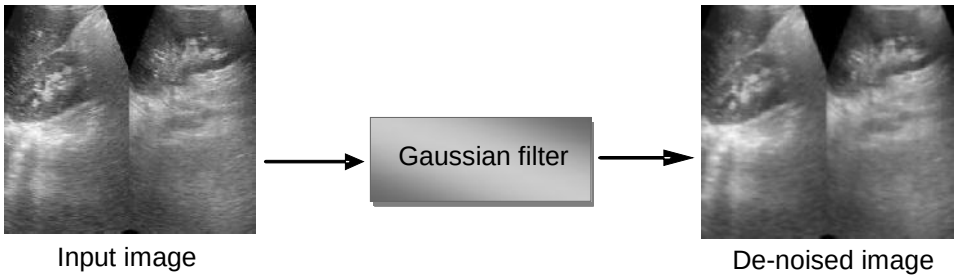


Figure 2. Process of preprocessing

b. Segmentation stage:

After preprocessing, the stone images are given to the segmentation stage. In this work, for segmentation stage, we have utilized the multi kernel k-means clustering algorithm. A lot of clustering algorithms have been used for segmentation. Among them, one of the most popular clustering algorithms is k-means algorithm, where groups are identified by minimizing the clustering error defined as the sum of the squared Euclidean distances between each data set point and the corresponding cluster center. This algorithm suffers from two serious limitations. First, the solution depends greatly on the initial positions of the cluster centers, ensuing in poor minima, and second, it can only just

discover straightly distinct groups. To conquer the issue, kernel k-means clustering algorithm was implemented. Kernel k-means is an extension of the standard k–means an algorithm that maps information points from the input space to a feature space through a nonlinear transformation and minimizes the clustering error in feature space. Consequently, non-linearly isolated clusters in input space are gotten, defeating the second restriction of k-means. To improve the proposed segmentation accuracy, we introduced hybrid kernel k-means clustering algorithm. Nowadays lot of kernels are used such as radial basis function (RBF) kernel, linear kernel and quadratic kernel. In this paper, we have hybridized the different types of kernels and the performance of the hybridization approach is analyzed in the result section. The individual kernel function formulas are given in equation (1-3).

Radial basis function:

$$K(p,q) = \exp\left(\frac{-\|p-q\|^2}{2\pm^2}\right) \quad (1)$$

Linear kernel function:

$$K(p,q) = p^T q + c \quad (2)$$

Quadratic kernel function:

$$K(p,q) = 1 - \frac{\| p-q \|^2}{\| p-q \|^2 + c} \quad (3)$$

In hybrid work, new hybridized kernel functions are taken and the clustering process is performed based on this hybrid kernel functions. Let $k_1 \; and k_2$ be two kernels. The hybrid kernel K-means algorithm under kernelization of the metric approach is an iterative two steps algorithm that gives a partition $S = \{S_1, \ldots, S_k\}$ of X into K clusters and their corresponding cluster centroids $Y_k \in R^p (K = 1, \ldots, K)$ which minimizes the objective function;

$$W = \sum_{k=1}^{K} \sum_{x_i \in P_k} \left\| \Phi(x_i) - \Phi(y_k) \right\|^2 \quad (4)$$

$$= \sum_{k=1}^{K} \sum_{x_i \in P_k} \left\{ K_{MK}(x_i, x_i) - 2K_{MK}(x_{i,} y_k) + K_{MK}(y_k, y_k) \right\} \quad (5)$$

$$K_{MK}(p,q) = K_1(p,q) + K_2(p,q) \quad (6)$$

Where; K_{MK} represents the hybrid kernel. In this paper, we used three types of hybrid kernel such as (linear + quadratic), (linear + radial basis kernel) and (quadratic + radial basis kernel).

- If K_1 is a linear kernel and K_2 is a quadratic kernel means

$$\left.\begin{aligned} K_1(p,q) &= p^T q + c \\ K_2(p,q) &= 1 - \frac{\| p-q \|^2}{\| p-q \|^2 + c} \end{aligned}\right\} \quad (7)$$

- If K_1 is a linear kernel and K_2 is a radial basis kernel means :

$$\left.\begin{aligned} K_1(p,q) &= p^T q + c \\ K_2(p,q) &= \exp\left(\frac{-\|p-q\|^2}{2\pm^2}\right) \end{aligned}\right\} \quad (8)$$

- If K_1 is a quadratic kernel and K_2 is a radial basis kernel means:

$$K_1(p,q) = 1 - \frac{\| p-q \|^2}{\| p-q \|^2 + c}$$
$$K_1(p,q) = \exp\left(\frac{-\|p-q\|^2}{2\pm^2} \right) \quad (9)$$

The center updated formula is shown in Equation (10):

$$y_k = \frac{\sum_{x_i \in p_k} K_{MK}(x_i, y_k) x_i}{\sum_{x_i \in p_k} K_{MK}(x_i, y_k)} \quad (10)$$

Once the centroid is updated for every cluster, the next step is to calculate the distance between centroid and the data point. Each data point is assigned to a cluster center whose distance is minimum. This process is repeated until the updated centroids of each cluster are similar in consecutive iterations. The algorithm for multi kernel k-means clustering is presented in Table 1.

Input:
Kidney stone image $I(i, j)$
Kernel matrix K, number of clusters k,
Output :
Segmented image
Start
1. Build hybrid kernel for clustering process using equation (6)
2. If K_1 is a Linear kernel and K_2 is a Quadratic kernel means use (7)
3. If K_1 is a Linear kernel and K_2 is a Radial basis kernel means use (8)
4. If K1 is a Quadratic kernel and K2 is a Radial basis kernel means use (9)
5. Initialize clusters (assign value for k)
6. For all points x_n n=1,….N do
7. For all clusters C_i i=1 to k do
8. Apply the hybrid kernel function to the data point
9. Compute the distance between data point and centroid $\|¤(x_i) - ¤(y_k)\|^2$
10. end for
11. find $c(x_i) = \arg\min_i \left(\|¤(x_i - y_k)\|^2 \right)$
12. end for
13. for all clusters C_i i=1 to k do
14. update cluster C_i
15. end for
16. if converged then
17. return final clusters $C_1, C_2, \ldots, C_k$
18. else
19. go to step 5
20. end if
21. end

Table 1: Algorithm For Multi Kernel K-Means Clustering

IV. Result and Discussion

In this section, we have discussed the result obtained from the proposed kidney stone segmentation. For implementing the proposed technique, we have used MATLAB version (7.12). This proposed technique was done in windows machine having Intel Core i5 processor with speed 1.6 GHz and 4 GB RAM. The proposed system has been tested on the data set available on the web. We have utilized the size of the image "512×512" which is publicly available.

a. Evaluation metrics

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

The system performance is analyzed by using the most common performance measure known as accuracy, sensitivity, and specificity. The metric values found are based on True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) with the option of segmentation and classification.

1) Sensitivity: The proportion of actual positives which are correctly identified is the measure of the sensitivity. It relates to the ability of the test to identify positive results.

$$Sensitivity = \frac{T_p}{T_p + F_n} \tag{11}$$

2) Specificity: The proportion of negatives which are correctly identified is the measure of the specificity. It relates to the ability of the test to identify negative results.

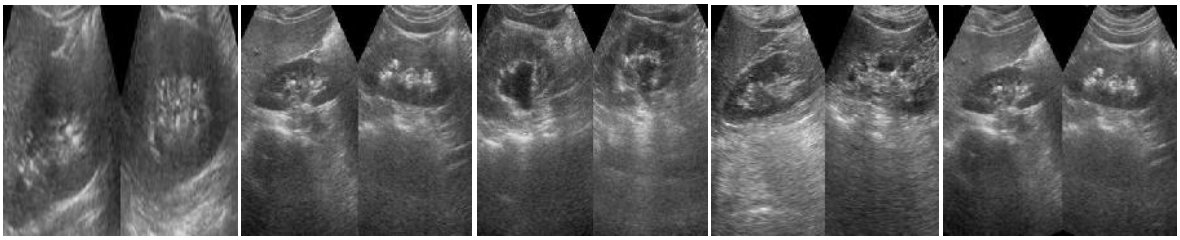
$$Specificity = \frac{T_n}{T_n + F_p} \tag{12}$$

3) Accuracy: We can compute the measure of accuracy from the measures of sensitivity and specificity as specified below.

$$Accuracy = \frac{T_p + T_n}{T_p + F_p + F_n + T_n} \tag{13}$$

b. Dataset description

The kidney image data set is effectively employed in the innovative image segmentation technique which is obtained from the publicly accessible sources. Figure 3 illustrates certain sample kidney images.



1.JPG 8.JPG 10.JPG 16.JPG 24.JPG

Figure 3. Input sample images

c. Performance evaluation

The basic idea of our proposed methodology is kidney stone segmentation using two stages. The performance is evaluated using accuracy, sensitivity and specificity measures. After preprocessing, we segmented the affected area using multi kernel k-means clustering algorithm. Here, we implemented the multi kernel based on three combinations. The visual representation of segmentation results and comparative analyses are explained in this section.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Figure 4. GUI representation of De-noised image

	Linear	Quadratic	L + Q	L + RBF	Q + RBF	RBF	Kmeans
Sensitivity	0.81308	0.76636	0.95327	0.53271	0.73832	0.71028	0.69159
Specificity	0.99982	0.99986	0.99983	0.99988	0.99908	0.99876	0.94013
Accuracy	0.99951	0.99948	0.99976	0.99911	0.99866	0.99829	0.93973

Figure 5. Performance of segmentation using various clustering techniques

Figure 5 shows the GUI results of various segmentation methods for stone images. The Figure 5 indicates the individual results of all the methods. From the results we clearly understood our proposed (Linear + Quadratic) multi-kernel k-means clustering algorithm achieved the better result compared to other methods.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

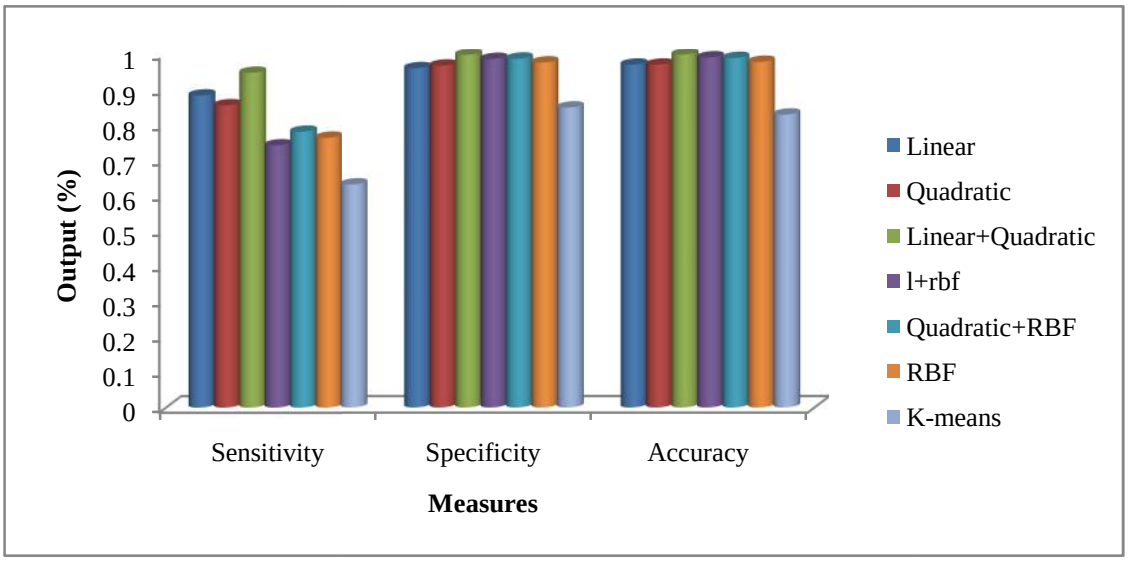


Figure 6. Comparative analysis for segmentation stage

Figure 6 shows the comparative analysis of segmentation stage. In this segmentation, we proposed a multi kernel k-means clustering algorithm (MKKM). Here, we have used three types of hybrid kernel such as (Linear + Quadratic), (Linear + Radial basis kernel) and (Quadratic + Radial basis kernel). In this section, we have analyzed the performance of segmentation using number of methods such as Linear kernel with K-means clustering, Quadratic kernel with K-means clustering, RBF kernel with K-means clustering, K-means clustering and three multi kernel K-means clustering. When analyzing Figure 6, Linear + Quadratic based segmentation achieved the maximum accuracy of 99.61%, Linear kernel based segmentation achieved the accuracy of 96.74%, Quadratic kernel based segmentation achieved the accuracy of 96.74%, Linear + RBF based segmentation achieved the accuracy of 98.77, Quadratic + RBF based segmentation achieved the accuracy of 98.61%, RBF based segmentation achieved the accuracy of 97.53% and K-means algorithm achieved the accuracy of 82.63%. From the result, we understood Linear + Quadratic kernel based segmentation approach achieved the maximum accuracy.

V. Conclusions

Our proposed kidney stone segmentation technique for ultrasound images was implemented in the platform of MATLAB version 7.12. The results were evaluated based on the performance measures such as Sensitivity, Specificity, and Accuracy. Our proposed work was evaluated with each metrics and then the results of each metrics were also analyzed. Different multi kernel functions were analyzed and compared. The proposed work achieved the maximum accuracy, sensitivity, and specificity value for all input sample images when compared with existing methods. In future, the researcher will have sufficient opportunities to perform various segmentation techniques in order to improve the segmentation accuracy and produce newer heights of excellence in performance.

References:

1. Ioannis Manousakas, Chih-Ching Lai and Wan-Yi Chang, "*A 3D Ultrasound Renal Calculi Fragmentation Image Analysis System for Extracorporeal Shock Wave Lithotripsy*", International Symposium on Computer, Communication, Control and Automation, Vol. 1, pp. 303-306, 2010.
2. Jie-Yu He, Sui-Ping Deng and Jian-Ming Ouyang, "*Morphology, Particle Size Distribution, Aggregation, and Crystal Phase of Nanocrystallites in the Urine of Healthy Persons and Lithogenic Patients*" IEEE Transactions On Nanobioscience, Vol. 9, No. 2, pp. 156-163, June 2010.
3. Ratha Jeyalakshmi and Ramar Kadarkarai, "*Segmentation and feature extraction of fluid-filled uterine fibroid-A knowledge-based approach*", Maejo International Journal of Science and Technology, Vol. 4, No. 3, pp. 405-416, 2010.
4. Tamilselvi and Thangaraj, "*Segmentation of Calculi from Ultrasound Kidney Images by Region Indicator with Contour Segmentation Method*", Global Journal of Computer Science and Technology Volume XI Issue XXII, 2011.
5. Akbari and Fei, "*3D ultrasound image segmentation using wavelet support vector machines*", Journal of Medical Physics, vol.39, no.6, pp.2972-84, 2012.
6. Anjit Rajan and Jennifer Ranjani, "*Segment based Detection and Quantification of Kidney Stones and its Symmetric Analysis using Texture Properties based on Logical Operators with Ultra Sound Scanning*", International Journal of Computer Applications (0975 – 8887), 2013.
7. Ashish K. Rudra , Ananda S. Chowdhury and Ahmed Elnakib, *"Kidney segmentation using graph cuts and pixel connectivity*", Pattern Recognition Letters 34, 1470–1475, 2013.
8. Prema T. Akkasaligar and Sunanda Biradar, "*Classification of Medical Ultrasound Images of Kidney*", International Journal of Computer Applications, 2014.

9. Viswanath K, Gunasundari R and Hussan SA., "*VLSI implementation and analysis of kidney stone detection by level set segmentation and ANN classi-fication*", Journal of Methods and Programs in Computer Science, 48: 612–22, 2015.
10. Mariam Wagih Attia, Hossam El-Din Moustafa and Abou-Chadi, "*Classification of Ultrasound Kidney Images using PCA and Neural Networks*", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 4, 2015.
11. Nikita Derle and Devidas Dighe, "*4D Image Analysis and Diagnosis of Kidney Disease Using DCE-MRI Images*", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 4, April 2015.
12. Divya Krishna, Akkala and Bharath, "*Computer Aided Abnormality Detection for Kidney on FPGA Based IoT Enabled Portable Ultrasound Imaging System*", Elsevier, 2016.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Personalization and Recommendation Issues: A Study

J.I. Christy Eunaicy[1], S.Suguna [2]
Research Scholar, Bharathiyar University, Coimbatore
[2] Asst.Prof, Dept of Computer Science, Sri Meenakshi Govt Arts college for Women, Madurai
[1]eunaicy@gmail.com
[2]kt.suguna@gmail.com

ABSTRACT *The World Wide Web is an interactive and popular platform to transfer information. Web Usage Mining is the type of web mining and it is application of data mining techniques. Web Usage Mining has become helpful for website management, and personalisation. Web usage mining is the process of extracting user patterns from the web usage. In web usage mining, pre-processing plays a key role, since large amount of irrelevant information are present in the web. In this paper, issues related with personalization and recommendation processes are analyzed towards prediction of user next request in web usage mining. This paper also deals with the importance of cleaning of weblogs.*
Keywords: *Preprocessing, Personalization, Recommendation, Data Cleaning, Clustering*

Introduction

Web mining has three distinct phases involved – content, structure and usage mining of web data. Mining the content involves extracting the relevant information, structure mining studies the structure and prototype and usage mining is the analysis of the discovered patterns. Web usage mining is one of the web mining techniques. Web users use the collection of web pages and web information are stored in web server. This usage of data is to provide leading paths to access the Web pages. This information is often gathered automatically by access web log through the Web server [1]. Web Usage Mining deals with understanding of user behaviour, while interacting with web site, by using various log files to extract knowledge from them. This extracted knowledge can be applied for efficient reorganization of web site, better personalization and recommendation, improvement in links and navigation, attracting more advertisement [2]. There are four main tasks for performing WUM: Preprocessing, Pattern Discovery, Pattern Analysis, Personalization and recommendation. This papers structured in four sections. Section-2 present survey of preprocessing and personalization works. Section-3 presents overview of related works done on personalization and recommendation area. Section-4 presents the proposed algorithm. Section 5 presents the performance analysis and rest of the article has Conclusion and future directions of the work.

II SURVEY OF RELATED WORKS

Sudheer Reddy et al [3] presented various details about data pre-processing activities that are necessary to perform web usage mining. Their work reduced the size of log file and also increased the quality of data available. Rachit Goel and Sandeep Jain [5] provided an improvised technique for performing the data cleaning technique on server log. That approach reduced the number of records and log files size and also increased the quality of the available data. Ankit Kharwar et al., [8] applied the data mining and knowledge discovery techniques to www server access logs in order to display the details of preprocessing data. Their experiments reduced log file size and also increased the quality of data available. Sukumar et al.,[4] analyzed the web server logs by using data preprocessing activities . The results showed the effectiveness and applicability of the heuristic based data preprocessing techniques and algorithm approaches. Aye [6] suggested that the data in the log file must be preprocessed to enhance the effectiveness and ease of the mining process and also the major task os removing noisy and unrelated data to lessen the data volume for the pattern discovery phase. Wasvand chandrama et al.,[7] proposed a model which enabled the administrator to access the web log file and to perform data preprocessing on it and also they used classification algorithm to identify interested web site. Mona et al., [9] proposed a system to handle websites where users are not comfortable to identifying them, due to this, backward reference took lesser time and also the session identification has higher precision. Many different techniques have been considered to discover patterns from web browsing logs for web personalization.

Table 1 shows the analysis of preprocessing and personalization issues with existing works:

S.No	Author	Function	Method	Issues Identified
1	Ramya et al., [10]	1. A complete preprocessing methodology for merging, data cleaning, user/session identification and data formatting and summarization. 2. Activities to improve the quality of data by reducing the quantity of data.	Relational database model	No solution for technical problems
2	Suguna and Sharmila [11]	1. Effectively performs preprocessing which support user interest level grouping. 2. Session and frequency values are considered as the key for identifying user interest level.	User Interest Level Preprocessing algorithm	No solution for time of user identification and existence of local caches and proxy servers and firewall.
3	Vijayashri Losarwar et al.,[12]	1. The importance of data preprocessing methods and various steps involved in getting the required content effectively. 2. A complete preprocessing technique is being proposed to preprocess the web log for extraction of user patterns.	Discussed about the preprocessing steps.	No efficient technique for distinct user identification and other issues.
4	Aldekhail[13]	1. Removes the irrelevant entries from web log and discards the uninterested attributes from log file. 2. User and sessions are identified.	Discussed the importance of data preprocessing methods and various steps involved in getting the required content effectively.	There is no specific method for overcoming the issues of session identification.
5	Sanjay Kumar Dwivedi et al.,[23]	Several data preprocessing techniques has been presented to prepare raw data suitable for mining and analysis tasks.	Discussed the data preprocessing steps.	Time consuming

Personalization is the ability to provide content and services that are tailored to individuals based on knowledge about their preferences and behaviour. Personalization requires collecting visitor information and gets knowledge that how is the user's behaviour on particular site which helps website administrator to decide "what information present to which user and how to present it"[14]. Personalizing web pages is one of the efficient parts of web-mining, which enables to understand user interests to offer services and enables them to discover web pages, text documents, multimedia files, images and other types of resources from web according to their choice [14]. Web personalization models include rules-based filtering, based on "if this, then that" rules processing, and collaborative filtering, which serves relevant material to customers by combining their own personal preferences with the preferences of like-minded others. Collaborative filtering works well for books, music, video, etc. However, it does not work well for a number of categories such as apparel, jewelry, cosmetics, etc. Recently, another method, "Prediction Based on Benefit", has been proposed for products with complex attributes such as apparel [16]. It recommends or predicts what kind of items the user may prefer. This can be made either directly based on the dataset collected in information collection phase which could be memory based or model based or through the system's observed activities of the user[17]. Recommender systems can be classified broadly into several categories depending on the information they use to recommend items. The use of efficient and accurate recommendation techniques is very important for a system that will provide good and useful recommendation to its individual users.

Table 2 shows the works related to presonalization and recommendation .

S.No	Author	Function	Method	Issues Identified
1	Blerina Lika et al., [18]	It incorporates classification methods in a pure CF system while the use of demographic data helps for the identification of other users with similar behaviour.	Classification Algorithm with similarity techniques.	1. Maintaining large amount of frequently changing information is missing. 2. No information about improving accuracy and relevancy of new recommendation.
2	Raju et al., [19]	1.An efficient web personalization approach based on user browsing time interval and utilizing web usage log. 2.Depends on the individual web usage pattern it can efficiently suggest resources that users will be most interested in over a period of time. 3.Web personalization is minimized and the run-time processing load over the server and improvise the user satisfactory level.	PR_Generation, Search_Periodic_ Activity	Lack of effectiveness in longer period in multiple intervals.
3	Anna Alphy et al.,[20]	1.It overcomes the information overload by handling dynamic behaviours of users. 2. It has higher precision, coverage, measure, and scalability than the traditional collaborative filtering systems.	WebBluegillRecom-annealing dynamic recommender uses swarm intelligence.	To track evolving user profiles.
4	Priyanga et al.,[21]	1.It provides a merging between the content semantics as well as the usage data that are stated as ontology terms. 2.This system supports the computation of the navigational patterns that are semantically improvised, so that constructive recommendations can be successfully engendered.	Discussed the recommendation process.	1.No idea about handling heterogeneous data sources. 2. Quality has to be improved.
5	S.Padmaja, et al[22]	Prediction of user behavior.	Improved K-Means Clustering	User has to specify the K values.

The following issues are identified based on this survey:

A. Preprocessing Issues

- More memory space needed.
- Timing Complexity
- Problem in Identifying distinct user

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

- Problem in Identifying different sessions belonging to different users.
- Caching problem in path completion.

B. Personalization & Recommendation Issues

- To improve relevancy of recommendation by filtering relevant information from large data collections that are apt for the user
- Clustering users into user communities.
- Handling heterogeneous data sources.
- Identifying new metrics to evaluate the performance of recommender systems.
- Maintaining large amount of frequently changing information.
- Improving accuracy and relevancy of new recommendation.
- The need for improvement in accuracy and handling large volume of heterogeneous information necessitates the use of specialized big data platforms with suitable mining techniques for personalization.
- Retrieval of information from huge volumes of data in diversified areas results in a tedious process.
- Accesses to cached web pages are not recorded in the server log and hence such information is missed

IV. PROPOSED ALGORITHM

In this section algorithm for Cleaning and User Identification process is proposed.

With the help of web logs we can predict user's next request. But the information available in the web logs is not suitable for further process. Hence the cleaning of web logs is necessary. Therefore data cleaning algorithm should be implemented with the server logs. After that only valid HTML documents remains. These cleaned data will be used to identify the user sessions.

Data Cleaning Algorithm

1.Remove logs not containing HTML documents from the user requested pages.
2. Delete logs having codes other than 200, 304 and 306 with GET method.
3. Delete logs generated for extremely long user sessions by search engine.
4. Delete entries related to web robots request and system request.
5. Pages that are accessed less frequently than the minimum threshold value can be deleted if they are not in the browser cache (304) or in the proxy cache (306).

B. User Identification

There are several ways to identify individual visitors. They are, using cookies, registration details and IP address. In general to identify the user, group the web logs based on the IPAddress and the browsing agents.

V. PERFORMANCE ANALYSIS

In this section, performances of proposed algorithms are analysed using simulated sports events related server logs.

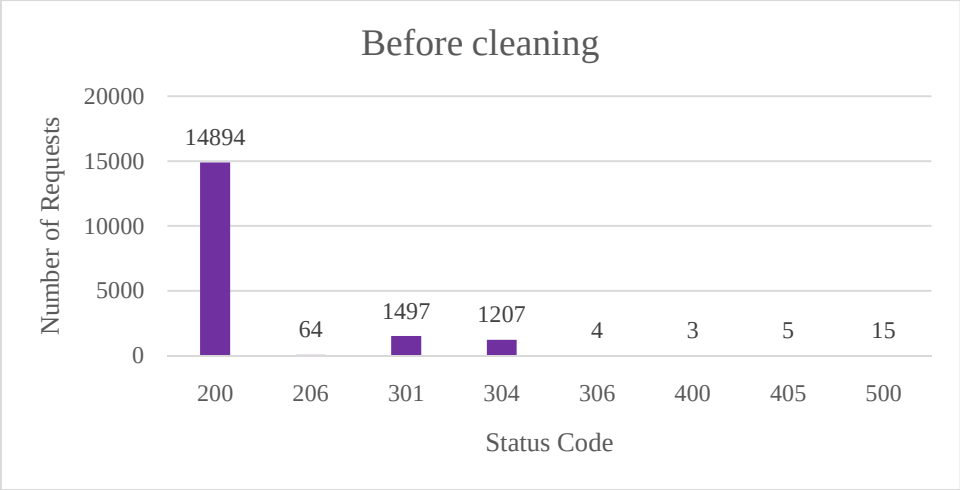


Fig 1: Analysis of Server logs based on Status Code – Before Cleaning

The fig. 1 shows the log file with unwanted detail, which has to be cleaned with the help of data cleaning algorithm.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

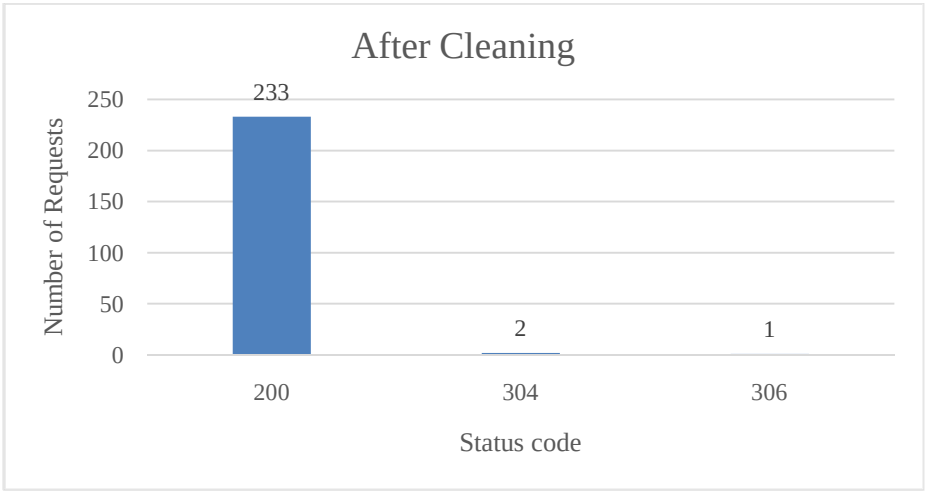


Fig 2: Analysis of Server logs based on Status Code – After Cleaning

The above graph shows that after applying the data cleaning algorithm, the unwanted files have been removed and the most frequently used web logs only shown. These web logs will be used for user session identification stage.

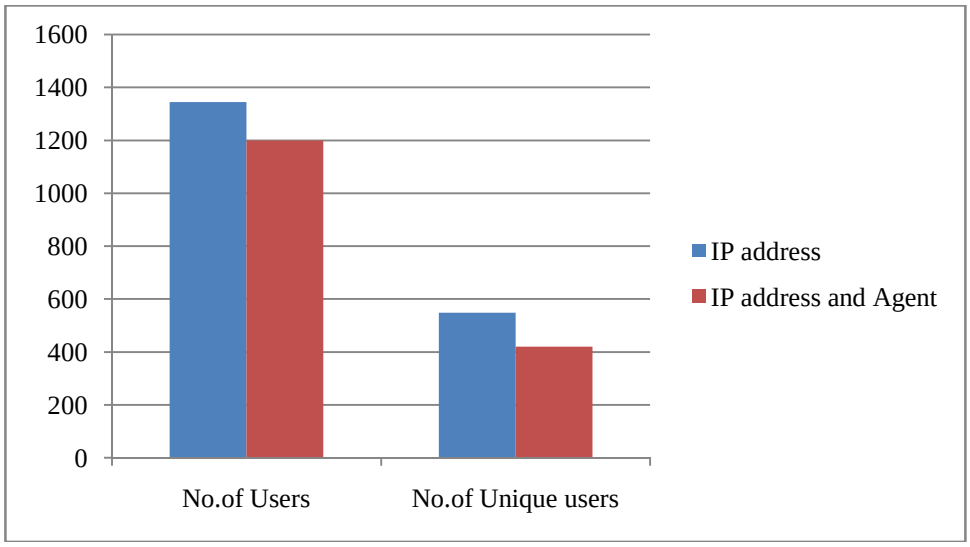


Fig 3: Processing of User Identification

The above graph shows that the cleaned data of simulated sports event are grouped based on the IPAddress and the browsing agents used by the users

VI. CONCLUSIONS

This review, investigated the web usage mining importance in many areas like E-commerce, education, medicine and website designing. These kind of work helps to web developers for improving websites, and site usability and accessibility. It has analyzed the significance of web usage mining phases such as preprocessing, personalization and recommendation information in detail. Also, it has described the various research issues and challenges. The importance of cleaning of web logs is also analysed.

VII. FUTURE WORK

In future the system could be extended to complete the rest of the preprocessing stage, so that to generate a learning graph to predict and prefetch the user‘s next request.

REFERENCES

1. ANITHA, ESAKKI,"A SURVEY ON PREDICTING USER BEHAVIOUR BASED ON WEB SERVER LOG FILES IN A WEB USAGE MINING", INTERNATIONAL CONFERENCE ON COMPUTING TECHNOLOGIES AND INTELLIGENT DATA ENGINEERING", IEEE, 7-9 JAN. 2016.
2. CHINTAN R. VARNAGAR, NIRALI N. MADHAK, TRUPTI M. KODINARIYA, JAYESH N. RATHOD, "WEB USAGE MINING: A REVIEW ON PROCESS, METHODS AND TECHNIQUES", INTERNATIONAL CONFERENCE ON INFORMATION COMMUNICATION AND EMBEDDED SYSTEMS(ICES), 22 FEB 2013, PUBLISHER IEEE.

3. Sudheer Reddy, Kantha Reddy, Sitaramalu, "An effective data preprocessing method for web usage mining", International conference on Information Communication and Embedded systems(ICICES), 2013, Publisher IEEE.
4. Sukumar, Robert and Yuvaraj, " Review on modern data preprocessing techniques in web usage mining (WUM), International conference on computational systems and Information systems for sustainable solutions, published by IEEE, 2016.
5. Rachit Goel and Sandeep Jain, " Improvisation in web mining techniques by scrubbing log files", International Journal of Advanced Research in Computer Science, vol 5, No 5, May-June 2014.
6. Theint Theint Aye , "Web log cleaning for mining of web usage patterns", Proceedings of the 3rd International Conference on Computer Resesarch and Development Mar 11 -13, IEEE Xplore Press, Shanghai, pp-490-494.
7. Wasvand Chandrama, Devale, Ravindra Murumkar, "Data Preprocessing method of web usage mining for data cleaning and identifying user navigational pattern", International Journal of Innovative Science, Engineering & Technology, Vol 1, Issue 10, Dec 2014.
8. Ankit R.Kharwar, Chandi A Naik, Niyanta K.Desai, "International Journal of Emerging Technology and Advanced Engineering, Vol 3, Issue 10, Oct 2013.
9. Mona Kamat, Bakal and Madhu Nashipudi, " Improved Data preparation technique in web usage mining", International Journal of Computer Networks and Communications Security, Vol 1, No.7, Dec 2013, 284 – 291.
10. Ramya, Shreedhara and Kavitha, "Preprocessing: A Prerequisite for Discovering Patterns in Web Usage Mining Process", International Conference on Communication and Electronics Information, IEEE Explorer, pp 317 - 321, 2011.
11. R. Suguna, D. Sharmila, "User Interest Level Based Preprocessing Algorithms Using Web Usage Mining", International Journal on Computer Science and Engineering (IJCSE), Vol. 5, No. 09, Sep 2013, pp. 815-822.
12. Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore.
13. Aldekhail, "Application and Significance of Web Usage Mining in the 21st Century: A Literature Review", International Journal of Computer Theory and Engineering, Vol. 8, No. 1, February 2016.
14. Rutvija Pandya, "Web Usage Mining with Personalization on Social Web", International Journal of Engineering Trends and Technology (IJETT) – Volume 29 Number 6 - November 2015 pp 325-328.
15. Anupama Prasanth, "Web Personalization Using Web Usage Mining Techniques", International Journal of Current Engineering and Scientific Research (Ijcesr), Volume-3, Issue-3, 2016, pp 45-49.
16. Raju , Suresh Babu, "A Novel Approaches In Web Mining Techniques In Case Of Web Personalization", International Journal Of Research In Computer Applications And Robotics", Vol.3 Issue.2, Pg.: 6-12 February 2015, pp 6-12.
17. Isinkaye et al., "Recommendation systems: Principles, methods and Evaluation", Egyptian Informatics Journal (2015) 16, 261–273/ www.elseiver/locate/eij.
18. Blerina Lika et al., "Facing the cold start problem in recommender systems", Expert Systems with Applications", 41 (2014), Elseiver. Volume 41, Issue 4, Part 2, March 2014, Pages 2065-2073.
19. Raju et al., " An Efficient Web Personalization Approach based on Periodic Accessibility and Web Usage Mining", Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, Number 8 (2017) pp. 2289-2308.
20. Anna Alphy et al.,"A Dynamic Recommender System for Improved Web Usage Mining and CRM Using Swarm Intelligence", The Scientific World Journal Volume 2015 (2015), Article ID 193631, 16 pages.
21. Priyanga, Naveen, "User Identification, Classification and Recommendation in Web Usage Mining – An Approach for Personalized Web Mining", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 4, April 2015.
22. Padmaja, Ananthi Shesasayee, "Clustering of user behaviour based on web log data using improved k-means clustering algorithm", IJET, vol 8 No , Feb – Mar 2016, 305-310
23. Sanjay Kumar Dwivedi, Bhupesh Rawat, "A Review Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process", International Conference on Green Computing and Internet of Things, IEEE Explore, 2015.
24. G. Arumugam, S. Suguna,"Optimal Algorithms for Generation of User Session Sequences Using Server Side Web Logs", International Conference on "Network and Service Security 2009, NS2'09" at Paris from june 24-26, Published in the IEEE Xplore Digital Library, pp. 1-6.
25. S. Suguna, V.Sundaravadivelu, I. Christy Eunaicy,"Predictive Prefetching of web pages using Server Side user logs", International Conference on Innovations in contemporary IT Research (ICITR 2012) 17-18 Feb 2012

A Novel Framework for Three Dimensional Craniofacial Reconstruction Based on Skin and Cranial Landmarks

Chitra Devi.M, [2] Pushpa Rani.M
[1,2]Department of Computer Science, Mother Teresa Women's University,
Kodaikanal, Tamilnadu, India

ABSTRACT *Face reconstruction is one of the most outstanding research areas in forensic medicine which has the great impact in medical field, in recent years. The aim of face reconstruction is estimate the unknown face from skull which support in reconstruction and recognition. Face reconstruction is the most challenging task since accurate face reconstruction is a complex technique. Hence we propose the craniofacial reconstruction method. Craniofacial reconstruction is the application of anthropology and forensic science. Craniofacial reconstruction is reconstruct the face based on shape and surface structure of skin and skull image. In this paper we reconstruct the face from 3D MRI human heads presented. Our proposed work reconstructs the face from 3D image by using the various features. Here we consider features are 3D image skull, face age, gender, Body Mass Index [BMI], and race values based on these features we reconstruct the face accurately. Our proposed method analyzes the accuracy and execution time of face reconstruction.*

Keywords: *Craniofacial reconstruction, 3D MRI image, Body mass index, race.*

Introduction

Facial reconstruction is the important technique in several research areas, especially in anthropology and forensic medicine. These two fields are widely research the face reconstruction. From now several facial reconstruction techniques are introduced [1, 7, 8 and 9]. This paper was proposed the forensic facial approximation technique which is reconstruct the face based on volume and shape of skull. Here facial approximation reconstructs the face by using both scientific standard and artistic skill. These methods classified in various reconstructive methods for reproduce the face. In [2] author proposed the Craniofacial reconstruction model which detect the face based on the land mark values on facial points. Here author reconstruct the face to build 3D statistical model of skull and face soft tissues from 3D CT images. The statistical model used for reproduce the face soft tissue from skull. Facial reconstruction is very important in various fields such as accident,

Terrorist attacks, genocide, large number of death result from wars. Here we implement the three dimensional reconstruction of same skull in each set tissue by the depth measurement. Facial reconstruction by using the computerized techniques which is based on the landmark interpolation procedure. Land mark interpolation constructed by using the single static facial surface template. By using this technique we reconstruct the facial point from skull [4].

Craniofacial reconstruction method reconstructs the face by two types of classifications such as traditional manual method and computer based approach. In traditional manual method face reconstruction consists of physically modeling face on skull replica with clay or pastime. These techniques take long time for facial reconstruction. Computer based approach reconstructs the face in short time [22]. In [23] face reconstructed by using the computerized technique. These types of techniques increase the system flexibility, speed and efficiency. Here reconstructs the face from either 2D or 3D images. Both methods reconstruct the face by using the various features such as Body Mass Index value, stature ethnic group or else race, age and gender. This method face reconstructs is differ from the other methods by the selection of land mark and skull features. Several methods are using the skull template for face reconstruction. Paper [5] was proposes the reconstruction method based on direct anthropometry using calipers which is the standard technique for Craniofacial surgical planning and outcome assessment. These methods have the several drawbacks such as time of reconstruction is high, limitation of data, training required and invasiveness. Normally humans are differentiated by the facial features; each person has the different skull structure and soft tissue depth. Here we consider lips, nose, eyes and ears are the facial features. Here we reconstruct the face by these facial features; these features are varied by each person [2]. In [6] author proposed face reconstruction by using evaluation of implicit critical social information. For example human face constructs the wide range of systematic information such as age, gender, expressions, race, pose etc., by using these feature we reconstructs the face. These method used in human computer interface and visual surveillance system [9].

Contributions

- We reconstruct the face by using the craniofacial reconstruction method which reconstructs the face by the feature extraction.
- We implement the craniofacial reconstruction from the 3D skull and skin image. Here we first find the facial points for each side of image and then we compare the facial points in skull with skin. Thus we reconstruct the face from 3D skull image.
- Our method finally analyses performance of face reconstruction with the parameter of the accuracy and execution time.

Remaining part of this paper is organized as follows. Section 2 describes related work of this research, section 3

illustrates the problem formulation, section 4 describes the detailed proposed system, section 5 includes experimental results and we conclude the proposed work and future work in section 6.

RELATED WORK

Several existing methods are involved for face reconstruction such as nonlinear deformable model, manual identification method, computerized three dimensional cranio face reconstruction technique, and computed tomography. Nonlinear deformable model constructed by author Adel kermi et al [12]. Deformable model proposes the computerized 3D MRI facial reconstruction method which is constrained the face by the information of soft tissue thickness in human heads at certain land marks. It is also known as the B- spline free from deformation model because it's fully based on non-linear registered technique. This model change the soft tissue thickness and locations are changed until reconstruct the accurate face reconstructed. Deformable model reconstruct the face by using the five stages such as generation of simplex meshes, automatic location of land marks, projection of land marks from skin surface to unknown skull, selection of land mark for reconstruction and add anthropometrical constraints and evaluate the deformable model. Here automatic land mark selection and insertion used as the four methods such as, Mean Curvature without uniform selection [MEA], Gaussian Curvature without uniform selection [GAU], using Mean and Gaussian without uniform selection [MEA-GAU] and using Mean and Gaussian with uniform selection [MEA-GAU-UNI]. Paper [15] proposed the manual identification method for facial reconstruction. Here we reproduce the set of real face points by using the comparison of the soft tissue points for each craniometric point conjunction with the skull points. In this method we increase the points until all the anatomical rules are used. Various existing methods are used the craniometric points for facial reconstruction, but it involves the various interpolation method and restriction technique.

Computerized three dimensional cranio face reconstruction technique which is used for human identification, this technique reconstruct the face without using the any facial template. This technique requires the age, gender and BMI values skull. This approach reconstructs the face from the 3D image of target skull [12]. Jimena et al [11] was proposed the tomography tool for facial reconstruction. Computed tomography is the powerful tool for face landmark reconstruction which reconstructs the face by estimation of the morphological variation. These variance calculated by the analyzing of measurement error. Here author identify the three types of land marks which is obtained by using following types, type1- observing the biological structures, that are easy to find successively. Type 2- is the skull geometry which defines as the local maxima and minima curvature. Type 3- defined as the external points for instance end points of breadth. MRI is the faster acquisition method than the manual measurement method. Authors [19] were proposed morphological feature based face reconstruction technique. Here MRI data sets are store the morphological features such as gender, race and age. Based on these features we reconstruct the face. The main problem in face reconstruction of these model is large number of photographs are stored for these characteristics. It requires the large amount of storage space for processing the single image. It leads the high computational complexity. Therefore we choose the head model choose from database based on the features.

Problem Formulation

Human face identification from 3D MRI skull image is the major issue. Several methods were proposed for Craniofacial reconstruction such as manual and statistical model. In manual method the skull points or image are physically move until the accurate face found. The statistical model we reconstruct the face by the depth measurement of soft tissue. Bruynooghe et al.[17] proposed computerized Craniofacial reconstruction model. This model consists of the Craniofacial template that is warped towards the unknown target skull. Here the skull transformed by the PCA based transformation model. In the previous works statistical model constructed by the facial shape variations and soft tissue depth measurement. Here we measure the soft tissue depth from the dense of head CT images. The major benefits of these model are requires the sufficiently large amount of data for morphological features such as age, gender, and race. This method has drawback for facial reconstruction, this model use the CT image as input image all CT images are lying down and faces seen in upward direction which leads the low accuracy.

Authors [18] were proposed the facial reconstruction for accurate identification of human face from skeleton. Here we correlate the skull and face points of target image by using canonical correlation analysis which mapping the skull and skin points. This method not like an existing method which mapping the skull and skin points only with the correlation points of statistical model. Here author introduce the region fusion strategy which improves the matching accuracy of the identification method.

Skull statistical shape model was constructed as follows:

$$S(a) = \overline{S} + \sum_{k-1}^{p} a_k U_k \quad (4) \qquad (1)$$

Face statistical shape model was constructed as follows:

$$F(c) = \overline{F} + \sum_{k-1}^{q} c_k V_k \quad (5) \qquad (2)$$

By the designing of above model we meets the several drawbacks such as the high execution time for face reconstruction, lower accuracy results, high uncertainty and high error rate. To overcome the above mentioned problem our proposed work design the novel frame work craniofacial reconstruction which provides the high accuracy and lower execution time for face reconstruction.

Proposed Work

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Our proposed work reconstruct the human face in more accurate and faster manner by using the method of cranio face reconstruction which reconstruct the face from the 3D MRI image.

3D Craniofacial Reconstruction

Craniofacial reconstruction hopes to estimate an individual's face look from its skull using the information of MRI datasets. MRI data set contains morphological characteristics of human such as age, gender, race and Body Mass Index. With the growth of three dimensional digitalization technologies, the research on both computers sided and manual sided craniofacial reconstruction has widely received interest. The valuation of the craniofacial reconstruction has a vital significance in improving the craniofacial reconstruction methods. The proposed approach it include the following inputs: 3D image of the target skull, 3D image of the skin, set of 72 landmarks placed on the both skull and skin surface where points are inserted by manually. The absolute input data required by the function is a set of characteristic metrics of the identified person: age, gender, race and BMI. Age, gender and race can be assumed from skull morphology that is always known by the user. However, Body Mass Index range will be an unidentified parameter, which will have to be estimated. Stages of proposed work are land mark insertion module and skin mesh generated module. Land mark module takes a charge of placing each landmark on skull and skin surfaces. In this phase 72 landmarks are inserted in both skull and skin surfaces [placed in all views such anterior view and lateral views]. Based on this information, skin mesh is generated.

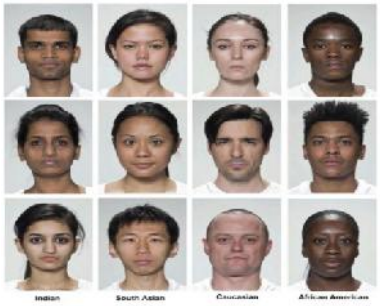


Fig.1. Racial groups of human faces in various countries [Indian, South Asian, Caucasian, and African American]

Land mark Insertion Module

Manual landmarks insertion was proposed in this approach. For accurate detection of human face, 72 set of landmarks inserted on both skull and skin surfaces in all sides. Set of landmarks inserted on skull surface is shown in fig.2 and the overall architecture for CFR is shown in fig.3.

Landmark insertion module is in charging of placing 72 landmarks on the skull and skin surfaces and gets the input data from MRI data set which contains set of parameters such as age, gender, race and BMI. Here two set of points are used for reconstructing the face. Furthermore, this module focuses on the selection of certain landmarks among the 72 landmarks based on the morphological characteristics. Once these landmarks are selected, we construct the 3D mesh belonging to the skull and skin landmarks. The selected landmarks (skull and skin) coincide with the positions of anatomical points. According to the results from the new desired mesh surface, reconstruction accuracy was evaluated.

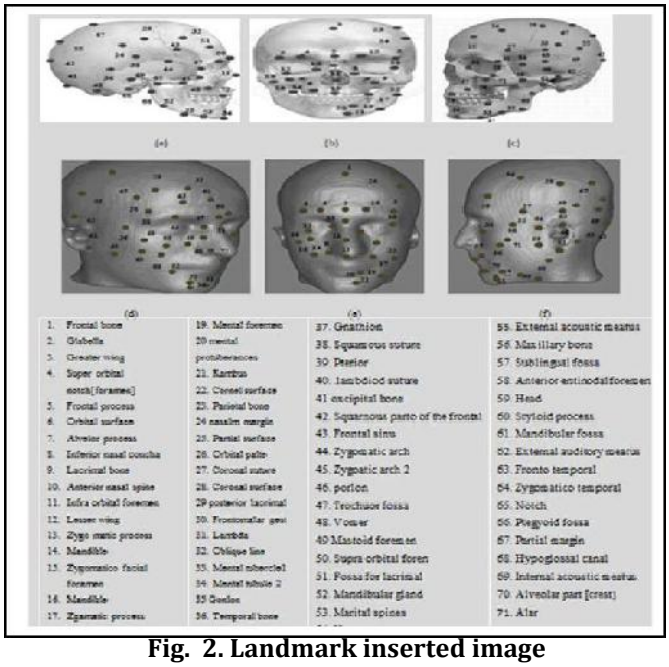


Fig. 2. Landmark inserted image

Figure 2 illustrates the location of land mark points in 3D skin and skull image, here figure (a) & (c) shows the side view of skull image, and (b) shows the front land mark points in skull. In figure 1 (d) & (f) shows the land mark

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

points in skin and (e) shows the front land mark points in skin image.

Algorithm 1: 3D-Craniofacial Reconstruction

Input: 3D skull, 3D skin, values from MRI data sets [Age, BMI, Race and Gender] and set of landmarks for both skull and skin [Ci and Si]
Output: Reconstructed Face {Rf}

1. Begin
2. Given input image of skull, skin, MRI Data sets and 72 landmarks
3. Landmark inserted for both skull and skin images
4. Build a full 3D Mesh from both set of points
   4.1 Consider set of landmarks {Ci & Si}
   4.2 for (Ci=1;Ci<=72;Ci++) {
   Select one skull landmark}
   for (Si=1;Si<=72; Si++) {
   Select one skin landmark (according to morphological characteristics)}
   4.3 if(Ci==Si) {
   Construct skull with face }
   else {
   Go to step 1 }
5. Repeat until (Ci==Si)
6. Reconstructed face {Rf}

Skin Mesh Generation Module:

The Skin Mesh Generation Module corresponds to the nucleus functional module in the 3D reconstruction application. From the results generated from 4.2 (Land Mark Insertion Module), it deal with to create a full 3D mesh denoting skin (soft tissue) belonging to the skull. The vital objective of this module is to make a set of intermediate points on the skull surface whose depth values can be interpose from thickness in reference points. The complete set of points such as land marks (72) and appropriate intermediate points will combine to construct the final skin mesh. For this reason the land mark insertion module obtains the set of 72 landmarks (positions and depths).

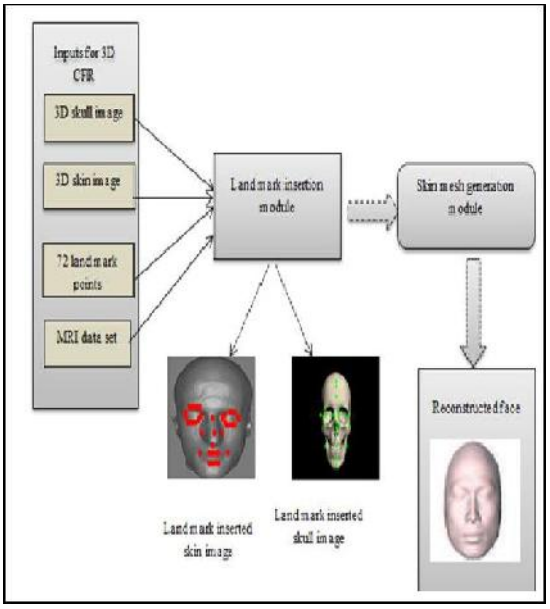


Fig. 3. Over all Architecture of CFR

Figure 3 shows the overall architecture of craniofacial reconstruction which reconstructs the face by the morphological characteristics which includes in the MRI data set. Here we take the age, gender, BMI and race from these parameter we insert the land mark and mesh generate for skin and skull of 3D image and finally we reconstruct the face.

Results And Discussion

Our proposed work simulated by using the matlab software. In this paper we implement the face reconstruction by the Craniofacial reconstruction method. In this section we have to discuss the result analysis of our

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

proposed work.

Results for Proposed Work

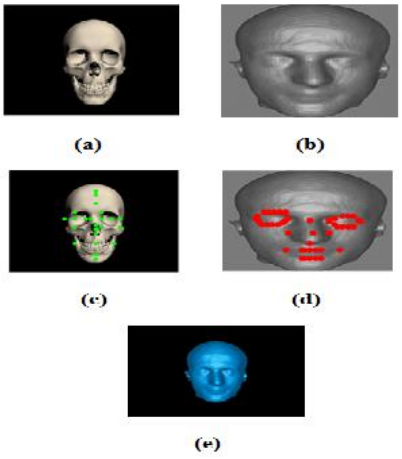


Fig. 4. Results for 3D- Craniofacial Reconstruction a. 3D skull, b. 3D skin image, c. Landmarks inserted image for skull d. Landmarks inserted image for skin e. Reconstructed face

Comparative Results

Our proposed work provides the excellent performance compared with existing methods. Here we compared the accuracy and execution time of face reconstruction with the other existing methods.

Table 1 : Comparison of the Accuracy Rate of Existing Methods and our Proposed Method

Methods	Accuracy rate
Xiaoming Liu et al [19]	20%
Yongli Hu et al [20]	72%
Peter Claes et al [4]	81.1%
Li Luo,1 Mengyang Wang et al [21]	86%
Proposed method	97%

Table 1 illustrates the accuracy rate analysis of proposed work with existing method. Here our proposed work provides the 97 percentage of accuracy.

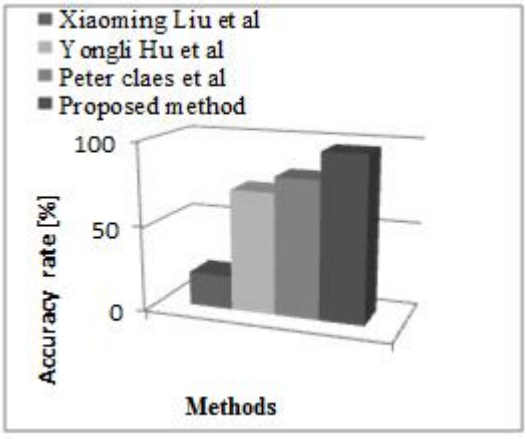


Fig. 5. Comparative results for 3D Craniofacial Reconstruction

Figure 5 shows the comparative results of accuracy for our proposed work; here graph clearly shows our

proposed work provides the better accuracy compared with the other previous works.

Table 2: Result of the Proposed Method

Method	Accuracy rate [%]	Execution time [sec]
Cranio face reconstruction	97%	12

Table 2 illustrates the accuracy time and execution time of our proposed work. Our proposed work take the 12 sec as the execution time and accuracy rate is 97% is relatively high compared with the other existing method.

Conclusion

We propose the craniofacial reconstruction method for reconstruct the face from 3D skull and skin image. Our proposed work reconstruct the face by using two modules of process such as landmark insertion module and skin mesh generation module.

First our proposed work inserts the landmark in skin and skull by using the morphology features such as race, age, gender and BMI values. These parameters are getting from the input MRI data set and then insert the land mark manually. From that landmark inserted image we reconstruct the face by the skin mesh generation module which generates the mesh for landmark inserted skin and skull image. Therefore our method reconstructs the face accurately. Here we consider the two performance metrics as accuracy and execution time. Our method achieve the higher accuracy and low execution time. Here we set the land mark manually In future direction we have to planned for reconstruct the face by the method of landmark insertion by the matching process. In addition we have planned for measure the error rate in face reconstruction and also improve the visual quality.

References

[1]. Á. Kustár, Z. Gerendás, I. Kalina, F. Fazekas, B. Vári, Sz. Honti5 & Sz. Makra "FACE-R – 3D skull and face database for virtual anthropology research" vol 105, pp 313- 319
[2]. M.Berar, M. desvignes, G.bailly, Y. payan" 3D statistical facial reconstruction ",springer, pp1-6
[3]. John M. Starbuck a, Richard E. Ward" The effect of tissue depth variation on craniofacial reconstructions" Forensic Science International, pp 130-136, 2007
[4]. Peter Claes a, Dirk Vandermeulen a, Sven De Greef b, Guy Willems b, Paul Suetens" Craniofacial reconstruction using a combined statistical model of face shape and soft tissue depths: Methodology and validation" Forensic Science International, pp 148-156, 2006.
[5]. Shu Liang, Jia Wu, Seth M. Weinberg, Linda G. Shapiro" Improved Detection of Landmarks on 3D Human Face Data" International Conference of the IEEE, pp 1-4, 2013.
[6]. Siyao Fu, Haibo He, Zeng-Guang Hou" race classification from face: a survey" IEEE transactions on pattern analysis and machine intelligence, pp 1-25, 2014.
[7]. Won-Joon Lee,M.Sc.; Caroline M. Wilkinson, and Hyeon-Shik Hwang," An Accuracy Assessment of Forensic Computerized Facial Reconstruction Employing Cone-Beam Computed Tomography from Live Subjects" Forensic Science, pp 318-328, 2012.
[8]. Caroline Wilkinson" Facial reconstruction – anatomical art or artistic anatomy?" journal of anatomy, pp235–250, 2010.
[9]. Caroline Wilkinson" Computerized Forensic Facial Reconstruction" Forensic Science, Medicine, and Pathology, pp 173-177,2005.
[10]. Dr.M.Pusha Rani, M.Chitra Devi, "Human Skull and Face Identification using SIFT Technique", International Journal of Applied Engineering Research, Vol.10, No.55 (2015) pp.2926-2930.
[12]. Jimena Barbeito-Andrésa,b, Marisol Anzelmoa,b, Fernando Ventricec, Marina L. Sardia,b" Measurement error of 3D cranial landmarks of an ontogenetic sample using Computed Tomography" Journal of Oral Biology and Craniofacial Research, pp. 77-82, 2012,
[13]. Adel Kermi, Sofia Marniche-Kermi, Mohamed Tayeb Laskri" 3D-Computerized facial reconstructions from 3D MRI of human heads using deformable model approach" IEEE, pp 1-7,2010.
[14]. Rafael Romeiro, Ricardo Marroquim, Claudio Esperanc¸a, Andreia Breda, Carlos Marcelo Figueredo" Forensic Facial Reconstruction using Mesh Template Deformation with Detail Transfer over HRBF" SIBGRAPI Conference on Graphics, pp1-8,2014.
[15]. Maria Vanezis" forensic facial reconstruction using 3-d computer graphics: evaluation and improvement of its reliability in identification" Molecular Pathology, 2007.
[16]. Donghua Huang, Fuqing Duan, Qingqiong Deng, Zhongke Wu, Mingquan zhou" Face Reconstruction from Skull Based on Partial Least Squares Regression" International Conference on Computational Intelligence and Security, pp1-4, 2011.
[17]. Miguel Salas Zúñiga" Extracting Skull-Face Models from MRI datasets for use in Craniofacial Reconstruction Miguel Salas" 2013.
[18]. E.Bruynooghe, J. Keustermans, D. Smeets, F. Tilotta, P. Claes, D. Vandermeulen CT-based robust statistical shape modeling for forensic craniofacial reconstruction" international conference, pp 1-7,2011.

[19]. Fuqing Duan, Yanchao Yang, Yan Li, Yun Tian, Ke Lu, Zhongke Wu, and Mingquan Zhou" Skull Identification via Correlation Measure Between Skull and Face Shape" IEEE transactions on information forensics and security, vol. 9,2014.
[20]. Peter Tu, Xiaoming Liu, Carl Adrian, Phil Williams" Automatic Face Recognition from Skeletal Remains" IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 1-8, 2007.
[21]. Yongli Hu, Fuqing Duan, Mingquan Zhou, Yanfeng Sun, and Baocai Yin" Craniofacial reconstruction based on a hierarchical dense deformable model" EURASIP Journal on Advances in Signal Processing, pp 1-15, 2012.
[22]. Li Luo, Mengyang Wang, Yun Tian, Fuqing Duan, Zhongke Wu, Mingquan Zhou, and Yves Rozenholc, "Automatic Sex Determination of Skulls Based on a Statistical shape Model" Computational and Mathematical Methods in Medicine, pp 1-7, 2013.
[23]. C.Wilkinson, "Forensic facial reconstruction" book reviews, pp 456-458, 2005.
[24]. Laura Verze, "History of facial reconstruction" publish on research gate pp 1-7.
[25]. Dr.M.Pushpa Rani, D.Sasikala "A Survey of Gait Recognition Approaches using PCA and ICA" Global Journal of Computer Science and Technology ISSN: 0975-4172, June 2012.
[26]. Leticia Carnero Pascual, Carmen Lastres Redondo, Belén Ríos Sánchez, David Garrido Garrido, Asunción Santamaría Galdón" Computerized Three-Dimmensional Craniofacial Reconstruction from Skulls Based on Landmarks" IEEE Computer Science and Information Systems, pp. 729–735,2011.
[27]. Dr.M.Pushpa Rani, Dr.G.Arumugam "Children abnormal Gait Classification using extreme learning machine" Global Journal of Computer Science and Technology, ISSN: 0975-4172, October 2010.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

A study on E-learning, M-learning and U-learning

Dr.K.Chitra[#1], R.Umamaheswari[*2]
[#] *Department of Computer Science, Govt. Arts College, Melur, * Department of CA & IT, Thiagarajar College, Madurai-625009*
[1]manikandan.chitra@gmail.com
[2]tcruma@gmail.com

ABSTRACT *This paper focuses on three types of EMU learning systems (E-Learning, M-Learning and U-Learning).The current trend is towards E-learning and M-Learning which is convenient for the users , promoted by the continuous development of new information and communications technology. The field of education also prefers to be Online. E-learning is a set of applications and processes, such as computer-based learning, Web-based learning, virtual classrooms and digital collaboration. Mobile learning is a learning system with small, portable computing devices. Using Mobile learning, a student can learn from any place at any time using portable learning devices. In Ubiquitous learning environment (u-learning) anyone can access anywhere, anytime or any device and it is supported by mobile. Ubiquitous computing technologies include mobile devices, embedded computer devices such as GPS, RFID tags, pads, and badges, as well as wireless sensor networks and devices.*
Keywords *: E-learning, M-learning , U-learning ,Technology, Distance learning.*

Introduction

A. E-Learning

We are in the world, where the technology itself often includes practical application of technological solutions in the development of software to teach, learn, and challenge based on a creative inquisition. E-Learning is a distance learning process using multimedia resources; this allows one or more persons to learn from their computer. Multimedia resources are the combination of text, graphics in two or three dimensions, sound, image, animation and even video. The Fig1 shows the model of E-learning system

Fig1: model of E-learning system

E-Learning [6] is a learning technology that use information technology and communication to all the levels of the training activities. The main objective of the training activities can be defined as independent learning, distance learning, individualized training courses and development of educational relationships online.

Usage of Internet plays an important role in developing all the sectors of education and particularly in academia. This new methodology of teaching facilitates distance education. The term e-learning is used to describe the use of the Internet as a part of information.

E-Learning[12] is the distribution of training through a network (Internet, Intranet). Any training in any field taught mainly based on the actors involved (learners, trainers, authors, etc.), the underlying educational field and teaching resources used for learning.

E-Learning is the ability to follow a distance learning program, self-paced or accompanied, individually or collectively. E-Learning[3] is based on the Internet and multimedia tools to offer short term training modules, progressive, adapted to the levels and needs of learners.

The applications of e-learning system[7] involved in various fields like Academics , Business, Health sectors, Research and development, Engineering, Security and Protection , Agriculture and so on

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

B. M-learning

Mobile learning happens on a smaller screen than conventional desktop e-learning and is often preferred by the user. This system encourages the instructional designers to hold close a bite-sized learning concept. The exact learning sessions which are easy to understand, than the conventional e-learning user. This approach gives m-learners greater flexibility in learning. Sometimes learners do not use desktops and the only alternative is mobile.

Mobile learning is a type of e-learning that uses mobile devices and wireless transmission. Mobile learning [5] is an advanced type of e-learning–distance education. The students, who use mobile devices for learning purposes become promoted and busy in learning, thus increasing their performance and success levels. The Fig2 shows the model of M-learning system.

Fig2: model of M-learning system

One way to achieve mobile learning is the use of smart devices, it includes smart phones, Personal Digital Assistants or PDAs [5] and other portable, handheld and palmtop personal computers.

The smart devices [4] will be less expensive and equipped with powerful processors, higher random access memory, screens with high-resolution and open operating systems. Therefore, the demand for smart devices is predicted to increase. Mobile technologies are increasingly used to facilitate learning process and the use of these technologies creates new opportunities and challenges.

M-Learning is always available whenever learners need to use it. The information can be retrieved immediately by learners. Learners can interact with peers, teachers, efficient and effective experts across different media. This provides adequate information to learners.

Access to information or knowledge anywhere and anytime we have mobile devices in lower prices than desktop PCs, which is in smaller size [4] and light weight than desktop PC. M-Learning is easy to handle, it supports multimedia , low cost applications are available and we get Instant solution of any problem

C. U-leaning

Ubiquitous learning, also known as u-learning is based on ubiquitous technology. The most significant role of ubiquitous computing technology in u-learning is to construct a ubiquitous learning environment, which enables anyone to learn at anyplace at anytime. Nonetheless, the definition and characteristic of u-learning is still unclear and being debated by the research community. Researchers have different views in defining and characterizing u-learning, thus, leads to misconception and misunderstanding of the original idea of u-learning.

According to Lyytinen & Yoo (2002), [16]"the evolution of ubiquitous computing has been accelerated by the improvement of wireless telecommunications capabilities, open networks, continued increases in computing power, improved battery technology, and the emergence of flexible software architectures".

Learning environment which uses the innovations of wireless technologies is called ubiquitous learning[2] (u-learning), it intelligence the situation of the learners, and offer more adaptive supports. a variety of embedded and invisible devices, as well as the corresponding software components, have been developed and connected to the Internet wirelessly.

The Fig3 shows the model of U-learning system

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

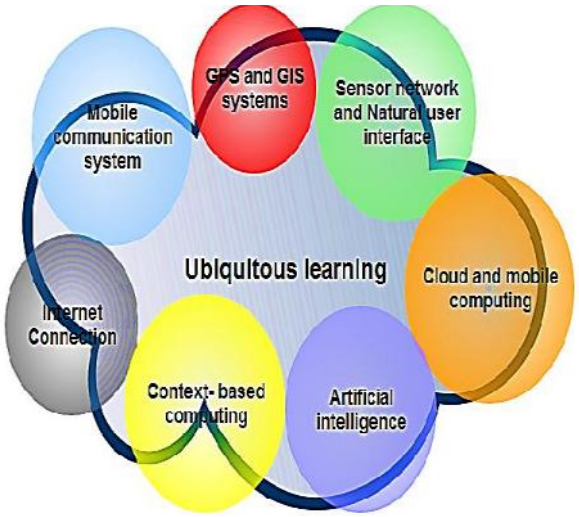


Fig3: model of U-learning system

This new Internet-ready environment[2] has been called a ubiquitous computing environment which makes many people to make use of huge amounts and various kinds of "functional objects" through network connections anytime and anywhere .

Another feature of the ubiquitous computing environment is the use of wireless communication objects with sensors, so that the system can sense user information and environmental information in the real world and provide personalized services accordingly. Such a feature is often called "context aware."

"Context-aware computing" [4] refers to the special capability of an information infrastructure to recognize and react to real-world context. Context, in this sense, includes any number of factors, such as user identity, current physical location, weather conditions, time of day, date or season, and whether the user is asleep or awake, driving or walking.

Perhaps the most critical aspects of context are location and identity. Location-aware computing systems[2] perform in response to a user's location, either abruptly or when activated by a user request. Such system also employ location information without the human attentive of it.

To develop context-aware built-in Internet environments, a variety of new techniques and products relating to ubiquitous computing have been developed in current years, such as sensors and actuators, RFID[1] tags and cards, wireless communication, mobile phones, PDAs, and wearable computers.

Physical combination and spontaneous interoperation are two main characteristics of ubiquitous computing systems. Physical combination means that a ubiquitous computing system involves some mixing between computing nodes and the physical world.

A ubiquitous learning environment is context aware that is, the learner's situation or the situation of the real-world environment in which the learner is located can be sensed.

Related work

This section presents some of the existing works related to e-learning, M-learning and U-learning system .

According to Joanne Gikas &. Michael M Grant (2013) they suggests that the student use mobiles to work together with each and share their information and skills via recording of videos or voice note to be uploaded to the course site and then it will be discussed by the whole class.

According to N.Mallikharjuna Rao and C.Sasidhar (2011) Mobile cloud plays an vital role in student learning system. Most of the cloud services providing security services to secure mobile cloud data within a cloud.

Student and teacher data have lot of importance because its usability and needs increased day by day. Today there are lot of direct applications for teaching and learning as opposed to simple platform independent tools and scalable data storage. [10]

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Dr. *M.* Thangaraj, Mrs. *S.* Vanathi (2014), recommended a system IRS-IEE find out the poor learned and well learned concepts among the students and suggest them to repeat the same if the result value is less than the threshold value.

The attractiveness of Internet along with the wide development of standard procedure and services creates a new measurement in the whole education situation. It makes the online education more smart. Everyday new approaches are coming and bringing new diagnosis in education and trying to refining the system towards modified self-learning [15].

S. Kakoty, and S.K Sarma, The benefits of E-learning are mainly the cost efficiency, convenience and elasticity. However, even as much has been made of the advancements to the organization of e-learning, there has been little, if any, qualitative examination into the attitudes and views of the users themselves .

Ardil proposed an e-learning two-way system for teachers and students to broadcast their research subjects/projects through a cyberspace where members can work together productively and freely to consider relevant issues, sharing persons experiences and providing and gaining support from each other.

Comparison of e-learning, m-learning and u-learning

This section presents the detailed comparison of the three learning systems. E-learning is a subset of the Distance Learning and m-learning is a subset of E-learning[6].

U-learning [1] is a learning environment that anyone can access anywhere, anytime or any device, it is the combination of E-learning and M-learning.

The Fig4 shows the relation between e-learning, m-learning and u-learning.

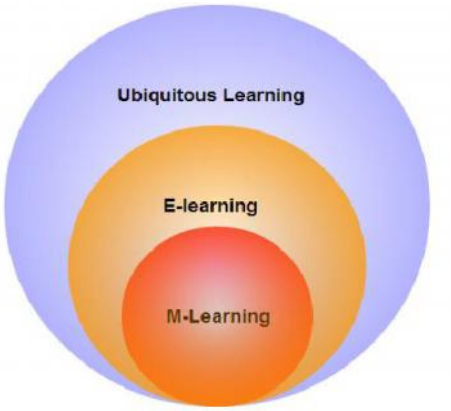


Fig4: Relation between e-learning, m-learning and u-learning

The Fig5 shows the evolution of learning

Fig5: Evolution of learning

Above discussed, three learning mechanism can be summarized on the basis of their specifications as followed in Table 1.

Table1

System specifications for e-learning, m-learning and u-learning.

Features	E-learning	M-learning	U-learning
Hardware requirements	Internet connection ,Compact Disk, desktops, mobile devices, webcam, television etc.	Mobile devices, internet connection.	Sensor networks, wearable computers, Geographical Information System, Virtual reality based projector, Radio frequency identification system (RFID)
Software requirements	Operating system, Suitable network protocols, internetworking communication technology and required applications	Mobile operating system like android iOS, windows 10 etc, WAP, GPRS, GPS, Bluetooth, Wi-Fi and wireless communication technology	Operating system, Location aware protocol, suitable software sensor network
Data Processing Technique	Cloud computing	mobile computing , Cloud computing	mobile computing , cloud computing and Context aware computing
Privacy concern	public	Personalize	high security methods and systems provides very low privacy.
Discovery and research support	supports discovery and research.	Usually use to access the information.	Support and promotes research activity
Type of device	wired	Wireless	Invisible
Accessibility	restricted to a region.	Anywhere, any time.	Everywhere, every time in a right way
Complexity	little	fair	very
Applications	distance learning	individual use	appliance oriented.

Conclusion

In this paper, comparison on E-learning, M-learning and U-learning such as definition, evolution, benefits, how it is used in present context and technologies that are used in these learning systems is discussed. The role of different networking technologies used for learning, comparative study on different architectures, design approaches, Accessibility, Complexity, Applications, device limitations and technologies are mentioned. The data processing techniques adopted in the learning systems are described. Their merits and demerits also described. To keep tempo with the technology as well as to learn these technologies it is essential to solve the question when it arrives. Among the three learning systems, U-learning is considered to be the best suitable way for the learners in future.

References

1. Gwo-Jen Hwang, 2006 "*Criteria and Strategies of Ubiquitous Learning*", *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC'06).*
2. Michael Friedewald, Oliver Raabe, 2011, "*Ubiquitous computing: An overview of technology impacts*", *Elsevier Telematics and Informatics 28 (2011),* pp 55–65.
3. Hsieh, S.-W., Jang, Y.-R., Hwang, G.-J., & Chen, N.-S. (2011).E*ffects of teaching and learning styles on students' reflection levels for ubiquitous learning. Computers & Education, 57*(1), 1194-1201. doi: http://dx.doi.org/10.1016/j.compedu.2011.01.004
4. Ogata, H., & Uosaki, N. (2012). A new trend of mobile an ubiquitous learning research: Towards enhancing ubiquitous learning experiences. *International Journal of Mobile Learning and Organisation, 6*(1), 64-78

5. Park, Yeonjeong (2011). *A Pedagogical Framework for Mobile Learning: Categorizing Educational Applications of Mobile Technologies into Four Types, The International Review of Research in Open and DistanceLearning*,Vol.12,No.2,February. http://www.irrodl.org/index.php/irrodl/article/view/791/1699
6. S.K. Behera, "*E-and M-Learning: A comparative study,*" International Journal on New Trends in Education and Their Implications, vol. 4(3), pp.65-78, 2013.
7. Desmond. K. (2010) *The future of learning: From E-learning to M-learning*. Available on line at http://learning.ericsson.net(2010)
8. Crescente, Mary Louise; Lee, Doris (2011). "*Critical issues of M-Learning: design models, adoption processes, and future trends*". Journal of the Chinese Institute of IndustrialEngineers.28(2): 111–123.
9. S. M. Jacob and B. Issac "*The Mobile Devices and its Mobile Learning Usage Analysis*". 2008 International MultiConference of Engineers and Computer Scientists 2008 Vol I IMECS 2008, 19-21 March, 2008, Hong Kong
10. Joanne Gikas a, Michael M. Grant. 2013 "*Mobile computing devices in higher education: Student perspectives on* learning with cellphones, smartphones & social media Internet and Higher Education 19 (2013) 18–26
11. N.Mallikharjuna Rao and C.Sasidhar, V. Satyendra Kumar "Cloud Computing Through Mobile-Learning". 2011
12. Dr. M. Thangaraj, Mrs. S. Vanathi, "An Intelligent Recommender System for Effective E-Learning "International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181Vol. 3 Issue 11, November-2014
13. C. Ardil, "*E-learning Collaborative Circles*" , International Journal of Humanities and social Services, Vol. 1, No. 4, 2007
14. S. Kakoty, and S.K Sarma. "*Expert System Applications in E-learning Environment: Analysis on Current Trends and Future Prospects*"International Journal of Internet Computing (IJIC), Vol. 1, 90 - 93, 2011
15. A.E. Blackhurst, and D.L. Edyburn, "*A brief history of Special Education Technology*" Knowledge by Design Inc., 2000
16. Liyytinen, K. & Yoo, Y. (2002). Issues and Challenges in Ubiquitous Computing. Communications of the ACM , vol.45, no.12, pp.62 – 65
17. Alsalem, A. (2004).Educational Technology and E-learning, Riyadh: lroshd publication.
18. Quinn C. "M-Learning: Mobile, Wireless, In-Your-Pocket Learning," http://www.linezine.com/2.1/features/cqmmwiyp.htm , 2000.
19. Kristiansen, T. "M-learning. from the use of WAP as a supplement in learning,"http://www.nade-nff.no/nadenff/konferanse/vettre02/TK010430%20Erfaringsr apport.pdf, 2009.
20. A Dye, B E Solstad. "J A K Odingo. Mobile Education-A Glance at The Future," http://www.nettskolen.com/forskning/mobile_education.pdf ,2009.
21. Desmond. K. (2010) The future of learning: From E-learning to M-learning. Available on line at http://learning.ericsson.net(2010)
22. Faqeeh. A. (2009).M-learning...A New vision by using wireless technology. Available: Math –Nablus,y007.com/search. Forum.
23. Moussa, A. A, E- learning: characteristics, Advantages, Obstacles. A research paper presented to the forum of future school,, KingSaud University,16-17/8/2010.
24. Ogata, H., & Yano, Y. (2004). Context-Aware Support for Computer-Supported Ubiquitous Learning. In Proceedings of the 2004 IEEE international Workshop on Wireless and Mobile Technologies in Education (pp. 27-34). New York: ACM.
25. Hwang, G. J., Tsai, C. C., & Yang, S. J. H. (2008). Criteria, Strategies and Research Issues of Context-Aware Ubiquitous Learning. Journal of Educational Technology & Society, 11(2), 81 - 91.
26. Michael Friedewald, Oliver Raabe, 2011, "Ubiquitous computing: An overview of technology impacts", Elsevier Telematics and Informatics 28 (2011), pp 55–65.
27. Tsai P-S, Tsai C-C, Hwang G-H. College students' conceptions of context-aware ubiquitous learning: a phenomenographic analysis. *The Internet and Higher Education*. 2011;14(3):137–141.
28. S. Yahya, E. A. Ahmad & K. A. Jalil, (2010) "The definition and characteristics of ubiquitous learning: A discussion", International Journal of Education and Development using Information and Communication Technology (IJEDICT), Vol. 6, No. 1
29. S. J. H. Yang, (2006) "Context Aware Ubiquitous Learning Environments for Peer-to-Peer Collaborative Learning", Educational Technology & Society, Vol. 9, No. 1, pp. 188-201
30. Hwang, G. J., Wu, T. T., & Chen, Y. J. (2007). Ubiquitous computing technologies in education. Journal of Distance Education Technology, 5(4), 1-4

Novel Approach of Noise Reduction in Images using Various Spatial Filtering Techniques

[1]Dr. PushpaRani.M, [2]Sudha.D
[1]Prof. & Head, Department of Computer Science,
Mother Teresa Women's University, Kodaikanal
[2]Research Scholar, Mother Teresa Women's University, Kodaikanal

ABSTRACT *Noise is a major problem in image processing domain, especially when the background image is subtracted or removed and the foreground binary image is obtained. The image enhancement is a life hold technique for the removal of noise from the processed binary image. The main goal of image enhancement is to process the original image into a more suitable form to the specific application. It is divided into two main categories such as spatial domain and frequency domain. The spatial domain involves the modification of the image area and the modification of the image's Fourier transform is known as frequency domain processing technique. The Spatial domain is concentrated in this approach. The spatial domain refers to the direct processing of pixels in the image. The binary image of the gait signal is taken as an input and the various filtering methods of linear and non- linear spatial domain filtering's such as average filter, weighted average filter, minimum filter and maximum filter and median filters are employed to enhance the images with the 3x3 mask. The output is evaluated by using different performance indicators like EME, PSNR, MSE, RMSE.*

Keywords: *gait, image enhancement, spatial domain filtering, linear filter, non-linear filter, average filter, weighted average filter, minimum filter, maximum filter, median filter, performance analysis, binary images*

Introduction

The main goal of image enhancement is to process the original image into a more suitable form to the specific application. The quality of the visual image is a completely subjective process [7]. To make a human perspective good view of an image the enhancement is required for the captured image or input image. The enhancement is a process involving the noise reduction most. This reduces the additional noise by calculating the pixel values and its neighborhood pixels. The old images or the damaged images by mishandling are enhanced through the different transformation process. The enhancement method is varying for different application and different images. For example, the satellite image and the medical image are treated differently. For example, best techniques for enhancement of X-ray image may not be best for enhancement for microscopic images [1]. The image like gait, iris, fingerprint and brain, skull and other biometric images are needed to be enhanced individually depends upon the application requirement. There are two aspects which require the image enhancement

Low level features can be extracted
Proper visualization of an image

Noise is a major problem in image processing domain, especially when the background image is subtracted or removed and the foreground binary image is obtained [3][11]. The image enhancement is a life hold technique for the removal of noise from the processed binary image.

```
folder = fullfile(matlabroot, '\Sudha\Gait\Samples');
baseFileName = 'sample1.png';
BaseFile = fullfile(folder, baseFileName);
grayImage = imread(BaseFile);
[rows columns numberOfColorBands] = size(grayImage);
subplot(2, 2, 1);
imshow(grayImage);
set(gcf, 'Position', get(0,'Screensize'));
noisyImage = imnoise(grayImage,'salt & pepper', 0.05);
subplot(2, 2, 2);
imshow(noisyImage);
I= noisyImage;
```

It is divided into two main categories such as spatial domain and frequency domain. The spatial domain involves in the modification of the image area and the modification of the image's Fourier transform is known as frequency domain processing technique. The Spatial domain is concentrated in this approach. The individual pixels in the image are concentrated and the different processing techniques are directly applied on those pixels is the main concept of spatial domain techniques. Spatial domain is denoted by

$$g(x,y)=T[f(x,y)]$$

where $f(x,y)$ is the image given as input, $g(x,y)$ is the enhanced image and T is an operator on f.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Basics of spatial filtering

This research deals with the various spatial domain filtering methods with the gait images. The spatial domain filtering is a process of moving a filter mask from pixel point to another pixel point. Using a predefined relation the value of the pixels is calculated at each point.

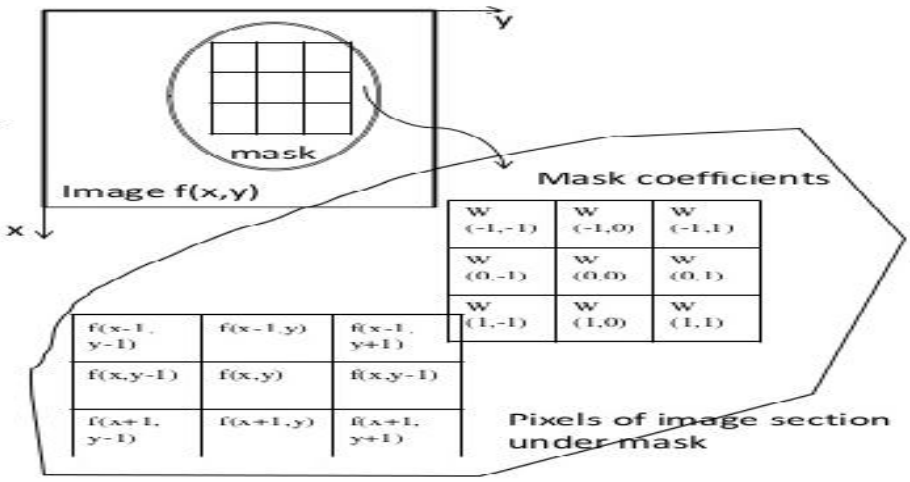


Fig. 1 Mechanics of Spatial Filtering with 3x3 mask

A. Smoothing Filters

Smoothing filters are used for blurring and noise reduction. Blurring is the process of removing a small detail from the pixel prior to feature extraction and adding some points to the pixel for smooth lines and curves[12]. Noise reduction is the removal of extra information which is not related to the neighborhood pixels. Smoothing filters are of two type linear filter and order statistics filter.

1) Linear Filter: The neighborhood pixels averaged in the pixel mask is called as the linear spatial filter. The given gait image processed under the linear filter and the output shows with the smooth edges. Linear filter R, with the filter mask at any point (x, y) in the image is

$$\boldsymbol{R = w(-1,-1)f\,(x-1,y-1) + w\,(-1,0)f\,(x-1,y) + w\,(0,0)f\,(x,y) + \ldots + w\,(1,0)f\,(x+1,y) + w\,(1,1)f\,(x+1,y+1)}$$

The above equation represents the sum of products of the coefficients of the linear mask and the pixels in the original image. The linear filter f of the image sixe M x N with the mask m x n is given by

$$\boldsymbol{g(x,y) = \sum_{s=-a}^{a}\sum_{t=-b}^{b} w(s,t)\mathbf{f}(\mathbf{x}+\mathbf{s},\mathbf{y}+\mathbf{t})}$$

where, a=(m-1)/2 and b=(n-1)/2, and x=0,1,2,...,M-1 and y=0,1,2,...,N-1.

```
c = [222 272 300 270 221 194 141 130 82];
r = [21 21 75 121 121 75 221 221 75];
BW = roipoly(I,c,r);
H = fspecial('unsharp');
J = roifilt2(H,I,BW);
```

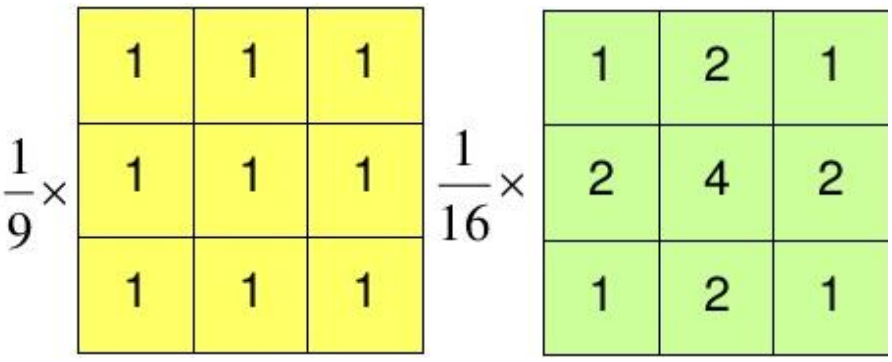


Fig. 2 (a) Averaging Filter mask (b) Weighted Average Filter mask

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

The linear filter has two kinds of masks: averaging filter and weighted average filter. In average filter the m x n mask has normalizing constant of M x N. It is known as Box Filter. A 3 x 3 mask has been used in this approach. The processed images are represented in the fig.3.

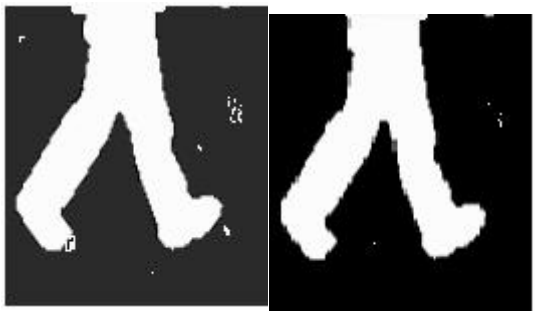


Fig. 3 (a) Original image (b) Enhanced image by average filter

Weighted average filter pixels are multiplied by different coefficients, the center pixel is multiplied by higher than any other values. The remaining pixels other than the center pixel are considered and their distance from the center of mask is averaged and weighted reversely [4]. The general implementation of this is as given by the following equation with the 3 x 3 mask in the fig.2. The processed images are placed in the fig. 4.

$$g(x,y) = \frac{\sum_{s=-a}^{a}\sum_{t=-b}^{b} w(s,t) f(x+s, y+t)}{\sum_{s=-a}^{a}\sum_{t=-b}^{b} w(s,t)}$$

Fig. 4 (a) Original image (b) Enhanced image by weighted average filter

2) Order Statistics filter: It is otherwise called as a non-linear spatial filter. The pixels which are surrounded by the filter area are processed by ranking each pixel. The ranking value is selected randomly and does not base upon the average values. By selecting the random values and providing the rank to each pixel the edges are preserved better than the linear filters. The salt and pepper noise reduction works better in this order statistics filter. Different types of order-statics filters are a) Minimum filter, b) Maximum filter, c) Median filter.

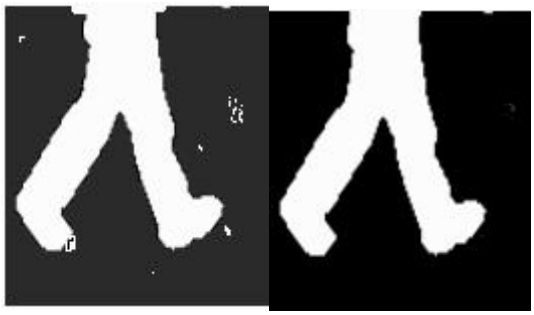


Fig. 5 (a) Original image (b) Enhanced image by minimum filtering.

a) Minimum filter: In the minimum filter, the center pixel value is replaced by the smallest value in the image area. This smallest value is selected by the pixel by pixel comparison of the entire image[6]. Using the 0th percentile results the dark portion of the image area is filtered perfectly. Since it is dealt with the dark areas of the image, the binary image in this research is enhanced using a minimum filter.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Fig. 6 (a) Original image (b) Enhanced image by maximum filter

b) Maximum Filter: The 100th percentile filter is the maximum filter. : In the maximum filter, the center pixel value is replaced by the largest value in the image area. This largest value is selected by the pixel by pixel comparison of the entire image. The bright portion of the image area is filtered perfectly. Since it is dealt with the bright areas of the image, the binary image in this research is enhanced using a maximum filter.

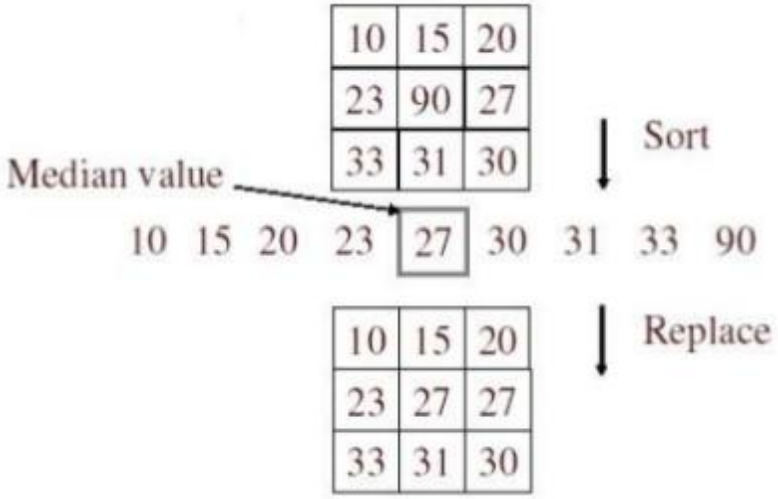


Fig. 7 Median filter example

c) Median Filter: The three steps involved in this median filter. The neighboring pixel values of a single pixel are selected and the values are arranged in ascending order. The median value among those pixel values is selected and it is replaced with the center pixel value. The isolated pixels of dark or light are eliminated by an n x n median filter.

```
medianFilteredImage = medfilt2(noisyImage, [3 3]);
                      noiseImage = (noisyImage == 0 | noisyImage == 255);
                      noiseFreeImage = noisyImage;
                      noiseFreeImage(noiseImage) = medianFilteredImage(noiseImage);
                      subplot(2, 2, 3);
                      imshow(noiseFreeImage);
```

Fig. 8 (a) Original image (b) Enhanced image by median filter

The adaptive median filtering can handle impulse noise with probabilities even larger than those [4]. The median filter gives the best result by comparing the other filtering techniques for the binary image. The fig.6 represents the processed images with high-quality human visual perception.

Performance evaluation and analysis

The above discussed spatial domain filtering methods were examined with the binary image. The performance levels were analyzed using the performance indicators like Measure of Image Enhancement - eme, Peak - to -Signal Noise Ratio - psnr, Mean Square Error - mse and Root Mean Square Error - rmse. These algorithms can be applied on n×n mask; the 3×3 mask has been taken in this research. The results are implemented using MATLAB12 using Image Processing Toolbox.

TABLE 1. PERFORMANCE ANALYSIS BY VARIOUS INDICATORS

Filtering domains	**Eme**	**Psnr**	**Mse**	**Rmse**
Average filter	2.4036	40.4552	5.8554	2.4198
Weighted Average filter	2.6538	47.3922	3.2200	1.4598
Minimum filter	3.8763	44.6505	5.9394	1.2684
Maximum filter	7.9600	52.8660	10.3686	1.0888
Median filter	13.0678	55.3817	0.1883	0.4340

Among the five different filtering methods, the median filter gives the improved result by analyzing the performance indicators. Among the other, the median filter has a higher psnr ratio. EME measure is suitable for images with the uniform background so it has been adopted as one of the performance indicator [5]. Thus the median filter shows the better result than the other filtering methods in the spatial domain filtering.

Results and discussion

The spatial domain filtering methods such as average filter, weighted average filter, minimum filter and maximum filter, median filter are discussed with its implementation methods. The gait binary image is obtained by background subtraction method, and it has been enhanced through different spatial domain filtering methods. In this paper, the image used has the following properties, uniform background, background and foreground containing uniform pixels distribution, background containing a large area of uniform intensity. The linear filter could not completely remove the noise. Non-linear filters are robust in this spatial domain filtering methods. As per the performance analysis indicators, the PSNR shows the best result with the median filter. While comparing the five different linear and non-linear spatial domain-filtering methods the median filter gives the improved result. The median filter is best suited for salt and pepper noise reduction.

References

[1] Krishan Kant Lavania, Shivali, “Image Enhancement using Filtering Techniques”, International Journal on Computer Science and Engineering (IJCSE), January 2012.
[2] Shapiro and Stockman, “Computer Vision”, 2000.
[3] MP Rani, G Arumugam, “An Efficient Gait Recognition System for Human Identification using Modified ICA”, International Jouranl of Computer Science and Information Technology, Voluem2, Issue 1, 2010.
[4] Rafael C Ganzalez, Richard E Woods, “Digital Image Processing”, Second Edition, 2002.
[5] Sunita Gupta, RabinsPorwal, “Appropriate Contrast Enhancement Measures for Brain and Breast Cancer Images”, International Journal of Biomedical Imaging, 2016.
[6] MP Rani, G Arumugam, “Children Abnormal Gait classification using extreme learning machine”, Global journal of Computer Science and Technology, 2010.
[7] TeadyMatius Surya Mulyana, “Reduce noise in the binary image using non linear spatial filtering mode”, IEEE Xplore Digital Library, 2016.
[8] D.Sudha, Dr.MPushpa Rani, “Gait Classification for ADHD Children Using Modified Dual Tree Complex Wavelet Transform”, IEEE Xplore Digital Library, 2017.
[9] https://www.slideshare.net/balamoorthy39/smoothing-filters-in-spatial-domain
[10] https://swappaz biz/code-de-reduction-special-t.html.
[11] M.Pushpa Rani, “Abnormal Gait Classification using hybrid ELM”, Electrical and Computer Engineering (CCECE), IEEE 27th Canadian Conference, 2014.
[12] MP Rani, D Sasikala, “A survey of Gait Recognition Approaches Using PCA and ICA”, Global journal of Computer Science and Technology, 2012.

A Survey on Multimedia Streaming Techniques over LTE Networks

[1]A.Valliammal, [2]Dr.E.Golden Julie
[1]Department of Embedded System Technologies, [2]Department of Computer Science and Engineering
[12]Anna University Regional Campus- Tirunelveli
[1]vallimahesh96@gmail.com [2]juliegolden18@gmail.com

ABSTRACT *The broad use of smart phones, with the availability of higher bandwidths in next generation mobile networks, have lead to increase the demand of live video broadcasting services that users can access through multimedia players operation on their mobile devices. The quality of services can be affected by service interruptions due to the mobile network transmission errors. The Evolved multimedia multicast broadcast services (E-MBMS) channel allows sending the same multimedia content into the multiple receivers. The forward error correction method and unicast recovery system are used to deal with the network errors occurs in the LTE multimedia broadcast services. These error recovery techniques are used to decrease the channel error. This paper provides a survey on multimedia broadcasting services. It gives an overview on techniques, technologies, architecture adopted in various environments. It also discusses about key challenges faced in broadcasting services. It also provides a summary of recent literatures on broadcasting services.*

Keywords: *Multimedia broadcasting, Buffering, LTE network, Latency, Playout frame rate.*

I. Introduction

Nowadays, video is the most leading application in the Internet. According to the recent learning and estimate, global Internet video traffic explained for nearly 14 PB per month in 2012,which is nearly 57% of all user traffic. In 2017, it is expected to reach 52 PB per month, which will then be 69% of the whole consumer Internet traffic. In live streaming services, an important quality of service parameter is service latency, defined as the difference between the live event occurs time and the time at which the live event is played on a receiving terminal.

The demand of multimedia streaming services is growing as a result of a wide adoption of more capable mobile devices, with the evolution of mobile networks. The evolution of mobile devices with higher multimedia capabilities together with the use of high capacity mobile network recently have made an vast growth in user data traffic particularly multimedia content. The mobile video consumption is stimulated with the increase number of Smart phones in the market.

The demand for wireless broadband bandwidth has increasing very fast. Users demand greater access capacity to utilize an increasing number of services and applications. At the same time, these applications happen to be more hungry for bandwidth. Wireless broadband exploits a limited resource and frequency spectrum.

The multimedia broadcast and multicast service standard defined by the Third Generation Partnership Project defines the workstation, radio network, core network, and user service aspects. MBMS is a point to multipoint service in which data are transmitted from a single source person to multiple recipients, multicast networks are used by TV channels or radio stations to broadcast their contents. The multicast network allows sending content to several clients using a single transmission operation. The multicast networks are used to transmit files.

Wireless video streaming is gaining more and more regards among mobile users. It is main, for both the content providers and wireless network designers of the future, to understand how to ensure a satisfactory quality of user experience. There are several video artifacts that have an effect on the QoE.

IP-BASED TV systems are increasingly being deployed to provide live and on-demand video streaming services in the Internet Protocol (IP)-based networks the demand for wireless broadband bandwidth has being increasing very rapidly. Users demand greater access capacity to utilize an increasing number of services and applications.

The rest of the paper is organized as follows: section 1, mainly provides an overview of broadcast services architecture. In section 2, provides the research on multimedia services techniques are discussed. In section 3, the available challenges of broadcasting services are investigated. Finally, summary of emerging research issues are identified and the future research directions are discussed

II. REVIEW OF BROADCAST SERVICES ARCHITECTURE

The broadcast services architecture, defines the different aspects of the LTE networks and multimedia broadcasting services.

E. Tan et.al [1] describes that AMP approach, which decrease the decoder playout frame rate to avoid buffer starvation. Depend on the encoder approach, the AMP techniques alter the playout frame rate which introduces more playout distortion than necessary. The frame is continuously move from the encoder buffer to decoder buffer it gradually affects the video continuity.

De Fez Lava et.al [12] describes the File Delivery over Unidirectional Transport (FLUTE) is the protocol used in unidirectional environments. It gives the reliability in the transmission of multimedia contents. This protocol is maintaining the File Delivery Table, which is the in-band mechanism used by FLUTE to inform clients about the files and their characteristics are transmitted within a FLUTE session. Clients require to receive the FDT in order to start downloading files.

C.M. Lentisco et.al [7] describes the architecture to support the distribution of video in wireless environments to the users. Raptor and RaptorQ are fountain codes, as needed many encoded symbols are generated on the encoder from the k symbols of the source block.The decoder is able to recover the block, if it receives a number of symbols slightly higher than k. In this context, the code rate is normally used .The solution standardized by 3GPP to ensure reliable transmissions over eMBMS is to use of Raptor codes as the Application Layer- Forward Error Correction (AL-FEC) scheme. Among other techniques, using Raptor and Raptor codes achieve the service data rate and the coverage .

J. F. Monserrat et.al [6] proposed the architecture to support growing the efficiency in the transmission of the same content to several users. It using the Wideband Code Division Multiple Access (WCDMA), was improved with Multimedia Broadcast Multicast Services. This combined transmission improves the received Signal to Interference plus Noise Ratio (SINR).

de Fez et.al [2] describes about the AL-FEC techniques which is increases the performance of multicast content download services. The implementation of adaptive LDPC for multicast content delivery based on the File delivery and unidirectional transport protocol. Adaptive AL-FEC codes represent a good alternative to improve the reliability of multicast connections over wireless channels.

de Fez et.al [3] proposed the architecture to supports the file delivery over unidirectional transport (FLUTE) protocol in multicast networks, which reduce significantly the bandwidth when there are many users interest in the same contents.

T. Hoßfeld et.al [10] proposed the architecture for QoE assessment methodology used in multimedia applications that are like online video, which is based on crowd sourcing. some inherent problems are improved by filtering based on additional test design measures. but difficult methods are required to reduce or avoid rejection of user results, by utilizing reputation systems of existing crowd sourcing platforms. crowd sourcing has high potential not only for testing online video usage scenario, but also used in QoE assessment of Internet applications like web.

C. Yang and Y. Liu [9] proposed the architecture new IP-based streaming framework. That is called fast IP-based TVsystem.It is used to achieve close-to-zero channel switching delay at the value of additional download bandwidth and better playback lags. Video segments are downloaded from the combination channel will be stored in a local buffer. When the client issues a channel request to a target channel, the client will immediately playback the most recently downloaded video of the target channel, To achieve close-to-zero SRT and good QoE performance for multichannel systems.

D. Lecompte and F.Gabin [11] describes about the Evolved Multimedia Broadcast Multicast Service (eMBMS) is a point-to-multipoint content delivery particularly designed for LTE standard. eMBMS can be used to enhance the network's capacity for providing good quality multimedia service in high user-density area is to validate and estimate the performance. The 3GPP standard implemented in the OpenAirInterface SDR platforms presented the implementation of eMBMS in LTE network.

III. REVIEW OF MULTIMEDIA BROADCASTING TECHNIQUES

This section of survey explains the different broadcasting techniques used in different papers are summarized as follows.

A long content data can be separated into one or more temporal content. Further, each alternative of a content can be divided into media segments using the MPEG DASH standard technique[4].

Evolved Multimedia Broadcast and Multicast Service (eMBMS) is a common channel that is used to send the same content of data into multiple receivers, thereby increasing the efficiency of the use of network resources[6]. The E-MBMS service transmissions to provide file delivery services as per different type of users and file sizes[7].

AL-FEC is the technique, which improves the performance of multicast content download services and also it is used to keep the information. The multimedia content are protected against errors[2].

Performed frame rate control by a joint adjustment of the encoder frame generation period and the playout frame interval. Lyapunov optimization method is used to derive the optimization policies that help to ensure the video continuity with the minimum network resource. It is useful for avoid buffer starvation [1].

Adaptive code rate is used to minimized the download time of all clients within the coverage area, with a realistic use of bandwidth[3].

An Adaptive HTTP streaming technique is used in live streaming, which maintain low delays. It provides the continuous live streaming(like sport events), using the media description file[11].

IV. KEY CHALLENGES IN BROADCASTING SERVICES

The video broadcasting in wireless environments has become a big challenge. It has some inherent limitations that impact on the streaming services. The limited bandwidth or the noisy channels are some of these limitations. It affects the service quality and also increasing the user's problems such as, high delay and video discontinuity during the playback time.

V. CONCLUSION AND FUTURE SCOPE

Multimedia communication over mobile networks is a new research area, with a limited, but, fastly growing set of research results. This paper surveys different architecture and techniques based on Multimedia communication over

mobile networks. However, most work fails to consider some features of the communication when evaluating the mobile network techniques. Many authors tried to solve these issues using many methods and produced acceptable solutions. However, some improvements have to be done in multimedia communication since mobile networks is limited. The future scope is to design a convenient and accessible way for communication over mobile networks.

REFERENCES

[1] Tan and A. C. Tung, "*A frame rate optimization for improving continuity in video streaming*," IEEE Trans. Multimedia, Jun.2012.
[2] I. de Fez, F. Fraile, R. Belda, and J. C. Guerri, "*Analysis and evaluation of adaptive LDPC AL-FEC codes for content download services*," IEEETrans. Multimedia, Jun. 2012.
[3] I. de Fez and J. C. Guerri, "*An adaptive mechanism for optimal content download in wireless networks*," IEEE Trans. Multimedia, Jun. 2014.
[4] T. Truong, H. Quang-Dung, J. Won, and A.T.Pham, "*Adaptive streaming of audiovisual content using MPEG DASH*," IEEE Trans. Consum.Electron.,Feb. 2012.
[5] U. Kumar and O. Oyman, "*QoE evaluation for video streaming over eMBMS,*" in Proc. Int. Conf. Comput., Netw. Commun., 2013, pp. 555–559.
[6] J. F. Monserrat, J. Calabuig, A. Fernandez-Aguilella, and D. Gomez-Barquero, "*Joint delivery of Unicast and E-MBMS services in LTE networks*," IEEE Trans. Broadcast, Jun. 2012.
[7] C. M. Lentisco et al., "*A model to evaluate MBSFN and AL-FEC technique sin a multicast video streaming service*," in Proc. IEEE Int. Conf.Wireless Mobile Comput., Netw. Commun., Oct. 2014.
[8] S.Wei and V. Swaminathan, *"Low latency live video streaming overHTTP2.0*," in Proc. Netw. Operating Syst. Support Audio Video, jan 2014.
[9] C. Yang and Y. Liu, "*On achieving short channel switching delay and playback lag in IP-based TV systems,"* IEEE Trans. Multimedia, vol. 17,no. 7, pp. 1096–1106, Jul. 2015.
[10] T. Hoßfeld et al. "*Quantification of YouTube QoE via crowd sourcing*," inProc. IEEE Int. Symp. Multimedia, Dec. 2011, pp. 494–499.
[11] D. Lecompte and F. Gabin, "*Evolved multimedia broadcast/multicast service(eMBMS) in LTE-advanced: Overview and Rel-11 enhancements*,"IEEE Commun. Mag., vol. 50, no. 11, pp. 68–74, Nov. 2012.
[12] DeFezLava,I.FraileGil,F.GuerrCebollada,JC,"*Effect of the FDT transmission frequency for an optimum content deliveryusingthe FLUTEprotocol*,"ComputerCommunications.36(12):1298130doi:10.1016/j.comcom.2013.04.08.

Anomaly Detection and Energy Efficient Multi Hop Routing (SEER) Protocol Design for Wireless Sensor Networks

[1]Komathi A, [2]Dr. M. Pushparani
[1]Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India
And Assistant. Professor , Department of Computer Science and Information Technology, Nadar Saraswathi College of Arts and Science , Theni
[2]Head and Professor, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, India

ABSTRACT *Lifetime optimization and security are two conflicting design issues for multi-hop wireless sensor networks (WSNs) with non-replenishable energy resources. In this paper, we propose a Anomaly Detection and Energy Efficient Routing (ADEER) protocol to address these two conflicting issues such as energy and anomaly detection in WSN. In this scheme, Extended Kalman filter mechanism to detect false injected data. Specifically, by monitoring behaviors of its neighbors and using EKF to predict their future states. . The node behavior is monitored by integration of Anomaly Node Observation and System observation module. SEER has the flexibility to support multiple routing strategies in message forwarding and to extend the lifetime while increasing routing security. Our analysis shows that we can increase the lifetime and the number of messages that can be delivered under the uniform energy deployment in the network.*
Keywords: *Wireless Sensor Network, Energy Balance Control, Security, Extended Kalman filter, Lifetime.*

Introduction

Many sensor systems are deployed in unattended and often adversarial environments such as battlefields. Hence, security mechanisms that provide confidentiality and authentication are critical for the operation of many sensor applications. Providing security is particularly challenging in sensor networks due to the resource limitations of sensor nodes. This multi-hop packet transmission can extend the network coverage area using limited power and improve area spectral efficiency. In developing and rural areas, the network can be deployed more readily and at low cost. We consider the civilian applications of multi-hop wireless networks, where the nodes have long relation with the network.

Fault tolerance in collaborative sensor network [1] introduced for collaborative target detection that are efficient in terms of communication cost, precision, accuracy, and number of faulty sensors tolerable in the network. Value fusion and decision fusion algorithm are used for identified node trust. The impact of hierarchical agreement on communication cost and system failure probability is evaluated and a method for determining the number of tolerable faults is identified.

Using message authentication code (MAC) [2] presented an interleaved hop-by-hop authentication scheme, which guarantees the base station detects injected false data packets, when no more than a certain number nodes are compromised. But this scheme requires each cluster has fixed nodes. Secure and privacy-preserving data communication [3] SDAP utilized the principles of divide-and-conquer and commit-and-attest is a general-purpose secure aggregation protocol applicable to multiple aggregation functions. The spirit of SDAP is similar to Merkle hash tree. Nevertheless, communication cost of SDAP is fairly high. Statistical en-route filtering mechanism [4] that can detect and drop false reports injected by compromised nodes.

The collaborative nature of Industrial WSN (IWSN) [5] plays a vital role in creating a highly reliable and self-healing industrial system that rapidly responds to real-time events with appropriate actions. In this scheme, technical challenges and design principles are introduced in terms of architecture designing, developing hardware and software. In [6] explains overview of the application of WSNs for electric power systems along with their opportunities and challenges and opens up future work in many unexploited research areas in diverse smart grid applications. Then, it presents a comprehensive experimental study on the statistical characterization of the wireless channel in different electric-power-system environments, including a 500-kV substation, an industrial power control room, and an underground network transformer vault.

In [7], a predictor-controller algorithm is used to mitigate the network delay in the network. Also this method reduces the computation complexity and enhances the system reacting time. Simulation results demonstrate the effectiveness of our proposed method. In [8] explains decentralized algorithm is used to compute the control signals in sensor networks. The estimation error of the new approach is one-eighth as large as that of the penalty method with one-fifth of its computation time.

Establishing stable and reliable routes in heterogeneous multi-hop wireless networks E-STAR [9] combines the payment and trust systems with a trust-based and energy-aware routing protocol. The payment system rewards the nodes that relay others' packets and charges those that send packets. The trust system evaluates the nodes' competence and reliability in relaying packets in terms of multi-dimensional trust values. The trust values are attached to the nodes and the public-key is used to identify the anomaly node in the network.

Proposed method

In this paper we develop a Anomaly Detection and Energy Efficient Routing protocol for sensor networks, named SEER. We design it to achieve a tradeoff between securities and reduce the energy utilization in the network. In ADEER, an EKF mechanism is suitable for WSN nodes because this mechanism may address those incurred uncertainties in a lightweight manner. The process of ADEER is divided into three phases: Registration phase, Anomaly Detection Phase, Energy Efficient Path phase and Data Transmission phase.

In Registration phase, every sensor nodes register to the Base Station (BS). It generates the random number to produce the private key for every sensor nodes. Then they find neighbors in one hop, and the process expands hop by hop from BS to the edge of sensor network. Every sensor node maintains the table. This table contains neighbor node ID, location, distance among sensors to BS.

While a normal sensor node is compromised through an anomaly, thus this anomaly node control of the compromised node. It may inject modified data readings into the WSN. As a result the modified original data can seriously disrupt aggregation operations.

In false inject data detection phase, Extended Kalman Filter (EKF) algorithm to detect false injected data. Anomaly Node Observation (ANO) and System Observation (SO) methods integrate to form the secure WSNs. The MNO is detects anomaly node and SO is using the EKF mechanism to monitoring the sensor behaviors for predicting future states. EKF can offer a relatively accurate prediction of neighbors' future states [10].

For example, when node A invokes an alert on node B due to some event E, to decide whether E is anomaly or emergent, A may initiate a further investigation on E by collaborating with existing SO. WSNs are usually densely deployed to collaboratively monitor some events around B and request from these nodes their opinions on the behavior of E. Because the majority of sensor nodes around the investigated event E are not compromised, after A collects the information from these nodes, if A finds that the majority of sensor nodes think. Collaboration between MNO and SO to differentiate anomaly events from emergency events that event E may happen, A then makes a decision that E is triggered by some emergency events.

Otherwise, if A finds that the majority of sensor nodes think that event E should not happen, A then thinks that E is triggered by either a anomaly node or a faulty yet good node. In this way, A can continue to wake up those nodes around event E and their opinions about the behavior of E. If A keeps finding that the majority of sensor nodes think that event E should not happen, A then suspects that E is anomaly. After A makes a final decision, A can report this event to base stations.

In Energy Efficient Path phase, the source selects the forwarder node based on node energy. The highest energy node is selection for data communication in the network. Thus it reduce the node dead problem also increase the energy efficiency in the network.

In Data Transmission phase, the source transmits the data via energy efficient path in the network. In this phase, the source S sends messages to the BS via intermediate nodes. Each node computes the signature by {*H(m), ID, TTL, intermediate hops and BS*}. Where, H(m) is refers to the hash message by private key. ID refers to the sensor ID, TTL denoted by Time to Live. The main aim of this signature is to ensure the message's authenticity and integrity. The source sends the data format is {*Source ID, intermediate hops, TTL, BS and sig*}.Every intermediate node verifies the source signature and adds the intermediate hop signature. This process is done until the destination reaches the data. The BS receives this message and check the every node signature, if it matches accepts the data otherwise, it will discard.

Performance evaluation

The performance of SEER scheme is analyzed by using NS2. The nodes are communicated with each other by using the communication protocol UDP. The traffic is handled using the traffic model Constant Bit Rate.

Packet Delivery Rate

It defined as the rate of data packets delivered to the destination node to the no of data packets sent by the source node. The PDR is calculated by the equation 1.

$$PDR = \frac{\sum_0^n Packets\ Received}{\sum_0^n Packets\ Sent} \quad (1)$$

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

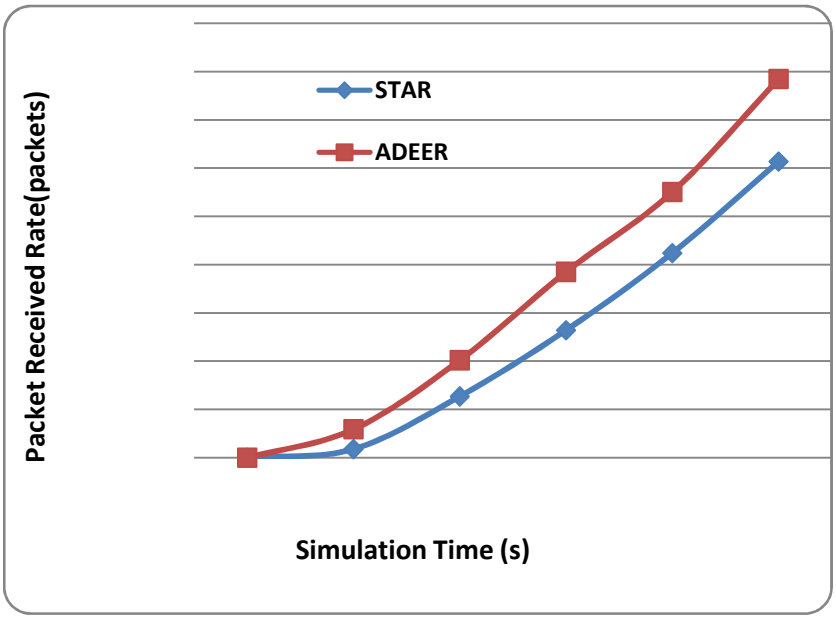


Figure.1 PDR of SEER and STAR scheme

The PDR of STAR and ADEER are plotted in Figure 1. It shows that the ADEER has better PDR while compared to the STAR.

Conclusion

We have proposed ADEER that uses hash based signature with energy aware routing protocol to established reliable routes in the networks. In ADEER, an Extended Kalman filter based mechanism to detect false injected data. Specifically, by monitoring behaviors of its neighbors and using Extended Kalman filter to predict their future states. ADEER has the flexibility to support multiple routing strategies in message forwarding and to extend the lifetime while increasing routing security. The result of Packet delivery rate demonstrates that the ADEER is highest packet delivery compare to the STAR scheme.

References

[1] T. Clouqueur and K. Saluja, "Fault tolerance in collaborative sensor networks for target detection," *IEEE Trans. Comput.*, vol. 53, no.3, pp. 320–333, Mar. 2004.

[2] Meghanathan, N. (2010). Impact of the Gauss-Markov Mobility Model on Network Connectivity, Lifetime and Hop Count of Routes for Mobile Ad hoc Networks. JNW, 5(5), 509-516.

[3] Zhu, L., Zhang, Z., & Xu, C. (2017). Secure and privacy-preserving data communication in internet of things. Springer Singapore.

[4] Ye, F., Luo, H., Lu, S., & Zhang, L. (2005). Statistical en-route filtering of injected false data in sensor networks. *IEEE Journal on Selected Areas in Communications*, *23*(4), 839-850.

[5] V. C. Gungor and G. P. Hancke, "Industrial wireless sensor networks: Challenges, design principles, and technical approaches," IEEE Trans. Ind. Electron., vol. 56, no. 10, pp. 4258–4265, Oct. 2009.

[6] V. C. Gungor, B. Lu, and G. P. Hancke, "Opportunities and challenges of wireless sensor networks in smart grid," IEEE Trans. Ind. Electron., vol. 57, no. 10, pp. 3557–3564, Oct. 2010.

[7] J. Chen, X. Cao, P. Cheng, Y. Xiao, and Y. Sun, "Distributed collaborative control for industrial automation with wireless sensor and actuator networks," IEEE Trans. Ind. Electron., vol. 57, no. 12, pp. 4219–4230, Dec. 2010.

[8] X. Cao, J. Chen, Y. Xiao, and Y. Sun, "Building-environment control with wireless sensor and actuator networks: Centralized versus distributed," IEEE Trans. Ind. Electron., vol. 57, no. 11, pp. 3596–3604, Nov. 2010.

[9] Choi, S., Baik, M., Kim, H., Byun, E., & Choo, H. (2010). A reliable communication protocol for multiregion mobile agent environments. *IEEE Transactions on parallel and distributed systems*, *21*(1), 72-85.

[10] Sun, B., Shan, X., Wu, K., & Xiao, Y. (2013). Anomaly detection based secure in-network aggregation for wireless sensor networks. *IEEE Systems Journal*, *7*(1), 13-25.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Analysis and Identification of cancer using Nanobots

[1]M.Pushpa Rani, [2]Padmaja Felix
[1,2]Department of Computer Science, Mother Teresa Women's University, Tamil Nadu, Kodaikanal

ABSTRACT *This paper will deal with the latest technology i.e. Nanobots in the field of cancer treatment.Generally, the ultimate result of cells that uncontrollably grow and do not die is CANCER. The process of programmed cell growth and death is APOPTOSIS. When this process breakdown then cancer cells begins to form. Which when grows uncontrollably leads to immediate death.There are so many processes in treating cancer. Such as Radiation Therapy, Stereotactic Therapy, Image Directed Therapy etc. Herein we are going to see how Nanobots involved in treating cancer. These Nanobots are designed to seek out and destroy cancer cells while leaving other normal cells unscathed. So far it been tested in cell culture and in animal studies.As per our traditional way of approach only 60% are been cured with much difficulty. So these comes Nano machine or Nanites or Nanobots are used to cure cancer more precisely and perfectly without damaging other health cells with limited time. In most patients, by the time cancer is detected, metastasis has already occurred. More than 80% of patients diagnosed with lung cancer, for example, present with incurable stage of disease. Nanotechnology is not alien to the clinic; more than 40 Nano therapeutics have reached patients, including anticancer drugs and imaging agents. Many current therapies are not reaching the sites of metastases. Nanomaterial's have the potential to combine multiple therapeutic functions into a single platform, can be targeted to specific tissues and can reach particular subcellular compartments. Primary targeting is the act of steering nanoparticles to the specific organ or organs in which the metastases reside. Secondary targeting is the direction of these delivered materials to the cancer cells and potentially to a specific subcellular location within the cancer cell. Many solid tumours exhibit the enhanced permeation and retention (EPR) effect through which nanomaterial may accumulate and be retained in the tumour. However, this effect is limited to tumours larger than ~4.6 mm in diameter, hindering its use for targeting small, vascularized metastases. To treat the complex problem of metastatic cancer, we must combine the expertise of engineers, biologists and clinicians.*
Keywords: *Nanobots, Cancer Cells, Thermal Ablation, Cancer Treating Drugs, Enzymes. Targeting signalling networks in cancer cells targeting the tumour microenvironment, Challenges for the future*

I. Introduction:

The National Cancer Institute estimates approximately 2.2 million of Americans. In today's history of cancer. We got 1 out of 2 has developed cancer in their lifetime. Though overall cancer rates is lower in India.Cancer survival rates can be greatly improved if scientists are successful in developing microscopic or Nano medical weapons in treating cells.In some areas probably in Universities such as UK have used Nanobots to drill into cancer cells and killing them in 60 seconds.Whereas these have been tested on animals before moving onto the rodents.Nanobots brought an optimistic approach in treatment of cancer. Since more precisely and perfectly these tiny Nanobots damage cancer cells without disturbing other healthy cells. Before going deep into the subject let us sees what **Nanotechnology?** It is the science and technology of small things in particular things that are less than 100nm in size. One nanometres is only 3 atoms long for a comparison a human hair is about 60-100 nm wide.

II. Nanobots:

A Nanobot is a tiny machine designed to perform a specific task repeatedly and with a precision at a Nano scale dimension of few nanometres (10^9) metres in size generally 0.1 to 10 micrometres. The size of Nanobots image is given below.

Figure 1. Original Nanobots Size

III. Features of Nanobots

Nanobots are essentially an adapted machine version of bacteria. They are designed to function on the same scale as both bacteria and common viruses in order to interact with and repel from the human system. Since they are small we cannot see them with our human eyes.IdealNanobots consists of a Transporting Mechanism, An Internal Processor and a fuel unit that enables or triggers them to function.

The Nanobots operate in a virtual environment comparing random, thermal and chemical control techniques. The Nanobots architecture model has Nano bioelectronics as the basis for manufacturing integrated system devices with embedded Nano biosensors (fig 2.)And actuators, which facilitates its application for medical target identification and drug delivery. The Nanobots interaction with the described workspace shows how time actuation is improved based on sensor capabilities. Therefore, our work addresses the control and the architecture design for developing practical molecular machines. Advances in nanotechnology are enabling manufacturing Nano sensors and actuators through Nano bioelectronics and biologically inspired devices. Analysis of integrated system modelling is one important aspect for supporting nanotechnology in the fast development towards one of the most challenging new fields of science: molecular machines. The use of 3D simulation can provide interactive tools for addressing Nanobots choices on sensing, hardware architecture design, manufacturing approaches, and control methodology investigation.

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Transporting Mechanism: Experts believe microscopic silicon called transducers can be successfully used for Nanobots legs since spider like body will work out (Now scientist has designed the new CAR genes to integrate into chromosomes to begin decoding the new gene and producing CARS within just one or two days.)They hope that spider like mechanism could help in fast moving and create a quick and efficient machine to move into human blood vessels.This design could rebuild tissue molecules, rebuild walls of veins and arteries and stop bleeding to save life.

Fuel unit: One possible solution is to adhere a fine film of radioactive particles to the Nanobots body. As the particles decay and release energy these Nanobots could be able to harness this power source. The radioactive film could be enlarged or reduced in any scale without a drop in efficiency.

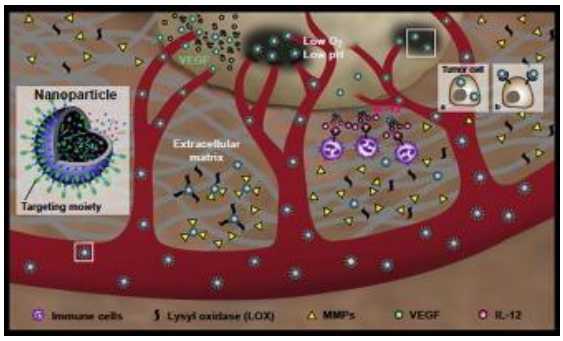


Figure2.embeded Nano bio sensors with fuel unit.

Nanobot's processor:At this time these tiny robots identifies 12 different types of cells in humans ranging from tumour cells to abnormal cells called Leukaemia.More work is yet to be done when it comes to processor development. [6]

IV. Technology behind Nanobots in Treating Cancer:

As these are too small they can safely travel into our Blood stream.Once it is been injected into our body it traverses and identifies and differentiates between Cancer cells and Normal cells.Then it binds itself in the Cancer or Tumour cells and creates a wall around cancer cells so that normal cells left undamaged.These Nanobots are then been exited by the fuel unit with certain range of Infrared Radiations.Where it uses Photo Thermal Ablation to heat cancer affected cells.(cancer cells die at 42'c whereas normal cells die at 46^c).And also these Nanobots have specially folded DNA that serves as a vessel for treating cancer cells without damaging other normal cells.Finally killing Cancer cells alone without disturbing normal cells.Still these biodegradable Nanobots used to genetically program immune cells to recognize and destroy cancer cells by it. Whereby our immune system will quickly mount a strong enough response to destroy cancer cells.[2]

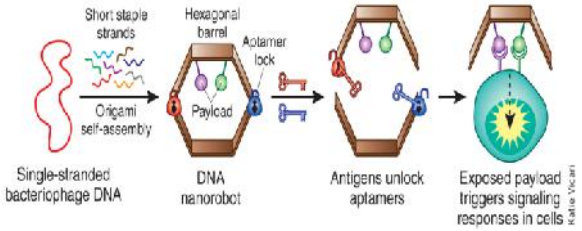


Figure 3.steps involved in curing cancer

In Short, (fig 3).
It destroys the cancer cells in human body.
Removes blocks in blood vessels.
Replacement of DNA Cells is also been done by educating our immune system.

Components of Nanobots used in cancer treatment:

The various components in Nanobots include power supply, fuel buffer tank, sensors, motors, manipulators, on-board computers, pumps, pressure tanks and structural support. The substructures in Nanobots include: 1. Payload- This void section holds a small dose of drug/medicine. The Nanobots could transverse in the blood and release the drug to the site of infection/injury. 2. Micro camera- The Nanobots may include a miniature camera. The operator can steer the Nanobots when navigating through the body manually 3. Electrodes- The electrode mounted on the Nanobots could form the battery using the electrolytes in the blood. These protruding electrodes could also kill the cancer cells by generating an electric current, and heating the cells up to death. 4. Lasers- These lasers could burn the harmful material like arterial plaque, blood clots or cancer cells 5. Ultra sonic signal generators- These generators are used when the Nanobots are used to target and destroy kidney stones. 6. Swimming tail- The Nanobots will require a means of propulsion to get into the body as they travel against the flow of blood in the body. The Nanobots will have motors for movement and manipulator arms or mechanical leg for mobility. The two main approaches followed in construction of Nanobots are Positional assembly and Self-assembly. In self-assembly, the arm of a miniature robot or a microscopic set is used to pick the molecules and assemble manually. In positional assembly, the investigators will put billions of

molecules together and let them automatically assemble based on their natural affinities into the desired configuration Nanobots Control Design is the software developed for simulating Nanobots in environment with fluids which is dominated by Brownian motion. The Nanobots have chemical sensors which can detect the target molecules. The Nanobots are provided with swarm intelligence for decentralization activity. Swarm intelligence techniques are the algorithms designed for artificial intelligence of the Nanobots. The swarm intelligence technique is been inspired by the behaviour of social animals such as ants, bees and termites which work collaboratively without a centralized control. The three main types of swarm intelligence techniques deigned are ant colony optimization (ACO), artificial bee colony (ABC) and particle swarm optimization (PSO) [8].

VI. Types of Nanobots:

The types of Nanobots designed by Robert A. Freitas Jr as artificial blood are:
Respirocytes.
Microbivores.
Clottocytes.

Reciprocytes are the Nanobots designed as artificial mechanical red blood cells which are blood borne spherical 1 μm diameter sized. The outer shell is made of diamonded 1000 atm pressure vessel with reversible molecule-selective pumps. Respirocytes carry oxygen and carbon dioxide molecules throughout the body. The respirocyte is constructed of 18 billion atoms which are precisely arranged in a diamondoid pressure tanks that can store up to 3 billion oxygen and carbon dioxide molecules. The respirocyte would deliver 236 times more oxygen to the body tissues when compared to natural red blood cells. The respirocyte could manage the carbonic acidity which will be controlled by gas concentration sensors and an on-boardNano computer. The stored gases are released from the tank in a controlled manner through molecular pumps. The respirocytes exchange gases via molecular rotors. The rotors have special tips for particular type of molecule.An artificial red cell—the respirocyte designed by Robert A.Freitas Jr Each respirocyte consists of 3 types of rotors. One rotor releases the stored oxygen while travelling through the body. The second type of rotor captures all the carbon dioxide in the blood stream and release at the lungs while the third rotor takes in the glucose from blood stream as fuel source. There are 12 identical pumps which are laid around the equator; oxygen rotors on the left, water rotors in the middle and carbon dioxide rotors in the left. There are gas concentration sensors on the surface of respirocyte. When the respirocyte passes through the lung capillaries, O2 partial pressure will be high and CO2 partial pressure will be low, therefore the on-boardNano computer commands the sorting rotors to load in oxygen and release the carbon dioxide molecules. The water ballast chambers aid in maintaining buoyancy. The respirocytes can be programmed to scavenge carbon monoxide and other poisonous gases from the body. The respirocyte works as an artificial erythrocyte by mimicking the oxygen and carbon dioxide transport functions. A 5 cc therapeutic dose of 50% respirocyte saline suspension containing 5 trillion Nanobots would exactly replace the gas carrying capacity of the patient's entire 5.4 litres of blood.

Microbivores are the Nanobots which functions as artificial white blood cell and also known as nanobotic phagocytes. The microbivore is a spheroid device made up of diamond and sapphire which measures 3.4 μm in diameter along its major axis and 2.0 μm diameter along minor axis and consists of 610 billion precisely arranged structural atoms. It traps in the pathogens present in the blood stream and break down to smaller molecules. The main function of microbivore is to absorb and digest the pathogens in the blood stream by the process of phagocytosis. The microbivore consist of 4 fundamental components: i. An array of reversible binding sites. ii. An array of telescoping grapples. iii. A morcellation chamber. iv. Digestion chamber. During the cycle of operation, the target bacterium binds to the microbivore surface via species-specific reversible binding site. A collision between the bacterium and the microbivore brings in the surface into intimate contact, allowing the reversible binding site to recognize and weakly bind to the bacterium. A set of 9 different antigenic markers should be specific and confirm the positive binding event confirming the presence target microbe. There would be 20,000 copies of the 9 marker sets distributed in 275 disk shaped regions across microbivore. When the bacterium is bound to the binding.

Clottocytes Haemostasis is the process of blood clotting when there is damage to the endothelium cells of blood vessels by platelets. These platelets can be activated by collision of exposed collagen from damaged blood vessels to the Biomedical Science and Engineering 45 platelets. The whole process of natural blood clotting can take 2-5 minutes. The nanotechnology has shown the capabilities of reducing the clotting time and reducing the blood loss. In certain patients, the blood clots are found to occur irregularly. This abnormality is treated using drugs such corticosteroids. The treatment with corticosteroids is associated with side effects such as hormonal secretions; blood/platelet could damage lungs and allergic reactions. The theoretically designed clottocyte describes artificial mechanical platelet or clottocyte that would complete haemostasis in approximately 1. It is spherical Nanobots powered by serum-ox glucose approximately 2 μm in diameter containing a fibre mesh that is compactly folded on-board. The response time of clottocyte is 1001000 times faster than the natural haemostatic system. The fibre mesh would be biodegradable and upon release, a soluble film coating of the mesh would dissolve in contact with the plasma to expose sticky mesh. Reliable communication protocols would be required to control the coordinated mesh release from neighbouring clottocytes and also to regulate multidevice-activation radius within the local clottocyte population. As clottocyte-rich blood enters the injured blood vessel, the on-board sensors of clottocyte rapidly detects the change in partial pressure,

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

often indicating that it is bled out of body. If the first clottocyte is 75 μm away from air-serum interface, oxygen molecules from the air diffuse through serum at human body temperature. This detection would be broadcasted rapidly to the neighbouring clottocytes through acoustic pulses. This allows rapid propagation of a carefully controlled device-enablement cascade. The stickiness in the fibre mesh would be blood group specific to trap blood cells by binding to the antigens present on blood cells (Figure 4). Each mesh would overlap on the neighbouring mesh and attract the red blood cells to immediately stop bleeding [22].

Blood clotting mechanism of clottocytes The clotting function by clottocyte is essentially equivalent to that of natural platelets at about 1/10,000th the concentration in the blood stream i.e. 20 clottocytes per cubic millimetre ofblood. The major risk associated with the clottocytes is that the additional activity of the mechanical platelets could trigger the disseminated intravascular coagulation resulting in multiple micro thrombi [6].

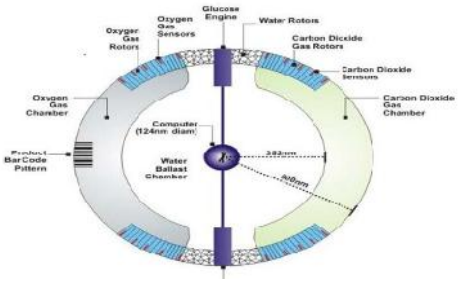


Figure4.Types of Nanobots

VII. Enzymes used in Nanobots to Treat Cancer:

RED65: This is an herbal formulation that uses an extract of Hirudin molecule from the salivary gland of Hiruorientalis, An Asian Medicinal Leech. The most effective anticoagulant agent for clearing toxins from blood streams and cleaning the blood fibrin so that it flows better.

P-A-L PLUS DIGESTIVE ENZYMES:A stack of digestive enzymes will get into blood stream and clean it up.

Papaya pro: This is Green Papaya powder and this is more aggressive than Pancreatic Enzymes in treating and destroying Cancer cells.

Endocarp Elixir: It stimulates cells to repair themselves and supports the body in several ways.

Foliate: This is second most important supplement you need to reverse catabolic waste and start gaining some weight.

VI. Statistics taken for Nanobots Treatment: If the Cancer cells in patient are high the usage of Nanobots reduce cancer cells by exactly treating the cancer cells alone with the enzymes. Whereas in normal therapy the rate of destroying the Cancer cells is very slow and not precise when compared to Nanobots .The graph below clearly states those Nanobots activating enzymes and its effective killing of tumour cells.It gives clear view of Nanobots compared to other technology.[1]

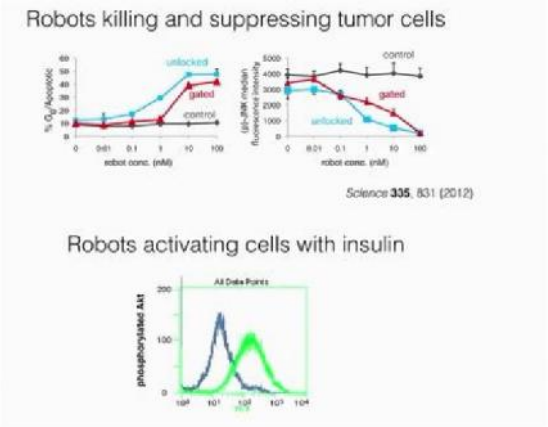


Figure 5: Activity of Nano boat in human cells

VIII. Various Medical applications of Nanobots:

Nanobots in drug delivery
Nanobots in surgery
Diagnosis and testing

Nanobots in gene Therapy
Parasite removal
Breaking up of kidney stones

IX.Conclusion:

Use of Nanobots in cancer treatment has a wider scope. It can be used anywhere in terms of human physiology. This would lead to new revolution in terms of medicine. With a swarm of Nanobots protecting from inside, we could realistically be free from diseases I the next few decades. [1] The Nano robots used in medicine are predicted to provide a wealth of promise. When the severe side effects of the existing therapies are been considered, the Nanobots are found to be more innovative, supportive to the treatment and diagnosis of vital diseases. The respirocytes would be 236 times quicker when compared to normal red blood cells. The Nanobots are found to exhibit strong potential to diagnose and treat various medical conditions like cancer, heart attack, diabetes, arteriosclerosis, kidney stones etc. The Nanobots can allow us a personalized treatment, hence achieving high efficacy against many diseases.[6].

Reference:

[1] Anjinkya Bhatt Nanobots," The Future Medicine".
[2] Brandon Tomlin,"Journal of a new method of treating cancer"
[3] Cambridge Network: Synthetic organs, Nanobots and DNA Scissors
[4] Cosmos01/18 – Super –strong cell size Origami Robots
[5] D.Karthick Raaja, V.Ajay, S.G.Jeyadev, Kumar, Karthikeyan, C.Ravi Chandra,"A Mini Review of Nanobots in Human Surgery hand cancer Therapy".
[6] Apoorva Manjunath, Vijay Kishore*
Department of Biotechnology, Sapthagiri College of Engineering, Bangalore, India the Promising Future in Medicine: Nanobots
[7] En.m.wikipedia.org.
[8] www.iflscience.com
[9] www.futureforall.org
[10] www.techopedia.com
[11] https://singulrityhub.com

Predictive Analytics: A Case Study on Multivariate Data

M.Thangaraj[1] ,P.Aruna Saraswathy[2], M.Sivakami[3]
[1,2,3]Department of Computer Science, Madurai Kamaraj University, Madurai
Tamilnadu State, India.

ABSTRACT *Connectivity ensured data accumulation and eventually produced Big data. This data is high in volume, variety and travels with higher velocity bearing unforeseen opportunities for knowledge discovery. Among other possibilities, the capability to predict an occurrence or probability is named as 'Predictive Analytics' (PA). Moreover, nowadays 'present' is only used to see the future. This aspect changed the perception of activities in all domains, Business, Education, Healthcare, Government, Entertainment and Banking to name a few. This paper discusses the usage of PA while handling unstructured data, it compares various techniques on a single dataset and describes the performance of each technique.*
Keywords: *Prediction, Machine learning, NLP, titanic data.*

Introduction

Predictive analytics is a highly scientific process that analyses past/current data to predict the future behavior of an entity based on the hypothesis, 'history has a tendency to repeat'. It comes under the field of "Advanced Analytics" of big data analytics. Advanced analytics is the superior adaptation of "Business Analytics", as the former uses more elegant tools like Machine Learning (ML). PA deals with the extraction of knowledge by identifying relationship between the entities from previous occurrences and utilizing the acquired knowledge to predict unknown trends. When the outcome is spread over a wider temporal span, such as, population over a period of five years or rainfall in the next ten years it is called "Forecasting". As the occurrence is a 'transition / evolution' of one from the other, it becomes safe to conclude that all forecastings are predictions but all predictions are not forecastings.
Data mining (DM) and PA intersect at their ability to deal with observable variables. The former uses variables to discover hidden-associations and the latter uses the same to invent outcomes. PA employs Statistical techniques to perform mathematical operations and test hypotheses on the data. Predictive analytics has its roots in many fields like data mining, statistics, modeling, machine learning and artificial intelligence. The combination of two or more of these techniques becomes the base for any prediction. This study aims to describe the predictive model building processes and its scope in various existing knowledge discovery approaches, particularly text analytics to investigate the existing problems and suggest directions for future exploration.

Related Research

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns [4]. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.
Predictive analytics is often defined as predicting at a more detailed level of granularity, i.e., generating predictive scores (probabilities) for each individual organizational element. It has various applications Analytical customer relationship management (CRM), clinical decision support systems, customer retention, marketing, fraud detection, project risk management etc [5]. Deep learning is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using a deep graph with multiple processing layers, composed of multiple linear and non-linear transformations [6].
Research in this area attempts to make better representations and create models to learn these representations from large-scale unlabeled data. Some of the representations are inspired by advances in neuroscience and are loosely based on interpretation of information processing and communication patterns in a nervous system, such as neural coding which attempts to define a relationship between various stimuli and associated neuronal responses in the brain [7]. When combined this model gains the capability to adopt the human learning handling with uncertainties in the examples. Stock market prediction has a very long research record in financial economics [1]. It is the act of predicting future price of a stock traded on an exchange.
When predicted correctly, it may result in a very high profit to the investors. There are various economics and statistical methods used by the analysts in the past to predict the price of stocks on the basis of market movements [20]. Various machine learning as well as deep learning concepts are used in data analytics for better predictive analytics. Deep neural networks are very successful models and are being popularly applied in the domain of machine learning. They have given better result in predictive analytics, however they have limitation in adapting the human reasoning process as they work at the low level of human thinking.

Predictive Modeling

The analytics processes could be generalized in two phases as given in figure 1. First phase deals with processing associated with raw data. Second phase deals with operations or techniques to learn the existing behavior of variables in

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

the data. The work flow is the same for all kind of analytical processes such as prescriptive, predictive and descriptive analytics. However, the techniques used to achieve 'Prediction' are entirely different from that of 'Prescription'.

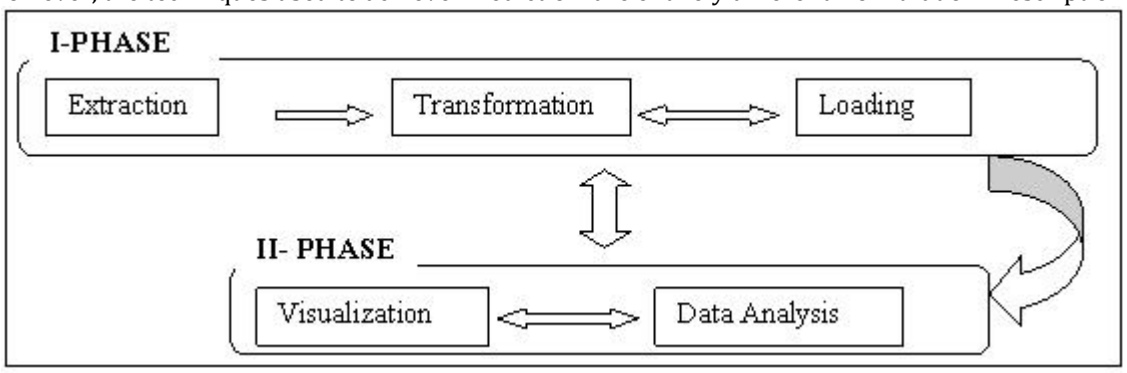


Figure 1 Predictive Modeling

First Phase

Data processing has reached a new level called 'specialized' extract, transform, load(s-ETL) operations [11] with the arrival of big data. Wherever the data source is unstructured or semi-structured, extraction/spidering undertaken to gather required data from a repository. Of the two approaches, [13]Full extraction (complete data) and incremental extraction (recently modified data) the latter is extensively followed. Transformation is appending metadata to the extracted data to give it a structure/classification. Normalization, Standardization, Scaling and Pivoting[12][14][15] are various forms [1] of transformation. Effective Feature selection and feature extraction [3] is mandatory for accurate classification. Loading is accumulating (copy-pasting) data in a database from various sources. When the data is large and distributed, experts suggest in-situ data analysis, external tables, selective tokenization and positional map indexing, multi-threads and speculative loading. According to **Wähner** other techniques to achieve high quality data include, batch processing[8], streaming ingestion[10] and data wrangling/data munging[5].

Second Phase

This is the stage of model building. Distributed analytical systems are adopted to analyze big datasets to reduce memory leakages. Social network analysis, Business analytics, Social media analytics, Business Impact analysis are various forms of analysis operations widely followed [2]. Data visualization is the abstraction of data to comprehend its behavior through visual aids. Multidimensional, Temporal and Tree map are some of the techniques used for visual analysis. Jens Rasmussen's abstraction-aggregation hierarchy [7] is used for interactive visualization. Major advantages of the analytic-pipeline are Multi-level analysis, Data integration and Model building. These models play an important role through extract-transform-analyze-visualize-loop until the optimal insight is attained. Misinterpretation of the results and subjective analysis are crucial problems to be solved in this regard [9].

Methodology

The experiments undertaken for this study were achieved using multivariate Titanic dataset from UCI repository. The datasets consist of 517 instances and 13 attributes. Based on the observations made, the following inferences have been obtained

Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in the data and other essential characteristics of the data. Below are the key factors that helps to select the right regression model:

Data exploration - identify the relationship and impact of variables

goodness of fit for different models - different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the Mallow's Cp criterion. This checks for possible bias in model, by comparing the model with all possible submodels.

Cross-validation -to evaluate models used for prediction. It divides data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give a measure for the prediction accuracy.

If data set has multiple confounding variables, it is not advisable to choose automatic model selection method because do not want to put these in a model at the same time.

It'll also depend on our objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.

Regression regularization methods(Lasso, Ridge and ElasticNet) works well in case of high dimensionality and multicollinearity among the variables in the data set.

Observations

In the titanic dataset, various algorithms where applied for various functions, mentioned in table 1. It gives a clear information about what were the algorithms employed for this empirical research.

Table 1. Techniques applied for analyzing the dataset

Algorithm	**Mining Function**
Naive Bayes	Classification
Generalized Linear Models	Classification, Regression
Support Vector Machine	Classification, Regression, Anomaly Detection
k-Means	Clustering
Non-Negative Matrix Factorization	Feature Extraction
Apriori	Association Rules
Minimum Descriptor Length	Attribute Importance

Performance Metrics of Various Algorithms

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. All the measures except AUC can be calculated by using four parameters such as true positives, true negatives, false positves and false negatives. True positive and true negatives are the observations that are correctly predicted. We have to minimize false positives and false negatives as they are the errors in classification.

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

False positives and false negatives, these values occur when actual class contradicts with the predicted class.

False Positives (FP) – When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells hat this passenger will survive.

False Negatives (FN) – When actual class is yes but predicted class is no. E.g. if actual class value indicates that this passenger survived and predicted class tells that passenger will die.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if higher the accuracy then the model is best. A good, accuracy is a great measure but only when we use symmetric datasets where values of false positive and false negatives are almost same. Therefore, it is safer to look at other parameters to evaluate the correct performance of the models. For our various models, the accuracy levels vary from 0.80 to 0.56 which means our models can be ranked in descending order starting from the k-means to minimum descriptor length .

Accuracy = TP+TN/TP+FP+FN+TN

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived?. High precision relates to the low false positive rate. According to figure 2, the precision of k-means is 0.9 which is a better score when compared to the value of non-negative matrix factorization.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class. The question recall answers is: Of all the passengers that truly survived, how many were labeled?. According to figure 3, the recall value of naive bayes is above 80% which describes the goodness of this model for this kind of binary classification problems.

Recall = TP/TP+FN

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. F1 is usually more useful than accuracy, especially if the class distribution uneven. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, F1 score is 0.6 for SVMs whereas that of generalized linear models if 0.9 indicating a better Fmeasure and goodness of fit

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

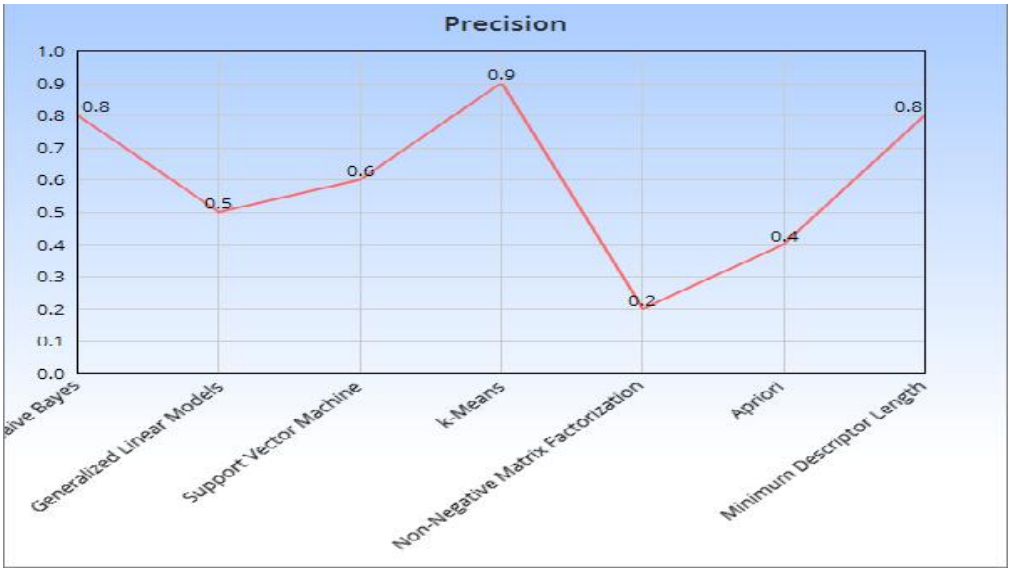


Figure 2 Precision Graph of various algorithms tested

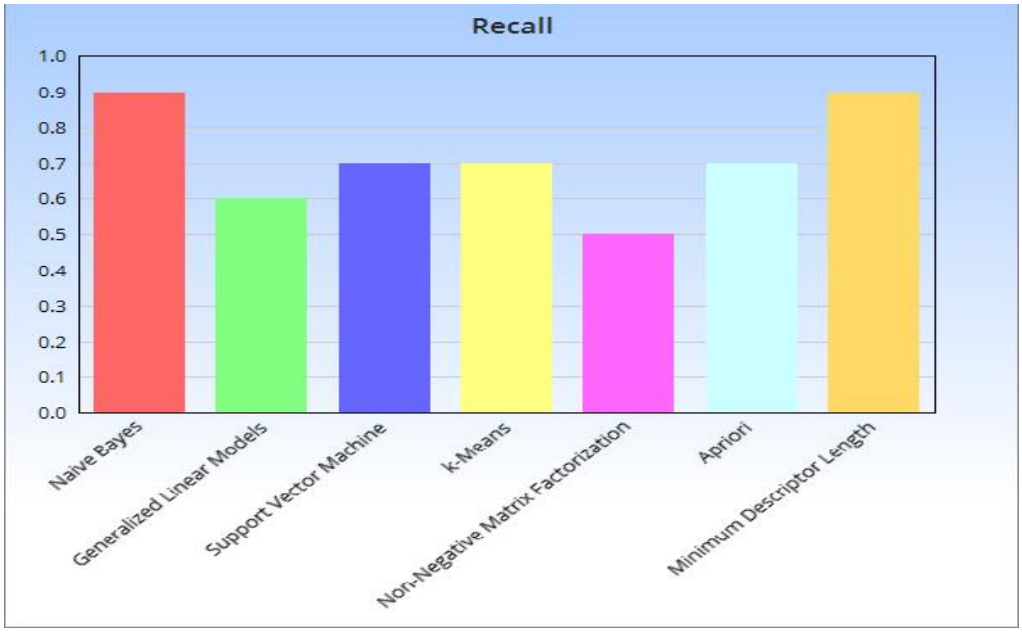


Figure 3 Recall values of various algorithms tested

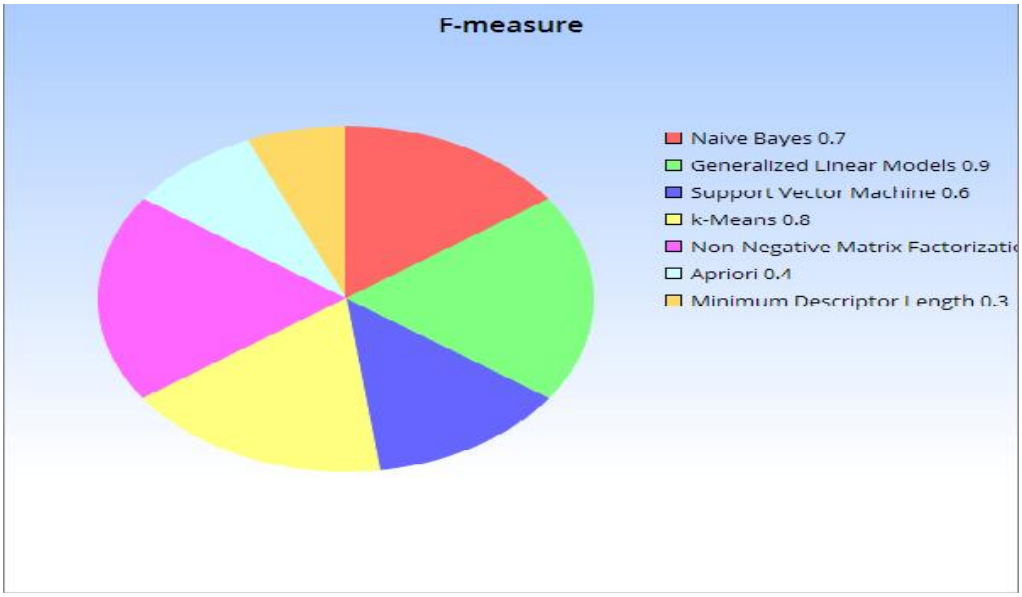


Figure 4 F-measure chart of various algorithms tested

Based on the performance of various algorithms on the given titanic dataset. It is concluded that for multivariate analysis k-means and generalized linear models are better options when compared to naive bayes and SVMs. Their precision, recall as well as f-measure scores show a better picture than other algorithms taken for consideration.

Conclusion

The existing ML/AI/Statistical algorithms have been showing conflicting results in different scenarios, stating that the performance of algorithms is relative, irrespective of its complexity and sophistication. By 2020 there will be 200 billion connected devices, hence, better frameworks for synchronizing data from multiple sources is required. Data collection will always be inconsistent, given the constant updation and search engine tweaks. Fetching only the relevant data for a particular analysis is another problem yet to be solved. Nevertheless, human intervention is still considered significant factor in prediction.

References

[1] Abdullah, M.F. & Ahmad, K. (2015). Business Intelligence Model for Unstructured Data Management. *Proceedings of the 5th International Conference on Electrical Engineering and Informatics.* DOI: 10.1109/ICEEI.2015.7352547. Denpasar, Indonesia. IEEE.

[2] Abiyeva, R.H., Günsela, I., Akkayaa, N., Aytaca, E., Çamana, A. & Abizada,S. (2016). Robot soccer control using behavior trees and fuzzy logic. *Procedia Computer Science, 102, 477-484.* IEEE.

[3] Alanazil, H., Abdullah, A. & Qureshi, K. (2017). A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care. *Journal of Medical Systems,* DOI: 10.1007/s10916-017-0715-6.

[4] Aledo, J.A., Gámez, J.A. & Molina, D. (2017). Tackling the supervised label ranking problem by bagging weak learners. *Information Fusion,* 35, 38–50.

[5] Al-Salemi, B., Noah, S. & Ab Aziz, M. (2016). RFBoost: An improved multi-label boosting algorithm and its application to text categorization*. Knowledge-Based Systems, 103,* 103-117.

[6] Aly, M., Yacout, S. & Shaban, Y. (2017). Analysis of Massive Industrial Data using MapReduce Framework for parallel processing. *Proceedings of Annual Reliability and Maintainability Symposium.*DOI: 10.1109/RAM.2017.7889681. Orlando, FL, USA. IEEE.

[7] Araque, O., Platas, I.C., Rada, J.F.S. & Iglesias, C.A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems With Applications*, 77, 236–246.

[8] Awan, A., Brorsson, M., Vlassov, V. & Ayguade, E. (2016). Micro-architectural Characterization of Apache Spark on Batch and Stream Processing Workloads. *Proceedings of International Conferences on Big Data and Cloud Computing, Social Computing and Networking, Sustainable Computing and Commuanications*. DOI 10.1109/BDCloud-SocialCom-SustainCom.2016.20. Atlanta, GA, USA. IEEE.

[9] Azimi, R., Ghayekhloo, M. & Ghofrani, M.(2016). A hybrid method based on a new clustering technique and multilayer perceptron neural networks for hourly solar radiation forecasting. *Energy Conversion and Management,* 118, 331–344.

[10] Barnaghi, P., Breslin, J.G. & Ghaffari, P. (2016). Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment. *Proceedings of the Second International Conference on Big Data Computing Service and Applications.* DOI: 10.1109/BigDataService.2016.36. Oxford, UK. IEEE.

[11] Bhasuran, B., Murugesan, G., Abdulkadhar, S. & Natarajan, J. (2016). Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases, *Journal of Biomedical Informatics*, *64*, 1–9.

[12] Bontempi, G.(2013). Machine Learning Strategies for Time Series Prediction. Lecturenotes, Université Libre de Bruxelles, Belgium. http://www.ulb.ac.be/di/map/gbonte/ftp/time_ser.pdf.

[13] Bradlowa, E.T., Gangwarb,M., Kopallec,P. & Voleti,S. (2017).The Role of Big Data and Predictive Analytics in Retailing. *Journal of Retailing,* 93, 79–95.

[14] Brownlee, J. (2016). A Gentle Introduction to XGBoost for Applied Machine Learning. Retrieved May 10, 2017 from http://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning

[15] Buia, D.T., Buib, Q.T., Nguyenc, Q.P., Pradhand, B., Nampakd,H. & Trinh, P.T. (2017). A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. *Agricultural and Forest Meteorology*, 233, 32–44.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

CLINICAL GAIT ANALYSIS ON AUTISM WITH CHILDREN

[1]M.Puspa Rani, Professor & Head & [2]B.Latha Kalyana Sunthari, M.Phil Scholar,
[1,2] Department of Computer Science, Mother Theresa Women's University
Kodaikanal,Tamil Nadu,India

ABSTRACT ***T****he manner of moving disorders of persons with an Autism Spectrum Disorder (ASD) have been gaining greater notice over modern lifetime. The paper has to be investigating gait patterns of children with autism using foot pressure variables.The main assessment of gait and upper-body postural types in autism. Abnormal and normal children are going to be compared according to age, height, weight, performance, and IQ and cadence, step length, step width will be wider, while cycle time, double support time, stance time is longer for the experimental group. The result of children with autism having abnormal gait compared with that of normal groups*
Keywords: *Gait recognition, Gait analysis, **A**utism spectrum disorders, GAITRite system.*

Introduction

Gait recognition is a famous technique for different field such as computer vision, machine learning, biomedical, forensic studying and robotics. Dr.M.Pushpa Rani, et al.,[1] described a similar pattern of gait recognition system for individual identification using Modified Independent Component Analysis (MICA).Gait model of autism children can be feel to pain, weakness, extra joint stress, which can affect a child's functional capabilities and an overall reduction in quality of life [3].The purpose of study, to build up an effective treatment specific to autistic children. At first, the background modeling is done from a video sequence. Then , the moving foreground objects in the individual image frames are segmented using the background subtraction algorithm [5].As a Final Point , when a video sequence is fed, the proposed system recognizes the gait features and thereby humans, based on same pattern of gait cycle measurements.

Review On Recent Literature On Gait Analysis

Biometric systems are important identification in various significant applications. Many biometric sources, for instance iris, fingerprint, palm print, hand geometry have been analytically studied and employed in many systems. These recognition suffer from two main disadvantages: 1) Failure to match in low resolution images, pictures taken at a distance and 2) Necessitates user support for accurate results [5]. For these reasons, innovative biometric recognition methods for human identification at a distance have been an urgent need for surveillance applications and gained immense attention most of the computer vision community researchers in recent years, the integration of human motion analysis and biometrics has fascinated several security sensitive environments such as military, banks, parks and airports etc and has turned out to be a popular research direction.

This paper deals with analysis of autism has to be investigated gait in newly diagnosed children. In early stage of children are hypothesized that motor symptoms indicative of basal ganglia and cerebella dysfunction would appear a cross the developmental trajectory of autism. The GAIT Rite gait way used to calculate the mean of gait and intra-walk measurements. Experienced physiotherapists define gait qualitatively based on the age, height, weight, and IQ; although not significant, IQ is minimum children with autism The children with autism of gait datas are compatible with results of autism with cerebellar ataxia: The greater difficulty walking along a straight line, and the coexistence of variable stride length and duration. Children with autism are less coordinated and rated as more variable and inconsistent (i.e. reduced smoothness) relative to the comparison group.

Disease on gait functions: only one study using instrumented measures focuses on this region in autism (see also [6]). The qualitative neurological assessment of individuals with autism by Hallet et al.[7] revealed irregular gait in four of the five participants, a finding imputed to cerebellar rather than basal ganglia dys-function. Consistent with qualitative observations, quantitative measures indicated elevated coefficient of variation (CoV; a quantitative measure of increased variability and irregularity in gait) scores for velocity and step length. Our more recent study of gait of 10 normally intelligent young people (6–14y) with autism indicated that although postural features of the upper body of children with autism resembled the pattern seen in early Parkinson's disease, the variable, rather than reduced, stride length suggested cerebellar involvement. While variable stride length has also been reported in patients with Parkinson's disease [8] , Ebersbach et al.[9] demonstrated that variability in stride length is more defining of cerebellar ataxia. Furthermore, unlike patients with Parkinson's disease, cerebellar ataxic patients also show increased variability in stride duration (also known as 'stride time').[9] Ebersbach et al. postulated that the coexistence of variable stride length with variable stride duration may underpin the clinical features of cerebellar disease, in particular the deficiencies in adjusting the relative movement of multiple joints (see also Stolze et al.[10]).

The aim of the present study was to explore further the extent to which the gait characteristics of newly diagnosed children with autism (i.e. 4–6y) may be consistent with the clinical features of cerebellar ataxia. This was achieved by using a more advanced gait analysis system (cf the Clinical Stride Analyzer used in Rinehart et al.[11] to assess whether individuals with autism show variable stride length together with variable stride duration. The GAITRite technology also enabled us to investigate whether individuals with autism showed deficits on a tandem walking task

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

similar to those reported for patients with cerebellar ataxia.[9] Stolze et al.[11] reported an increased number of missteps, increased step width, and higher ataxia ratios (i.e. the pathway of the foot during tandem walking was very variable) when patients with cerebellar disease tandem walk.

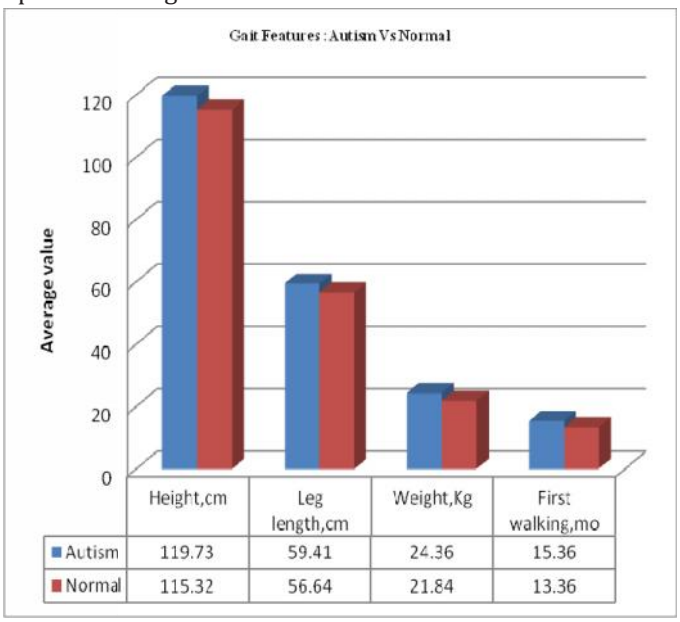
A two phase algorithm is proposed and implemented to detect the abnormal gait. In the first phase, ranking is performed to determine the top gait features. This paper uses T-Test techniques for this purpose. In the second phase, Machine learning algorithms are used for training and testing the occurrence of abnormal gait. For this purpose, this paper uses a modified version of Extreme Learning Machine called Hybrid Extreme Learning Machine (HELM). HELM uses the Analytical Network Process (ANP) for choosing the input weights and hidden biases. The proposed technique is evaluated using CGA Normative Gait database. Experimental results prove that the proposed technique for gait classification results in better accuracy compared to the existing techniques.[2]

METHOD

Comparison participants were children of the authors' colleagues, but not known to the research staff involved in data collection and analysis. Groups were matched according to sex, age, and IQ (Table I). Clinical patients were recruited through their involvement in a Monash University Early Intervention Parent Education Program.[12] Children were recruited during the month after diagnosis from consecutive referrals to two metropolitan and two rural regional assessment services for young children suspected of having autism. The children had a strict *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition (DSM-IV) diagnosis of autistic disorder [17] based on a standardized clinical interview, the Developmental Behavior Checklist [13] for children with developmental disability, and the Autism Diagnostic Interview – Revised.[14] Each diagnosis was confirmed using a screen observation or video record of the interview by an independent clinician. Inter-rater reliability of the diagnosis of autis- tic disorder was high: percentage of agreement 0.98 (95% confidence interval [CI] 0.93–0.99), calculated on the sample of 108 initially referred to the Early Intervention Program.

PROCEDURE

The GAITRite is an electronic walkway, 830cm long and 89cm wide, with pressure sensors embedded in a horizontal grid. The recordable area of the mat is approximately 732cm long X 61cm wide. Sensors are separated at a distance of 1.27cm, with a frequency of 80Hz and temporal resolutions of 11ms. The walkway is connected by a serial interface cable to a desktop computer running MS Windows XP.



Graph I: Gait Features: Autism Vs Normal [17]

For each individual trial, the participant walked along the length of the GAITRite walkway. Participants completed five trials of preferred gait followed by tandem walking. For each condition, participants were given a demonstration and were then required to show their understanding of the instructions by walking down the mat. No participant required more than one demonstration and practice trial.

Preferred gait

Participants walked using their preferred (normal) gait. Participants were instructed to 'walk to the line at the end of the mat, keeping in the middle as you go' and told to 'keep in the middle of the pink mat'.

Tandem walking

A white strip of elastic (width 20mm) was placed along the centre of the mat, from beginning to end. Participants were instructed to walk along the line, placing one foot in front of the other.

All participants completed the preferred gait condition. The data for one comparison participant were removed from the final analysis of the Tandem walking condition after it was ascertained that their data represented an extreme

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

outlier (e.g. above three SDs from the mean); however, re-analysis of the demographic data showed that there was no group difference after this participant's data were withdrawn from this condition.

In addition to the GAITRite quantitative gait measures, walking trials were videotaped (sagittal and coronal plane views), to allow qualitative observational analysis of gait motion.. A sub-group of six children from each group were selected for the qualitative analysis based on tape availability (e.g. not all parents consented to their children being videotaped and video recording malfunctioned on two occasions).

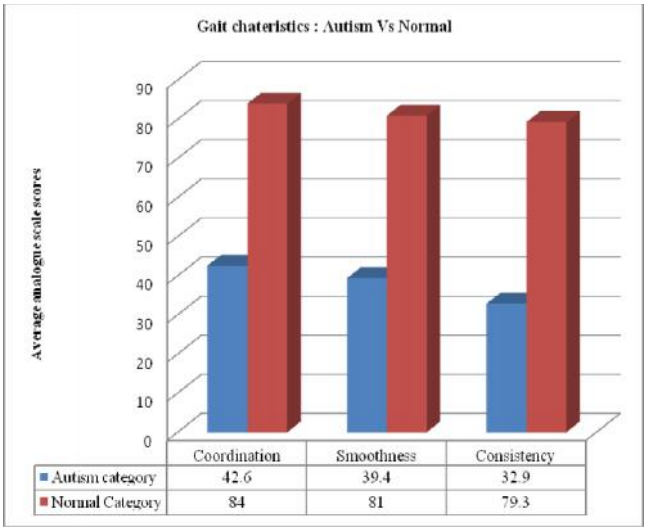
Three videotaped walks at their preferred pace from each of the 12 children were edited onto a single videotape with the participants presented in random order.

Three experienced physiotherapists with experience in gait analysis were recruited as observers to evaluate key qualitative movement characteristics during walking (clinical experience 10–20y; mean 16y). Observers were blind to participant diagnosis. Like our previous study,[18] specific aspects of qualitative movement and postural control during walking were observed and rated. Observed variables included coordination, smoothness, consistency, and head and trunk posture. These items were derived from previous descriptions of gait abnormalities associated with autism and Asperger's disorder. Observers scored the gait characteristics on visual analogue scales (VAS) similar to those used for observational movement analysis by Bernhardt et al.,[17] and reported in our earlier study.[18] Data were extracted from each VAS, and mean scores derived from the three observer scores for each item (see Graph II). Inter-rater reliability was high (mean interclass correlation coefficient [ICC; 2,1]=0.73; two-way random effects).

Results

INTER-WALK ANALYSES

The variables are compared between groups by using independent samples *t*-tests: velocity (cm walked per sec), cadence (steps taken per min), stride length (combined length of
left and right steps in full gait cycle), double sup- port (% of time that both feet are grounded in a complete gait cycle), and heel-to-heel base of support (width [cm] between grounded left and right heel during gait cycle). Average values are presented in Graph II.



Graph II : *Gait Characteristics : Autism Vs Normal [17]*

Consistent with our research examining gait function in older children with autism,[17] there is no significant difference between the mean values for the autism and comparison groups across velocity, cadence, stride length, double support, and heel-to-heel base of support. In our inter-walk analyses, we were also interested in the ability of children with autism to walk in a straight line; some components of ataxic gait, including lateral veering, might be indicative of cerebellar dysfunction.[18] .

We examined three variables:

(1) Missteps in the tandem line walking condition;
(2) Range of sensors activated on the y-axis (width) of the gait mat during each walk; and
(3) An adjusted ataxia ratio.[11]

Missteps

Missteps (i.e. footsteps that did not fall on the line) were calculated across all trials for the line condition. As all participants took a different number of steps, the percentage of missteps across all line trials was calculated. A misstep was defined as a step in which the midpoint for both the heel and toe placement was further away from the edge of the line than half of an average child's shoe width. There was a strong statistical trend for individuals with autism to have a greater percentage of missteps (mean 10.91% [SD 8.10]) than participants in the comparison group (mean 4.52% [SD 4.89]; t[18]=2.07, p=0.052).

y-axis range

The GAIT Rite provides an index of where each footfall occurred in relation to the length and width (x- and y-axis) of the mat. The sensors along the y-axis number sequentially from 1 (left edge of mat) to 48 (right edge of mat).The Average range of sensors activated on the y-axis (width) of the gait mat during a walk was calculated for each individual. This is a measure of an individual's ability to maintain a straight line when walking, with a higher value indicating that more of the mat's width was covered by an individual (i.e. deviation from a straight-line walking pattern). By contrast, a smaller range is analytical of less lateral veering (i.e. walking in a straight line). The range of sensors covered on the y-axis was converted to centimeters. Participants with autism demonstrated a greater y-axis range (mean 27.21cm [SD 3.93]) than comparison group participants (mean 22.42cm [SD 4.32]; t[20]=2.73, p=0.013), suggesting a compromised ability to walk in a straight line.

Ataxia ratio

Stolze et al.[16] calculated an ataxia ratio, which provided a measure of three-dimensional (length, width, and height) stride regularity: (SD of step length + SD of step width + SD of step height)/3. As GAIT Rite only allows for two-dimensional measurement, we calculated an adjusted ataxia ratio: (SD of step length + SD of step width)/2. Participants with autism had a significantly larger adjusted ataxia ratio (mean 19.98 [SD 8.07]) than comparison group participants (mean 13.35 [SD 3.93]; t[20]=2.45, p=0.024).

INTRA-WALK ANALYSES

The GAITRite enables us to compute CoV using within-trial Standard deviation values. Thus, the final CoV value for each condition is the mean of the five CoV values from each of the five trials for that condition. CoV compared between groups, using independent samples *t*-tests, for the following variables: velocity, stride time (time taken to complete full gait cycle), stride length, double support, and heel-to-heel base of support.Children with autism display greater CoV than participants in the comparison group for velocity (t[20]=2.08, p=0.05), stride time (t[14]= 3.27, p=0.033), and stride length (t[14]= 2.30, p=0.037) for the preferred gait condition. No differences in gait variability emerged for the line condition.

The subjects started barefoot walking 5 m before they stepped onto the GAITRite pressure mat and finished 5 m beyond. Each trial the subjects were encouraged by the coach to maintain their most natural gait pattern and speed. For two subjects who didn't have a consistent gait pattern the coach held their hand lightly and walked beside them. The average of three trials for the right foot was calculated for analysis. The temporal-spatial and pressure distribution variables were calculated by the GAITRite software (version 3.2b).

All dependent temporal- spatial and pressure distribution variables were entered into SPSS (version 18.0). To investigate the differences between the two groups means an independent t-test was performed with a significance level of 0.05 applied. The outcome of statistical implications of the temporal-spatial variables are shown in the Graph 1 .

Abnormal Children Gait Classification

Analyzing human gait has earned significant interest in recent computer vision researches, as it has enormous use in deducing the physical well-being of people[19]. Detection of unusual movement patterns can be performed using Support Vector Machines classification with T-Test pre-normalization. Support Vector Machine classifiers are powerful tools, specifically designed to solve large-scale classification problems. Almost all recent works broadly uses SVM method for gait analysis because of its remarkable learning ability. But when dealing with time complexity there exists a limitation with the SVM. As the computation cost for the SVM is high, the recently developed Extreme Learning Machine (ELM) is being used for the gait classification as a better option in this paper . ELM ignore the problems like improper learning rate and over fitting commonly faced by previous iterative learning methods and completes the training very fast. The multi category classification performance of ELM with T-Test is evaluated with the Virginia gait dataset. The conclusion of ELM produces better classification accuracies with reduced training time and implementation complexity when compared to SVM.

CONCLUSION

We analyzed that Children with autism can be very problematic, It is important to improve social communication between the children and their caretakers.The portable GAITRite system can be advantageous as it can be transported to where the subjects are in a familiar environment. The paper has to be investigating gait patterns of children with autism using temporal-spatial and foot pressure variables.

The gait has similar characteristics with elderly gait i.e. a reduction in cadence, gait velocity, step length and an increase in step width. Any implement of the treatment prescribe for autistic children should focus on improving the control and strength of the plantar flexors. Even though, it may do well to perform more studies focusing on the other factors such as age and fitness level that affect autistic children's gait.

REFERENCES

[1] M. Pushpa Rani 1 and G.Arumugam, An efficient gait recognition system for human identification using modified ICA, International Journal of Computer Science and Information Technology

[2] MP Rani , Abnormal GAIT classification using hybrid ELM , Electrical and Computer Engineering (CCECE), 2014 IEEE Canadian 27th

[3] Calhoun, M., Longworth, M., & Chester, V. L. (2011). Gait patterns in children with autism. *Clinical Biomechanics*, 26 (2), 200-206.

[4] Gait-Based Emotion Detection of Children with Autism Spectrum Disorders: A Preliminary Investigation Nursuriati Jamil ,or Haniza Mohd Khia , Marina Ismail, Fariza Hanis Abdul Razak,Faculty of Computer and Mathematical Sciences UiTM Shah Alam 40450 Selangor Malaysia

[5] Pushparani M, Sasikala D, A Survey of Gait Recognition Approaches Using PCA and ICA Global Journal of Computer Science and Technology.

[6] Haas RH, Townsend J, Courchesne E, Lincon AJ, Schreibman L, Yeung-Courchense R. (1996) Neurological abnormalities in infantile autism. *J Child Neurol* 11: 84–92.

[7] Hallet M, Lebiedowska MK, Thomas SL, Stanhope SJ, Denckla MB, Rumsey J. (1993) Locomotion of autistic adults. *Arch Neurol* 50: 1304–1308.

[8] Blin O, Ferrandez AM, Serratrice G. (1990) Quantitative analysis of gait in Parkinson patients: increased variability of stride length. *J Neurol Sci* 98: 91–97.

[9] Ebersbach G, Sojer M, Valldeoriola F, Wissel J, Muller J, Tolosa E, Poewe W.(1999) Comparative analysis of gait in PD, cerebellar ataxia and subcortical arteriosclerotic encephalopathy. *Brain* 122: 1349–1355.

[10] Stolze H, Klebe S, Petersen G, Raethjen J, Wenzelburger R, Witt K, Deuschl G. (2002) Typical features of cerebellar ataxic gait. *J Neurol Neurosurg Psychiatry* 73: 310–312.

[11] Rinehart NJ, Tonge BJ, Bradshaw JL, Iansek R, Enticott PG, McGinley J. Gait function in high-functioning autism and Asperger's disorder: evidence for basal-ganglia and cerebellar involvement? *Eur Child Adolesc Psychiatry*. (Forthcoming)

[12] Tonge BJ, Brereton AV, Kiomall M, MacKinnon A, King N, Rinehart NJ. (2006) Effects on parental mental health of an education and skills training programme for parents of young children with autism: a randomised controlled trial. *J Am Acad Child Adolesc Psychiatry* 45: 561–569.

[13] Einfeld SL, Tonge BT. (2002) *Manual for the Developmental Behaviour Checklist – Primary Carer Version (DBC-P) and Teacher Version (DBC-T).* 2nd Edn. Melbourne and Sydney: Monash University Centre for Developmental Psychiatry and Psychology, and School of Psychiatry, University of New South Wales.

[14] Lord C, Rutter M, Le Couteur A. (1994) Autism Diagnostic Interview – Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord* 24: 659–685.

[15] Wechsler D. (1989) *Wechsler Preschool and Primary Scale of Intelligence–Revised*. San Antonio, TX: The Psychological Corporation.

[16] Wechsler D. (1991) *Wechsler Intelligence Scale for Children*. 3rd edn, Australian adaptation. Sydney, Australia: The Psychological Corporation.

[17] American Psychiatric Association. (1994) *Diagnostic and Statistical Manual of Mental Disorders – Text Revision.* 4th edn. Washington, DC: American Psychiatric Association.

[18] Stolze H, Klebe S, Petersen G, Raethjen J, Wenzelburger R, Witt K, Deuschl G. (2002) Typical features of cerebellar ataxic gait. *J Neurol Neurosurg Psychiatry* 73: 310–312.

[19] Titianova, E. B., Mateev, P. S., & Tarkka, I. M. (2004). Footprint analysis of gait using a pressure sensor system. *Journal of Electromyography and Kinesiology,* 14, 275-281.

[20] MP Rani, G Arumugam, Children abnormal Gait classification using extreme learning machine, Global journal of computer science and technology, 7,2010

Ontology Based Healthcare System for Dengue Awareness

M.Thangaraj[1], P.Aruna Saraswathy [2], M.Sivakami[3]
[1,2,3,]Department of Computer Science, Madurai Kamaraj University, Madurai
Tamilnadu State, India.

ABSTRACT *The need for knowledge integration in an era of big data popularized the concept of ontology in data mining. Nowadays, by virtue of its data assimilation capabilities ontologies act as a template for knowledge representation aiding both the practitioner and the machine in various application domains such as education, healthcare, banking, security etc., This work intend to utilize the expertise of ontology in the healthcare domain with respect to the effects of 'dengue fever', given the difficulty of communicating medical knowledge and patient related information in layman terms. Tamilnadu registered the most number of dengue cases among other Indian states in the year 2017 with more than 50 dengue related deaths in a period of 30 days. It has been found that ninety percent of deaths resulted due to late referrals. The aim of this paper is to educate the user about the genuine nature of dengue fever and necessary precautions to be followed by constructing dengue ontology for tamilnadu scenario. This ontology reflects the various effects of dengue fever, its types, vulnerability criteria, critical factors and appropriate remedies put together to equip the citizens for a better future.*
Keywords: *Ontology, dengue, inference, data mining.*

Introduction

Ontology is the explicit representation of concepts related to a particular domain [1]. Knowledge representation and integration are two main applications of ontology in the field of data mining. It follows a logical structure to establish knowledge in the format understandable by the machines and men. Knowledge integration in the biomedical field is the need of the hour, given the enormity of technical terms and necessity of critical information among lay users. According to the WHO report 2015 Dengue is the highly contagious mosquito-borne viral disease in the world. It is estimated that one third of world population lives in dengue endemic countries. In India three states such as, Karnataka, Kerala and Tamilnadu account for 87,018 confirmed dengue cases and 151 deaths in 2017 alone.

People's ignorance towards importance of cleanliness, viral complexity, complicated epidemiology through vector and the opportunities available to prevent infection unlike other fevers like zika and chikungunya are some of the motivation behind this work to carry out a dengue knowledge representation framework for the Tamilnadu scenario.This dengue fever ontology will provide a platform to access information regarding the epidemiological characteristics of the disease, which enlighten an ordinary person about the remedies and precautions available, in case of infection. One can identify the dengue hotspots effortlessly and stay prepared for the impending outbreak.

The rest of the paper is organized as follows, section 2 provides an overview of related research on dengue ontologies, section 3 explains the methods used in developing the dengue ontology, section 4 describes the contents and inferences of the ontology and finally section 5 concludes the study with suggestions for future research.

2. Related Research

Automated reasoning through ontology is an active area of research [3]. Initially, the role of printed dictionaries containing vocabularies and terminologies was to assist education. The introduction of biomedical research again required the support of dictionaries, for undertaking bibliographic searches, coding public health data and integrating databases. For example, Medical Subject Headings (MeSH) vocabulary for indexing MEDLINE database, International Classification of Diseases (ICD) for coding diagnostic information for billing, SNOMED for documenting pathology, Gene Ontology (GO) for annotating gene and gene-product data to help interoperability between diverse databases, etc. This integrated knowledge helps in decision support for pathogen surveillance, text mining on various diseases to obtain quick diagnostic information and disease evolution pattern mining.

ENVO represents environments in biomedical fields, based on respective communities [4]. It is developed in OBO-Format using the OBO-Edit ontology development tool an alternative for Web Ontology Language (OWL). Dengue has unpredictable clinical evolution [5]. In some serious cases ordinary fever worsens to plasma leakage results in death of individuals. Undifferentiated fever, dengue fever (DF) and dengue haemorrhagic fever (DHF) are three variants of dengue based on different symptoms [7]. A structural representation of knowledge about dengue will be helpful carry out clinical management and disease control, especially when it became a global threat [6].

In the absence of a standard vaccination, deeper knowledge about the disease can also encourage research towards discovering/inventing appropriate vaccines. The dengue virus is of four types of serotypes such as, DENV-1, DENV-2, DENV-3, and DENV-4 [7]. This kind of translational research on infectious diseases offers unified query access aggregated content and semantic indexing of the models. In Tamilnadu, the infection is due to the antigen DENV-1, a moderate variant compared to other types, however, untreated cases prove to be fatal. Ontologies enhance interoperability and optimized search engines [2]. IDODEN, ontology describes dengue fever, enabling better decision making support independent of software architecture.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

3. Methodology

The aim of this work is to utilize the role of ontology in knowledge extraction to attain disease intelligence. The prime motive of ontologies is to model real world concepts explicitly known as domain conceptualization. The dengue ontology constructed for this work is named as DOTAM, Dengue Ontology for Tamilnadu. It consists of the generic concepts such as 'Disease', 'ImpactArea', 'DiseaseRemedies', 'DiseasePrecautions' and 'DiseaseSymtpoms'. This concept articulates the knowledge of the domain in reality through a process called domain abstraction. For eg., taking in lot of fluids is one of the details under the concept DiseaseRemedies. To represent all such details in an unambiguous manner ontology language are used. Predicate Calculus, KIF, Ontolingua, UML, EER, LINGO, ORM, CML, DAML+OIL, F-Logic, OWL are some of the languages to represent ontologies [8]. This work is carried out using the language OWL 2 using protégé-4.3. Final stage in preparing for building any ontology is selecting the best practices for ontology engineering. DOTAM is built in four phases, these phases are common to all kind of ontolgies.

3.1 First phase – This is the phase of 'Purpose Identification', where the clear-cut plans of the ontology are decided based on the intended usage. In this study, creating awareness about the dengue fever is primary aim and educating the masses is the intended usage. Therefore, the ontology should reflect the domain information in a form suitable to be understood by both the machine and man.

3.2 Second Phase - It deals with constructing the ontology and is achieved through 'ontology capture' procedures. Ontology capture identifies important concepts and relationship between them in dengue fever data. Formal definitions are articulated for each concept, relationships and other technical terms to refer to them. The concepts are analysed and the most important ones are selected for generic representation and others are hierarchically placed to imitate the reality through the model. The generic concepts used here are, 'Disease', 'ImpactArea', 'DiseaseRemedies', 'DiseasePrecautions' and 'DiseaseSymtpoms'. There is vast amount of data related to dengue fever infections, its impact on endemic areas, remedial and preventive measures in various databases, medical blogs, scientific reports, and infectious diseases statistics. However, they are not coded or formally structured and standardized among them, so as to extract the complete information by linking the data.

3.3 Third Phase – The requirement specification and competency questions are formulated for the ontology so that the resulting ontology fulfils the aim of the study. The fact plus plus (fact ++) pluggin imported to Protege 4.3 software accomplishes the task of classification and derives the inference model. Inferred ontology is the formal structure with relationships belonging to the disease information identified by the reasoner.

4. Fourth Phase – The constructed ontology is documented according to their resources and type. For this work, the entity annotations in other words, the human-readable comments made on the entity are implemented to some extent.

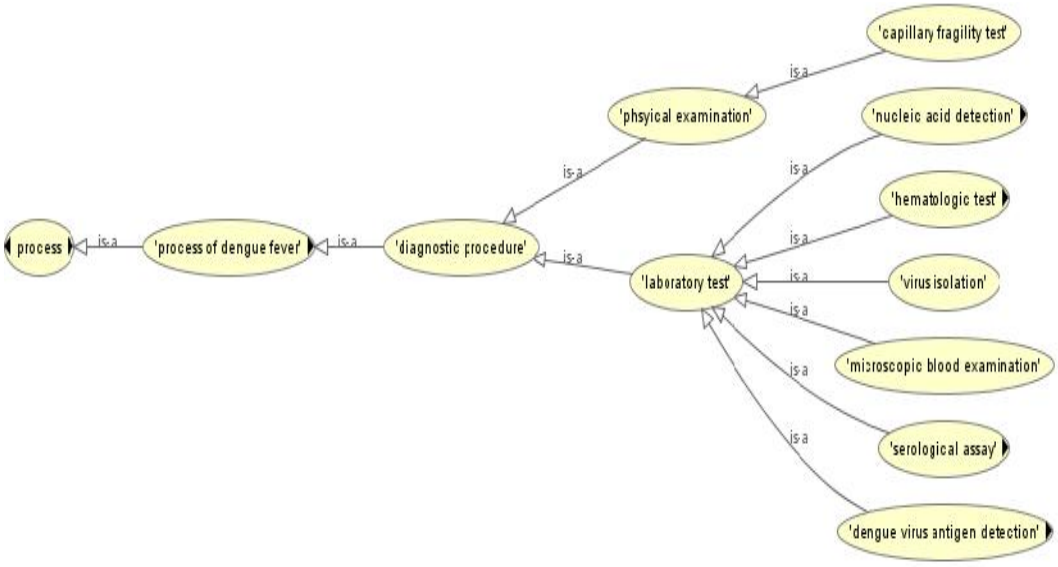


Fig.1 Dengue fever ontology

Following naming convention is used. Class names are capitalized and when there is more than one word, the words are run together and capitalize each new word. All class names are singular. Properties have prefix "has" or "is" before property name or "of" after property name when a verb is not used for a property. All properties begin with a simple letter and when there are more than one words, words are run together with capitalized first letter from second word.

This DOTAM ontology clarifies the user regarding causes of dengue, primary part of infection, early identification, home remedies for prevention, effect of drugs and other environmental effects of dengue.

Autoimmune and Infectious are two catgories of disease variants identified. Disease symptoms are also identified and results of diagnosis are recorded to confirm disease existence. These information may be used by different types of users such as patients and doctors. It can also be updated based on new records and also exceptional cases are

identified. DiseasePrevention class stores data about the preventive methods and measures of diseases. OWL classes are interpreted as sets of individuals (or sets of objects). The class Thing is the class that represents the set containing all individuals. Because of this all classes are subclasses of Thing. The proposed Disease ontology has the following tree structure shown in Figure 1. Figure 1: is a class hierarchical structure of the DOTAM, physicalexamination class tells about the cause of the disease based on symptoms on human body. Thus that is interconnected with disgnosticprocedure by 'is-a' relationship. So this relationship makes it possible to map disease symptoms to disease.

4. Results and Discussion

From the implementation of ontology the following findings have been made and discussions have been expressed. The Disease class has the most important place in this ontology and it is named with intention that other disease ontologies exist in the web may be imported to this ontology in future. It contains information regarding origin of the disease and all its data are logically arranged. It has two sub classes called infectious and autoimmune. Under infectious the diseases related to five most common infectious organisms/ agents namely Virus, Fungus, Prion, Bacteria and Protozoa are placed as sub classes. OrganismStructure subclass has Micro level details of organisms related to diseases by creating the relationship hasStructure between DiseaseStructure class and Disease class. It is a unique feature of this ontology because this disease ontology has molecular level details of both humans and most of other organisms.

The other sub class of Disease class is Autoimmune class and it has three sub-classes called Debilitating, Chronic and Lifethreatening. Among these, the most successful candidate for having other ontologies incorporated into it is Chronic class. The reason behind this is the mostly discussed topic among these three categories on the web is chronic disease. Then the class DiseaseStructure has two sub concepts/ classes called AreaStructure and OrganismStructure. AreaStructure class is for describing the affected area of the disease. Here only the details regarding structural changes at cellular level and below (molecular and sub molecular level) and functional changes are stored. So it shows about what kind of disease occur in that place.

There may be issue regarding placing such a kind of class here and separate DiseaseArea class. There exist some subtle and vital difference between those two classes and it is better to have them as separate classes rather than as a single class. Once ontology has been fully developed, the two classes can be merged, without difficulty. The reason behind the class to be allowing to be existed as a separate class is that it provides a unique way to represent micro level details of the humans in separate place. It is not necessary to identify the disease to place such details in this class as it is not directly derived from Disease class. The OrganismStructure contains details about the organism structure both in micro and macro level. Even the details available about organisms which are not yet associated with disease would also be placed with this class. DiseaseArea has two sub classes called Internal and External. Internal has details about diseased internal parts of human body describing the internal parts both with respect to disease and not with respect to disease, if disease details are not available.

This is basically about the human body parts and not the micro structure of the disease area. This identifies where the disease attacks and how sensitive the disease to that particular area of human body. Even statements given by patients about those areas can be stored here. So this class can be considered as some general class to store information about the disease. External class is same, except that it discusses external body parts of humans such as surface of skin, limbs, face, and hair and so on. Internal class and External classes are not disjoint as some parts may be discussed both in Internal and External classes. DiseaseSymptoms class is responsible for explicitly storing whatever symptoms there regarding a disease. In real world, cause of the disease can be found only by its symptoms on human body. Disease class interconnected with DiseaseSymptoms class (by hasSymptoms relationship) makes it possible to map disease symptoms to disease. Sometimes disease may not be known but identify the abnormality in body as a kind of disease symptoms.

So this class which is not under Disease class and act as separate class facilitates information regarding such kind of symptoms to make its way through to this class. DiseaseSymptoms has two classes called Inside and Outside. They are responsible for symptoms of inside and outside of the human body respectively. Other class is DiseasePrevention and it contains information regarding disease prevention. It will have most results out of research work carried out by doctors, scientists, researchers, individuals etc. all over the world about disease prevention. In Infectious class, all subclasses are made disjoint to each other as no organism is fall into more than one class in this domain, i.e. the Infectious class cannot have any instances in common. The same is done for subclasses of Autoimmune, subclasses of DiseaseArea and subclasses of DiseaseStructure.

The proposed Disease ontology has some notable properties / slots / relations. Two of them are hasStructure and hasSymptoms with inverse properties isStructureOf and isSymptomsOf respectively. Although storing the information 'in both directions' or with inverse properties is redundant from the knowledge acquisition perspective, it is convenient to have both pieces of information explicitly available. This approach allows users to fill in the Disease in one case and the DiseaseStructure in another. When disease is not known disease structure can still be stored and described in relation to unknown disease. Also the knowledge-acquisition system could automatically fill in the value for the inverse relation ensuring consistency of the knowledge base, if the other value exists. There are sub properties as well in the proposed Disease ontology. The hasOrganismStructure is a sub property of hasStructure. The hasAreaStructure is a sub property of hasStructure. The proposed Disease ontology has defined domains and relevant ranges as well. For example, the domain and range for the hasSymptoms property are Disease and DiseaseSymptoms classes respectively. The domain and range for isSymptomsOf is the domain and range for hasSymptoms swapped over. Although the

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

domains and ranges of hasSymptoms and isSymptomsOf properties are specified, it is not advisable to doi it over other properties of the Disease ontology without further studying those properties and classes covered by them. The reason behind this is that domain and range conditions do not behave as constraints. So they can cause 'unexpected' classification results which lead problems and unexpected side effects. Also the proposed Disease ontology has restrictions. If a disease is there, at least a symptom should be there to indicate that the disease exists. Here an 'existential restrictions' is used to describe individuals in Disease class that participate in at least one relationship along a hasSymptoms (some) property with individuals that are members of the DiseaseSymptoms class. These restrictions are applied to the properties depicted by the dotted arrows in Figure 2.

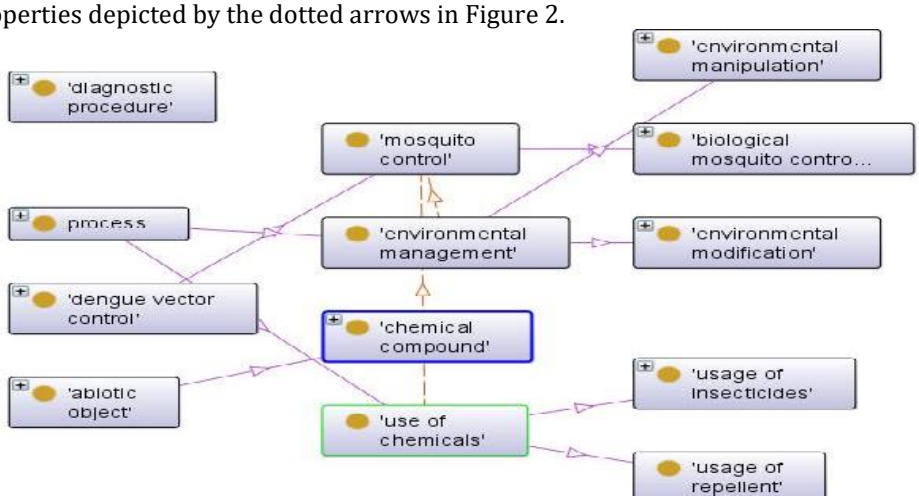


Figure 2: Class Hierarchy with Properties of the Disease

The proposed Disease ontology has primitive classes as well as defined classes to enable the reasoner to classify the ontology. For example, OrganismStructure class under DiseaseStructure class has the individual Giardia lamblia with data property 'locomotion' with the value 'Flagellates'. So this individual has some relation to a disease which is an individual assigned to the Disease class and so acquired by the hasOrganismStructure sub-relationship between the disease and the organism. #Giardia_lamblia'. It should be noted that all the members of the OrganismStructure class are also the members of other super classes of it namely DiseaseStructure and Thing. OrganismStructure class should be used to populate the proposed disease ontology with millions of organisms existing in the world either by importing ontologies which contain those individuals or adding those individuals by communities under the OrganismStructure class. If the Disease ontology designed here is helps to assist in natural language processing of articles in healthcare, health research and medical magazines / journals, it may be important to include synonyms and part-of-speech information for concepts in the Disease intelligence. This is little bit discussed when naming conventions are discussed. In addition to that, annotation which can be incorporated with the concepts will facilitate this.

Conclusion

Disease intelligence is the core aim of this work carried out with regard to dengue impacts. The proposed DOTAM ontology, provides information about rapidly spreading and changing diseases in the tamilnadu area. The proposed Disease ontology should be further developed by the community, updating with new information at regular intervals. This system should be made available for lay users to make them aware of effects of dengue beforehand and enable minimizing fatal cases. As a future work this ontology could be further extended to include concepts like other defects of patients, hereditary information, and if possible genetic data. This new ontology will be effective for other disease intelligence tools in healthcare domain.

References

[1] Herdiani, A., Fitria, L. Hayurani, H. Wahyu, C. Wibowo & Sungkar, S.(2012, July.). Hierarchical Conceptual Schema for Dengue Hemorrhagic Fever Ontology. *IJCSI International Journal of Computer Science Issues*, Vol. 9, *Issue 4*, No 3.

[2] Mitraka E, Topalis P, Dritsou V, Dialynas E, Louis C (2015, Feb.). Describing the Breakbone Fever: IDODEN, an Ontology for Dengue Fever. *PLoS Negl Trop Dis* 9(2): e0003479. doi:10.1371/journal. pntd.0003479.

[3] Cowell, L.G & Smith, B.(2010). Infectious Disease Ontology. *Infectious Disease Informatics.*DOI 10.1007/978-1-4419-1327-2_19, Springer Science+Business Media.

[4] Buttigieg P.L ., Morrison, N. Smith, B. Mungall, C.J & Lewis, S.E and the ENVO Consortium. (2013). The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics*. http ://www.jbiomedsem.com/content/4/1/43.

[5] WHO(2015). Dengue guidelines for diagnosis, treatment, prevention and control. *A joint publication of the World Health Organization (WHO) and the Special Programme for Research and Training in Tropical Diseases (TDR).* ISBN 978 92 4 154787 1.

[6] Whitehorn, J & Farrar, J. (2010, July). Dengue. *Published Online.* British Medical Bulletin . 95: 161–173. DOI:10.1093/bmb/ldq019.

[7] Thisyakorn, U & Thisyakorn, C. (2014). Latest developments and future directions in dengue vaccines. *Ther Adv Vaccines.* 2(1) 3–9 . DOI: 10.1177/2051013613507862.

[8] Rajapakse, M., Kanagasabai, R. Ang, W.T. Veeramani, A. Schreiber, M.J & Baker, C.J.O.(2008). Ontology-centric integration and navigation of the dengue literature. *Journal of Biomedical Informatics*. 41. 806–815.

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

Enhanced Algorithm for Scheduling Task in Cloud Computing

[1]R.Barani,[2] Suguna Sangaiah
[1]Assistant Professor, Department of Computer Science, Sri Sarada Niketan College of Science for Women, Karur-5. Email: baraniraj77@gmail.com
[2]Assistant Professor, Department of Computer Science, Sri Meenakshi College for Women (A), Madurai. Email:kt.suguna@gmail.com

ABSTRACT *Cloud computing is an infrastructure which is suitable for handling large sizes tasks. Scheduling becomes challenging issue in cloud computing. Task scheduling is one of the types of scheduling. To improve overall performance of the system, task scheduling must be done effectively. In this paper, we discuss on several task scheduling algorithms available in the cloud. We have compared our proposed Task scheduling algorithm with existing Heuristics algorithms Improved Max-Min and Enhanced Max-Min algorithms using metrics of waiting time, Throughput and Turnaround time. We have found that our algorithm performs well when compared to other scheduling algorithms.*
Keywords: *Cloud computing; Task scheduling; Heuristics algorithms; Enhanced Task Scheduling Algorithm; Performance analysis*

Introduction

Cloud computing is a distributed computing Environment which handles a large amount of data. Cloud computing is used to store, manage and process data. The three service models available for cloud are Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). The characteristics, service models and deployment models of cloud are shown as in figure1.

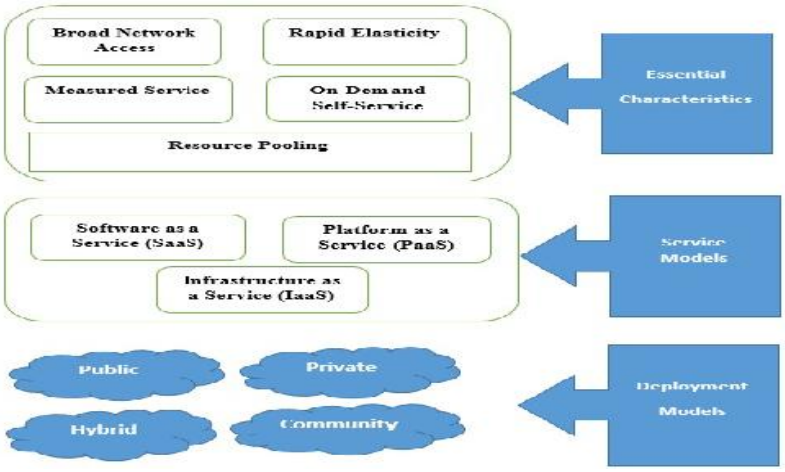


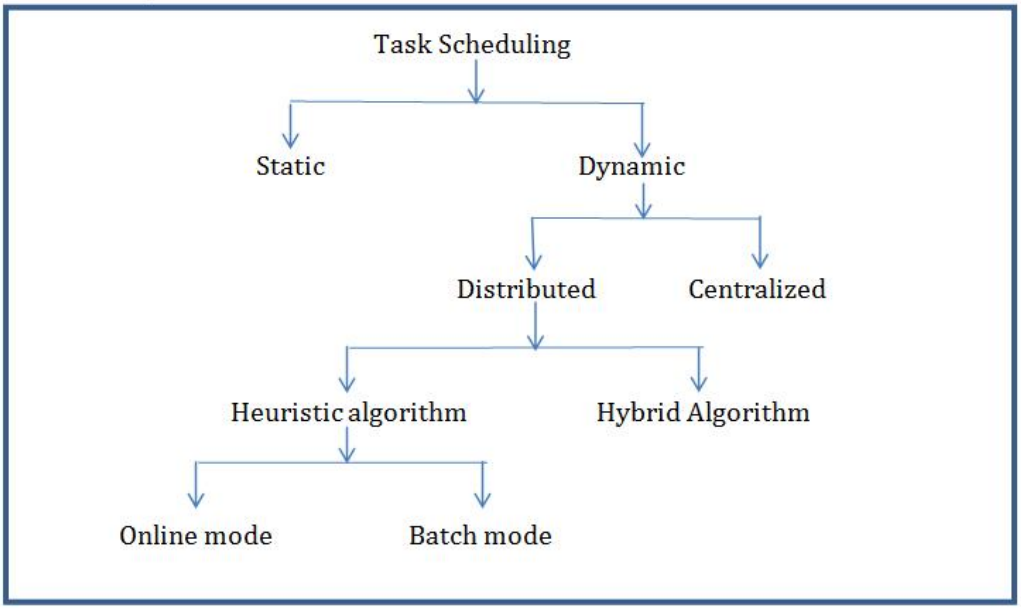
Figure 1.Cloud computing model

There are several issues in cloud computing. One of them is cloud load balancing. Load Balancing is the process of allocating incoming tasks to available resources such that no node can be overloaded or under loaded. In order to balance the load for cloud, we need scheduling methods. Scheduling algorithms are required to improve CPU utilization, to increase turnaround time and to improve throughput of the system. There are three common scheduling methods available. Resource scheduling, workflow scheduling and Task scheduling, Resource scheduling is mapping of virtual resources among machines, Workflow scheduling is scheduling of workflow in suitable order and task scheduling is scheduling of available tasks to virtual machines.

A Task is a piece of work executed in a unit of time. The task can be of independent or dependent. Independent tasks are those whose execution time is known in advance before execution starts. Dependent tasks are those whose execution time is known during run time. Task scheduling can be done in homogeneous or heterogeneous environments. Task Scheduling can be static or dynamic in nature. Static scheduling is done in a homogeneous environment. In static scheduling, we can schedule the task which is known in prior and assign to virtual resources. Dynamic scheduling can be done in heterogeneous environments. Here the task is scheduled instantly as they arrive on the system. [1] Dynamic scheduling mechanism performs better than static mechanism.

Dynamic task Scheduling algorithms can be Centralized or Distributed. In centralized, Scheduling decisions are made by the single node called central node. In this type of scheduling, time gets reduced as decisions are made by single node, but it creates overhead on centralized node. There is no fault tolerance. In Distributed scheduling, no node is responsible for scheduling therefore no overhead occurs. It is fault tolerance and load is balanced. The Types of scheduling algorithms are shown in Figure 2 as

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236



In Next section we present the study of various task scheduling algorithms in the cloud. In next section, existing Heuristic methodologies are discussed. Then, a new Enhanced Algorithm for Task scheduling is proposed and in next section performance analysis of scheduling algorithms is made.

Survey of Related Works

Task scheduling is performed on different parameters in different ways. The tasks can be allocated to available resources at compile time or at run time. The scheduling algorithms are needed to achieve Quality of Service between cloud providers and cloud users, to balance the load in the cloud and to increase the throughput of the system. There are many algorithms developed by researchers for scheduling based on different parameters. They are

Aarti Singh et al [2] proposed an Autonomous Agent Based Load Balancing algorithm for dynamic load balancing. This algorithm has maximum resource utilization, throughput and minimum response time. But they considered only scalability and reliability. Medhat A et al [3] proposed an algorithm to improve cloud task scheduling. This algorithm is used to find the optimal resource allocation. It considers tasks in batch in the dynamic cloud system. This algorithm results in minimizing the make span of the tasks on the entire system. In this algorithm performance is improved but it has low load balancing.

Gziqian Dong et al [4] proposed a greedy scheduling of tasks with time constraints for energy-efficient cloud computing data center. They used Greedy-task scheduling scheme, most efficient-server-first scheduling to reduce energy consumption of DC servers. This algorithm minimizes average response time and server related energy expenditure. This algorithm has high response time. Jay Patel et al [5] proposed an algorithm which is the combined approach in the improvement of load balancing techniques. It compares clustered algorithm, Throttled algorithm and Equally Spread Current Execution algorithms. This algorithm produced improved response time and the load is distributed and handled in a more efficient way. But it lacks performance.

Santanu Dam [6] proposed an Ant colony based Load Balancing in which an Ant is created to find the under load VM to allocate job. This algorithm results in better response time. But fault tolerance is not considered and all jobs are predicted with same priority. Yingchi Mao et al [7] proposed Max-Min scheduling algorithm where they select task with longest execution time and assign it to VM with minimum completion time. It results in improved resource utilization and reduce response time of tasks. Here starvation occurs for longer tasks. Siva Theja Maguluri et al [8] proposed a load balancing and scheduling algorithm that is throughput optimal. But here job sizes are not considered. Shaowei Liu et al [9] proposed a task backfill based scientific overflow scheduling algorithm with reduced cost effectively bit reliability is not considered here.

Amir Nahir et al [10] proposed a Replication based scheme for distributed load balancing in the context of large data centers. This algorithm improves the expected queuing overhead, but there occurs inter-server signal propagation delay. In [11] Genetic algorithm is used to improve resource allocation. The Genetic Algorithm is an optimization algorithm which uses the computerized searching technique which is based on natural selection and genetics. It reduces the execution time by handling vertical elasticity of resources. In [12] the public cloud is divided into partitions and uses switch mechanism for selecting different strategies for different situations. This algorithm improves throughput, response time and latency, but suffers from deadlock problem.

In [13] Bio-inspired algorithms Honey Bee Optimization (HBO), Ant colony Optimization (ACO) and Random biased are analyzed for scheduling. ACO is better that HBO with minimal processing cost and there is no specified key element to select a specific behavior of scheduling. In [14] Probabilistic provisioning and scheduling in uncertain cloud, an Optimization problem is formulated whose object is to identify scheduling patterns that minimize overall monetary

cost. In this with given probability, deadline associated with the application is satisfied. But it handled only fluctuating workload patterns.

In [15] an algorithm is proposed based on estimating end of service time in heterogeneous cloud. This results in improved processing time and response time, but there occurs a power consumption problem. In [16] a DPSO algorithm is proposed which altered the task selection of PSO for ensuring deterministic load balancing and combined with ORCHID algorithm, this algorithm outperforms in naïve and classical Load Balancing approach but lacks scalability.

Existing Heuristics Methodologies for Task Scheduling

Max-Min Algorithm

Yingchi Mao et al[7] proposed a Max-Min algorithm where a task with maximum execution time is selected and is assigned to resource with minimum execution time. The priority is given to larger tasks. It first assigns time consuming jobs to resources. The early execution of the larger task increases the total response time of the system.

Improved Max-Min Algorithm

O. M. Elzeki et al [10] proposed an Improved Max-Min algorithm which selects a task with maximum execution time and assign it to resource with minimum completion time. This algorithm is an improvement of Max-Min algorithm. In Max-Min algorithm always largest task will be assigned to best available resource and do not consider the completion time. In improved Max-Min algorithm completion time is considered. This reduces overall make span of tasks.

Enhanced Max-Min Algorithm

Upendra Bhio et al [11] proposed an Enhanced Max-Min algorithm which is a modification of Improved Max-Min algorithm. Here we find the average of execution time of jobs. Then select task greater than the average execution time and assign it to slowest resource and then execute improved max-min algorithm. Selecting average sized tasks results in better make span and average utilization of resources.

An Enhanced Task Scheduling Algorithm (Proposed)

In order to improve the performance of the system, we need to increase the response of the tasks (i.e. the waiting time of tasks is reduced) which in turn improves the turnaround time of the tasks. We have proposed an algorithm which increases the response time and turnaround time of tasks without increasing the make span of tasks.

In this algorithm we are finding the mean of execution time of tasks. We are dividing the tasks into two lists as L1 and L2. The first list L1 has tasks which are less than themean value of tasks and other list L2 has tasks which are greater than or equal to the mean value of tasks. The two lists are sorted in ascending order of execution time. If the number of tasks is odd, then the largest value from L1 are selected. If number of tasks are even then smallest value from L2 are selected and is assigned to minimum completion time resource. Once the task completes its execution, it is removed from the list. This process is repeated until all the tasks from the L2 list are executed. Select the maximum execution time task from L1 and is assigned to available resource. Once the tasks complete its execution, it is removed from the list. This process is repeated until all the tasks are executed. The algorithms are shown as

Algorithm :

1. For all submitted tasks T_i in Meta tasks list
2. Calculate Arithmetic mean M of Execution Time Ei for all Tasks in Tasks list.
3. Divide the tasks into two lists L1 and L2.
4. L1 has T_i<M and L2 has T_i>=M
5. For all tasks in L2 execute the algorithm2
6. For all tasks in L1 execute the algorithm1

Algorithm 1:

1. For all submitted Tasks T_i in L1
2. Arrange T_i such that T_i>T_{i+1} >T_{i+2}...Tn for i=1 to n.
3. For all resources R_j
4. Compute C_{ij}
5. While tasks in L1 not empty
6. select first task from L1
7. assign to available resource R_j
8. Remove the task T_k from L1
9. Update R_j and C_{ij}
10. Update L1

Algorithm 2:

1. For all submitted Tasks T_i in L2
2. Arrange T_i such that $T_i<T_{i+1}<T_{i+2}..T_n$ for i=1 to n.
3. for all resources Rj
4. Compute C_{ij}
5. While tasks in L2 not empty
6. Select first task from L2
7. Assign to minimum completion time Resource R_j
8. Remove the task T_k from L2
9. Update R_j , C_{ij}.
10. Update L2

Performance Evaluation of Scheduling Algorithms

The main aim of scheduling algorithm is to improve the overall performance of the system. In order to improve the performance of the system some metrics are to be considered. They are

1. Cost: To improve performance of the system, the cost of resources used to be reduced.
2. Make span: It is the total time taken to complete all tasks in the system. It should be minimized.
3. Waiting Time: It is the difference between Start time of task and submitted time of task.
4. Turnaround Time: It is the time taken for task to complete its execution. It is calculated as
   Turnaround Time= C_i-A_i where C_i–Completion time of Tasks ,A_i - Arrival time of Tasks
5. Resource Utilization : It is keeping the resource as busy as possible.
   Average Resource Utilization=(∑time taken to finish all jobs)/(Makespan * n) Where n is a number of resources
6. Throughput: It is the number of jobs completed per unit time. It should be high.

In order to have Quality of Service ,we need to consider the Execution cost, deadline, Performance, Cost,Makespan,etc.

Consider there are 5 tasks with Instruction Volume (in MI) and Data Volume (in Mb) as shown in Task Specification Table

Table 1. Task Specification Table

Tasks	Instruction Volume(MI)	Data Volume(Mb)
T1	200	400
T2	1000	1500
T3	500	800
T4	400	1000
T5	100	500

The resources R1 and R2 with their processing speed (in MIPS) and Bandwidth (in Mbps) are shown in table as

Table 2. Resource Specification Table

Resource	Processing Speed (in MIPS)	Bandwidth(in Mbps)
R1	100	70
R2	200	100

The Execution time of task on each resource is shown as

Table 3. Execution of Tasks on each Resource

Task	Resources	
	R1	R2
T1	2.0	1.0
T2	10.0	5.0
T3	5.0	2.5
T4	4.0	2.0
T5	1.0	0.5

The results of proposed algorithm is shown graphically as

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

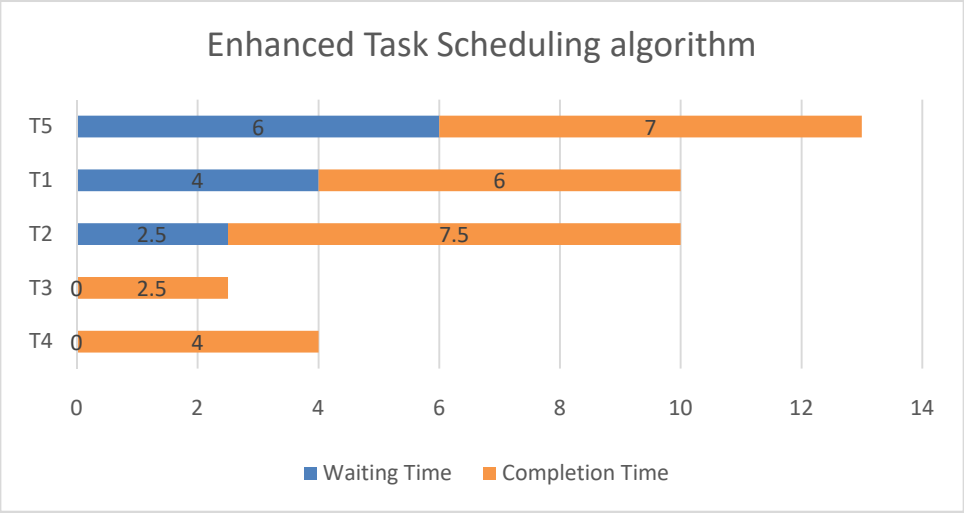


Figure 3: Graphical representation of Enhanced Task Scheduling Algorithm (Proposed)

In proposed algorithm, average waiting time of Tasks are 2.5ms and Average Turnaround Time are 5.4ms.
The Results of using Improved Max-min and Enhanced Max-Min algorithm for the same data are shown graphically as

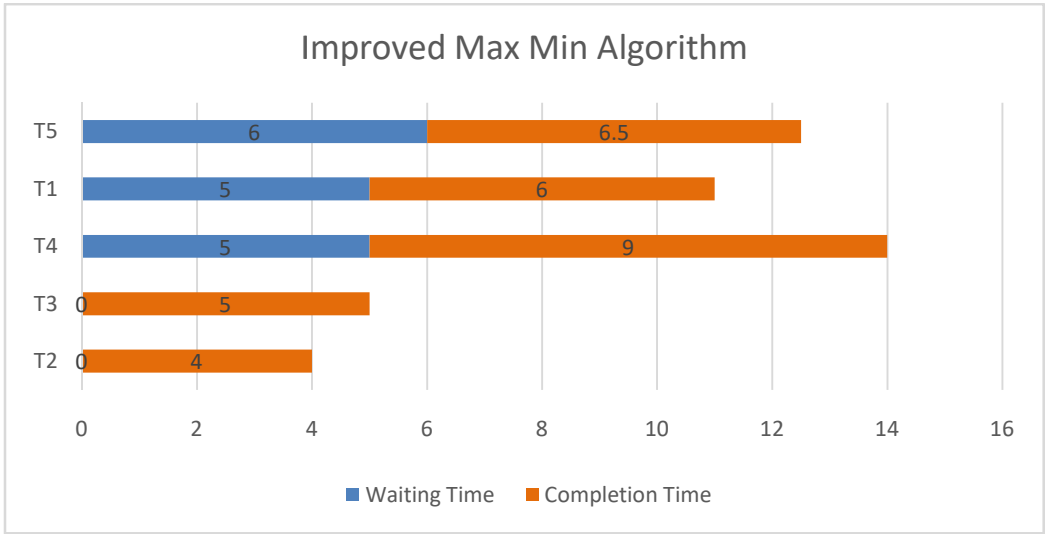


Figure 4. Graphical Representation of Improved Max-Min Algorithm.

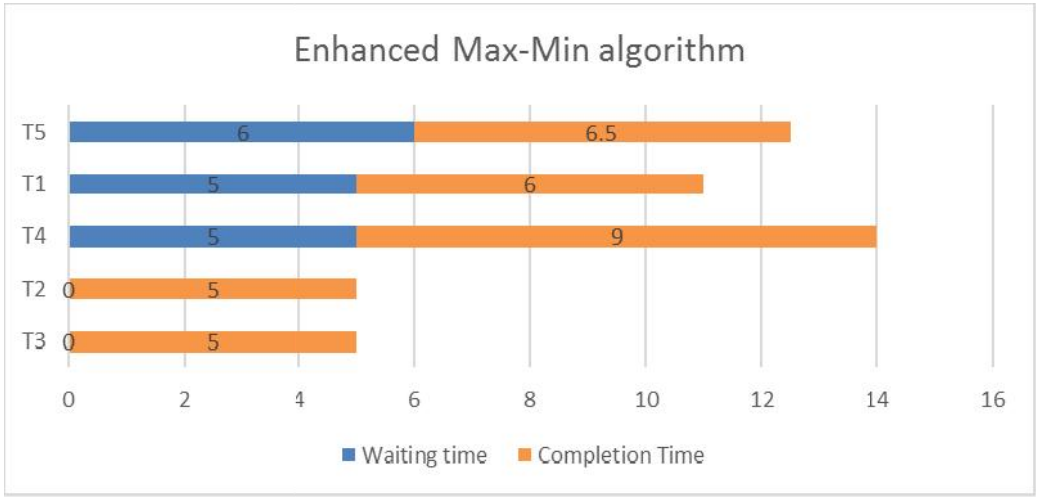


Figure 5. Graphical Representation of Enhanced Max-Min Algorithm

The comparison of these three Heuristics algorithms are as follows

E ISSN 2348 –1269, PRINT ISSN 2349-5138
Cosmos Impact Factor 4.236

Table 4. Comparison of Algorithms

Algorithm/Metrics	Average Waiting Time	Resource Utilization	Throughput
Improved Max-Min	3.2	0.86	0.322
Enhanced Max-Min	3.2	0.86	0.322
Enhanced Task Scheduling	2.5	1.03	0.344

From these results it is clear that in proposed algorithm, the average waiting time of tasks are reduced which increases the response time of tasks. The resource Utilization increases and also throughput are increased. This in turn increases the performance of the overall system. When the number of tasks increases, the waiting time of tasks are decreased and throughput is increased. This shows that the algorithm is scalable.
The results of Average Waiting time of tasks using 5 Tasks, 10 Tasks and 20 tasks in 2 resources are shown as

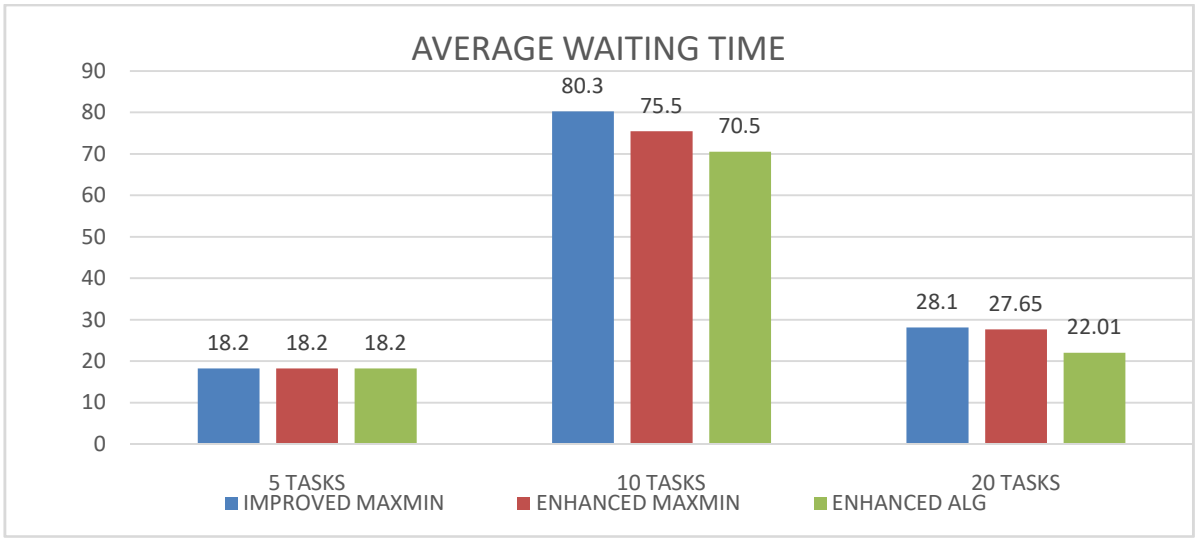


Figure 6. Graphical representation of Average Waiting time of tasks when executed in 2 resources.

From the graph it is clear that average waiting time is reduced when the number of tasks are increased.

The average Turnaround time of 5 tasks, 10 tasks and 20 tasks when executed in 2 resources are shown as

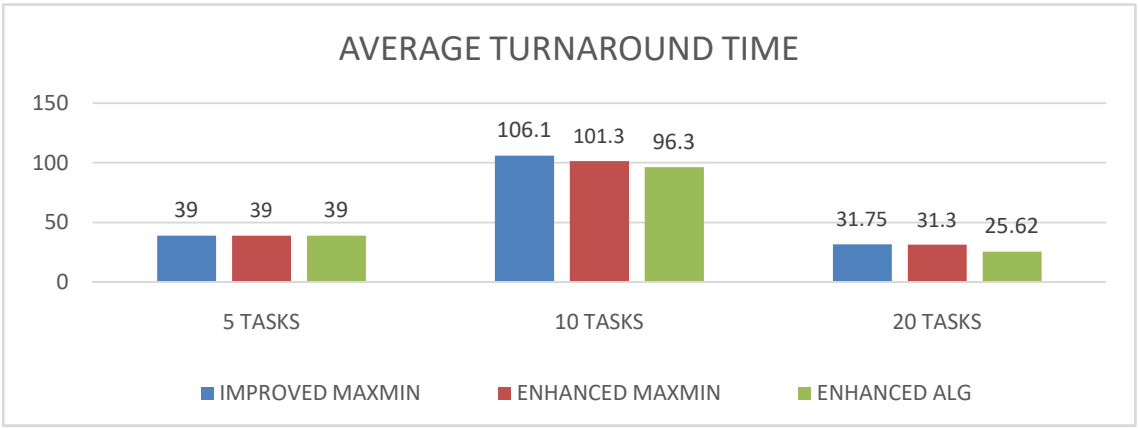


Figure 7. Graphical representation of Average turnaround time of tasks when executed in 2 resources

From the graph it is clear that average turnaround time is reduced when the number of tasks are increased.

Conclusion

There are a number of algorithms that exist for scheduling the task to balance the load in cloud computing. The aim of good scheduling algorithm is to increase CPU utilization, increase throughput of the system. The scheduling is to be said to be effective when it reduces the operational costs of the system, reduce queue waiting time, and decrease turnaround time and to increase resource utilization. In this paper, we have compared Improved Max-Min and Enhanced Max-Min algorithm with our proposed algorithm and observed that our Enhanced Task Scheduling algorithm performs well when compared to other two algorithms.

Future Enhancement

According to our manual calculations our algorithm performs well compared to Improved Max-Min and Enhanced Max-Min algorithm. Our future plan is to simulate it using cloudsim in cloud environment to check for the desired results.

References

[1] Teena Mathew, Chandra Sekaran, John Jose," Study and Analysis of Various Task Scheduling Algorithms in the Cloud computing Environment", 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI).

[2] Aarti Singh,Dimple Juneha and Manisha Malhotra, "Autonomous Agent Based Load balancing in Cloud Computing", International International Conference on Advanced Computing Technologies and Applications (ICACTA-2015).

[3] MedhatA.Tawfeek,Ashraf E1-Sisi,Arabi E.Keshk,FawzyA.Torkey, " An Ant Algorithm for Cloud Task Scheduling", International Workshop on Cloud Computing and Information Security(CCIS 2013).

[4] Gziqian Dong, NingLiuaand Roberto Rojas-Cessa,"Greed Scheduling of tasks with time constraints for energy-efficient cloud-computing data centers.", Journal of cloud computing: Advances, Systems and Applications. Springer open Journal-2015

[5] Jay Patel,Chirag S Thaker and HardikChaudhan,"Task Execution Efficiency Enrichment in Cloud based Load Balancing Approaches.",ICTCS – Nov 14-16 2014,ACM

[6] Santanu Dam, GopaMandal, KousikDasgupta, and ParamarthaDutta,"An Ant Colony Based Load Balancing Strategy in Cloud Computing", Advanced Computing, Networking and Informatics- Volume 2, Smart Innovation, Systems and Technologies 28, Springer 2014

[7] Yingchi Mao, Xi Chen and Xiaofang Li," Max–Min Task Scheduling Algorithm for Load Balance in Cloud Computing:,Proceedings of International Conference on Computer Science and Information Technology, Advances in Intelligent Systems and Computing 255,Springer 2014

[8] Siva ThejaMaguluri*, Student Member, IEEE*, and R. Srikant*, Fellow, IEEE,"* Scheduling Jobs With Unknown Duration in Clouds", IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 22, NO. 6, DECEMBER 2014

[9] Shaowei Liu,Kaijun Ren, Kefeng Deng and Junqiang Song, "A Dynamic Resource Allocation and Task Scheduling Strategy with Uncertain Task Runtime on IaaS Clouds",6th International conference on Information Science and Technology, Dalian, China, May 6-8,2016 IEEE.

[10] Amir Nahir, Member, IEEE, Ariel Orda, Fellow, IEEE, and Danny Raz, Member, IEEE," Replication – based Load Balancing" IEEE Transactions on Parallel and Distributed Systems, Vol 27, No.2, February 2016.

[11] M.Durairaj and P.Kannan,"Improvised Genetic Approach for an Effective Resource Allocation in cloud infrastructure",International Journal of Computer Science and Information Technologies,Vol 6(4),2015.

[12] NehaGohar Khan, Prof. V. B. Bhagat ,"Cloud Partitioning Based Load Balancing Model for Performance Enhancement in Public Cloud",International Journal of Science and Research(IJSR),Volume 3 ,Issue9,September 2014.

[13] Ali Al Buhussain, Robson E. De Grande, AzzedineBoukerche," "Performance Analysis of Bio- inspired Scheduling Algorithms for Cloud Environments", 2016 IEEE International Parallel and Distributed Processing Symposium Workshops.

[14] Marco L.DellaVedova,DanieleTessera and Maria Carla Calzarossa, "Probabilistic Provisioning and Scheduling in Uncertain Cloud Environments", 2016 IEEE Symposium on Computers and communication(ISCC).

[15] Nguyen Khac Chien, Nguyen Hong Son, Ho Dac Loc,"Load Balancing Algorithm Based on Estimating finish time of services in Cloud Computing",International Conference on Advanced Communication & technology(ICACT),Jan 31-Feb3,2016.

[16] Mohammed Ahmed Sherif and Axel-Cyrille Ngonga Ngomo, " An Optimization Approach for Load balancing in Parallel Link Discovery", SEMANTiCS , Sep 15-17,2015,ACM

[17] YingChi Mao, Xi chen and Xiaofang Li,"Max-Min Task Scheduling Algorithm for Load Balancing in Cloud Computing" Journal of Springer (2014)

[18] O.M.Elzeki, M.Z.Reshad and and M. A.Elsoud,"Improved Max-Min Algorithm in Cloud Computing", *International Journal of ComputerApplications (0975 – 8887), Volume 50-No. 12, July 2012.*\

[19] Upendra Bhoi, Purvi N. Ramanuj, Enhanced Max-Min Task Scheduling Algorithm in Cloud Computing, International Journal of Application or Innovation in Engineering & Management, Volume 2, Issue 4, April 2013, pp 259-264.

Human Abnormality Detection using Iris Features based on Fuzzy Support Vector Machine Classifier and Genetic Algorithm

[1] M.Pushpa Rani, [2] R.Subha
Professor & Head, Dept. of Computer Science,
Mother Teresa Women's University, Tamil Nadu, India
Ph.D Scholar, Dept. of Computer Science,
Mother Teresa Women's University, Tamil Nadu, India.

ABSTRACT *Abnormality detection in human beings play an important role in many health care applications. A significant deviation from the well-known standards in human nature is regarded as abnormal. An accurate and reliable method of abnormality detection in biometric feature will helps improving the health-care screening process. Existing efforts in abnormality detection of human beings have been focused on several algorithms databases and had some limitation. Iris is one of the major unique identifier and stable biometric trait is employed. In this paper a new classifier is implemented for abnormality detection of human beings using iris features. The proposed Genetic Algorithm and Support Vector Machine classifier detects the human abnormalities based on the obtained result, the proposed classifiers identify the normal and abnormal conditions of humans. The performances of the proposed classifier are analyzed and compared with the traditional classifiers.*
Keywords:*Iris Recognition, Abnormality Detection, Feature Extraction, Fuzzy Logic andSupport Vector Machine*

Introduction

Biometrics which refers to identifying an individual by his or her physiological or behavioral characteristics has the capability to distinguish between authorized user and an imposter. Biometric used many methods of recognizing a person based on the characteristic such as eye, iris, fingerprint, hand gesture, face detection, iris detection, and vein. In which eye is the most significant features in a human face and it is mostly preferred to find the abnormality detection easier. This paper aims to develop a detection system that automatically detects the iris to show the human condition as normal or abnormal condition.

The iris is ensured securable organ and externally visible whose pattern will be steady for the duration of the life. The iris is unique because of the many-sided quality of the fundamental ecological with hereditary autonomy, contains to a great degree data-rich physical structure and one of a kind surface example and in this way is very sufficiently unpredictable to be utilized as a biometric signature. Since the iris patterns don't modify essentially amid a person‘s lifetime, it is thought to be a standout amongst the most stable and precise personal identification biometric. Statistical analysis reveals that irises have an outstandingly high-level of-flexibility up to 266 (fingerprints appear around 78) and consequently are the most scientifically extraordinary component of the human body, more one of a kind than fingerprints. Henceforth, the human iris guarantees to convey an abnormal state of uniqueness to confirmation applications that different biometrics can't coordinate.

Problem Statement

The unpredictability of iris patterns has higher dimensionality, recognition, decisions are made with high certainty levels enough to support reliable and large database. Detecting abnormal activities is a vital assignment in security checking and medicinal services applications. The Support Vector Machine suffer from certain flaws such as if the points on the boundaries are not informative (e.g., due to noise) and computationally expensive. So it is important to present the hybrid classifiers for investigating the feature extraction and classifier to identify the abnormalities in human beings. It provides a good tradeoff for abnormality detection and allows abnormal activity models to be automatically derived by scanning iris pattern.

The outline of the paper is as follows. Section 2, deals Related work in the areas of abnormality detection and various classifiers. Section 3 gives methodology to solve this problem. Section 4 shows the experiment results. Finally, conclude in Section 5.

Related Works

Galbally et.al proposed novel probabilistic approach based on genetic algorithms to reconstruct iris images from binary templates and analyzes the similarity between the reconstructed synthetic iris image and the original one. A genetic search algorithm are the nature of the search space is unknown. Mainly this method is employed to minimize the probability of a successful attack even when a template is compromised. This could be accomplished by using biometric-based countermeasures to distinguish synthetic images from real iris images or to employ liveness-detection techniques. The performance results indicate that an eventual attack against iris matchers using such reconstructed images would have a very high chance of success.

Chenet.al (2016) assessed another structure for Detection of Diabetic from Iris image. For clinical feature analysis, enhancement is essential for extraction of deep layer features. An eye image is procures and stored into

database, filtered that image of eye by using median filter, found iris pattern from eye image, normalized and enhanced that iris pattern and extracted some features like mean, variance, standard deviation etc. Several classification methods can be used for training and classification purpose.

Yu et.al (2004) concentrated on the usage of MultipleInstance Learning (MIL) in the area of medical image mining, especially to hard exudates detection in retinal pictures from diabetic patients. To extract relevant image features, utilize the algorithm described in that naturally finds relative invariant connections among objects in an arrangement of pictures. This approach deals with the highly noisy images that are common in the medical area, improving the detection specificity while keeping the sensitivity as high as possible. Experiments on real-life data set of diabetic retinal screening images show that we are able to achieve some improvement upon a previous system without any additional time cost.

Recently, a human-in-the-loop iris recognition system was created, in view of detecting and matching iris crypts. Chen et.al (2016) proposed a detection method to catch iris crypts of different sizes consequently.

Menyhartet.al (2016) gives a short depiction of the historical backdrop of the Support Vector Machine (SVM) method and fuzzy logic and their main parameters. The main focus of this method is to find the most optimal hyperplane. It is possible to find solutions to different types of classification problems with these methods. The system has to check the telemetry data in real time and it is supposed to be able to define which hyperplane is the best in practice, in different real situations.

Proposed Methodology

Abnormality in human beings can be detected by the classifying the various abnormal regions based on their severity.This paper basically engaged, to identify abnormalities of human beings by automatic detection of iris crypts. The fundamental point is to recognize the human abnormalities by iris scanning. The procedure of iris process incorporates following steps and the extracted iris images are identified by proposed Genetic Algorithm with SVM to distinguish the abnormalities.At first, a subset of the images is prepared to get the relevant features for learning the characteristics of the abnormalities in the images and makes the observation that in images, more often than not, it is the relative relationships among the image features that are meaningful for interpretation. The following are the steps involved in iris-based abnormality detection. It starts from mage acquisition, preprocessing, feature extraction, feature selection and abnormality classification.

A. Classifiers

Certain classification using a single visual descriptor fails to achieve satisfactory and robust results since a certain feature may be present and dominant in more than one classes.The Support Vector Machine have some limitations such as due to the large data set the required training time is higher and also it doesn't perform well, when the data set has more noise i.e. target classes are overlapping to overcome this flaws Fuzzy SVM are implemented.Support Vector Machines are learning machines based on statistical learning theory that can be used for pattern classification or regression. They provide high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. Fuzzy Systems are those systems whose variables have as domain fuzzy sets. They encode structured, empirical (heuristic) or linguistic knowledge in a numerical framework . They are able to describe the operation of the system in natural language with the aid of human-like IF-THEN rules. However, they do not provide the highly desired characteristics of learning and adaptation.

The abnormality detection of human beingsbased on iris pattern is implemented by hybrid classifiers as Genetic Algorithm with SVM classifier.

B. Proposed GA/SVM Classifier

The popular classification technique as hybrid classifiers combination of Genetic Algorithm and Support Vector Machine is employed. To sum up, the GA selects some features as an individual and SVM evaluates them by classification, and the result is used for estimating the fitness of the individual and to determine the abnormality detection of iris. The possible choices of feature pools FPidefine the evolutionary search space. Figure 1 shows the flowchart for Proposed GA and SVM. This is carried out separately on each FPi. At the start of the search, a population of individuals (i.e., feature subsets) is randomly initialized from the feature pool FPi. Each individual of the current population is evaluated according to a fitness function. Each time the fitness is evaluated, an SVM classifier is built and tested on the feature subset under investigation. Then, a new population is generated by applying genetic operations (selection, crossover and mutation) and the fitness is again evaluated until a pre-specified number of generations G is reached. This evolution process results in a best individual that further refine by initializing from it a new population that is used as a starting point of a new evolution process. The refinement is iterated until a pre-specified stopping criterion is met. When the entire round of search is completed, the final feature subset is returned. The basic components of our GA are as follows.

C. Representation of Individuals

Generally, a Genetic Algorithm represents the individual as a string or a binary array. Considering the large number of iris pattern and it is represent as a binary vector, this results in a verylong chromosome. Since the pre-processing step reduces the dimensionality of initial iris set, to limit the maximum size of each individual, that is, the length of chromosome, to a predetermined parameter size M ∗T that denotes the maximum cardinality of a feature pool.

The individuals are encoded by n-bit binary vectors. If a bit is "1" it means that the corresponding feature is included in the gene subset, while the bits with value 0 mean the opposite.

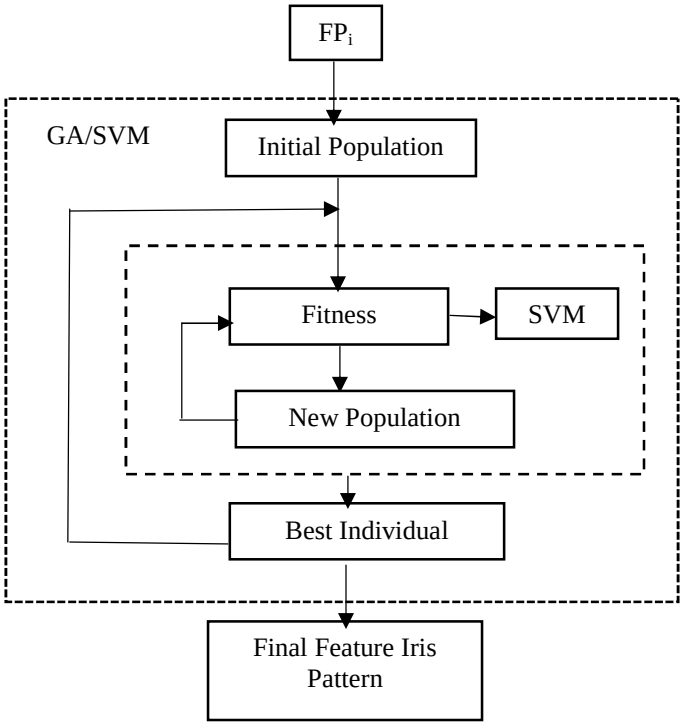


Figure 1 Proposed GA/SVM

Fitness Function

The fitness function is a key factor which affects the performance of GAs. The main aim is to define a function to scale the merit of a feature subset in terms of both classification accuracy and degree of dimensionality. The main idea is to achieve a tradeoff between the accuracy and the size of the obtained feature subsets. As a compromise between these two evaluation criteria, the fitness is defined as follows:

$$F = w.C(x) + \frac{1-w}{S(x)}$$

where w is a parameter between 0 and 1, x is a feature vector representing an individual, C(x) is the classification accuracy of a classifier built on x, and S(x) is the x size, that is, the number of genes included into x. Here, the first term measures the weighted classification accuracy from a classifier and the second one evaluates the weighted size of the feature subset x. The parameter w is a fitness scaling mechanism for assessing the relevance of each term. Increasing the value of w will give more relevance to accuracy and reducing it will set more penalties on the size. This multi-objective fitness makes it possible to obtain diverse solutions of high accuracy, while conventional approaches tend to be converged to a local optimum.

The obtained output shows the automatic recognition of normal and abnormality of iris pattern and it is used for further process. The performance of proposed GA/SVM are evaluated by certain performance metrics to achieves a good classification accuracy.

Experimental Results

The performance results of Proposed Genetic Algorithm - SVM are evaluated using MATLAB. The iris database are derived from The Chinese Academy of Sciences - Institute of Automation (CASIA) eye image dataset of (version 1.0) consist of total 115 images.

The abnormality detection of human beings can be identified from the obtained classifier results based on iris which shows the normal and abnormal datasets. Table 1 shows overall detection rate which is a sufficient result for abnormality detection system.

Table 1Abnormality Detection

	Number of Images	Detection Rate
Normal	40	(35/40) 87.5%
Abnormal	75	(75/75) 100%
Total	115	93.5% (Total Abnormality Detection)

e ISSN 2348 –1269, Print ISSN 2349-5138
Cosmos Impact Factor 4.236.

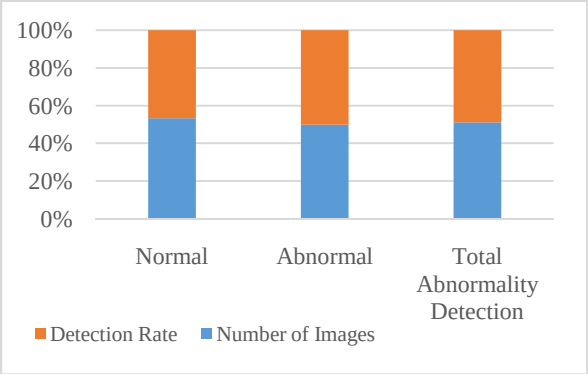


Figure 2 Abnormality Detection

From the Figure 2 it is clear that the abnormality detection rate is high for implemented proposed GA/SVM Classifier. It shows 93.5% for total number of images. The overall sufficient detection rate is sufficient for proposed abnormality detection system.To validate the efficient performance of proposed GA/SVM, it is necessary to compare its performance with other classifiers such as K-Nearest Neighbor , Navies Bayes, Back Propagation Network (BPN) , Probabilistic Neural Network (PNN) ,Support Vector Machine and Fuzzy SVM in terms of accuracy, sensitivity and specificity.

Table 2 shows the performance comparison of Various Classifiers for abnormality Detection.

Table 2 Comparison of Various Classifiers for Abnormality Detection

Classifiers	Accuracy (%)	Sensitivity (%)	Specificity (%)
KNN	73	62	89.5
Navies Bayes	79	68.5	83.2
BPN	82	73.5	81.4
PNN	87.5	76.5	78.6
SVM	94.7	86.1	74.2
Fuzzy SVM	96.4	92.3	72.6
GA-SVM	98.6	95.4	71.6

Table 3 EER Comparison of Various Classifiers

Classifiers	Equal Error Rate(%)
KNN	1.47
Navies Bayes	1.22
BPN	1.10
PNN	1.02
SVM	0.86
Fuzzy SVM	0.77
GA-SVM	0.75

Table 3 shows the Equal Error Rate for different classifiers. It is the measurement parameter to monitoring the abnormality detection. In general, the smaller the EER is, the more accurate the proposed method. It shows that the proposed GA-SVM has less EER when compared to various classifiers.It shows that Proposed GA-SVM has less EER when compared to various classifiers.The Abnormal detection for 75 images based proposed GA-SVM shows error rate of 0.75. The performance results classify the iris dataset as normal and abnormal and further it is used to find the abnormality detection in human health condition. The experimental results prove that the proposed GA-SVM outperforms well in classification accuracy when compared to other classifiers.

Conclusion

The iris has a fine texture and it remains stable as protected internal organ. So it used for detection of human abnormalities. A new approach for human abnormality detection based iris are described. The Genetic Algorithm and Support Vector Machine classifier are employed to identify the abnormality detection in human beings. In this paper iris image is acquired, stored into database, preprocessed, normalized, enhanced that iris pattern and extracted some

features. These features are forwarded to proposeGA-SVM for abnormality detection. The performance results shows that the proposed GA-SVM outperforms well when compared to other techniques. The obtained results are used in various applications to identify the human abnormalities.

References

1. Galbally, J., Ross, A., Gomez-Barrero, M., Fierrez, J. and Ortega-Garcia, J., 2013. Iris image reconstruction from binary templates: An efficient probabilistic approach based on genetic algorithms. Computer Vision and Image Understanding, 117(10), pp.1512-1525.
2. Karthikeyan, R. and Alli, P., 2012. Retinal image analysis for abnormality detection-an overview. Journal of Computer Science, 8(3), p.436.
3. Chen, J., Shen, F., Chen, D.Z. and Flynn, P.J., 2016. Iris recognition based on human-interpretable features. IEEE Transactions on Information Forensics and Security, 11(7), pp.1476-1485.
4. Chen, W.S., Chih, K.H., Shih, S.W. and Hsieh, C.M., 2005. Personal Identification with Human Iris Recognition based on Wavelet Transform. In MVA (pp. 351-354).
5. Dessì, N. and Pes, B., 2009. An evolutionary method for combining different feature selection criteria in microarray data classification. Journal of Artificial Evolution and applications, 2009, p.3.
6. Ibrahim, A.A., Khalaf, T.A. and Ahmed, B.M., 2016. Design and Implementation of Iris Pattern Recognition Using Wireless Network System. Journal of Computer and Communications, 4(07), p.15.
7. Mansouri, A., Affendey, L. and Mamat, A., 2008. Named entity recognition using a new fuzzy support vector machine. IJCSNS, 8(2), p.320.
8. Masek L., "Recognition of human iris patterns for biometric identification", 2003.
9. Menyhárt, J. and Szabolcsi, R., 2016. Support Vector Machine and Fuzzy Logic. Acta PolytechnicaHungarica, 13(5), pp.205-220.
10. Mhaske, M.M., Manza, R.R. and Rajput, Y.M., Detection of Retinal Images as Normal or Abnormal using Texture Feature Analysis to identify the Diabetic Retinopathy.
11. Sachdeva, G. and Kaur, B., Iris Recognition Using Fuzzy SVM Based on SIFT Feature Extraction Method.
12. Spyrou, E., Stamou, G., Avrithis, Y. and Kollias, S., 2005. Fuzzy support vector machines for image classification fusing MPEG-7 visual descriptors.
13. Yu, X., Hsu, W., Lee, W.S. and Lozano-Pérez, T., 2004. Abnormality detection in retinal images.
14. Andrews, S., Hofmann, T. and Tsochantaridis, I., 2002, July. Multiple instance learning with generalized support vector machines. In AAAI/IAAI (pp. 943-944).
15. Galvan, J.R., Elices, A., Munoz, A., Czernichow, T. and Sanz-Bobi, M.A., 1998, November. System for detection of abnormalities and fraud in customer consumption. In Proc. of the 12th Conference on the Electric Power Supply Industry.
16. Jha, R.S., Kumar, S., Kumar, I. and Borah, S., Brain Abnormality Detection from MR Images using Matrix Symmetry Method.
17. Mankar, B.S. and Rout, N., Automatic Detection of Diabetic Retinopathy using Morphological Operation and Machine Learning.
18. Vallabha, D., Dorairaj, R., Namuduri, K. and Thompson, H., 2004, November. Automated detection and classification of vascular abnormalities in diabetic retinopathy. In Signals, Systems and Computers, 2004. Conference Record of the Thirty-Eighth Asilomar Conference on (Vol. 2, pp. 1625-1629). IEEE.
19. M Pushparani, D Sasikala, A Survey of Gait Recognition Approaches Using PCA and ICA, Global Journal of Computer Science and Technology.
20. M.Pushpa Rani, Abnormal GAIT classification using hybrid ELM, Electrical and Computer Engineering (CCECE), 2014 IEEE 27th Canadian.
21. MP Rani, G Arumugam, Children abnormal Gait classification using extreme learning machine, Global journal of computer science and technology.