Sukkur IBA Journal of Computing and Mathematical Sciences

E-ISSN: 2522-3003

P-ISSN: 2520-0755

Volume: 1 | No: 2 | July - December | 2017

Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS) is the bi-annual research journal published by **Sukkur IBA University**, Pakistan. **SJCMS** is dedicated to serve as a key resource to provide practical information for the researcher associated with computing and mathematical sciences at global scale.

Disclaimer: The opinions expressed in **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)** are those of the authors and contributors, and do not necessarily reflect those of the journal management, advisory board, the editorial board, **Sukkur IBA University** press or the organization to which the authors are affiliated. Papers published in **SJCMS** are processed through double blind peer-review by subject specialists and language experts. Neither the **Sukkur IBA University** nor the editors of **SJCMS** can be held responsible for errors or any consequences arising from the use of information contained in this journal, instead errors should be reported directly to corresponding authors of articles.

Publisher: **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)**
Office of Research, Innovation & Commercialization – ORIC
Sukkur IBA University - Airport Road Sukkur-65200, Sindh Pakistan
Tel: (09271) 5644233 Fax: (092 71) 5804425 Email: sjcms@iba-suk.edu.pk URL: sjcms.iba-suk.edu.pk

Mission Statement

The mission of **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)** is to provide a premier interdisciplinary platform to researchers, scientists and practitioners from the field of computing and mathematical sciences for dissemination of their finding and to contribute in the knowledge domain.

Aims & Objectives

Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS) aims to publish cutting edge research in the field of computing and mathematical sciences.

The objectives of **SJCMS** are:

1. To provide a platform for researchers for dissemination of new knowledge.
2. To connect researchers at global scale.
3. To fill the gap between academician and industrial research community.

Research Themes

The research focused on but not limited to following core thematic areas:

Computing:

- Software Engineering
- Formal Methods
- Human Computer Interaction
- Information Privacy and Security
- Computer Networks
- High Speed Networks
- Data Communication
- Mobile Computing
- Wireless Multimedia Systems
- Social Networks
- Data Science
- Big data Analysis
- Contextual Social Network Analysis and Mining
- Crowdsource Management
- Ubiquitous Computing
- Distributed Computing
- Cloud Computing
- Intelligent devices
- Security, Privacy and Trust in Computing and Communication
- Wearable Computing Technologies
- Soft Computing
- Genetic Algorithms
- Robotics
- Evolutionary Computing
- Machine Learning

Mathematics:

- Applied Mathematical Analysis
- Mathematical Finance
- Applied Algebra
- Stochastic Processes

Publisher: **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)**
Office of Research, Innovation & Commercialization – ORIC
Sukkur IBA University - Airport Road Sukkur-65200, Sindh Pakistan
Tel: (09271) 5644233 Fax: (092 71) 5804425 Email: sjcms@iba-suk.edu.pk URL: sjcms.iba-suk.edu.pk

Patron's Message

Sukkur IBA University has been imparting education with its core values merit, quality, and excellence since its inception. Sukkur IBA University has achieved numerous milestones in a very short span of time that hardly any other institution has achieved in the history of Pakistan. The distinct service of Sukkur IBA University is to serve the rural areas of Sindh and also underprivileged areas of other provinces of Pakistan. Sukkur IBA University is committed to serve targeted youth of Pakistan who is suffering from poverty and deprived of equal opportunity to seek quality education. Sukkur IBA University is successfully undertaking its mission and objectives that will lead Pakistan towards socio-economic prosperity.

In continuation of endeavors to touch new horizons in the field of computing and mathematical sciences, Sukkur IBA University publishes international referred journals. Sukkur IBA University believes that research is an integral part of modern learnings and development. **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)** is the modest effort to contribute and promote the research environment within the institution and Pakistan as a whole. SJCMS is a peer-reviewed and multidisciplinary research journal to publish findings and results of the latest and innovative research in the fields, but not limited to Computing and Mathematical Sciences. Following the tradition of Sukkur IBA University, SJCMS is also aimed at achieving international recognition and high impact research publication in the near future.

Prof. Nisar Ahmed Siddiqui
Sitara-e-Imtiaz
Vice Chancellor
Sukkur IBA University
Patron **SJCMS**

Publisher: **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)**
Office of Research, Innovation & Commercialization – ORIC
Sukkur IBA University - Airport Road Sukkur-65200, Sindh Pakistan
Tel: (09271) 5644233 Fax: (092 71) 5804425 Email: sjcms@iba-suk.edu.pk URL: sjcms.iba-suk.edu.pk

Editorial

Dear Readers,

It is a great pleasure to present you the second issue of **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)**. As the mission statement of Sukkur IBA University clearly demonstrates the importance of research, SJCMS is a remarkable effort of Sukkur IBA University towards its commitment for creating a research-based community. The SJCMS provides an interdisciplinary platform to researchers, scientists, practitioners and academicians for publishing their contributions to the recent technological advances and innovations in the area of Computing and Mathematics for dissemination to the largest stakeholders.

This issue contains the double-blind peer-reviewed articles that address the key research problems in the specified domain as the aim of SJCMS is to publish original research findings in the field of Computing and Mathematics. The SJCMS adopts all standards that are a prerequisite for publishing high-quality research work. The Editorial, Advisory and Reviewers Board of the Journal is comprised of academic and industrial researchers from technologically/academically advanced countries. The Journal has adopted the Open Access Policy without charging any publication fees that will certainly increase the readership by providing free access to a wider audience.

On behalf of the SJCMS, I welcome your submissions for upcoming issue (Volume-2, Issue-1, January-June 2018) and looking forward to receiving your valuable feedback.

Sincerely,

Ahmad Waqas, PhD
Chief Editor
SJCMS

Publisher: **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)**
Office of Research, Innovation & Commercialization – ORIC
Sukkur IBA University - Airport Road Sukkur-65200, Sindh Pakistan
Tel: (09271) 5644233 Fax: (092 71) 5804425 Email: sjcms@iba-suk.edu.pk URL: sjcms.iba-suk.edu.pk

Publisher: **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)**
Office of Research, Innovation & Commercialization – ORIC
Sukkur IBA University - Airport Road Sukkur-65200, Sindh Pakistan
Tel: (09271) 5644233 Fax: (092 71) 5804425 Email: sjcms@iba-suk.edu.pk URL: sjcms.iba-suk.edu.pk

Contents

Publisher: **Sukkur IBA Journal of Computing and Mathematical Sciences (SJCMS)**
Office of Research, Innovation & Commercialization – ORIC
Sukkur IBA University - Airport Road Sukkur-65200, Sindh Pakistan
Tel: (09271) 5644233 Fax: (092 71) 5804425 Email: sjcms@iba-suk.edu.pk URL: sjcms.iba-suk.edu.pk

Volume 1 | No. 2 | July – December 2017

A Remotely Deployable Wind Sonic Anemometer

Muhammad Awais[1], Syed Suleman Abbas Zaidi[1], Murk Marvi[1], Muhammad Khurram[1]

Abstract:

Communication and computing shape up base for explosion of Internet of Things (IoT) era. Humans can efficiently control the devices around their environment as per requirements because of IoT, the communication between different devices brings more flexibility in surrounding. Useful data is also gathered from some of these devices to create Big Data; where further analysis assists in making life easier by developing good business models corresponding to user needs, enhancing scientific research, formulating weather prediction or monitoring systems and contributing in other relative fields as well. Thus, in this research a remotely deployable IoT enabled Wind Sonic Anemometer has been designed and deployed to calculate average wind speed, direction, and gust. The proposed design is remotely deployable, user-friendly, power efficient and cost-effective because of opted modules i.e., ultrasonic sensors, GSM module, and solar panel. The testbed was also deployed at the roof of Computer & Information Systems Engineering (CIS) department, NED UET. Further, its calibration has been carried out by using long short-term memory (LSTM), a deep learning technique; where ground truth data has been gathered from mechanical wind speed sensor (NRG-40 H) deployed at top of Industrial & Manufacturing (IM) department of NED UET. The obtained results after error minimization using deep learning are satisfactory and the performance of designed sensor is also good under various weather conditions.

Keywords: *IoT, Big Data, Anemometer, LSTM.*

1. Introduction

Anemometer is a device that is used to measure wind speed and direction. It is a core component, of weather station, for monitoring and analysis of air quality. It is used for wind turbine control, air flow measurement and observation in tunnel, airports, and chemical plants. The sensed data by this sensor can also be used to predict storms to generate alert for governments, help engineers and climatologists. Aerodynamic effects on cars and ballistic missiles are also important set of parameters for engineers to develop efficient design. Apart from this, wind speed and direction are few of the key factors behind growth of crops. Therefore, with the help of real time data gathered from anemometers, an expert system for agriculture sector can be trained which makes decision in real time. Hence, automation in agriculture sector can be introduced which ultimately leads to efficient utilization of precious resources i.e., water and improvement in overall food quality and quantity.

In literature, different types of anemometers i.e. cup anemometer, vane anemometer, ultrasonic anemometer, laser Doppler and hot-wire are available [1] - [6]. Vane and cup anemometer are mechanical type and mostly used on commercial scale because they are known for ages [1]. However, due to dependency of vane anemometer on mechanical gears their life time is less. They need constant maintenance because dust particles in air jam the gears and affect its accuracy, especially during wind gusts and lulls. That is why ultrasonic anemometers are preferred, to calculate wind data and capture

[1] Computer Information and Systems Engineering, NED University of Engineering and Technology, Karachi, Pakistan
Corresponding author: m.awais1231919@gmail.com

gusts and lulls properly, because there is not any involvement of physical moving parts.

Word ultrasonic basically refers to speed of sound, Thus it is clear that ultrasonic anemometer calculates wind speed and direction with the help of sound waves generated through transducers. Various designs and mathematical formulation are available in literature for design of ultrasonic anemometers each having different set of pros and cons. In [2] an ultrasonic anemometer with three piezoelectric transducers has been used to form a rectangular tetrahedron, along with phase meter coupled with receiver to calculate wind velocity. However, in extreme weather condition this design does not give satisfactory results. The proposed solution in [2] also causes air turbulence in the wind, thus reducing further accuracy. However, a modified design proposed in [3], with three transducers forming a horizontal equilateral triangle minimizes this turbulence of wind through its frame. Yet, in [3] the effect of temperature and humidity on ultrasonic waves was not eliminated completely. Therefore, in [4] ultrasonic anemometers with four transducers were used to eliminate the effect of temperature on ultrasonic waves. Transducers were also operated in low frequency and average of wind velocity was also taken to attain good accuracy especially in frost. This four-transducer orthogonal design, proposed in [4], is recognized as 2D ultrasonic anemometer. A 3D ultrasonic anemometer in [5] provides less aerodynamic turbulence by having complex design and mathematical equations. However, the design and implementation of ultrasonic anemometer mentioned is still very expensive. Recently, a new type of solution has entered anemometry world called Laser Doppler Anemometer [6] in which wind velocity components are calculated through the Doppler shifts, wind produces in laser rays but this anemometer is still not commercially available. In [14] an innovative ultrasonic anemometer has been designed for marine applications and calibrated using wind tunnel, however, authors have used deep learning for calibration. In [15] another low cost ultrasonic anemometer is designed using FPGA.

In terms of accuracy ultrasonic anemometers are far better than vane anemometers because no physical moving part is present. However, cost of ultrasonic anemometer varies between US $1500 and US $2500 which makes them difficult to function on commercial scale. That is why, a cost-effective solution has been proposed [7], as an initial study of the research proposed in this paper. In [7], three different design models of ultrasonic anemometer have been presented and final solution has been tested extensively by deploying testbed at Computer and Information Systems Engineering Department of NED UET for over week duration. Data collected from it has been compared with NRG-40H cup anemometer deployed at Industrial and Manufacturing Engineering Department of NED UET. It has been justified, through comparisons, that results obtained by designed sensors are satisfactory. Furthermore, calibration of designed ultrasonic anemometer has also been carried out using simple machine learning techniques. Although, the solution proposed in [7] was based on Internet of Things (IoT) technology and providing direct link to cloud through Wi-Fi communication link. However, it was powered through direct power source which was a major drawback since at remote locations IoT enabled devices need to be power efficient and self-rechargeable as well. Apart from that, at remote locations, especially in Pakistan Wi-Fi communication is hardly available. Thus, from commercialization point of view, in this research the authors have proposed a modified solution of one presented in [7] by integrating additional features i.e., solar panel, battery, general packet radio service (GPRS) module which makes it possible to be deployed at any remote location for long period of time. Apart from that, final design has been properly cased with the help of fiber base frame for maintaining better line of sight (LoS) communication between ultrasonic sensors

and their fine operation in harsh or rainy weather conditions. Enhanced machine learning techniques have also been exploited for calibration of final design. The cost of this enhanced design for anemometer is around US $60 which is extremely low as compared to existing solutions [1] - [6]. It is unique in its kind with eight pillars supporting four transmitters and receivers of HC-SR04 sensor and makes it one off the cheapest ultrasonic anemometer.

2. Sensor Design

Three different designs models of ultrasonic anemometer and analysis of their accuracy with respect to mathematical equations have been presented in [7]. Basic principle is to consider the effect of wind velocity on the speed of sound waves generated by ultrasonic sensors in different directions. A wind velocity component in the direction of the propagation of sound supports the speed, thus leading to an increase in it but decrease in duration and vice versa. Thus, different wind velocities provide different propagation times at fixed distance between sensors. This change in duration of speed of sound is inversely proportional to the speed of wind. Analysis and pros and cons of each design have been discussed in [7] with details. Equations for calculating wind speed and direction are as follows.

$$V_x^{wind} = \frac{d_{TR}}{2} * \left(\frac{1}{t_1} - \frac{1}{t_2}\right) \quad (1)$$

$$V_y^{wind} = \frac{d_{TR}}{2} * \left(\frac{1}{t_3} - \frac{1}{t_4}\right) \quad (2)$$

$$V^{wind} = \sqrt{{V_x^{wind}}^2 + {V_x^{wind}}^2} \quad (3)$$

$$\theta = \left(\tan^{-1} \frac{V_y^{wind}}{V_x^{wind}}\right) \quad (4)$$

Equation (1) and (2) are used for measuring horizontal and vertical component of wind velocity. d_{TR} is distance between transmitter and receiver, t represents time-of-flight (tof) calculated by respective HC-SR04 sensors and V^{wind} in (3) is speed of wind.

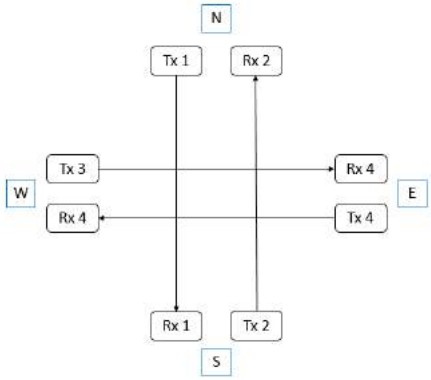

Figure 1: Design of ultrasonic anemometer with four modified HC-SR04

In [7], four sensors have been placed in front of each other to calculate horizontal and vertical component of wind. Transmitter of modified HC-SR04 sensor is kept in perfect line-of-sight with its receiver as shown in Fig. 1, 2, and 4 where Tx-1 and Rx-1 represents transmitter and receiver of sensor 1 respectively, same notation has been followed for other sensors as well. As speed of sound depends on many factors such as temperature and humidity which could reduce the accuracy of ultrasonic anemometer but this dependency has been removed in [7] by using (1) and (2). Following that conversion of components from rectangular to polar coordinate i.e., by using (3) and (4) wind speed and its direction has been obtained. If wind blows from East to West, please refer to Fig. 1, then duration $t_3 > t_4$. Hence, vertical component will have some negative value, however, t_1 and t_2 will remain same due to which horizontal component approaches to zero. Thus, 270° represents that wind is blowing in East. Further in-depth analysis has been presented in Table 1 of [7].

2.1. Problem Statement

Initial version, in [7], of ultrasonic anemometer had following problems. 1) It cannot withstand harsh weather condition and rain can easily damage its electronic circuitry due to simpler and wooden type of casing. 2) Adapter cable is required to provide power through direct source which makes it difficult

to deploy at remote sites. 3) ESP module requires Wi-Fi router and internet service provider (ISP) to upload data on cloud. Thus, causing an overhead of arranging router and keeping ultrasonic anemometer within range of router signals. In addition to it, no such internet facility is available at remote locations. 4) The software and hardware is not optimized in terms of power efficiency. Thus, in this research the authors have tried to overcome the mentioned problems by proposing additional features into existing ultrasonic anemometer given in [7].

2.2. Components Used

Main electronic components used for complete design of IoT enabled ultrasonic anemometer are; AVR Atmel 328 microcontroller, ESP8266 module, SIM 900, LM2596 module, solar panel and cheapest sonar sensor HC-SR04. Interfacing of respective components with ATMEGA328 microcontroller has been done to collect and upload wind data on cloud. Ultrasonic sensor, microcontroller and SIM-900 functions at *5V* however for SIM-900 a higher current rating close to *1A* is required. HC-SR04 sensor has accuracy of ±29.1µs which makes it very difficult to design ultrasonic anemometer because a few µs of fluctuation in Time-of-flight (tof) would result in incorrect measurements for wind velocity, hence, calibration of ultrasonic sensors has been carried out through proper averaging rules. Fine-tuning of error has also been carried out to achieve accuracy of ±1µs.

2.3. Prosposed Solution

2.3.1. Frame design with minimum air turbulence

The casing of sensor has been enhanced by laser cutting acrylic sheets in place of wooden pillars used in [7]. This new acrylic based structure has been designed in a way to keep electronic circuitry safe in a box, please refer to Fig. 2 and 4. Thus, the enhanced casing of sensor makes it possible to survive rain and withstand harsh weather conditions while keeping transmitter and receiver in open atmosphere without causing air turbulence.

2.3.2. Improving communication

The communication through Wi-Fi module has been replaced with SIM-900 module. SIM-900 module has its own set of AT commands to configure its operation. Any mobile network operator, with internet package can be used with SIM-900 to upload data on cloud. Since cellular networks are already providing coverage at large extent and to most of the locations including urban and rural. In future, more feature rich communication is expected to be provided by cellular networks in the form of 5G. Therefore, the issues due to Wi-Fi unavailability at remote locations has been resolved through use of cellular technology.

Figure 2: Frame and circuit of advance version

2.3.3. Solar Circuitry

A separate PCB for solar charging circuit has been designed as shown in Fig. 2. It charges lithium-ion battery which provides power to whole circuitry. Charging circuit has been designed using two 3904-NPN transistors, TIP-127 Darlington bipolar power transistor and a Zener diode. Zener voltage of D1, as shown in Fig. 3, should be equal to voltage of battery to enable charging. Simple logic behind working of circuit is that, when voltage of battery is less than Zener voltage then low logic is passed to base of Q1 through anode of Zener which turns on Q1. Now Q2 gets OFF because of high logic passed by collector of Q1 to its base thus, Q2 sends low logic to base of TIP-127 transistor. TIP-127,

being a PNP transistor, is capable of handling inputs from two transistors, gets ON and shorts the positive pins of solar panel and battery connected to its emitter and collector respectively. Charging is continued unless the voltage of battery gets more than Zener voltage which will pass high logic to base of Q1. Thus, making Q1 OFF, Q2 ON and TIP-127 OFF to stop charging.

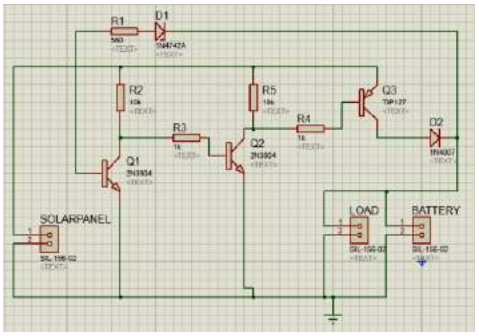

Figure 3: Charging circuit for Solar Panel and lithium-ion battery

Figure 4: Testbed deployment of advance version of ultrasonic anemometer

Figure 5: Real time wind data uploaded on cloud

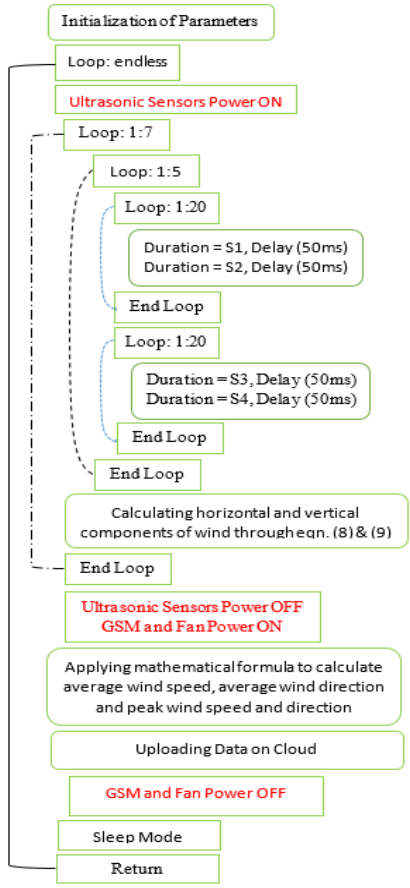

Figure 6: Software flow of advance vision of ultrasonic anemometer

2.3.4. Improving power efficiency

Power to all the ultrasonic sensors and cooling fan has been provided through transistor based switches that are controlled by microcontroller. High rating current and power to SIM-900 is provided through LM2596 module; its enable pin is connected with microcontroller to turn on/off power for SIM-900 module. Software flow shown in Fig. 6 reveals that power in all components has been utilized efficiently. Since the components interfaced with micro-controller draw power only when they are performing some operation, otherwise they are kept off. Furthermore, sleeping modes in software have been enabled which consumes least amount of

power between successive readings of designed anemometer sensor.

3. Error minimization using machine learning

In [7], regression model of machine learning has been used to calibrate ultrasonic anemometer due to which certain bias was observed in data. Data from NRG-40H was used as ground truth, and we were able to minimize error to some extent on pre-processed data of ultrasonic anemometer [7].

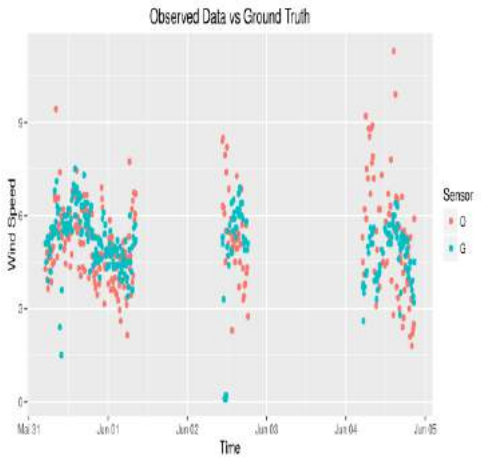

Figure 7: Observed data from ultrasonic anemometer vs ground truth data from NRG 40H

3.1. Data

ESMAP Tier2 meteorological station deployed at NED UET has NRG-40H sensor for measuring wind velocity, shown as blue dots and red dots represent data from ultrasonic anemometer in Fig. 7. Time stamp of data was 20 minutes and around thousand readings were used for training models. Before any network the data contained a mean squared error of 3.285. Fig. 7 shows that ultrasonic anemometer is capturing the general trend of wind velocity; however there is certain nonlinear bias present in the data. To account for this bias term, authors tried following models with an objective to reduce the mean squared error.

3.2. Linear Regression models

Four linear regression models were tested using R language Table. 1. However, any significant improvement in mean squared error was not achieved.

TABLE I. Linear regression models with their features and mean squared error values.

Model	Features	Mean Squared Error
1 Linear	Observed wind speed	3.0484
2 Linear	Observed wind speed, temperature	3.0864
3 Polynomial	Observed wind speed, (Observed wind speed)^2	3.3233
4 Polynomial	Observed wind speed, (Observed wind speed)^2, temperature, (temperature)^2	3.2852

4. Error minimization using deep learning

On closer observation of machine learning approaches, we found that, the error did not only depend on observed value at that particular time stamp but also on the sequence of readings observed in the past. Thus, a model capable of somehow incorporating this sequence dependence is likely to produce better results. Hence, state-of-the-art deep learning techniques called Long Short Term Memory (LSTM) was used to minimize error of ultrasonic anemometer. To account for this memory element, we replaced the regression based learning models with a (LSTM) network [8], as shown in Fig. 9. LSTM network is a recurrent neural network (RNN) which uses LSTM cells instead of simple perceptron.

4.1. LSTM cell

The key difference between LSTM and other RNN networks is its cell. An LSTM cell

consists of a state and three gates as shown in Fig. 8; forget gate, input gate and output gate. These gates are sigmoid layers which decide what information to pass through the cell state by returning a value in between 0 to 1 and its product with the cell state. The forget gate decides what information to throw away from the previous state. The input gate separates out useful information from the current input and the output states, decides what to produce as the output of the network. It is because of these gates that an LSTM is able to remember information obtained in a distant past but converge quickly avoiding the vanishing gradient problem [9].

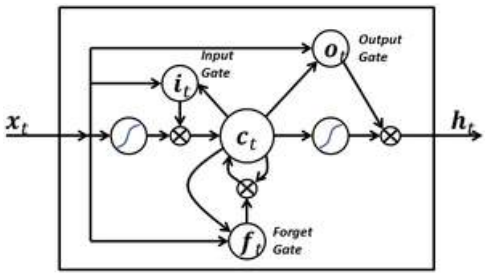

Figure 8: A typical LSTM cell

4.2. LSTM for regression

LSTM networks are most commonly used for sequence prediction problems. Due to this, they have found their way in areas like Speech Recognition [11], Optical Character Recognition [10] and etc. The output layer of an LSTM network is a softmax layer containing k labels, where k denotes number of outputs. However, for the problem under consideration in this research, outputs are not limited to a set of labels that the network can predict. To solve this problem, we used two approaches. In the first approach, we modified training set in order to train a typical LSTM model on it. Whereas, in the second one we modified the LSTM network for required results. In both approaches a network with one hidden layer having hundred nodes was used. However, different libraries for implementing LSTM were used in each approach.

4.2.1. Approach 1

The data set was rounded off to one decimal place. Since wind speed typically does not exceed 36 miles per hour (mph), the readings were divided in between 0 to 36, in the form of 360 labels. After this, with the help of simple pre-processed data from deployed ultrasonic anemometer, the LSTM was trained. This method solved the problem of error using input as a sequence of labels and predicted the output as another sequence of labels. Fig. 9; shows the model of the network for this approach. OCRopus, which is a free document analysis and optical character recognition (OCR) tool and uses bi-directional LSTM for OCR was used for implementing the model.

There are, however, several limitations in this approach. One limitation is in the number of outputs such a network can predict. Since the output layer has only 360 elements, it can only produce 360 different values as outputs. In case of a cyclone or an extremely powerful gust, this method will fail to produce any output. Also, since there are too many classes there is a high probability that many classes will not appear too much and hence will be easy to classify whereas, the ones which appear regularly may become confusing [12]. Thus, instead of modifying our training parameters, we made a slight change in the LSTM network which is discussed under. Different input sequences were used to train the LSTM network Fig. 11; and their root mean square values were compared Table. 2. Even with all the limitations mentioned, this network produced great results on input data of sequence length 8, please refer to Table. 2.

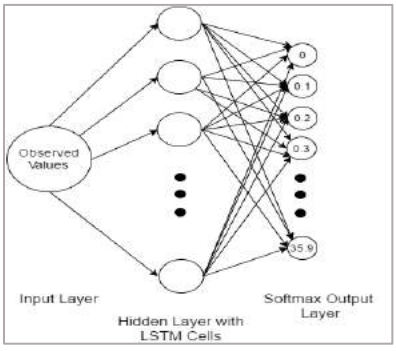

Figure 9: LSTM Network used in approach 1.

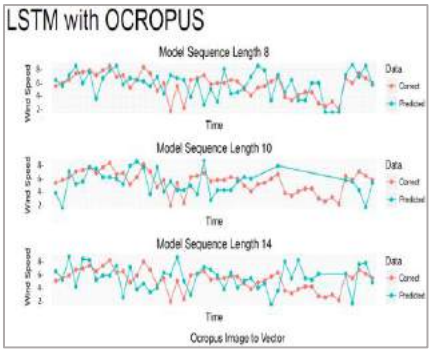

Figure 10: Visualization of results using approach 1 performed with ggplot2 R

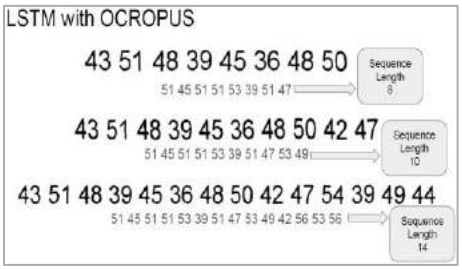

Figure 11: Different input sequences used in approach 1

TABLE II. Results of approach 1.

Model	Mean Squared Error
Sequence length 8	1.776
Sequence length 10	5.597
Sequence length 14	2.763

The network Fig. 9; will take a sequence of labels as input and give a vector of 360 values as output with probability for next label in a sequence. The label with highest probability will be finally chosen as output. Fig. 10; shows that model is fitting curve more properly for model sequence of length 8, and hence has the least mean squared error of 1.776.

4.2.2. Approach 2

In this approach, authors trained LSTM to perform regression instead of classification i.e. Neural Network as a function approximator. Instead of using a 360 unit softmax layer with sigmoid units as outputs, we used a single cell with a Rectifier Linear Unit (ReLu) activation. Advantage of ReLu function is that, unlike a sigmoid unit it can be used to model positive real numbers. It also helps speeding up the process in backpropagation and reduced the computation time of a single classification as well [13]. Rectifier activation function is defined as:

$$f(x) = \max(0, x) \qquad (5)$$

Using this approach, we were able to reduce the mean squared error to 0.572 on the test set, please see results given in Fig. 12. Keras python library was used to implement LSTM network with theano backend.

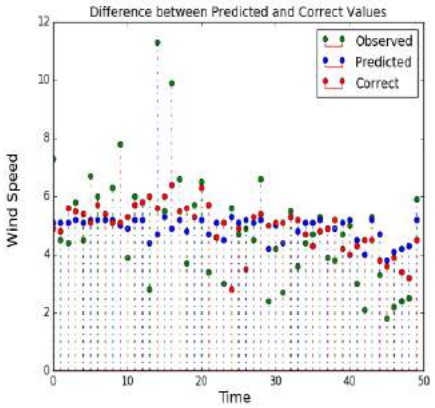

Figure 12: Visualization of results from Approach 2 matplotlib python

5. Conclusion

Observing the significance of weather data in scientific research, artificial Intelligence (AI) and other related fields a cost effective and commercially deployable ultrasonic anemometer has been designed, and calibrated in this research. Real time readings have also been obtained through cloud to make it a complete IoT enabled device. The testbed of anemometer was deployed at roof of CIS dept. NED UET and rich amount of data has been collected. Its calibration has also been carried out using LSTM, a deep learning technique. Addition of solar panel, battery, and GSM provides smooth operation and connectivity at remote locations. Through GSM, loss rate of

data during uploading on cloud has also been reduced. Future directions include; 1) to enable two-way communication through GSM to perform actuation in real time when required. 2) Integrating rain, dust, altitude, pressure, temperature and humidity sensors along with designed ultrasonic anemometer to make a complete weather station device. 3) Exploiting more powerful machine learning techniques for making weather prediction. 4) Integrating Bluetooth module and observing data on android app without any internet connection.

ACKNOWLEDGMENT

The author would like to thank the chairperson of Mechanical department, Dr. Mubashir Ali Siddique for providing access to readings of mechanical sensor.

REFERENCES

[1] J. Wyngaard, "Cup, propeller, vane, and sonic anemometer in turbulence research," Annual Review of fluid Mechanics, vol. 13, no. 1, pp. 399-423, 1981.

[2] J. Nicoli, "Ultrasonic anemometer," Nov. 20 1979, uS Patent 4,174,630. [Online]. Available: https://www.google.com/patents/US4,174,630.

[3] S. Ammann, "Ultrasonic anemometer," Sep. 6 1994, uS Patent 5,343,744. [Online]. Available: https://ww.google.com/patents/US5343744.

[4] B. Pincent, P. Journe, and G. Brugnot, "Ultrasonic anemometer," Jan. 2 1990, uS Patent 4,890,488. [Online]. Available: https://www.google.com/patents/US4809 0488.

[5] García-Ramos, F. Javier, et al. "Analysis of the air flow generated by an air-assisted sprayer equipped with two axial fans using a 3D sonic anemometer." Sensors 12.6 (2012): 7598-7613.

[6] Schotanus, P., FTMf Nieuwstadt, and H. A. R. De Bruin. "Temperature measurement with a sonic anemometer and its application to heat and moisture fluxes." Boundary-Layer Meteorology 26.1 (1983): 81-93.

[7] Muhammad Awais, Murk Marvi, S.S.Abbas Zaidi and Muhammad Khurram, "A Cost Effective Solution for IoT Enabled Ultrasonic Anemometer Sensor design", NED-ICONICS 2016.

[8] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[9] The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions Sepp Hochreiter International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 1998 06:02, 107-116

[10] Liwicki, Marcus, et al. "A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks." Proc. 9th Int. Conf. on Document Analysis and Recognition. Vol. 1. 2007.

[11] Graves, Alex, Santiago Fernández, and Jürgen Schmidhuber. "Bidirectional LSTM networks for improved phoneme classification and recognition." International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, 2005.

[12] Gupta, Maya R., Samy Bengio, and Jason Weston. "Training Highly Multi-class Linear Classifiers." (2014).

[13] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[14] B. Allotta, L. Pugi, T. Massai, E. Boni, F. Guidi and M. Montagni, "Design and calibration of an innovative ultrasonic, arduino based anemometer," 2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Milan, 2017, pp. 1-6.

[15] R. Chandran, R. Bhakthavatchalu and P. P. Kumar, "An FPGA based low cost receiver for ultrasonic anemometer,"

2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kumaracoil, 2016, pp. 729-73.

Volume 1 | No. 2 | July – December 2017

Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network

Shamshad Lakho[1], Akhtar Hussain Jalbani[1], Muhammad SaleemVighio[1], Imran Ali Memon[1], Saima Siraj Soomro[1], Qamar-un-Nisa Soomro[2]

Abstract

Medical judgments are tough and challenging as the decisions are often based on deficient and ambiguous information. Moreover, the result of decision process has direct effects on human lives. Act of human decision declines in emergency situations due to complication, time limit and high risks. Therefore, provision of medical diagnosis plays a dynamic role, specifically in preliminary stage when a physician has limited diagnosis experience and identifies the directions to be taken for the treatment process. Computerized Decision Support Systems have brought a revolution in the medical diagnosis. These automatic systems support the diagnosticians in the course of diagnosis. The major role of Decision Support Systems is to support the medical personnel in decision making procedures regarding disease diagnosis and treatment recommendation. The proposed system provides easy support in Hepatitis disease recognition. The system is developed using the Bayesian network model. The physician provides the input to the system in the form of symptoms stated by patient. These signs and symptoms match with the casual relationships present in the knowledge model. The Bayesian network infers conclusion from the knowledge model and calculates the probability of occurrence of Hepatitis B, C and D disorders.

Keywords: Decision support system, Diagnosis, Diagnosticians, Probablistic Model, Knowledge Model

1. Introduction

Diagnosis is the process of recognition the cause of a problem and disease. The process of diagnosis is performed on the basis of information acquired from past and check-up of a patient. Since the advancement of Information Technology, there are significant improvements in the development of computerized medical diagnosis systems. These improvements have led to advances in medical aid. The integration of Information Technology (IT) in health care centers is not only limited in administrative applications like patients' registration, billing, record keeping and payroll, but also it plays an important role in the assistance of the physicians in the diagnosis of different diseases. In this regard, decision/assistance support system has evidenced to be a suitable tool that helps physicians in solving complicated medical issues such as disease diagnosis [1] [2]. Medicine, engineering, business and science etc. are the fields that work on diagnosis [3]. The stipulation of decision making plays a chief role in the field of medical science. A good doctor diagnoses a disease by his practice, knowledge and talent on the basis of symptoms reported by a patient. A recent practice is that patients consult specialists for better diagnosis and treatment. Other general physicians may not have sufficient expertise in controlling some high risk diseases. On the other hand, it is very hard to get an appointment from a specialist; it may take some days, week or months as well. Most likely, the disease may have affected the

[1] Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology Nawabshah, Pakistan
[2] Department of Examination, Quaid-e-Awam University of Engineering, Science & Technology Nawabshah, Pakistan
Corresponding email: Shamshad.lakho@quest.edu.pk

patient most before the patient refers to the specialist for diagnosis [4]. Most of the high-risk disorders could be treated only in the initial stage. Therefore, the computer-based approaches for the diagnosis of the diseases are important. By using the computer-based disease diagnosis systems, the death ratio and the waiting time could be reduced [5].

Recently, many researchers have used different AI techniques to identify the correct disease [6] - [7]. Bayesian Belief Network is also commonly used AI technique in the field of Biomedical Science for the diagnosis of different high-risk diseases [8].

2. Related work

Decision support systems are communicating and computer based programs that assist users in decision making and judgment [9]. They do not substitute humans but enhance their limited capability to solve complex problems [10]. DSSs are mostly used in military, health care and business areas where complex decision making situations will encounter [11]. The diagnosis system uses Bayesian network in identifying the particular category of Hepatitis disease. A Bayesian belief network is also known as acyclic graphical model. It is a probabilistic model that denotes a set of arbitrary variables and their conditional independencies through a directed acyclic graph [12]. A Bayesian Network could denote the probabilistic connections between disorders and symptoms. The Bayesian Network could calculate the likelihoods of the existence of the numerous diseases on the basis of the signs or symptoms stated by the patient. A Bayesian Network is a bidirectional diagram that permits the stream of information from causes to effects and vice versa. A Bayesian Network can manage with partial and ambiguous data. Artificial intelligence is the art of converting human intelligence in machines and deals with enlarging such intellectual machines which could support and assist humans in problem solving and decision making domains. The manipulation of AI in medical applications has been increasing since the last decades. The application area of AI may be in educational systems, diagnostic systems, expert systems and as well as in machine learning systems. Thus there are significant developments in the field of computer-based diagnosis. AI machine learning techniques play a vital role in the development and support of computer based disease diagnosis systems as diagnosis requires reliable accuracy and performance [13]. Velikova et al [5], have proposed a Bayesian network approach for constructing probabilistic disease progress models built on clinical principles and developed a system for remote pregnancy care. Alonzo et al [7], have proposed a diagnosis system using Bayesian Network to diagnose ENT (Ear, Nose, Throat) diseases based on the specified signs and symptoms. The author claimed that the system is capable of diagnosing a patient with ENT disease and also able to detect a person if he/she is not infected of ENT disorder. Mahesh et al [8], worked on Hepatitis B disease and built a system for the Hepatitis B diagnosis which comprises the neural network. The system is able to classify the patient infected with immune category. Ratnapida and druzdzel [14], have developed an expert system named as Marilyn that is capable of diagnosing and identifying different problems (like computer, heart disease, breast cancer and lymphographic) based on previous cases cured. They constructed a diagnostic model that performs diagnosis based on communication between diagnostician and diagnostic system. Yan et al [15], designed a decision support system for the diagnosis of heart disease. The decision support system is based on multilayer perceptron. They claimed on the basis of assessment that the diagnosis accuracy of the system is greater than 90%. A probabilistic model named as Hepar is developed for the hepatic disease diagnosis. The proposed system is founded on expert knowledge and patients data. The authors claimed that the diagnostic accuracy of the proposed model reaches nearly 80% [13]. Hussain et al [16], proposed a three-layered Bayesian Network model for analyzing and diagnosing the mental states of students from their gestures.

3. Proposed system

The Figure 1 shows the design framework of the proposed system that comprises the following components:

1) Graphical User Interface
2) Medical Database
3) Knowledge Model
4) Bayesian Network
5) Ranking
6) Disease Recognition
   - HBV (Hepatitis B Virus)
   - HCV (Hepatitis C Virus)
   - HDV (Hepatitis D Virus)

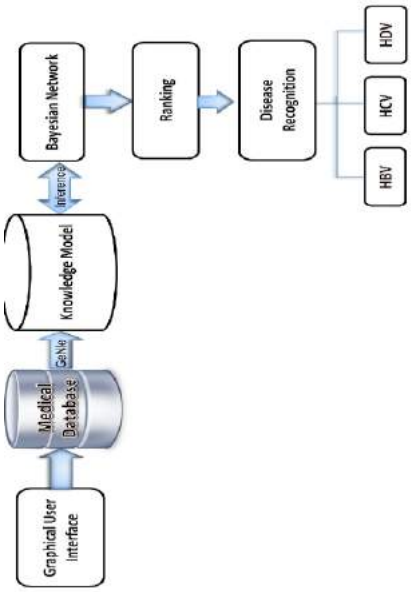

Figure 1: Design Framework

The physician selects the symptoms described by the patient through the graphical user interface. The submitted query is stored in the Medical Database that contains patients' cases and disease information. The query is then compared with the casual relationships present in the knowledge model. The knowledge model is created by learning the network.

The Bayesian network applied inference on knowledge model to calculate the posterior probability distribution for a query provided by the user. The Bayesian network has been used for classification and making decisions to recognize the particular category (i.e. B, C or D) of Hepatitis disease in our case. The categories of hepatitis are classified based on the probability distribution. Finally, the particular category of hepatitis is diagnosed i.e. either the patient is infected of hepatitis B, C or D virus. We use GeNIe/SMILE[3] software package for training and testing the Bayesian network.

3.1. Diagnostic interface

Figure 2 shows the screenshot of the user interface that is used to interact with the Hepatitis Diagnostic system.

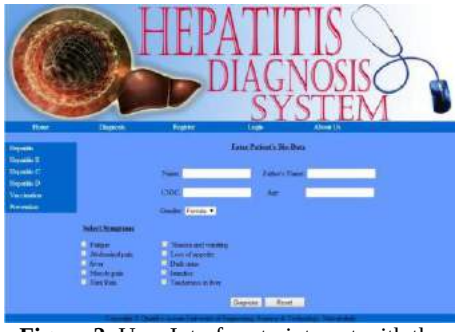

Figure 2: User Interface to interact with the Diagnostic System

Figure 3 shows a snapshot of the backend interface of the system. The right hand side of the window comprises a list of all possible signs and symptoms present in the Bayesian network model. The upper right part of the window holds a list of those possible observations that have not yet been instantiated by the user. Those symptoms that have been observed are taken over to the lowest part of the window.

[3]available at http://genie.sis.pitt.edu

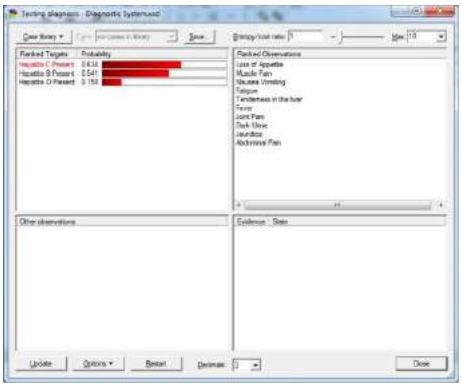

Figure 3: Diagnostic Interface

3.2. Data description

The data was collected time by time from different online and offline resources like mayoclinic.com and Peoples Medical College & Hospital Nawabshah for the training purpose and interviews were also conducted with hepatologists. The collected data contained one year record of Hepatitis patients consisting of 275 cases. It is difficult to collect medical records from hospital due to their privacy however despite of that difficulty, we tried our best to collect 275 records of hepatitis patients from well-known hospital and train our system to drive the results. The system´s decision making capability is based on the physical exam of the patient. The patient´s cases belonged to Hepatitis B, C & D categories. We have used 200 records for training of the Bayesian Network and 75 records for testing of the system.The statistics of trained data set is that among 200 cases, 79 cases are of HBV patients, 95 cases are of HCV patients and 26 cases are of HDV patients as shown in Figure 4.

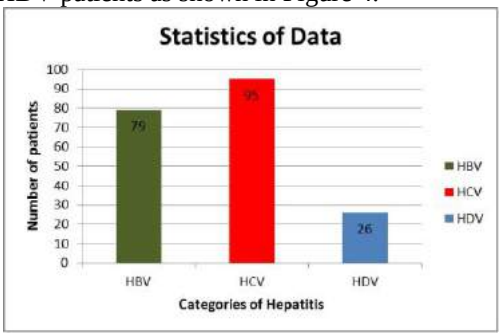

Figure 4: Statistic of trained data set

3.3. Two layered Bayesian network model

Figure 5 shows the Bayesian network that reveals the arrangement of problem domain. The network models 13 variables to diagnose hepatitis disease including 10 symptoms and 3 disorder nodes. The arcs depict the direct probabilistic associations between couple of nodes, so the arc between Jaundice and Hepatitis B node denotes the fact that the existence of Jaundice increases the chance of Hepatitis B disorder. The arrangement of the model is the illustration of the relationships between components of diagnosis process i.e. cause and effect. We have created the structure based on medical fiction and discussions with hepatologists and the numerical factors of the model i.e. prior and conditional probability distributions are mined from a medical database containing cases of Hepatitis patients.

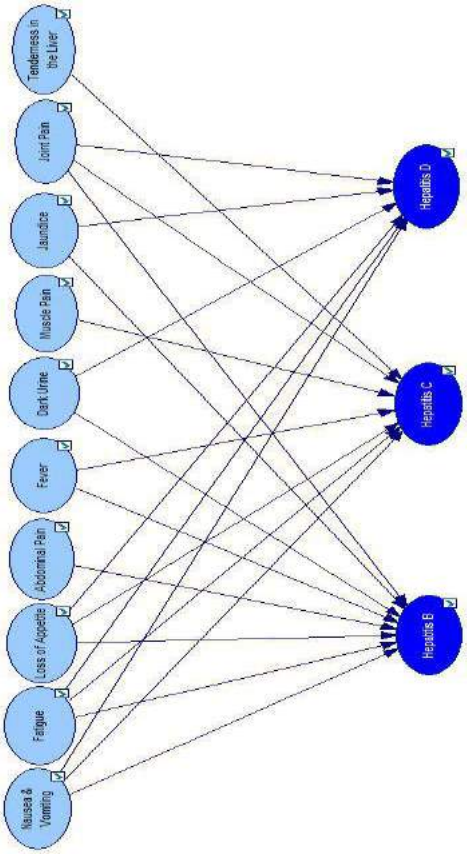

Figure 5: Two Layered Bayesian Network Model

4. Results & Discussion

Calisir and Dogantekin [17], proposed PCA-LSSVM classifier for recognition of hepatitis disorder. The authors claimed that the proposed technique produced valuable results with that the proposed technique produced valuable results with high accuracy rate of 96.12%. Kaya and Uyar [18], proposed RS and ELM based decision support system for identifying the hepatitis disease. The authors claimed that the RS-ELM method is better than other classification methods. Bascil and Temurtas [19], proposed hepatitis disease diagnostic network and LevenbergMarquardt training algorithm and compared the derived results with the other studies focusing on the similar UCI database. The medical diagnosis systems are gaining popularity in disease diagnosis and providing timely medical aid. The diagnostic accuracy of such systems is rather promising and reliable. This research study also shows that the developments of decision-support systems of high-risk diseases using AI techniques proved beneficial in providing inexpensive and timely diagnosis. Chen et al. [20], offered the LFDASVM method for the hepatitis disease diagnosis. The proposed method examined the condition of the liver and distinguished the live liver from the dead one. Other researchers worked on image based features to diagnose the hepatitis disease while our proposed system works on the symptoms stated by patient. However if both image features and physical features are combined together the system will provide good results that help physicians to take appropriate decisions.

4.1. Diagnosis result of trained data set

The Figure 6 shows the diagnosis result of the proposed system performed on trained data set. Selecting the possible values of the symptoms, the associated probabilities of the disorders are updated spontaneously and a new well-organized list of the disorders is presented. In the following result Hepatitis B is diagnosed with the probability value of approximately 95%.

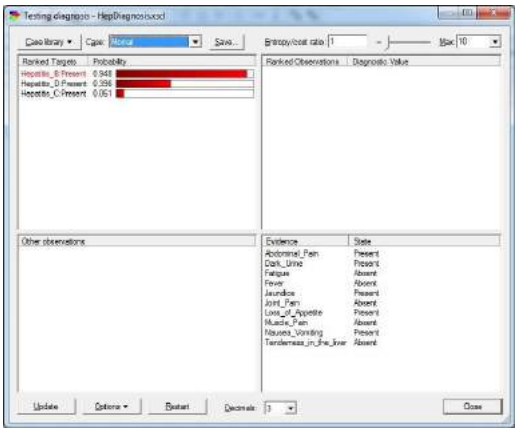

Figure 6: Diagnosis of Hepatitis B Disease

4.2 Diagnosis rate of trained data set

The diagnosis rate of the trained data is calculated with the help of hit and trial method. Twenty five trained cases of each category of Hepatitis are tested on the system respectively. Out of 25 cases of Hepatitis B, system diagnosed only 21 cases correctly, remaining 4 cases are diagnosed incorrectly, 2 cases are identified as HCV and 2 as HDV. Similarly out of 25 cases of HCV, system diagnosed only 23 cases accurately and the remaining 2 cases were diagnosed inaccurately as HBV cases. Likewise, out of 25 cases of HDV, system recognized only 18 cases properly and the remaining 7 cases were identified improperly. Five cases are diagnosed as HBV and 2 cases are diagnosed as HCV.

Table-I. Diagnosis Rate of Trained Data Set

	HBV	**HCV**	**HDV**	**Producer Accuracy**
HBV	21	2	2	84%
HCV	2	23	0	92%
HDV	5	2	18	72%
Overall Accuracy			**82.667%**	

Therefore, on trained data set, the success rate of HBV diagnosis is 84%, the success rate of HCV diagnosis is 92% and the success rate of HDV diagnosis is 72% as shown in the Figure 7.

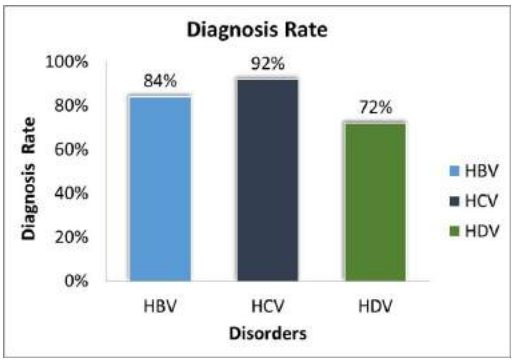

Figure 7. Diagnosis Rate of Trained Data Set.

4.3. Diagnosis result of test data set

Figure 8 shows that in this particular case Hepatitis B is diagnosed with the probability value of 71%.

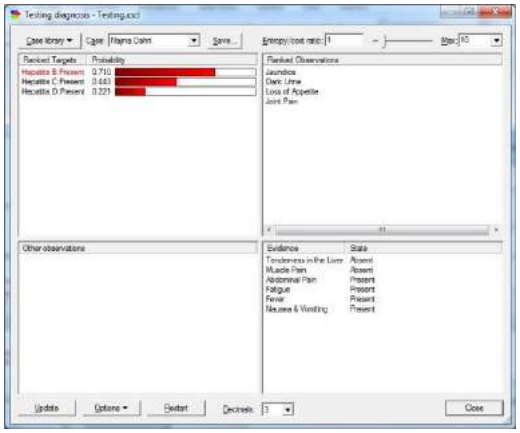

Figure 8. Diagnosis of Hepatitis B Disease.

4.4. Diagnosis rate of test data set

The diagnosis rate of the test data set is also calculated with the help of hit and trial method. Twenty five new cases of each category of Hepatitis are tested on the system respectively. Out of 25 cases of Hepatitis B, system diagnosed only 20 cases correctly, remaining 5 cases are diagnosed incorrectly, 2 cases are identified as HCV and 3 as HDV. Similarly, out of 25 cases of HCV, system diagnosed only 22 cases accurately and the remaining 3 cases were diagnosed inaccurately, 2 cases as HCV cases and 1 case as HDV. Likewise, out of 25 cases of HDV, system recognized only 13 cases properly and the remaining 12 cases were identified improperly. Seven cases are diagnosed as HBV and 5 cases are diagnosed as HCV.

Table II. Diagnosis Rate of Test Data

	HBV	**HCV**	**HDV**	**Producer Accuracy**
HBV	20	2	3	80%
HCV	2	22	1	88%
HDV	7	5	13	52%
Overall Accuracy		**73.33%**		

Therefore, on test data set, the success rate of HBV diagnosis is 80%, the success rate of HCV diagnosis is 88% and the success rate of HDV diagnosis is 52% as shown in the Figure 9.

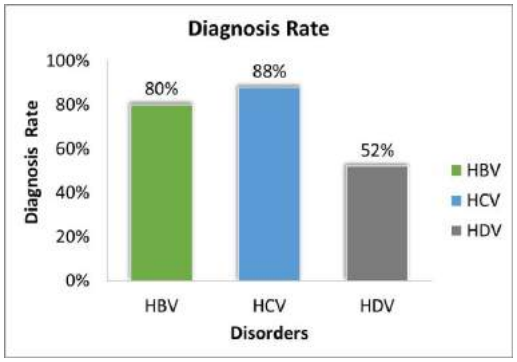

Figure 9. Diagnosis Rate of Test Data Set.

4.5. Accuracy of trained and test data sets

Consequently the overall diagnosis accuracy of trained data is 83% and test data is 73% as shown in the figure 10, is calculated on the system.

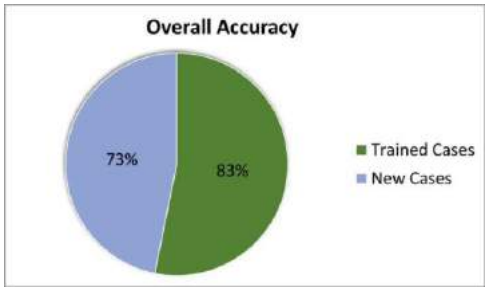

Figure 10. Overall Accuracy of the System.

5. Conclusion & Future work

Medical diagnosis is one of the challenging tasks. Computerized decision support systems are becoming familiar and useful in assisting the physicians in the diagnosis process. The proposed system is the first step towards the diagnosis of hepatitis disorder; it's not the full fledge system with complete functionality. Yet the system is especially designed for the emergency medical care in ruler areas where there is no any facility of laboratory testing and other necessary examinations for the diagnosis of hepatitis due to that death ratio is very high in those areas. The main objective is to facilitate the ruler areas people as a first aid to save their lives in emergency situations. The proposed model is the prototype of the Hepatitis disease recognition system which may be helpful in the future for practical implementation in real situations in different areas. The model is trained by using the information of prior probabilities, extracted from the collected data of different patients. Bayesian Network Model has become an effective method for avoiding the over fitting of data and collected samples can be "smoothed" so that all available data can be used properly without preprocessing of the information. Bayesian Network model provides the rich facility to handle the uncertain situation. Therefore, we proposed to use it for the recognition of Hepatitis disease; however many researchers proposed the different machine learning techniques to classify the Hepatitis, but in our best knowledge there is no any work has been proposed yet to use Bayesian Network for Hepatitis disease recognition. In the future, the system can be enhanced with different modalities for the recognition of the disease. In this research a diagnosis system has been developed using the decision theoretic approaches to diagnose the Hepatitis B, C & D diseases and tested on a data set of new cases. The system works well for calculating the influence of various symptoms on the probability of different categories of Hepatitis and assists in the process of diagnosis and classification of various categories of Hepatitis. The system could be used as a training tool for the practice of trainee physicians as well. The main application of the system is in helping, serving and supporting in the diagnosis of Hepatitis. System is capable of maintaining a depository of patient cases. The diagnosis accuracy of the system is reasonable. The diagnosis accuracy of those disorders that are greater in number in the medical database of cases ranges almost more than 80%. The overall accuracy of the system seems to be much better and satisfactory than that of trainee diagnosticians. The diagnosis support systems would improve the diagnosis power of the physicians.

In the future, system would be enhanced:

- To increase the number of patients and incorporate more laboratory test results and risk factors in the domain model to improve the accuracy of the system.
- To diagnose all the categories of Hepatitis i.e. A to E.
- To diagnose different Hepatic disorders i.e. PBC (Primary Biliary Cirrhosis), Liver Cancer and Alcoholic Liver disease etc.

ACKNOWLEDGEMENT

The authors would like to thank Information Technology department of Quaid-e-awam University of Engineering, Science & Technology Nawabshah for supporting this research work and civil hospital, Nawabshah for providing medical records. This research work is taken from the MS (IT) thesis written by Shamshad Lakho.

REFERENCES

[1] S. Andreassen, D. Karbing, U. Pielmeier, S.Rees, A. Zalounina, L. Sanden, M. Paul, and L. Leibovici, "Model-based medical decision support–a road to improved diagnosis and treatment?" in 15th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC 2011). Springer, 2011, pp. 257–260.

[2] E. S. Berner, Clinical Decision Support Systems. Springer, 2007.

[3] G. S. Uttreshwar and A. Ghatol, "Hepatitis b diagnosis using logical inference and generalized regression neural networks," in Advance Computing Conference, 2009. IACC 2009.IEEE International. IEEE, 2009, pp. 1587–1595.

[4] M. Neshat, M. Sargolzaei, A. NadjaranToosi, and A. Masoumi, "Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization," ISRN Artificial Intelligence, vol. 2012, 2012.

[5] A. B. Mrad, V. Delcroix, S. Piechowiak, and P. Leicester, "From information to evidence in a bayesian network," in Probabilistic Graphical Models. Springer, 2014, pp. 33–48.

[6] A. Oni´sko and M. J. Druzdzel, "Impact of bayesian network model structure on the accuracy of medical diagnostic systems," in Artificial Intelligence and Soft Computing. Springer, 2014, pp. 167–178.

[7] A. L. D. Alonzo, J. J. M. Campos, L. L. M. Layco, C. A. Maratas, and R. A. Sagum, "Entdex: Ent diagnosis expert system using bayesian networks," Journal of Advances in Computer Networks, vol. 2, no. 3, 2014.

[8] C. Mahesh, V. Suresh, and M. Babu, "Diagnosing hepatitis b using artificial neural network based expert system," infection, vol. 3, no. 6, 2013.

[9] N. M. Sharef and H. Madzin, "Ims: An improved medical retrieval model via medical-context aware query expansion and comprehensive ranking," in Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on. IEEE, 2012, pp. 214–218.

[10] D. Dinh and L. Tamine, "Towards a context sensitive approach to searching information based on domain specific knowledge sources," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 12, pp. 41–52, 2012.

[11] I. Niˇzeti´c, K. Fertalj, and B. Milaˇsinovi´c, "An overview of decision support system concepts," in Proceedings of the 18th International Conference on Information and Intelligent Systems/Boris Aurer and MiroslavBaˇca (ur.). Varaˇzdin, 2007, pp. 251–256.

[12] A. Oni´sko and M. J. Druzdzel, "Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems," Artificial intelligence in medicine, vol. 57, no. 3, pp. 197–206, 2013.

[13] H. Wasyluk, A. Onisko, and M. Druzdzel, "Support of diagnosis of liver disorders based on a causal bayesian network model," Medical Science Monitor, vol. 7, no. 1; SUPP, pp. 327–332, 2001.

[14] P. Ratnapinda and M. J. Druzdzel, "Does query-based diagnostics work?" 2011.

[15] H. Yan, Y. Jiang, J. Zheng, C. Peng, and Q. Li, "A multilayer perceptronbased medical decision support system for heart disease diagnosis," Expert Systems with Applications, vol. 30, no. 2, pp. 272–281, 2006.

[16] A. Hussain, A. R. Abbasi, and N. Afzulpurkar, "Detecting & interpreting self-manipulating hand movements for students affect prediction," Human-Centric Computing and Information Sciences, vol. 2, no. 1, pp.1–18, 2012.

[17] D. C¸ alis¸ir and E. Dogantekin, "A new intelligent hepatitis diagnosis system: Pca–lssvm," Expert Systems with Applications, vol. 38, no. 8, pp. 10 705–10 708, 2011.

[18] Y. Kaya and M. Uyar, "A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease," Applied Soft Computing, vol. 13, no. 8, pp. 3429–3438, 2013.

[19] M. S. Bascil and F. Temurtas, "A study on hepatitis disease diagnosis using multilayer neural network with levenbergmarquardt training algorithm," Journal of Medical Systems, vol. 35, no. 3, pp. 433–436, 2011.

[20] H.-L. Chen, D.-Y. Liu, B. Yang, J. Liu, and G. Wang, "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis," Expert Systems with Applications, vol. 30, no. 2, pp. 272–281, 2006.

Effects of Icon Design & Styles On Human-Mobile Interaction: Case Study on e-Literate vs. Non e-Literate user

Zulfiqar A. Memon[1], Rakhi Batra[2], Jawaid A. Siddiqi[2], Javed A. Shahani[2]

Abstract

Cell phones have turned out to be the most central communication gadget in our daily life. This results in an enormously intense competition between almost all the mobile phone vendors. Despite manufacturer's diverse types of advertising strategies such as exceptional price cut offers or modern attractive functions, what really matters is whether this everyday communication gadget has been designed according to the preference and requirements of all types of users. The miniature type screen interface design is one of the recent research themes of the Human-Computer Interaction domain. Because of the restricted screen size, "icons" have been considered as the prevailing style in the functional course of action of a cell phone. This article investigates the effects of icon designs and styles employed by different vendors on the perception of both the e-literate users and non e-literate users. We have explored various articles from the literature, summarizing their results of experimental validations, and a comparative analysis is described at the end. It was found that designers of mobile phone icons have to balance a trade-off between the need, requirements and understanding of both e-literate and non e-literate users.

Keywords: *Mobile phones, Icon designing, Human-mobile interaction.*

1. Introduction

With the enhancement of technology, the communication devices like mobile phones are not limited to common applications of calling and texting, they provide much more facilities now than ever before. These applications are represented through icons at the interface of small screen of mobile phones and smart phones to facilitate users to perform their everyday jobs. Visual facets, such as graphics display on the screen and icons, are fundamental rudiments of human-mobile interaction; they have been used in interface design in broader sense on the supposition that visual icons are adequate for handling impediments like language and present information in summarized form.

Literature has abundant evidence of analyzing the graphic illustrations by using icons for portable devices. Investigating the level that any icon symbolizes the sense of the purpose for which it has been intended to design, selected and configured by the cell phone maker and designer, has attracted the researcher community at large. A large proportion of older adults (the non e-literate users) due to their aging, grow worse in many of their natural abilities, like perception, motor and abilities being or relating to or involving cognition, which limit the quality of moving freely and their independence, and hence requiring more support [1]. Cell phones can support non e-literate adults for staying connected online; remembering important information by the help of memory aids and portable games stimulates mental exercises and can even provide them fun [2]. However, as being non e-literate they have find these devices more difficult to use and slower to adopt mobile computer technologies. The reluctance of these non e-literate users to adopt mobile devices can be explained by the modern HCI investigations that has inspected various diverse usability issues [3], [4].

On the other hand, the situation from the perspective of younger users (e-literate users)

[1] Department of Computer Science, National University of Computer and Emerging Sciences, Karachi
[2] Department of Computer Science, Sukkur IBA University
Corresponding email: zulfiqar.memon@nu.edu.pk

is quite opposite. They are very enthusiastic and motivated to use these mobile devices for almost all tasks of their daily life. And, as they don't share any of the disabilities like the older users mentioned above, they interact with these mobile devices in a quite easy way and in a fast pace.

The literature has evidence a very scarce work related to the investigations of the effect of icons made up of graphics on e-literate as well as on non e-literate user's use of portable equipment, even though the icons have a profound impact on most interfaces of the user cell phones. The ability of adults who are too aged to interpret icons made up of graphics, is effected by the gradual weakening in cognitive as well as perceptual abilities complemented with gradual decline in the age. The strength of interpreting the icons in older non e-literate users has also been affected by their low and narrow interaction with present-day smart equipment coupled with the low familiarity with the icons of the device and its various apps.

This article discusses that in order for non-e-literate users to be able to use the interfaces and icons many icons need to be redesigned. Below, in Section 2, various articles have been explored from the literature, summarizing their results of experimental validations and a comparative analysis is explained in Section 3. At the end, Section 4 concludes that a trade-off among the need, requirements and understanding of both e-literate and non e-literate users must be balanced by the designers of the icons of the smart phones.

2. Case Studies

Literature have evidenced much promising work on designing computer interfaces for non e-literate users (e.g. [5], [6], [7]), but less work has appeared specifically at the usability of computer icons. The literature has identified that many of the characteristics related to users have affected the usability of computer technology also affect the usability of icons for the group of non e-literate users. These user characteristics include attention, the capacity to learn and remember new information and associations, verbal and visual abilities. In addition to above, the icon usability may also be affected due to less experience with software interfaces by this age group.

The authors in [8] suggest a concrete icon design methodology for mobile base interface for the naïve low literate user segment. They also try to identify the key constructs under cognitive absorption which may have significant effect on behavioral intentions of low literate users. The authors revealed the relationships that exist between icon characteristics, and different dimensions come under cognitive absorption. The author has advocated for metaphor driven icon design methodology for designing icon design interface for the low literate target. As practical contribution, they offer clear design strategies for crafting coherent sequence effective user interactions which will facilitate self-initiated learning and usage of mobile base application.

In [9], guided by the two major goals the authors have conducted a research. To determine icons characteristics in the sector of mobile phone that ensures high semantic transparency was the first goal. The aim of this goal was to scrutinize Icons' visual complexity and concreteness. They also try to identify the key constructs under cognitive absorption which may have significant effect on behavioral intentions of low literate users. The authors revealed the relationships that exist between icon characteristics and different dimensions come under cognitive absorption [7, 10]. The author has advocated for metaphor driven icon design methodology for designing icon design interface for the low literate target. As practical contribution they offer clear design strategies for crafting coherent sequence effective user interactions which will facilitate self-initiated learning and usage of mobile base application. To determine icons characteristics in the sector of mobile phone that ensures high semantic transparency was the first goal. The aim of this goal was to scrutinize Icons' visual complexity and concreteness.

It is suggested in [11], that as far as the established functions/objects are concerned, such as messaging, address book, calls log, mobile internet etc., the icons of the smart phone, in relation to different age groups, have been standardized or customized. As for as safety applications [12] are concerned, high motivation has been evidenced however, to standardize the icons of the smart phone as compared to other applications. High performance variation across handsets of different sorts will be experienced by the users, if standardization will not be achieved in the icons of mobile phones. The designers of the icons must make them as learnable as well as understandable as possible, if in case the standardization is not feasible as indicated in the literature. For such popular device for interaction such as mobile phone, the performance could be problematic in absolute terms as such indicated in the literature. As for as safety applications [12] are concerned, high motivation has been evidenced however, to standardize the icons of the smart phone as compared to other applications. High performance variation across handsets of different sorts will be experienced by the users, if standardization will not be achieved in the icons of mobile phones. The designers of the icons must make them as learnable as well as understandable as possible if the standardization is not feasible as indicated in the literature. For such popular device for interaction such as mobile phone, the performance could be problematic in absolute terms as indicated in the literature. Therefore, for further evaluations of alternatives and for redesigning purposes, "calls log" is seemingly a good candidate. Just to mention, that this is actually the main objective and holds true irrespective of any group of age, and was holding the main scope of their study.

As far as safety applications [13] are concerned, high motivation has been evidenced however, to standardize the icons of the smart phone as compared to other applications. High performance variation across handsets of different sorts will be experienced by the users, if standardization will not be achieved in the icons of mobile phones. The designers of the icons must make them as learnable as well as understandable as possible, if in case the standardization is not feasible as indicated in the literature. For such popular device for interaction such as mobile phone, the performance could be problematic in absolute terms as such indicated in the literature. The authors included 54 icons in their study and their results related to functions suggest 6 advices for future research on the workings of icons and icon design practices:

- As far as safety applications are concerned, high motivation has been evidenced however, to standardize the icons of the smart phone as compared to other applications.
- High performance variation across handsets of different sorts will be experienced by the users, if standardization will not be achieved in the icons of mobile phones.
- The designers of the icons must make them as learnable as well as understandable as possible, if in case the standardization is not feasible as indicated in the literature.
- For such popular device for interaction such as mobile phone, the performance could be problematic in absolute terms as such indicated in the literature.
- The users frequently interpreted correctly the Icons with concrete imagery.

The authors in [14] have shown that despite widespread use and acceptance in diverse computing environments, there are still several issues in the design of icons. They have found that there are significant differences in the performance of different offerings. In particular, the authors have found that:

- As far as safety applications are concerned, high motivation has been evidenced however, to standardize the icons of the smart phone as compared to other applications.

- High performance variation across handsets of different sorts will be experienced by the users, if standardization will not be achieved in the icons of mobile phones.
- The designers of the icons must make them as learnable as well as understandable as possible, if in case the standardization is not feasible as indicated in the literature.
- For such popular device for interaction such as mobile phone, the performance could be problematic in absolute terms as such indicated in the literature.
- The users frequently interpreted correctly the Icons with concrete imagery.

In [15], the authors after conducting an empirical experiment claim that not all of the cell phone users could identify an alarm clock, despite the fact that they are using on daily basis the phones they possess. Many of them replied, when asked further, that for them the mobile phone is used mainly for sending messages and receiving or making calls. As far as safety applications [16] are concerned, high motivation has been evidenced however, to standardize the icons of the smart phone as compared to other applications. High performance variation across handsets of different sorts will be experienced by the users, if standardization will not be achieved in the icons of mobile phones. The designers of the icons must make them as learnable as well as understandable as possible, if in case the standardization is not feasible as indicated in the literature. For such popular device for interaction such as mobile phone, the performance could be problematic in absolute terms as such indicated in the literature.

3. Comparison and Conclusion

Some researchers identify more categories than others; however, in general, many researchers employ similar principles to classify icons. Due to researcher's inclusion criteria some differences also exist. Due to its text element, some authors would consider exemplar to the icon depicting a knife & fork for 'restaurant' and the icon depicting a rubbish bin for 'trash' would most likely be considered a mixed icon by few authors. Some authors have found a recognition rate of more than 66.7%, whereas others found below 40%. Literature has also evidenced, among age groups, many prominent differences in recognition rates; the recognition rate displayed by males being only 4% higher than that displayed by females; there is no significant difference between genders; and with recognition rate decreasing as age increases. We have concluded and presented a number of reasons why interface usability and good icon is particularly important on mobile devices used by this population. We have also concluded that in order for non e-literate users to be able to use the interfaces and icons many icons need to be redesigned. Based on our results, we suggest that icons incorporate commonly used symbols or concrete objects. Significantly, we suggest, using familiar metaphors, reducing semantic distance by choosing icon objects semantically close to the icon meaning, allowing users to choose an icon from a set of potentially suitable icons and using labels. Our empirical results are consistent with many existing icon design guidelines. Nevertheless to improve a non e-literate user's initial usability of icons on mobile devices and other computer interfaces, our results highlight related guidelines that should be followed. Future and existing mobile devices offer non e-literate users diverse and rich opportunities to increase their independence, get connected and to remain active. We expect that by making mobile device icons easier for non e-literate users will have a better chance of being adopted and to use the overall device will be more usable.

REFERENCES

[1] Goodman, J., Brewster, S., and Gray, P., 2004. Older people, mobile devices and navigation. In: J. Goodman and S. Brewster, eds. HCI and the older population, workshop at the HCI 2004, Leeds,UK, 13–14. Available from: http ://www.dcs.gla.ac.uk/~stephen/ research/utopia/ workshop/goodman.pdf

[2] Inglis, E.A.,et al., 2003. Issues surrounding the user-centred development of a new interactive

memory aid. Universal Access in the Information Society, 2 (3), 226–234.

[3] Jacko, J.A.,et al., 2002. Macular degeneration and visual icon use: deriving guidelines for improved access. Universal Access in the Information Society, 1 (3), 197–206.

[4] Ziefle, M. and Bay, S., 2005. How older adults meet complexity: aging effects on the usability of different mobile phones. Behaviour & Information Technology, 24(5), 375–389.

[5] Docampo Rama, M., de Ridder, H., and Bouma, H., 2001. Technology generation and age in using layered user interfaces. Gerontechnology, 1 (1), 25–40.

[6] Gregor, P., Newell, A.F., and Zajicek, M., 2002. Designing for dynamic diversity: interfaces for older people. In: Proceedings of the fifth international ACM conference on assistive technologies (ASSETS), 8–10 July 2002, Edinburgh, Scotland. New York: ACM Press, 151–156.

[7] Fisk, A.D., et al., 2004. Designing for older adults: principles and creative human factors. London: CRC Press.

[8] Sengupta, Avijit and Chang, Klarissa Ting Ting, "Effect of Icon Styles on Cognitive Absorption and Behavioral Intention of Low Literate Users" (2013). PACIS 2013 Proceedings. Paper 184. http://aisel.aisnet.org/pacis2013/184

[9] Sabine Schroder, Martina, "Effects of Icon Concreteness and Complexity on Semantic Transparency: Young vs. Older Users, In: K. Miesenberger et al. (Eds.): Proceedings of 11th International Conference, ICCHP 2008, LNCS 5105, pp. 90-97, 2008.

[10] Park, D., Schwarz, N.: Cognitive Aging. Buchanan, Philadelphia (1999)

[11] Charalambos Koutsourelakis, Konstantinos Chorianopoulos, "Icons in mobile phones: Comprehensibility differences between older and younger users", In: Information Design Journal, 2010, ca. 100, pp. 22-35, 2010.

[12] Hancock, H. E., Rogers, W. A., Schroeder, D., & Fisk, A. D. (2004). Safety symbol comprehension: Efects of symbol type, familiarity, and age. Human Factors, 46, 183–195.

[13] C. Gatsou, A. Politis and D. Zevgolis, "The Importance of Mobile Interface Icons on User Interaction", IJCSA, vol. 9, no. 3, 2012, pp. 92-107.

[14] Koutsourelakis, C. & Chorianopoulos, K., In: The Design Journal, vol. 13, no. 3, 2010, pp. 313-328, 2010.

[15] Restyandito, Chan, Alan H. S., Mahastama, Aditya W., Saptadi, Tri S., "Designing usable icons for non e-literate user", In: The Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol II, IMECS 2013.

[16] Z. Lalji, and J. Good, "Designing new technologies for illiterate populations: A study in mobile phone interface design" in Interacting with Computers, Vol.20, Issue 6, December 2008, Elsevier B.V, 2008. pp. 574-586.

Volume 1 | No. 2 | July – December 2017

Enhancing Cognitive Theory of Multimedia Leaning through 3D Animation

Zeeshan Bhatti[1], Abdul Waheed Mahesar[1], Ghullam Asghar Bhutto[1], Fida Hussain Chandiio[2]

Abstract

Cognitive theory of Multimedia learning has been a widely used principle in education. However, with current technological advancements and usage, the teaching and learning trend of children have also changed with more dependability towards technology. This research work explores and implement the use of 3D Animation as a tool for multimedia learning based on cognitive theory. This news dimension in cognitive learning will foster the latest multimedia tools and application driven through 3D Animation, Virtual Reality and Augmented Reality. The three principles, that facilitate cognitive theory of multimedia learning using animation, addressed in this research are temporal contiguity principle (screening matching narration with animation simultaneously rather than successively), personalization principle (screening text or dialogue in casual form rather than formal style) and finally the multimedia principle (screen animation and audio narration together instead of just narration). The result of this new model would yield a new technique of educating the young children through 3D animation and virtual reality. The adaptation of cognitive theory through 3D animation as a source of multimedia learning with various key principles produces a reliable paradigm for educational enhancement.

***Keywords*:** *Multimedia Learning, Animation, Cognitive Theory*

1. Introduction

People tend to learn better and accurate from combination of words and pictures, rather than just from words alone [1]. A famous slogan used to highlight the power of pictures states "A Picture is worth a Thousand Words" [2]. This slogan alone emphasis that through a single picture, thousand words worth of information can easily be conveyed to the learner. Multimedia Learning deals with teaching instructions given with aid of Graphics/images along with verbal words. Multimedia instruction deal with presenting the teaching materials using both Words or Text and Pictures or Graphics, with prime goal of promoting student learning [3]

The basic principle behind multimedia learning is that the students tend to understand and learn more effectively, the topic being taught is presented in words and pictures combined as compared to only words. Multimedia based communications can be grounded on following [1], [3].

- **Delivery Media**; this involves having a projector screen with multimedia speaker system.
- **Presentation Mode;** this involves the way in which teaching is communicated, for example words and illustrations or pictures.
- **Sensory Modalities**; the third type involves human senses that are able to capture the information such as auditory senses and visual senses.

Based on these modes, the Instructional multimedia based information and text can be centered around **technology-centered** approach that specifically focus on the usability and implications of the advance technological tools and trends. Whereas, the

[1] Institute of Information and Communication Technology, University of Sindh Jamshoro
[2] Institute of Mathematics and Computer Science, University of Sindh Jamshoro
Corresponding email: zeeshan.bhatti@usindh.edu.pk

second approach can be directed towards the **learner-centered** that is primarily targeted towards the human cognitive system and its natural behaviors [3]. Various studies show that every student can gain knowledge and understanding from a multimedia based simulated environment. They can also apply that knowledge and information in some real world scenarios.

2. Similar Work

There has been several research work in the field of Instructional multimedia learning. A computer based brake mechanism has been effectively described using animated illustrations of the entire process as shown in Figure 1. Whereas, an interactive multimedia (IMM) program has been developed that demonstrates young children about the rules and do's and don'ts of safe pedestrian skills [9], shown in figure 2. Similarly, in Figure 3, Virtual reality and multimedia technology tools are again used to train the children in safe street-crossing skills [10].

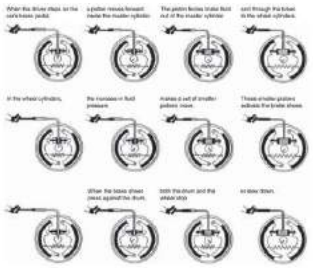

Figure 1: Frames from the narrated animation for the computer-based brakes lesson (source [3]).

Figure 2: An interactive multimedia (IMM) platform that educates young children about the safe pedestrian skills (source [9]).

Figure 3: Using virtual reality to train children in safe street-crossing skills [10]

Whereas, another pictorial animation is created to develop a multimedia based teaching lesson on how a tire pump Works as illustrated in figure 4, [8] [11]. The role and importance of education and respecting the tools of education are discussed in [13], as shown in figure 5.

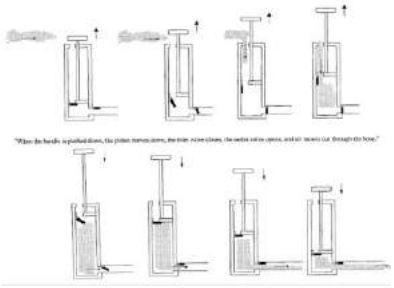

Figure 4: Selected animation frames and corresponding narration from a multimedia lesson on how a tire pump Works [8] [11].

Figure 5: Be-Educated: Multimedia Learning through 3D Animation [13]

The simple process of six segments of static illustrations of the flushing toilet tank is again taught through illustrations as shown in figure 6, [12]. Whereas, a Research Framework was purposed for FIQIR Road Safety system, to teach children about road safety using multimedia tools as shown in figure 7, [14].

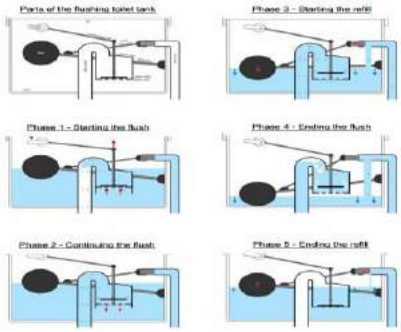

Figure 6: Six segments of static illustrations of the flushing toilet tank [12]

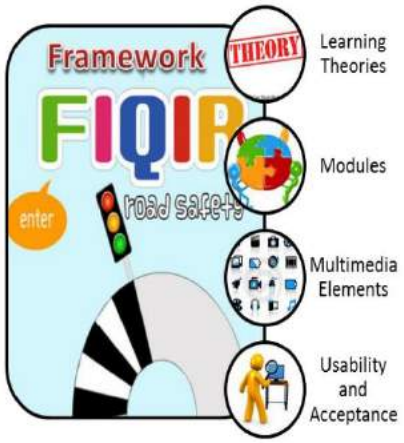

Figure 7. Research Framework for FIQIR Road Safety [14]

Similarly, Application of Interactive Multimedia Tools in Teaching Mathematics was discussed in [15] as illustrated in Figure 8.

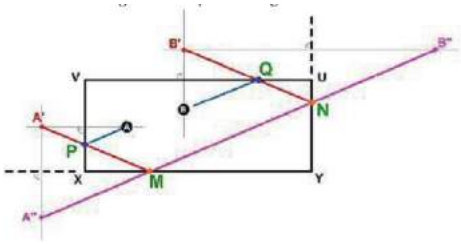

Figure 8: Lessons from Geometry [15]

3. Multimedia Learning

Multimedia learning is generally viewed as having three basic principle objectives. First being **Response strengthening**, under which the responses of the learners are tested and made stronger through drill-and–practice system. Second principle is **Information Acquisition**, where multimedia technology and tools are used to transfer information and text data to the learner. Finally, the third principle is **Knowledge Construction**, through which an understanding and logical sense is developed about the subject matter in order to increase the retention of given knowledge.

3.1 Cognitive Theory of Multimedia Learning

The cognitive theory of multimedia learning (CTML) is centered on three cognitive science principles of learning [1][4][5]. First Principle, for **processing of information** two separate - Audio and Visual data, channels are to be used. Second Principle says that the **channel capacity** will always be limited, exposing the fact that there will be only a partial and restricted amount of information that can be actively processed at any single given time. Third Principle states that the **learning** is an active process concerning and dealing with filtering, selecting, organizing, and integrating information. These principles and their relational working are illustrated in

Figure 9.

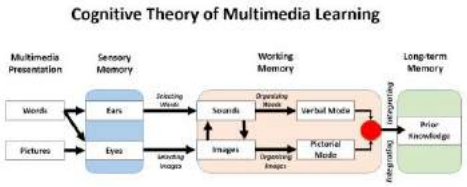

Figure 9: Illustrates the basics of cognitive theory of multimedia learning (Source: [1]).

4. Research Question

Q1. HOW to Improve Students Learning?

The central question of research is that: How can we design multimedia instructions that improves students understanding of the presented material? The two qualitative variables to be considered are learning and retention rate. How learning can be achieved effectively, that would yield a better and higher retention rate.

4.1 Animation to the rescue

Through this project, an approach to combine the cognitive multimedia learning principle with Animation based Learning techniques is designed. Generally, multimedia learning involves, images and video along with text, still 3D animation techniques are employed to attain a level of sophistication through which a strong message and lesson on education can be portrayed.

It has been discussed and proven by many researchers that animation can encourage and enhance the ability of the learner and viewer to understand and gain the message. Especially, when it has been used within the principles of cognitive theory of multimedia learning [1][4][5][6].

a. What is Animation?

Computer Animation refers to sequence of frames or images in which the subject/object is changing its position or orientation with respect to time in each subsequent frame in succession, that yield the perception of motion [7]. Animation can promote learner understanding when used in ways that are consistent with the cognitive theory of multimedia learning. [6]

4.3 Animation as an Aid to cognitive Multimedia Learning

In order to address the research question discussed above, we purpose a modified version of Cognitive theory that uses a new dimension of Animation and Regional language, to be incorporated in multimedia learning as highlighted in figure 10. Through these two new parameters, i.e. Regional Language based instructional media and use of 3D Animation to describe the complex scientific phenomena's, the current Hypothesis is that it would significantly increase the learner's ability to understand and grasp the knowledge. While the same would result in increasing the retention rate of the learner. In our Model of Animation based Cognitive Theory of Multimedia Learning (ACTML), we Introduce animation as the third channel of instructional multimedia learning and propose a new model for ACTML [16].

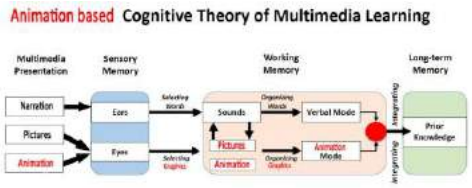

Figure 10: Purposed and Modified version of Cognitive Theory of Multimedia Learning.

5. Principles of using Animation with cognitive Multimedia Learning

In order to use animation with cognitive theory, we have devised the following three principles in conjunction with multimedia learning

A) Animation principle (screen animation and audio narration together instead of just narration).

B) Temporal contiguity principle (screening matching narration with

animation simultaneously rather than successively)

C) **Personalization principle** (screening text or dialogs in casual form rather than formal style).

5.1 Animation principle

The first principle we propose and emphasis on is that the students are expected to understand and remember the taught topic easily and with profound accuracy, if that topic is taught with specially created 3D animation principles, along with text and voice based narrations. This principle is based on simple rationale fact that the students are able to respond much better towards moving illustrations and 3D animations along with textual captions and voice narrations.

5.2 Temporal contiguity principle

The second principle that we have used is derived from the basic rule that students understand and learn the topic more profoundly and easily when the multimedia instructions containing test/words/narration is presented synchronously with the animation. This principle simply relies on the theoretical validation that each student is better able to focus and follow when the narrations, textual caption and animation are all synchronized together seamlessly [6].

5.3 Personalization principle

The third principle we adopted was, that the students understand and learn the topic more profoundly and easily from animated and narration specially when the voice based recitation is in general casual conversational form rather than based on official technical wording with formal style.

6. Conclusion

Multimedia learning is a technique that enhances and facilitates the learner's internal information system that comprises visual and verbal channels of processing information. This system requires a cognitive processing at each level for ability to learn and gain knowledge from taught topic, effectively. In order to enhance this ability, we introduced a third channel of Animation, into the Cognitive theory of multimedia learning. Animation is a very powerful means and mode of communication, whereby the knowledge of deep and complicated phenomenon can be easily taught to students. The use of animation in multimedia learning is structured on three codes of cognitive theory, these principles are Multimedia principle, temporal contiguity principle and personalization principle, which will produce new dimensions of multimedia learning.

ACKNOWLEDGMENT

This research work was carried out in Multimedia Animation and Graphics (MAGic) Research Group at Institute of Information and Communication Technology, University of Sindh, Jamshoro.

REFERENCES

[1] Mayer, R. E. (2014). Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning*, *43*.

[2] Pinsky, L. E., & Wipf, J. E. (2000). A picture is worth a thousand words. *Journal of general internal medicine*, *15*(11), 805-810.

[3] Mayer, R. E. (2001)."Multimedia Learning" Chapter 1, Cambridge University Press, 2001.

[4] Mayer, R. E. (2005). "Cambridge Handbook of Multimedia Learning" , Chapter 1 and Chapter 2. Cambridge University Press, 2005.

[5] Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of educational psychology*, *91*(2), 358.

[6] Mayer, R. E., & Moreno, R. (2002). Animation as an aid to multimedia learning. *Educational psychology review*, *14*(1), 87-99.

[7] Thalmann, N. M., & Thalmann, D. (1990). Computer Animation. In *Computer Animation* (pp. 13-17). Springer Japan.

[8] Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions?. *Educational psychologist*, *32*(1), 1-19.

[9] Glang, A., Noell, J., Ary, D., & Swartz, L. (2005). Using interactive multimedia to teach pedestrian safety: An exploratory study. *American journal of health behavior*, *29*(5), 435-442.

10] Schwebel, D. C., & McClure, L. A. (2010). Using virtual reality to train children in safe street-crossing skills. *Injury prevention*, *16*(1), e1-e1.

[11] Mayer, R. E., & Anderson, R. B. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of educational Psychology*, *84*(4), 444.

[12] Paik, E. S., & Schraw, G. (2013). Learning with animation and illusions of understanding. *Journal of Educational Psychology*, *105*(2), 278.

[13] Abro, A., Bhatti, Z., Gillal, A.R., Mahesar, A.W., Karbasi, M., (2016), Be-Educated: Multimedia Learning through 3D Animation. Pakistan Journal of Engineering and Applied Sciences (submitted)

[14] Rawi, N. A., Mamat, A. R., Deris, M. S. M., Amin, M. M., & Rahim, N. (2015). A Novel Multimedia Interactive Application to Support Road Safety Education among Primary School Children in Malaysia. *Jurnal Teknologi*, *77*(19).

[15] Milovanovic, M., Obradovic, J., & Milajic, A. (2013). Application of Interactive Multimedia Tools in Teaching Mathematics-Examples of Lessons from Geometry. *TOJET: The Turkish Online Journal of Educational Technology*, *12*(1).

[16] Bhatti, Zeeshan and Waqas, Ahmed and Mahesar, Abdul Waheed and Chandio, Fida and Bhutto, Ghullam Asghar, (2017) "*Use of Multimedia, Temporal Contiguity & Personalization principles in Cognitive Multimedia learning through 3D Animation*", in proceedings of International Conference on Computing and Mathematical Sciences, at Sukkur IBA, 25-26 February, 2017.

Volume 1 | No. 2 | July – December 2017

Reflections of Practical Implementation of the Academic Course Analysis and Design of Algorithms Taught in the Universities of Pakistan

Faryal Shamsi[1], Muhammad Irshad Nazeer[1], Raheel Ahmed Memon[1]

Abstract

The Analysis and Design of Algorithm is considered as a compulsory course in the field of Computer Science. It increases the logical and problem-solving skills of the students and make their solutions efficient in terms of time and space. These objectives can only be achieved if a student practically implements what he or she has studied throughout the course. But if the contents of this course are merely studied and rarely practiced, the actual goals of the course are not fulfilled. This article will explore the extent of practical implementation of the course of analysis and design of algorithm. Problems faced by the computer science community and major barriers in the field are also enumerated. Finally, some recommendations are made to overcome the obstacles in the practical implementation of analysis and design of algorithms.

***Keywords*:** *Analysis and Design of Algorithms, Teaching and Practice, Analysis in Pakistan.*

1. Introduction

The core objective of this course is to introduce the tools and techniques for the problem solving and decision-making skills of a programmer. The question is, whether the objectives are practically achieved? How many people implement what is actually learned in the course? To answer these questions in appropriate way, a research was conducted. The contents of the research will be explored in further subsections.

- **Analysis and Design of Algorithms**

Algorithms are the basis for the solution of computational problems and identification of computational complexities [2]. The algorithm is defined [1] as the formal sequence of steps which converts an input into the output. We can say that the [2] analysis of algorithm is to evaluate the effectiveness of the algorithm and number of resources required to use implement it.

- **Core objectives of the field**

The field targets to reduce the overall cost of solving a problem, the cost both in terms of time and money. The course of Analysis and design of algorithms has the following objectives –

- Estimates the running time of the algorithm
- Estimates the storage space or memory capacity required
- Reduces the number of resources required
- Identifies different methods of solving the same problem
- Determines the best solution for a computational problem

- **Why we analyse the algorithm today?**

With the advent of technology we have enough computational power and storage not only available but wasted in most of the cases. In such environment where abundant computational resources are available, one can take the importance of analysis and design of algorithm for granted.

The literature [1] suggests that even if we consider that we have infinite resources available, which is not actually the case.

[1] *Department of Computer Science, Sukkur IBA University, Sukkur, Sindh, Pakistan*
Corresponding email: faryal.mscs16@iba-suk.edu.pk

Even then, we need to study this course and understand the importance of analysis of algorithms. The resources may be infinite but will never be free [1]. So, we still need to analyze to the algorithms to –

- Identify the solution that generates the required output.
- Look for the easiest solution and implement it
- Ensure to comply with the software engineering standards.

- **Teaching Methodology for Analysis and Design of Algorithms**

Conventional teaching methodology was questioned in the research conducted by Xuemin Zhang in 2014 [3]. Some teaching reforms are suggested in literature to decrease the difficulty and increase the motivation of students [3]. A research based teaching methodology should be followed which can facilitate the learning environment by incorporating theory and practical. Combination of blackboard and multimedia is also prescribed by the study [3]. The course content comprises one-way communication most of the time and students are not able to interact and contribute in the classroom. According to Zhang the multimedia support encourages student to be creative and innovative [3].margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

2. Methodology

This research is based on an online survey where 200 respondents with the background of computer science participated. 10 experts of the field were also interviewed. Some focused group discussion sessions were also conducted with some students of the course. The whole community of computer science could not be reached with the limited time and budget. So, conveniently stratified sampling was done to divide the population in three strata –

- Students
- Teachers
- Professionals

The practical implementation of analysis of algorithm and its barriers cannot be quantified. For this purpose a scaling instrument was developed for quantification. To measure the extent of practical implementation of the course consensus based factor scaling was used. For making consensus a penal of ten experts of the field was developed.

- **Factors to be demonstarted**

Following factors were suggested by the experts, to be demonstrated, in regard to the practical implementation of analysis and design o algorithm –

- Resource Management (CPU, RAM and so on.)
- Attempting various techniques
- Code Optimization
- Improved Problem Solving
- Improved Decision Making

- **Major Distractors**

To measure the major distraction from the analysis, following factors were suggested be analyzed, by the expert panel–

- Mathematical Background
- Personal Interest
- Ability to relate with expressions with the realworld problems
- Teaching Methodology

3. Results

200 respondents took part in an online survey. One hundred respondents were the students of 2 reputable institutes of Sukkur city. Forty were the teachers of same institutes. And 60 were the computer science professionals from the various organizations of Pakistan. The questionnaire was designed using the semantic differential factors scaling

technique. The factors of practical implementation of the course of analysis and design of algorithm were identified by consensus of a panel of 10 experts of the field. The questions, responses from Teachers, Students and Professionals are indicated in Table 1, 2 and 3 respectively.

TABLE I. Feedback from 40 Teachers

S. no	Questions and Responses		
	Question Contents	***YES***	***NO***
1	Do you check the available resources of the system such as speed of the processor or capacity of RAM before programing?	30	10
2	After solving a problem, do you try it again to find a better solution?	15	25
3	Do you think that the course of analysis & design of algorithm has improved your decision -making skills?	40	0
4	Do you think that the course of analysis & design of algorithm has improved your problem solving skills?	32	8
5	Do you find the course of analysis and design of algorithms interesting?	27	13
6	Do you have weak mathematical foundation?	33	7
7	Do you find it difficult to relate the real world problems with mathematical expressions?	9	31
8	Are you satisfied with the teaching methodology followed for the analysis & design of algorithms?	35	5

TABLE II. Feedback from 100 Students

S. no	Questions and Responses		
	Question Contents	***YES***	***NO***
1	Do you check the available resources of the system such as speed of the processor or capacity of RAM before programing?	8	92
2	After solving a problem, do you try it again to find a better solution?	14	86
3	Do you think that the course of analysis & design of algorithm has improved your decision making skills?	37	63
4	Do you think that the course of analysis & design of algorithm has improved your problem solving skills?	56	44
5	Do you find the course of analysis and design of algorithms interesting?	76	24
6	Do you have weak mathematical foundation?	12	88
7	Do you find it difficult to relate the real world problems with mathematical expressions?	78	22
8	Are you satisfied with the teaching methodology followed for the analysis & design of algorithms?	90	10

TABLE III. Feedback from 60 Professionals

S. no	Questions and Responses		
	Question Contents	*YES*	*NO*
1	Do you check the available resources of the system such as speed of the processor or capacity of RAM before programing?	11	39
2	After solving a problem, do you try it again to find a better solution?	23	27
3	Do you think that the course of analysis & design of algorithm has improved your decision making skills?	35	25
4	Do you think that the course of analysis & design of algorithm has improved your problem solving skills?	21	39
5	Do you find the course of analysis and design of algorithms interesting?	41	19
6	Do you have weak mathematical foundation?	36	24
7	Do you find it difficult to relate the real world problems with mathematical expressions?	10	50
8	Are you satisfied with the teaching methodology followed for the analysis & design of algorithms?	33	27

To understand the trends of the feedback from the teachers, students and professionals, the responses are illustrated in the following figures along with the speculations that can be made. Each figure from 1 to 8 corresponds to the questions that were asked.

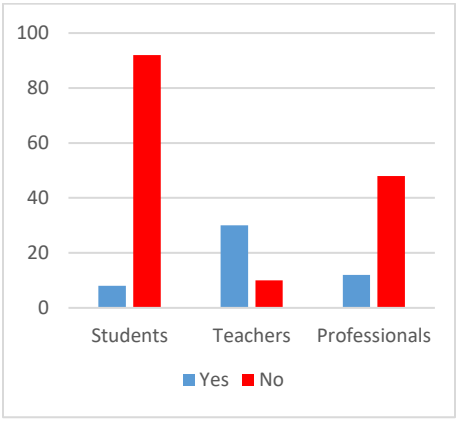

Figure 1: Awareness of Resources

Figure 1 illustrates the trend that almost all of the students and most of the professionals do not even check the available resources, although the teachers are mostly aware of the resources.

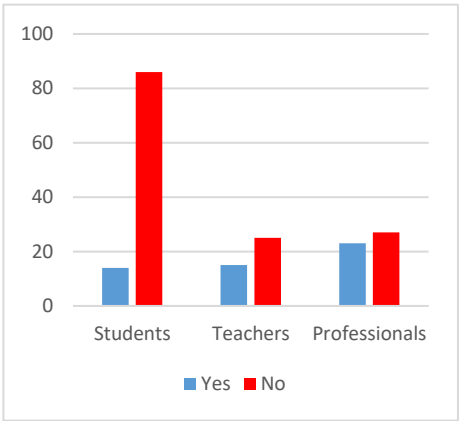

Figure 2: Optimization Practice

Figure 2 clearly indicates that most of the people do not try to optimize the solution. This problem is at its peak in case of students. Furthermore, the teachers and even the professionals do not practice the optimization of problems.

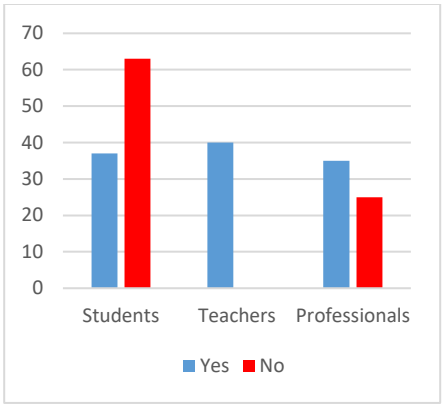

Figure 3 - Improved Decision Making

Figure 3 shows that all of the teachers expressed that the course has made a significant improvement in the decision making. Many professionals but a small proportion of students consider that the course has made impact on the decision making skills.

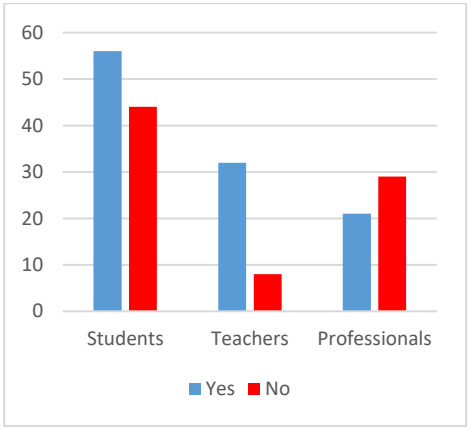

Figure 4 - Improved Problem Solving

Figure 4 surprisingly indicates that most of the professionals think that the course of analysis and design of algorithms has no major impact on their problem-solving skills. On the other hand, students and teacher were found to be optimistic about the impact of the course on their problem -solving skills.

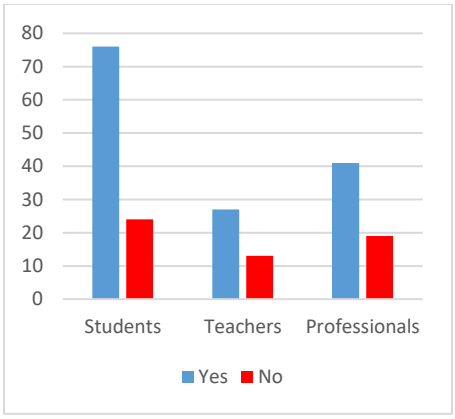

Figure 5 – Have interest in the course

Figure 5 reveals the fact that lack of interest is not actually the core issue here. The majority of students, teachers and professionals responded that they find the course interesting.

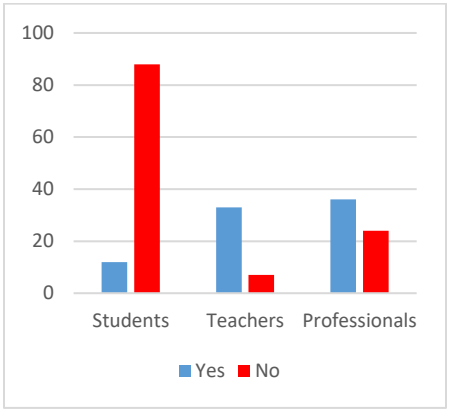

Figure 6 - Weakness in Mathematics

Figure 6 removes the misconception that weak mathematical foundation is the barrier in the practical implementation of the course. A large number of students reported to have strong mathematics.

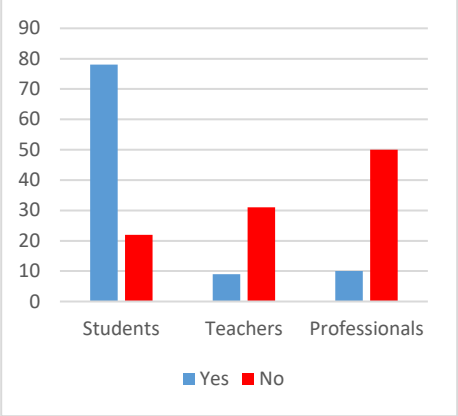

Figure 7 - Difficulty in relating mathematical expressions with real world problems

Figure 7 diagnoses the first major distraction and barrier for students in the way of practical implementation of the course concepts. This is the difficulty to relate the real world problems with the mathematical expressions.

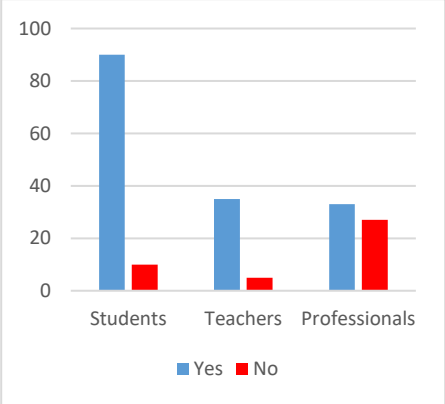

Figure 8 - Problematic Teaching Methodology

Figure 8 points out a major barrier towards the practical implementation of analysis and design of algorithms reported by all the respondents – Students, Teachers and Professionals. That's the Teaching Methodology.

For further investigation, the score of individual with respect to practical implementation was calculated. The correlation of each distraction with the practical implementation was calculated. The values for mathematical background, interest in the course, difficulty in relating with expressions and teaching methodology were found to be -0.23, 0.03, 0.51 and -0.67 respectively.

The data clearly summarizes that the one major barrier in the practical implementation of the academic course of analysis and design of algorithms specifically in the Universities of Pakistan is the teaching methodology. Although the result suggest that, there is another barrier that is the difficulty in relating the mathematical statements to the real world analysis of algorithm issues. It can be assumed that the other barrier is also related to teaching methodology. It is the inappropriate teaching style that causes the students the stated problem. On the basis of this assumption it is necessary to investigate that if the teaching methodology a major barrier for all three populations – (i.e. Students, Teachers and professionals. For that purpose the Chi-Square test to homogeneity was applied to the results of the three populations.

H_0: Teaching Methodology is not a barrier to Students, Teachers and Professionals
H_A: Teaching Methodology is a barrier to Students, Teachers and Professionals

Populations	**Satisfied with Teaching Methodology**		Total
	YES	*NO*	
Students	90	10	100
Teachers	35	5	40

	Professionals	33	27	60
	Total	158	42	200

By using the above contingency table, the degree of freedom is 2, and the estimated Chi-square value is 29.86793 which is significant. The p-value 5.991 is much lower than the Chi-square value, so we can reject the null hypothesis and accept the alternative. This proves that the teaching methodology being used in the universities of Pakistan is not suitable for the course of analysis and design of algorithms and requireS some instantaneous reforms.

4. Speculations

- The results suggest that most of the people in the computer science community are not performing the analysis of algorithms.
- It is obvious that most of the people do not even check the resources. But even considering the people who claim to check the resources, does not suggest that they manage the resources effectively, or even try to manage at all.
- Limited number of people try to optimize the solution while the rest just fix the problem at any cost and forget it forever.
- All the teachers believe that the course has made a significant impact on their overall potential in terms of decision making and problem solving skills. The impact on other respondents was found to be insignificant.
- Lack of interest and poor mathematics is not the greatest barrier to analysis. Most of the respondents were found to have interest in the field as well as good mathematical skills but were not satisfied with the teaching methodology.
- The major barrier to analysis and design of algorithm was proved to be the teaching methodology. Even if the students are unable to relate the mathematical expressions with real world analytical problems, it reflects the problematic teaching methodology.

5. Recommendations

- The analysis and design of algorithms must be taught as an advanced course. Most students study it in the initial phase of the computer science education, while this course demands a great deal of prior knowledge. For Example –
  - Major programing techniques and different languages.
  - Major Mathematical Course –
    - Calculus
    - Discrete Mathematics
    - Computational Maths
    - Probability Statistics
    - Inferential Statistics
    - Data Structures
    - Theory of Automata
- Traditional teaching methodology is a major distraction. There must be some innovative techniques to maintain students' interest and motivation.
- The significance and importance of analysis should be communicated effectively and convincingly to retain the motivation level of the students.
- Students must be provided with limited computational resources to encourage optimization.
- Benchmarks must be used in the classroom to evaluate the

performance of different computational solutions

- Simulators and graphical objects must be used connect the real world analytical problems with complicated mathematical expressions.
- As the programing and computational techniques have evolved from linear programing to visual programing and automated application development. The analysis techniques must also evolve by use of artificial intelligence, quantum computation or Data Mining etc.

6. Conclusion

The purpose of the course of Analysis and Design of Algorithms is to equip the computer science community with the abilities of problem solving, resource management, code optimization. Unfortunately, it is contemplated that the essence of the curriculum is not extracted by the computer science community.

After a thorough survey based study, it was scrutinized that course is studied compulsorily but applied rarely by the computer science community. The article proves the claim and pinpoints some barriers of analysis of algorithms. The recommendations in reforms required in the teaching methodology to overcome such barriers are also specified.

ACKNOWLEDGMENT

We are thankful to all the experts, professionals, our teachers and other teachers who supported us throughout the research and provided with their feedback and precious comments.

REFERENCES

[1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein, "Introduction to Algorithms", Massachusetts Institute of Technology, Cambridge, London, England. (2009)

[2] https://en.wikipedia.org/wiki/Analysis_of_algorithms, Wiki Encyclopedia

[3] Xuemin Zhang, Zenggang Xiong, Yaomin Ding, Guangwei Wang, Conghuan Ye, "Research-oriented teaching of Design and Analysis of Algorithm", esearch-oriented teaching of Design and Analysis of Algorithm (2012)

[4] HEC Course Schema. Course Outline for Analysis and Design of Algorithm 2013, http://hec.gov.pk/english/services/universities/RevisedCurricula/Documents/2012-2013/computer-science%202012-13.pdf

Volume 1 | No. 2 | July – December 2017

Role of GIS in Crime Mapping & Analysis

Iqra Shafique[1], Syeda Ambreen Zahra[1], Tuba Farid[1], Madiha Sharif[1]

Abstract

In most recent years, crime analysis has turned into a broad- spectrum term that needs a considerable measure of research on crime investigation and crime mapping. Study about crime with respect to its spatial and temporal distribution is important because data about crime incident is one of the most urgent need to fight against crime. Crime mapping and analysis play an integral role in essentially advanced form of crime representation, visualization and respond satisfactorily to the problem of criminality. It also lets the analysts to figure out how crimes are spread evenly over the zone. GIS plays an effective role in mapping of crime. This paper puts on the diverse utilities of GIS to recognize the hotspots in addition to encourage the advancement of investigation inclination strategy for policing. The functional approach in the present investigation for crime mapping can be successfully applied for improvement of user-interfaces stage for the advancement of safe city strategies.

***Keywords*:** Crime mapping, Spatial, GIS, Hotspot, Spatial Temporal analysis, RTM, Reggresion, Correlation, GAM, GEM.

1. Introduction

For the last few years, a new worldwide social order leads to the growing ratio on the criminal activities and raises the need to investigate latest methods to deal with information about criminality. (Jorge Ferreira, et al, 2012). Crime mapping and spatial analysis of crimes are acknowledged as strong methods for the learning and control of crime because crime maps help one to investigate crime data and enhanced perceptive not only why a crime raises, but where it is taking place.(Gupta, Rajitha, et al, 2012).

"Hurtful work or need against the masses as the State needs in similarity with stop yet which, above conviction, is culpable by means of fine, detainment, as well as death. No organization constitutes a crime unless it is pronounced pecan inside the lawful rules on the nation. Crime is unlawful attempt up to desire is denied by method for the law. Blame is habitually called an 'offense'. A few people put on shirts so much discourse 'it's exclusively unlawful if ye reach got'."[34]

Straightforward maps, as show the areas the place of violations or centralizations of wrong doings hold came to fruition perform keep matured as per assist endorse watches in similarity with areas those are generally required. Approach producers inside police divisions may utilize more convoluted maps in congruity with inspect patterns among criminal movement, and maps may also demonstrate importance between settling destructive cases.

For instance, analysts can likewise utilize maps as per better catch the looking examples on progressive hoodlums then as per speculate the place this guilty parties may live. Utilizing maps so help people envision the geographic parts on wrongdoing, be that as it may, is presently not limited as per law requirement. Mapping is capable outfit particular records with respect to blame or criminal conduct in impersonation of lawmakers, the press, yet the general open.

Crime mapping answer three main subprograms within crime investigation.

- It uses visual and statistical analyses of the spatial conducting crimes and other types of actions.

[1] *CS & IT, University of Lahore, Gujrat, Pakistan*
Corresponding email: Msit.5504@gmail.com

- It allows analysts to associate spatial and non-spatial data.
- It furnishes maps that are helped to put across analysis results.

2. Crime Mapping Today

Colleague effect neither provides careful, true and adequate matter not far from protect nor does it help in development goals and decision support. Spatial data analysis help one investigates crime data and enhanced perceptive not only why a crime rises, but where is taking place. . (Gupta, Rajitha, et al, 2012).

An acronym in light of Geographic records Systems that alludes after current portable workstation transcription up to desire catches, records, stores and examinations information regarding utilizations of earth's floor (James, O. 2014). It is additionally portrayed by method for paying for actualities alluding to as indicated by applications and in that place areas of floor surface sure as roadway, video show units exercises as much she happen, recovery or show of uncommon information, as pleasantly as mapping.

It additionally involves geographic profiling the place areas are carefully entered by method for address, GIS thrived along the upward push concerning automated pc mechanical insurgency or has subsequently some separation measured as per remain completely phenomenal into settling many entangled social, monetary then politic inconveniences of humankind. Effectively, such has settled much injustice issues in the predominant world.

2.1 Getting Guilt to a Map

It is nonetheless workable in imitation of leading easy offense mapping through occupying pins between maps; alternatively crimes facts (both entire into entire or exclusively) contain a multiplicity on spatial-transient data.

Unless the records are mechanized then examined utilizing fitting programming, substantial tests also, clear procedures, so statistics intention remain usually inaccessible to both specialists also, professionals. The excellent programming arrangements are usually eluded in accordance with so geographic information frameworks, or GIS. GIS maintain spatial data of 3 essential ways: records are eking out away namely focuses, traces then polygons.

While spatial data are last as like focuses, traces and polygons, characteristic records are critical proviso the spatial records is in accordance with hold extra than shallow honor crimes records are mapped by a procedure called Geocoding. (Jerry Ratcliffe, 2010).

Geographic statistics frameworks (GISs) improve PCs in conformity with speak according to and observe spatially associated wonder. All GISs bear twin functions:

(i) To show maps then geographic components, because example, obliqueness locations (focuses), streams (lines), yet assessment tracts (polygons)

(ii) To utilize a database supervisor so arranges and relates faith facts to specific information highlights.

A GIS uses an advanced information database in conformity with interface spatial records in accordance with wise data. Several varieties over coordinating calculations possess a GIS in imitation of connect then preserve upon spatial connections amongst geographic yet enlightening data. The potential according to interface or keep above spatial connections among statistics units characterizes a GIS. (Philip R. Canter). The uncovering was undiminished close to handsome after goals

- To pick out warm spots as nicely as much using army because of specific sorts on crime.
- To help police in conformity with take strong measures kind of expanse regarding legion in area Inclined according to crime.
- To build over a methodological law because of wrongdoing mapping making use of GIS.

3. Geographical Information System and Crime Mapping

GIS plays an essential role in crime mapping and analysis. The ability to contact and proceed information quickly, whereas displaying it in a spatial and visual means allows agencies to deal out assets rapidly and more successfully. The mainly dominant beat in law enforcement is information technology. Geographical information system is an information system that describes the objects with location.

A geographical information system converts physical elements of the real world such as roads, rivers, mountains, buildings into forms that can be visualized and analyzed, such as banking system, climate system, oceans, traffic, educational system police information about crimes,. GIS utilizing two sorts of information model vector and raster information. Vector deal with the discrete objects and raster deal the continuous objects. Both vector and raster are not the same as every other. (Paulo Joao, et al, 2012).

After collection, edition and approving this data spatial analysis permits the assessment of these attributes and with the following space, it gives a geographical value to any geographical wonders. The usage of geographic data framework for wrongdoing mapping maps, envision and analyze crime hot spots along with other vogues and forms. It is a basic constitute of offence judgment and the protect deposit. A GIS applies pair types of make to suggest objects and locations in the real world. (Jose Martin, et al, 2012) These are denoted to as polygon, point, line and image features.

The spatial data may be the location. GIS not only permits consolidation and spatial analysis of the data to discover, capture and indict mistrust, but it also helps more positive measures in the course of helpful allowance of resources and better policy setting. In the next section of the paper we provide a framework of crime mapping that include background, methodologies and conclusion.

4. Background

Crime mapping has long been the fundamental past of crime analysis today. The use of maps in criminology has its shadow back to at least years 1900. Most up-to-date nations have moved from the "pin on maps" to the use of GIS. Sorrowfully, many developed nations are still using outdated file system and installing analog. Several times police operations are brought on the simple "trial and error". Moreover when the crimes are committed the outdated systems were useful to point out but they also had a noticeable constriction because the preliminary crime figure was missing in latest version. The analog maps were not child's play. Additionally, when certain kinds of crimes are shown by pins of different colors which are mingled in map are difficult to understand. Pin maps have short comings i.e. space, wastage of time and lacking the ability of evolving a logical national database. Having these shortcomings in existing system, it is the demand of time to research in better way regarding criminology. Therefore, efforts have been made to move on GIS. [22]

5. Literature Review

Till our study for this proposal shows that despite the advancement in the technological field still outdated methods are being used for mapping and tracking the crime in the society. This practice of outdated methods produces a major gap between the response from police or action bodies and the criminals. This happens due to the use of slow method (i.e. pin maps) to detect the crime location (Toju Francis balogun, 2016). The outdated method makes a huge waste of manpower and time resources, and produces frustration in the police department. Hence the latest technology like Geographical Information System (GIS) must be used to map and detect the crime location in order to make quick response to the crimes. (Pukhtoon Yar, 2016)

Overview of Methdologies

After reviewing several papers, we have summarized different methodologies that help

Authors to understand spatial analysis of crime.

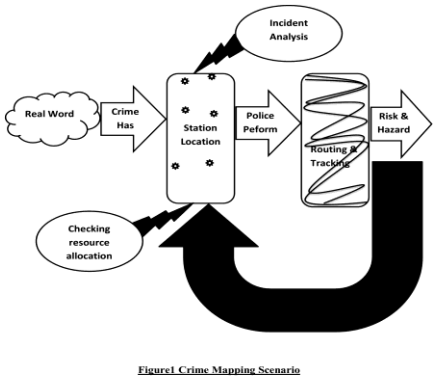

Figure1 Crime Mapping Scenario

Figure 1: Survey based analysis

1. Survey based analysis

To effectively understand the nature and location of criminal activities, a reliable data is necessary. For this purpose criminologists have developed two methods for obtaining information on the extent and pattern of crime and victimization in society

- Crime Survey
- Victim Survey

After understanding the spatial patterns associated with crime and victims in the area, the criminologists can determine the origin of criminal activities in the zone and made possible solutions of the problems. [1]

2. (Balogun et al, 2014) proposed a "**questionnaire method"** to obtain information from police and the general public. Studies reveals that people have no confidence on police. Due to the outdated resources and methods, police departments are restricted in reducing crime. But by utilizing GIS resources, crime analysists can create:

- Digital land use maps for showing crime location
- Crime Geo-spatial database to reduce redundancy of records
- Spatial analysis for processing spatial queries

GIS can provide both printed and soft copy navigated maps to help analyzing crime incidents planted on their location, type or proceeding time.

3. Identifying Hotspots

Hotspot technique is basically used to identify areas where crime levels are high. The Hotspot analysis tool identifies spatial clusters of statistically significant high or low value attributes. Different methods for hotspot detection are as follows:

- **Spatial analysis method**:

This method checks the location, attributes and connection of features in spatial data among overly and other technical techniques.

- **Interpolation Method:**

This process used to predict values at alternate unknown points

- **Mapping cluster:**

Also known as spatial Autocorrelation. This process used an amount of degree to which a set of spatial features and the data values are associated with it. [4]

4. According to (Shahebaz m Ansari, 2014), hotspot consists of the following Six steps:

Data collection is the basic step of implementation. After data collection **Geo-referencing** of dataset takes place. Then spatial features are **digitized**. Then police station and their boundaries are **mapped**. After that **crime data base** is created and the hotspot method is applied on the data set and in the end **result and analysis** are done.

5. Risk Terrain Model (RTM)

RTM displays a map that describe the spatial properties of land [11]. In RTM, separate map layers are created, that represents the spatial phenomena of a land with GIS. Then thematic layers are joined to produce a blended risk terrain map with values that demonstrate the spatial impact of all features at every place throughout the landscape. RTM not only identify the hotspot areas but also it explains the "Spatial patterns"

that are associated with crime and offer possible solutions to overcome risk factors.
The disadvantage of RTM is that, "it do not create obsolete scenarios where crime will occur" [11]

6. According to (Yar & Nasir, 2016), there is "relationship between **crime and weather"**. Most crime occur in summer season because weather effects on the mentality of a man. To prove this, the data should be collected from both internal and external resources. The collected data was evaluated by the following criterion:

- Police station wise circulation of crime
- Crime committed on the basis of weather
- Determination of hotspots

Analysis shows that when weather gets hot, people became mad and loss their temper.

7. Newer Technologies:

To address the changing circumstances in crime, Law enforcement agencies find out new technologies to reduce crime. These are described below:

- From Feet to the ground to data:

To check where there is a need to open a new police department

- From guts to analytics:

Departments rely on professional resources that can present a new geographical way to produce meaningful impact.

- From reaction to anticipation:

In the past, police departments reacted to crime by responding when something happened. Through the use of information and analytics, a more proactive approached can be embraced in which issues are identified and resources are utilized to prevent criminal activity throughout the community. This requires three analytical techniques:

- Identifying Hot spots

To figure out where crime levels are high and where crimes are occurring frequently

- Correlation

When attempting to comprehend crime through GIS, it is critical to analyze how two separate components are connected. Correlation demonstrates the association between components on land.

- Regression

Shows the corporation between variables. It is a method that locates the regular connection among attributes. (Corporal & Beaty, 2013)

8. ArcGis

Software that provide us the facility to generate Maps. [8] To perform analysis on spatial data, manage large amount of spatial data and produce cartographically appealing maps that help us in decision making, ArcGis will give one common platform to meet your GIS needs.

9. There are two ways to automate the geographical analysis of crime:

- **GAM (Geographic Analysis Machine)**

Demonstrate that without having detailed statistical and geographic knowledge, we can analyze the crime hotspots areas and draw patterns of crime on map that are found in that particular area.

- **GEM (Geographic Explanation Machine)**

Demonstrate that after getting information from GAM, we can investigate crime data. GEM explain the geographic variables on ground so that it can fully support the crime analysts to deal with the hotspots found. [9]

10. Sensor technologies

To investigate crime, GIS depend on "Sensor Technologies" (Umer Shehu, 2015). In GIS a technology, named as SPAGHETTI can be used for map digitizing. For crime detection, this technique comprise with Ariel Sensor technologies that identifies objects on earth. It additionally includes geographic graphs in which areas are digitally entered by address, interpret with calculation that

delivers a likelihood surface indicating the possibility where crime incidents are high.

6. Study Area

The crimes data were collected from the Faisalabad City Police Department. There are forty police stations in the Faisalabad district, and 18 police stations fall within the study area. The paper based crime reports were used for marking the crime events. The law enforcement agencies in Pakistan fall behind in using the modern technology and the electronic crime recording is not in practice. Since there was no standardized crime reporting system, we had to digitize the paper based crime reports and mark the crime incidents on google maps. The crime reports of 2015 and 2016 were geocoded with the XY locations.. Only the network constrained crime events were included in the analysis. House robberies, burglaries, crimes against persons, and other petty crimes are not part of this analysis. A total number of 2059 crime events in 2015 and 2016 were geocoded. The points were snapped to the lines by using the RTW tools for ArcGIS developed by Wilson (2012).

7. Geo Referencing of the Dataset

Raster data is commonly obtained by scanning maps or collecting aerial photographs and satellite images. Thus, to use some raster datasets in conjunction with our other spatial data, we may need to align or geo reference them to a map coordinate system.

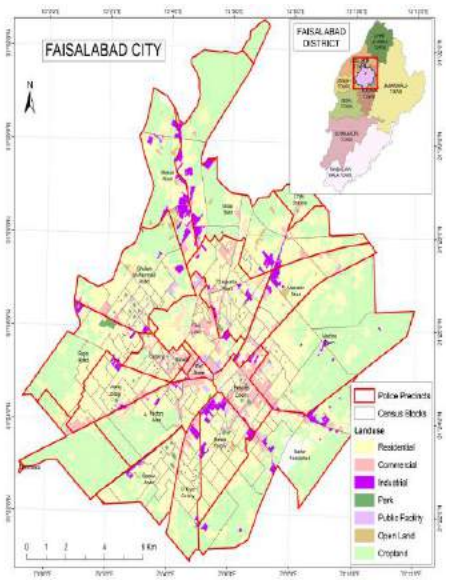

Figure 2: Study area Faisalabad

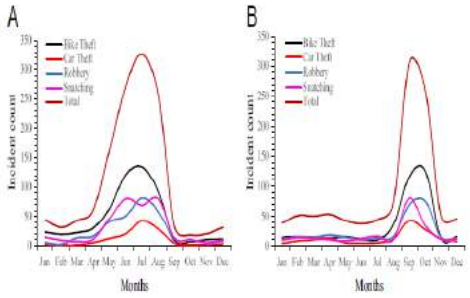

Figure 3: Crime incident 2015(a) and 2016 (b)

Table 1 A comparison of crime events *2015-2016*

Months	Bike Theft		Car Theft		Robbery		Snatching		Total	% age
	2015	*2016*	*2015*	*2016*	*2015*	*2016*	*2015*	*2016*		
January	23	14	2	4	5	10	14	0	72	3.50
February	19	15	1	0	2	4	9	0	50	2.43
March	22	4	1	0	13	4	7	0	51	2.48
April	30	3	5	0	16	8	11	0	73	3.55
May	64	5	12	0	41	4	47	0	173	8.40
June	121	1	20	2	51	0	80	2	277	13.45
July	134	10	42	2	81	3	69	6	347	16.85
August	92	35	28	11	58	15	80	10	329	15.98
September	10	116	4	42	5	69	13	80	339	16.46
October	7	126	1	27	0	75	10	34	280	13.60
November	10	3	2	0	3	4	5	1	28	1.36
December	11	5	3	2	8	1	9	1	40	1.94
Total	543	337	121	90	283	197	354	134	2059	100
% age	41.74	44.46	9.30	11.87	21.75	25.99	27.21	17.68	100%	

8. Mapping of Police station

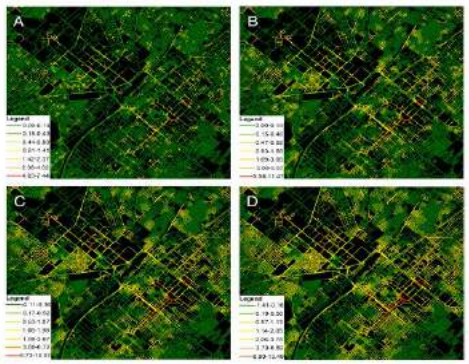

Figure 4: Police Stations and there boundaries (a)100m (b) 200m (c) 300m (d) 400m

9. Conclusion

Forecast about crime is a tall order; we are not on final stage where we define specific events by a special offender at specific movement in the crime. Using GIS into law enforcement has been an important dramatic for crime analyst and criminal justice researchers [28].To keep crime analysis and decision-making, we need to recognize the complex (spatial) clustering (block) analysis. Maps provide crime analyst graphics representation of crime-related problems. Perceiving where and why crimes occur, can improve the struggle to fight for crime.

Using good management techniques we can reduce crime rates. We need to follow new technology in the 21st century to prohibit crime [20]. Eventually, mapping and GIS support regional and complicated oriented policing GIS and Mapping can show the comprehensive correlation between the crime, the victim and the offenders. Some important facts of GIS and Crime mapping are: showing the probability and people changes, help in resourcing allocation, combining data from the government resources and community, providing effective communication tools.

Whatever approaches makes sense for you, applying and studying GIS into law and defense is a twice successful choice [21]. As you advance, your own career should make important addition for social freedom and order. Think of it as two benefits for one effort.

ACKNOWLEDGMENT

We authors acknowledge with thanks the assistance rendered by Dr. Javed Anjum Sheikh, University of Lahore, Gujrat Campus for providing crucial insight during the course of the research work which greatly improved the manuscript.

REFERENCES

[1] Y. Bello et al., "Principal Component Analysis of Crime Victimizations in Katsina Senatorial Zone", International Journal of Science and Technology, vol. 3, no. 4, pp. 192- 202, 2014.

[2] P. Yar and J. Nasir, "GIS Based Spatial and Temporal Analysis of Crimes, a Case Study of Mardan City, Pakistan", International Journal of Geosciences, vol. 7, no. 19, pp. 325- 334, 2016.

[3] T.F. Balogun et al., "Crime Mapping in Nigeria using GIS" , Journal of Geographic Information System, vol. 6, no. 5, pp. 453- 466, 2014.

[4] S.M. Ansari and Dr. K.V. Kale, "Methods of Crime Analysis using GIS", International Journal of Scientific and Engineering Research, vol. 5, no. 12, pp. 1330- 1336, 2014.

[5] U.S. Usman, "The Role of Geographic Information System (GIS) in Effective Control of Terrorism in Nigeria", International Journal of Economics, Commerce and Management, vol. 3, no. 4, pp. 1- 9, 2015.

[6] 2016. [Online]. Available: http://us.corwin.com/sites/default/files/upm-binaries/6244_Chapter_4_Boba_Final_PDF_3.pdf. [Accessed: 23- Nov- 2016].

[7] Crimemapping.info, 2016. [Online]. Available: http://crimemapping.info/wp-content/uploads/2015/07/CMN3PDFv4.pdf. [Accessed: 23- Nov- 2016].

[8] Esri.com, 2016. [Online]. Available: http://www.esri.com/software/arcgis/arcgisonline. [Accessed: 11- oct- 2016].

[9] "Crime Mapping and spatial Analysis", Itc.nl. [Online]. Available: http://www.itc.nl/library/papers_2003/msc/gfm/ahmadi.pdf. [Accessed: 27- Nov- 2016].

[10] F. Fajemirokun et al., "A GIS Approach to Crime Mapping and Management in Nigeria: A Case Study of Victoria Island Lagos", Nigeria.

[11] Crime Mapping & Analysis News", Crime Mapping and Analysis News, 2016. [Online]. Available: https://crimemapping.info/wp-content/uploads/2016/12/CMAN-Issue-5.pdf. [Accessed: 17- Oct- 2016].

[12] N. Levine, "Crime Mapping and the Crimestat Program", Wiley Online Library. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.0016 7363.2005.00673.x/full. [Accessed: 22- Nov- 2016].

[13] A. Ahmed & R. Saliha, "spatiotemporal pattern of crime using GIS approach in Dala L.G.A. Kano state, Nigeria", American Journal of Engineering Research, vol. 2, no. 3, 2013

[14] G.O. Igbaekemen & S.U. Umar, "A Purview into the historical Development of Terrorism in Nigeria", Journal of Developing country Studies, vol. 4, no. 14, 2014

[15] O. James, "GIS and Crime Management in Nigeria", International Journal of Sociology, Management and Security Studies, Maiden Edition, Kano, Nigeria, 2013

[16] F. Onu, "Corrupt Practices in Nigeria Society", A Search For Causes and Remedies. IJMSSS, Kano State, Nigeria, vol. 1, no. 1, 2014

[17] J. Shekwo, "Juvenile Delinquency in Mararaba, Karu L.G.A. of Nasarawa State, Nigeria", International Journal of Sociology, Management & Security Studies, Maiden Edition, Kano, Nigeria, 2013

[18] S.A. Yelwa & Y. Bello, "Complimenting GIS and Cluster Analysis in Assessing Property Crime in Katsina State, Nigeria", American International Journal of Contemporary Research, Vol. 2, no. 7, 2012

[19] S. Pires"crime mapping and analyses news",International Journal of Science and Technology, vol. 4, no. 5, pp. 1-30 , 2012.

[20] M.A.P chamikara.,"GIS in crime analysis", International Journal of Science and Technology, vol. 3, no. 6, pp. 3 , 2014.

[21] J. Bueermann.,"Crime analysis",Journal of Environment and Earth Science, vol. 2, no. 3, pp. 1-6 , 2012.

[22] M.brokmaan et al.,"Crime Mapping and Analysis in the Dansoman Police Subdivision", Journal of Environment and Earth Science, vol. 4, no. 3, pp. 1-11,2014.

[23] C.D.J beaty.,"GIS for Crime Analysis, Law Enforcement, GIS and Public Safety", Journal of Environment and Earth Science, vol. 4, no. 3, pp. 1-17,2012.

[24] T.Fransic et al.,"Crime Mapping in Nigeria",scrip.org.[Online].Available: http://www.scirp.org/journal/PaperInformation.aspx?PaperID=50296. [Accessed: 20- Nov- 2016].

[25] j.corso et al.," Toward Predictive Crime Analysis via Social Media, Big Data, and GIS", Journal of Environment and Earth Science, vol. 2, no. 3, pp. 1-6 , 2015.

[26] S.Muhammad et al.," Mapping and Analysis of Crime in Aurangabad City using GIS", IOSR Journal of Computer Engineering (IOSR-JCE), vol. 2, no. 4, pp. 67-76 , 2014.

[27] F.Wang.,"Why Police and Policing need GIS",Journal of Environment and Earth Science, vol. 2, no. 4, pp. 67-76 , 2012.

[28] T.Balogan et al.," Crime Mapping in Nigeria Using GIS",Journal of Geographic Information System, vol.6, no. 4, pp. 453-466 , 2014.

[29] J. H. Davis and J. R. Cogdell,"Crime in GIS," Elect. Eng. Res. Lab., Univ. Texas, Austin, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.

[30] R. E. Haskell and C. T. Case, "Hotspot in Crime mapping," USAF Cambridge Res. Labs., Cambridge, MA, Rep. ARCRL-66-234 (II), 1994, vol. 2.

[31] P. Diament and W. L. Lupatkin,"GIS in Crime Mapping," Dept. Elect. Eng., New York, Sci. Rep. 85, Aug. 1991

[32] Crime with GIS, 3rd ed., Western Electric Co.,Winston-Salem, NC, 2010, pp. 44–60.

[33] GIS in Crime mapping., 1st ed.,Spencer Chainey, Jerry Ratcliffe., 2010,pp.1-421.

[34] "Hartsfield-Jackson Atlanta International Airport," [Online]. Available: Http://Www.Esri.Com/Software/Arcgis/ Arcgisonline. Accessed: Nov. 25, 2016.

Volume 1, No. 2 | July – December 2017

Initiative for Thyroid Cancer Diagnosis: Decision Support System for Anaplastic Thyroid Cancer

Jamil Ahmed Chandio[1], M. Abdul Rehman Soomrani[1], Attaullah Sehito[1], Shafaq Siddiqui[1]

Abstract

Due to the high level exposure of biomedical image analysis, Medical image mining has become one of the well-established research area(s) of machine learning. AI (Artificial Intelligence) techniques have been vastly used to solve the complex classification problems of thyroid cancer. Since the persistence of copycat chromatin properties and unavailability of nuclei measurement techniques, it is really problem for doctors to determine the initial phases of nuclei enlargement and to assess the early changes of chromatin distribution. For example involvement of multiple transparent overlapping of nuclei may become the cause of confusion to infer the growth pattern of nuclei variations. Un-decidable nuclei eccentric properties may become one of the leading causes for misdiagnosis in Anaplast cancers. In-order to mitigate all above stated problems this paper proposes a novel methodology so called "Decision Support System for Anaplast Thyroid Cancer" and it proposes a medical data preparation algorithm AD (Analpast_Cancers) which helps to select the appropriate features of Anaplast cancers such as (1) enlargement of nuclei, (2) persistence of irregularity in nuclei and existence of hyper chromatin. Proposed methodology comprises over four major layers, the first layer deals with the noise reduction, detection of nuclei edges and object clusters. The Second layer selects the features of object of interest such as nuclei enlargement, irregularity and hyper chromatin. The Third layer constructs the decision model to extract the hidden patterns of disease associated variables and the final layer evaluates the performance evaluation by using confusion matrix, precision and recall measures. The overall classification accuracy is measured about 97.2% with 10-k fold cross validation.

Keywords**:** *biomedical image;algorithm;classification;dicision support system*

1. Introduction

Recently biomedical image inference of DICOM (Digital communication in medicine) images have been witnessed one of the active research area(s) of machine learning and AI (Artificial Intelligence) base techniques have shown significant impact upon the diagnostic process. Various CAD (computer added diagnosis) systems have been proposed to solve the classification problems of malignant diseases such as lung, breast, head & neck, lymphatic system, thyroid and other cancers. Some of the very nice approaches [1],[2],[3],[4] were proposed to solve the classification problem of cancer disease.

Infact, it is really one of the challenging field to identify the object of interest for different organs of human body as stated above because the classification of poorly differentiated, well differentiated and undifferentiated cancers have been found with lots of variations and divergent properties. For example, thyroid Anaplast cancer is one of the aggressive malignancy and its growth rate is higher than the other type of cancers.

Since the DICOM images of thyroid disease provides sufficient information to diagnose the Anaplast cancers but due to the use of poor staining material such as H and E, it may deceive doctors while examining the

[1] *Department of Computer Science Sukkur IBA University Sukkur, Pakistan*
Corresponding author: Jameelahmed.phdcs@iba-suk.edu.pk

mimic chromatin properties either it is hyper chromatin or not. Secondly, due to the unavailability of nuclei measurement techniques, at very early stage, irregular set of nuclei may produce point of confusion for doctors to determine the initial stage of enlargement from concerned group of nuclei. Involvement of multiple transparent overlapping nuclei may also be prone to one of the causes confusion to infer the growth pattern of nuclei grooves and early eccentric properties of nuclei are considered one of the difficult decisions because detachment of cell wall may be initial phase. In-order to solve all the above stated problems this paper proposes a system "Decision Support System for Anaplast Thyroid Cancer" to assist the doctors to diagnose the Anaplast cancer at early stage. Methodology of proposed system comprises over four layers and it proposes an algorithm for data preparation for Anaplast cancer [Algorithm 1]. In first layer noise reduction techniques have been used to reduce the noise of DICOM image by using adaptive threshold method and edges of objects have been detected by using canny edge detection algorithm whereas regional clusters have been tagged by using watershed algorithm. In second layer our proposed data preparation algorithm detects the properties and behaviors of Anaplast cancer such as enlargement of nuclei, irregularity of nuclei and hyper chromatin related features. In the third layer, random forest AI base technique has been used to construct the decision model to decide about the existence of cancerous material. In final layer performance evaluation have been conducted by observing confusion matrix, precision and recall measures. The overall performance has been shown by using AUC (area under curve). The measured accuracy of proposed system is about 97.20%. A real world dataset has been used for Anaplast cancers received from SMBBMU (Shaheed Muhtarma Benazir Bhutto Medical University) Pakistan as such datasets are unavailable in literature.

Rest of paper is organized in five sections. Introduction is presented in section one. Related works in section two. Proposed methodology is shown in section three. Results in section four and conclusion are described in section five.

2. Related Works

Basically this paper falls into the category of productive mining and deals with the classification problem of DICOM (Digital communication in medicine) images of biopsy. Specially; this paper proposes a preprocessing algorithm for Thyroid related Anaplast cancers which are most aggressive type of malignancy comprises over dissimilar growth patterns in terms of shape, size and other morphological properties, which are significantly important to diagnose at proper stage. Following related works have been seen in the recent past.

SVM (support vector machine) and AdaBoost machine learning techniques were compared for breast cancer classification problem [1]. The best classification accuracy was approximated as 87.42%. This paper presents a novel algorithm in data preparation phase and uses random forest AI base technique to construct the classification model for cancerous and non-cancerous thyroid Anaplast malignancies.

A system [2] using Convolutional neural network based machine learning technique was proposed for thyroid disease classification. The DICOM images need significant pre-processing techniques for every individual class of disease such as well-differentiated, poorly differentiated and others. All the cancer types may not be pre-processed with the same procedure because the key building blocks nuclei are considered most important micro architectural components, which usually have been found with different properties for every histo-pathological class. In data preparation phase proposed algorithm selects enlargement, irregularity and hyper chromatin related features. A comparative [3] study was conducted by using various segmentation algorithms based on clustering i.e. K-means and watershed than a supervised learning approach was used to construct the decision model obtained accuracies for template matching strategy were measured as

72% and 87%. Since the histopathology focuses on dissimilar micro structures of different human cells and tissues, whereas in a single organ various type of cancer diseases may be appeared because every disease has its own properties which needs significant efforts in data preparation phases.

On the concept of Neural Network [4] different AI base algorithms i.e. Scaled Conjugate Gradient, BFGS Quasi-Newton, Gradient Descent method and Bayesian regularization algorithms were used and best approximated accuracies for thyroid disease datasets were recorded as respectively 90.5%, 86.30% and 83.50%. Since the DICOM images of thyroid disease provide sufficient information to diagnose the Anaplast cancers, it deceives doctors when hyper chromatin may be observed with nearby properties due to the use of poor staining material and at very early stage irregular nuclei could not be seen properly due to the unavailability of nuclei measurement techniques, whereas enlargement of cells could easily be picked by the doctors but due to the involvement of multiple transparent overlapping nuclei in deceiving nature may create confused situations. It is direly needed to propose an efficient system with the support of systematic data preparation algorithm which recommends deepest decisions about the evidences of cancerous material at early stages.

3. Methodlogy

This paper proposes a system which deals with the predictive mining in the field of machine learning. The system comprises over four layers, in the first layer noise reductions have been done and in the second layer several disease related properties have been selected under the umbrella of behavior detection and feature selection. In the third layer classification models have been constructed and in the final layer performance evaluation have been done. Since the DICOM images of Anaplast cancers need significant methodology to pre-process, the nuclei behaviors of Anaplast cancers have dissimilar behaviors that are very difficult to select object of interest in proper way. For example enlargement of nuclei, irregularity in the nuclear patterns and variations in chromatin distribution are very difficult to interpret as digital set of objects. This paper also contributes a data preparation algorithm AD (Analpast_Cancers) which performs all the above stated tasks effectively and every observation is recorded carefully with the assistance of expert medical panel.

Algorithm 1: Anaplast_Diagnosis
Input: DICOM Dataset as D
Output: Enlargement, Irregularities, Hyperchromatin and Class Label Cancerous = Yes/No
$Datset \leftarrow Anaplast_{Cells} \|\Delta f\| \leftarrow \sqrt{\left(\frac{\partial f}{dx}\right)^2 + \left(\frac{\partial f}{dy}\right)^2}$
$Goto = every\ unit\ of\ dataset\ DasD_n\ (x_i, y_i)$
$Find\ \sigma^2 w(D) \leftarrow x1(D)\sigma_1^2 + y1(D)\sigma_1^2(D)$
$foreach P_i \in x_i(D) do$
  $Enlargement \leftarrow P \leftarrow g(x,y)^{x1(D)} \leq 0_n(P) \leq 1$
  $if\ g(x,y)^{x1(D)} = 1$
    Heterogeneity $\leftarrow g_h(\tau) = h_b(x) * g(x)$
  $Count \leftarrow \sum_{j=n}^{i=1} Size\ of\ p_i\ 1 + +||Enlargment||$
  $if g(x,y) \leftarrow Size\ of p_i = 0$
  $Count \leftarrow \sum_{j=n}^{i=1} p_i\ 0 + +||irragularity||$
  $if g(x,y) \leftarrow Size\ of p_i = p_i$
  $Count \leftarrow \sum_{j=n}^{i=1} p_i = p_i + +||$hyperchromatism$||$
  $endif$
$Return \leftarrow Enlargment, irragularity,$ hyperchromatism, Class Label $\leftarrow Cancerous = Yes|\ No_{Cells}$

Algorithm1: AD (Anaplast Diagnosis)

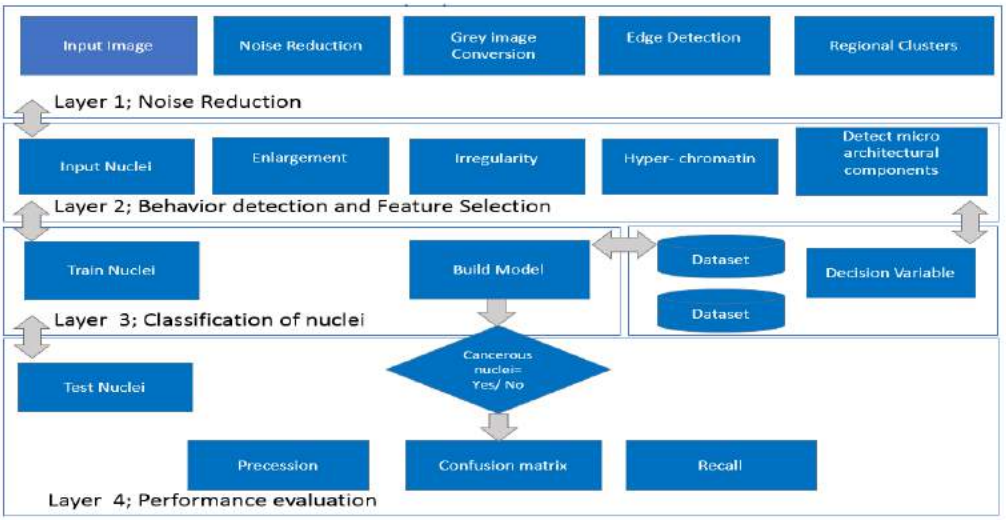

Figure 1: Decision Support System for Anaplast Thyroid Cancer workflow

3.1. Layer 1: Noise Reduction

Proposed algorithm has to perform three numbers of tasks. Firstly to detect the enlargement of cells. Secondly to find out the irregularity and third the hyper chromatin. Let us consider a medical image of biopsy consisting of the several number of nuclei and each nuclei occupies size, shape, color and many features. Computationally, the algorithm reduces the noise by applying the image pixel derivation as per eq. (1).

$$\|\Delta f\| \leftarrow \sqrt{\left(\frac{\partial f}{dx}\right)^2 + \left(\frac{\partial f}{dy}\right)^2} \qquad (1)$$

Medical image $M = \{x+h_1, x+h_2, x+h_3 \dots\dots\dots x+h_n\}$ is consisting upon the h spaces in vertical and $N = \{y+h_1, y+h_2, y+h_3 \dots\dots\dots y+h_n\}$ horizontal vectors. If complexity of huge vectors is selected as features of large size medical images, it may require high computational power to compute the correlation of the associated pixels. Proposed algorithm reduces the unnecessary information of medical image and detects the nuclei by eq. (1), where homogenous intensity based pixels regions are formed as foreground of second derivation. Since $g(x, y)$ is position of particular pixel p which represents the each nucleus considering homogenous threshold eq. (2).

$$g(x,y) \quad = (x_i, y_i) \leq 0 \quad (P) \leq 1 \; (2)$$

Proposed algorithm visits each pixel and selects appropriate set of weighted pixels p designated as regions and tagged as Enlargement of nuclei where shape of various nuclei behaviors are considered with same visual properties. the task of this subsection of algorithm is to extract the statistical morphological features. Additionally, canny edge detection algorithm is used to record the shape related changes between the set of nuclei in medical images.

3.2. Canny edge detection:

Nuclei edges provide additional information about the selected regions of enlarged objects where shape sizes could be recorded for further analysis and eq.(3) describes that all the edges of medical image objects are the lies between the $(G) = \sqrt{G(x_i, y_i)^2}$, since the G is targeted connected set of lines to be formed around the (x_i, y_i) spatial locations. G_x and G_y Angles could also be used to measure the directions of objects which are involved in expected expansion over multiple regions eq. (4). This

helps to quantify the max and min distances for overlap objects and by $(\frac{G_x}{G_y})$ dividing the spatial positions logical separation could be done.

$$Edge_{Nuclei}(G) = \sqrt{G(x_i, y_i)^{\ 2}} \quad (3)$$

$$Angle_{Nuclei\ (\theta)} = tan^{-1}(\frac{G_x}{G_y}) \quad (4)$$

3.3. Layer 2: Behavior detection and feature selection:

- Enlargement:

$$\boldsymbol{Enlargment} = \sum_{j=n}^{i=1} \boldsymbol{Sizeofp} \ = \mathbf{1} \quad (5)$$

Let's consider (x_i, y_i) are the spatial locations residing in the vector space of $f(G\ )$ where shape are not equal to the same size eq. (5). It is said to enlarge shapes that are counted with the assistance of auto image separation techniques as per eq (6). The image particles are considered as seed and each seed is cropped with its mean value and surrounded area is mapped with edges that the necessary information of particular object can be obtained.

$$\boldsymbol{seed}(\boldsymbol{\mu x}, \boldsymbol{y}) = \{(\frac{\boldsymbol{\pi_x}}{\boldsymbol{\pi_y}})\boldsymbol{Sum}(\boldsymbol{a}(\boldsymbol{x}, \boldsymbol{y})\} \quad (6)$$

- **Irregularity:**

Mostly the enlargement detection deceives in terms of object detection, since the proposed preprocessing technique measure the center of enlarged sequences of nuclei and rings acquired through canny edge detection are transformed around the large objects$(x_i + h_i, y_i + h_j)$ where (h_i, h_j) are transformed as boundaries of multiple objects.

$$\mathbf{X_{circle\ c}} = \frac{\mathbf{1}}{\mathbf{M}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i}\ \mathbf{m_i}(\frac{\boldsymbol{\pi}_{\mathbf{x+h}}}{\boldsymbol{\pi}_{\mathbf{y+h}}}) \quad (7)$$

The next task of our algorithm is to encircle the closed objects $X_{circle\ c}$ where every point is denoted by M and numbers of associated pixels have been calculated with aggregated values eq. (7). Since the locations $\left(x_i + h_i(x_i + h_i, y_i + h_j)\right)$ are encircled. The irregular margins M have been counted and recorded in each observation of training and testing datasets.

- **Hyper chromatin:**

The hyper chromatin is also known as access of chromatin could be found in cells. The cell wall of nuclei has irregular quantities of the chromatin fluid. It is a condition where nuclei could not able to maintain its shape related properties due to the detachment of surrounding nuclei walls and loss of nucleus properties at DNA levels. Computationally proposed subsection of algorithm records a high concentrated mean value of chromatin color movement based features and color deviation is quantified on the basis of mass value and recorded into the observation.

$$Hyperchromatin = \sqrt{\frac{\sum_{i=1}^{n^*}(y_i - \hat{y}_i)^2}{n^*}} \quad (8)$$

Hyper chromatin is one of the key feature in which each nuclei has to lose the chromatin. Since the loss of chromatin may incurred due to the eccentric nucleolus of nuclei or the expansion in the size of nuclei may be affected to nuclei shape, size and behaviors. Aggregated Color movements were measured with the assistance of distance matrix, where color spectrum is shown in [Figure 2] to represent the behaviors of chromatin in each nuclei with sounding set of nuclei. Let us suppose every DICOM image D is represented as collection of pixels contain the information based upon $X = \{X = x_1, x_2 \dots \dots \dots . x_n\}$ and$Y = \{y_1, y_2 \dots \dots \dots . y_n\}$ where each set of pixels $P(X, Y) = \{m_1, m_2 \dots \dots \dots . m_n\}$ qualifies an object with distinct set of features$F = \{f_1, f_2 \dots \dots \dots . f_n\}$ on$H = \{h_1, h_2 \dots \dots \dots . h_n\}$.

3.4. Layer 3: Classification of nuclei:

Random forests machine learning algorithm is widely used to predict the different classification problems due to

several advantages to regress the rank relationships of important variables. Deep decision trees have ability to aggregate more than one decision related to every class label attribute.

Let's consider DICOM dataset D consists having several no of important variables$D = \{(X_i, Y_i)\}\ _{i=1}^{n}$. the decision model can be constructed as aggregated decision variable by fitting by measuring the J_{th} feature in training stage, since the decision tree J_{th} features are distressed datasets because of aggregated permutation operations (out of bag error) conducted for overall trees.

$$\hat{y} = \sum_{i=1}^{n} W(x_i, \acute{x}) y_i \quad (9)$$

Let's suppose $\hat{y}$ is a class of functions in x′set of points which formalized by weighted function known as W. thus following weighted decision tree can be constructed as per eq. (9) and (10).

$$\hat{y} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} W_j(x_i, \acute{x}) y_i = \sum_{i=1}^{n}\left(\frac{1}{m}\sum_{j=1}^{m} W_j(x_i, \acute{x})\right) y_i \quad (10)$$

3.5. Layer 4: performance evaluation:

A total number of 20 biopsy images were used for training and testing purposes and 1829 number of nuclei were detected from those images. The confusion metrics [Table 1] shown for cancerous class which classified 948 observations and non-cancerous classified a number of 830 instances. . The precision eq. (11) and recall eq. (12) was approximated for cancerous classes 98.63% and about 97.73% was measured for non-cancerous classes.

$$Precision = \frac{NumberofTrePositives}{NumberofTruePositives + FalsePositives} \quad (11)$$

$$Recall = Sensitivity = \frac{TruePositive}{FalsePositive} \quad (12)$$

$$Sepecify = \frac{TrueNegetive}{TruePositive + FalsePositive} \quad (13)$$

4. Results

In [Figure 2] there are five columns. The column one is presented as an image input where all the images are belonging to the Anaplast thyroid cancer with different histo-pathological material H and E. In second column the points of individual cells have been shown, where each point denotes number of enlarge nuclei. In column number three the irregular nuclei are placed where each set of overlapping of nuclei are represented. The system generated circles show that there is several numbers of nuclei which are eccentric due to the loss of cell wall. In column number four hyper chromatin, level is shown by distance matrix in a color spectrum which reveals that the all those nuclei are considered as hyper chromatin where red color entropies are found with highest aggregated values. The ratio of chromatin is said to be hyper when the surrounded features are present with low red color ratio using color spectrum properties. Column five is placed to show the summarization of each observation. All these detected features have been recorded in a dataset for training and testing purposes.

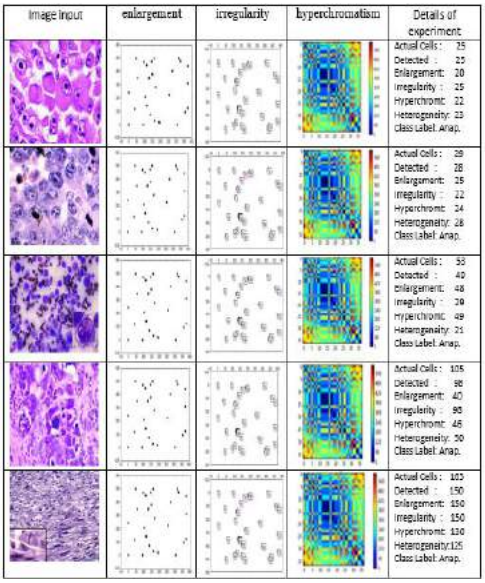

Figure 2: Results of preprocessing algorithm Anaplast Diagnosis

Table-I: Confusion matrix

	Cancerous	***Non-Cancerous***
Cancerous	948	31
Non-Cancerous	20	830

Table-II: Overall performance of proposed methodology

	Raw images	***No of Extracted nuclei***	***No of classified Nuclei***	***No of miss-classified Nuclei***	***Precession***	***Recall***
Cancerous	10	979	948	31	98.63%	98.83%
Non-Cancerous	10	850	830	20	97.73%	97.64%

Table-III: Comparison of our system with literature

Approaches	***Image Type***	***Cancer Type***	***Technique***	***Accuracy***
1	Ultrasound Image	Follicular	SVM	97.50 %
			AdaBoost	87.42%.
2	FNAC Images	Follicular	NN	91.00 %
3	FNAC Images	Medullary	Templated matching strategy	87.00%
4	FNAB Images	Papillary	Scaled Conjugate Gradient	90.5%,
			BFGS Quasi-Newton,	85.40%
			Gradient Descent method	86.30%
			Bayesian regularization	83.50%
Our Proposed approach	FNAB Images	Anaplast	Random forest	97.20%

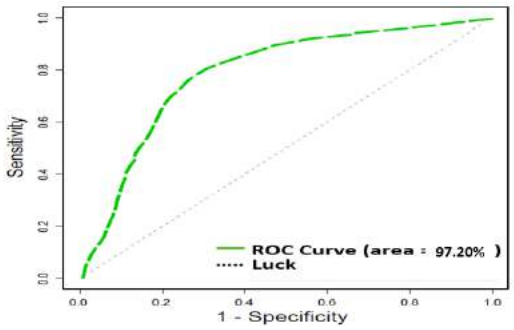

Figure 3: Estimated ROC Curve for the proposed

5. Conclusion

This paper contributes a novel data preparation algorithm AD (Anaplast Diagnosis) for Anaplast cancers, since the feature selection is one of the difficult tasks in medical images diagnosis. It needs significant efforts from the perspective of image feature engineering to select the related features from DICOM (Digital Communication in Medicine) images of thyroid biopsy such as enlargement of nuclei, irregular nuclei as individual & also in sequences and different chromatin distribution level features.

The methodology of proposed system comprises over four interconnected layers, where layer one has been assigned the job to reduce the noise and the layer two has been assigned core tasks to select appropriate features by measuring the aggregate values of involved pixels as color movements of selected objects. The layer three has been used to construct the classification model by using the random forest algorithm and the final layer has to perform the performance evaluation of the system.

The numbers of 1829 nuclei were detected from 20 images of cancerous and non-cancerous classes. The confusion matrix show that a total number of 948 instances out of 979 observations were classified as cancerous class and out of 850 instances 830 observations were classified for non-cancerous class.

The overall accuracy of the system was recorded as 97.2% with 10-k fold cross validation. Since such datasets are unavailable in literature, used dataset was received form SMBBMU, Pakistan. By observing the experiment in this research, we conclude that special data preparation algorithms are required to be developed for each histo-pathological medical image classification problem distinctly, because every DICOM image can be identified with its own properties which are mostly different in nature from each other.

ACKNOWLEDGEMENT

This research work is carried out due to motivational attitude of my beloved father. He remained seriously ill for ten years and left me alone with wet eyes, but even in tight situations he encouraged me to carry on my studies. I am also highly thankful to my teachers and medical partners, who guided and motivated me to keep my step forward to complete my studies

REFERENCES

[1] Lashkari, AmirEhsan, and Mohammad Firouzmand. "Early Breast Cancer Detection in Thermogram Images using Supervised and Unsupervised Algorithms." The Journal of Engineering Vol 7, No 3, pp. 114-124, (2016)

[2] Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M. and Chang, E.I. , May. "Deep learning of feature representation with multiple instance learning for medical image analysis". In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference pp. 1626-1630. (2014)

[3] Cheng Chen, Wei Wang, John A. Ozolek, and Gustavo K. Rohde., "A flexible and robust approach for segmenting cell nuclei from 2D microscopy images using supervised learning and template matching". Journal of Cytometry A. 2013 May ; 83(5): p.p 495–507, (2013)

[4] Pourahmad, S., Azad, M., Paydar, S., & Reza, H. "Prediction of malignancy in suspected thyroid tumour patients by three different methods of classification in data mining". In First International Conference on advanced information technologies and applications pp. 1-8, (2012).

[5] Glotsos, D, Tsantis, S. Kybic J, Daskalakis, A. "Pattern recognition based segmentation versus wavelet maxima chain edge representation for nuclei detectionin microscopy images of thyroid nodules". Euromedica medical center,

Department of Medical Imaging, Athens, Greece (2013).

[6] Gopinath, B, Shanthi, N.. "Support Vector Machine Based Diagnostic System for Thyroid Cancer using Statistical Texture Features", Asian Pacific J Cancer Prev, 14 (1), pp. 97-102. (2013).

[7] Gopinath, B, Shanthi, N. "Computer-aided diagnosis system for classifying benign and malignant thyroid nodules in multi-stained FNAB cytological images", Australas Phys Eng Sci Med Vol (36), pp. 219–230, (2013).

[8] Stavros T., "Improving diagnostic accuracy in the classification of thyroid cancer by combining quantitative information extracted from both ultrasound and cytological images". 1st IC-SCCE-Athens, pp. 8-10, (2004).

[9] http://www.thyroidmanager.org/chapter/fine-needle-aspiration-biopsy-of-the-thyroid-gland/ accessed on 15-01-2014.

[10] Thyroid Cancer: Web Site, http://www.medicinenet.com/thyroid_cancer/article.htm accessed on 15-01-2014

[11] Han J., Kamber M., and Pei P. "Data Mining: Concepts and Techniques",. 3rd (ed.) The Morgan Kaufmann Series in Data Management Systems (2011).

[12] Rafeal, C. G., Richard, E. W., (2007). "Digital Image Processing" , 3rd (ed). Prentice Hall. (2007).

[13] Hussein, Mohamed, Amitabh Varshney, and Larry Davis. "On implementing graph cuts on cuda." First Workshop on General Purpose Processing on Graphics Processing Units. Vol. (1). (2007).

[14] Boykov, Y., & Funka-Lea, G. "Graph cuts and efficient ND image segmentation". International journal of computer vision, 70(2), pp. 109-131, (2006).

[15] Malik, J., Belongie, S., Leung, T., & Shi, J. "Contour and texture analysis for image segmentation". In Perceptual Organization for artificial vision systems Vol(1) pp. 139-172. (2000).

[16] Al-Kofahi, Y., Lassoued, W., Lee, W., & Roysam, B. "Improved automatic detection and segmentation of cell nuclei in histopathology images".Biomedical Engineering, IEEE Transactions on, 57(4), pp. 841-852. (2010).

[17] Liu, X., Huo, Z., & Zhang, J. "Automated segmentation of breast lesions in ultrasound images", In Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the IEEE pp. 7433-7435, (2006).

[18] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350. (1963).

Volume 1 | No. 2 | July – December 2017

Schema Integration of Web Tables (SIWeT)

Nayyer Masood[1*], Amna Bibi[2], Muhammad Arshad Islam[1]

Abstract:

Schema integration has been mainly applied in database environment whether schemas to be integrated belonged to a single organization or multiple ones. Schema extraction is a relatively new area where schema is extracted from a web table. The extracted schema is not as much concretely defined as in a typical database environment. The work in this paper brings two areas together where extracted schemas from multiple web tables are integrated to form a global schema. The data are also extracted from web tables and placed into global table. This creates a large repository of data of the same domain extracted dynamically from websites which is then available for different types of ad-hoc queries. This work also imposes challenges on schema integration to be studied in the context of schema extraction and other way round.

Keywords*: Schema Extraction, Schema Iintegration, Semantic Heterogeneities, Web Tables.*

1. Introduction

Web is a tremendous source of huge volume of data which is in structured, semi-structured and unstructured formats. Structured data include lists and tables. A table is a collection of rows containing data in one or multiple columns. Each column represents an attribute. Web tables are a simple, meaningful, effective and popular way of representing data on web. Data from these sources can be accessed either using search engines or navigating through the web pages. However, structures (or schemas) of these web tables are not available, so they cannot be queried efficiently; hence their data mostly remain inaccessible to the users.

Extracting data from these tables and storing it in a database could be very useful for many value added services, like in business and social web domains. For example, in business domain, data extraction techniques help in reducing time and manpower and increase the business efficiency. This data can help the analysts and manager to revise or change the business strategies and plans. Context-aware advertising, customer care, comparative shopping, Meta query, opinion mining and database building are the major applications of web data extraction techniques [1].

Schema extraction is the process of extracting schema from the structured data (e.g. tables, spreadsheets) on the web. The process is followed by fetching the also data from these web tables. The extracted schema is used to create tables in a database, which are populated with the extracted data. The database tables can then be used for better and efficient querying.

Many sites on the web can be found that belong to the same domain. For example, sites from the banking domain, education, entertainment etc. We can apply schema extraction on multiple sites of the same domain and can store tables extracted from these sites at a single place for efficient querying. However, even in this case the query will be applied to individual tables if data are to be accessed from multiple sites. It will be more beneficial if schema integration could be applied on schemas extracted from the multiple sites of the same domain. Schema integration is the process of merging/combining same or similar data items to build a canonical schema. For example,

[1] Department of Computer Science, Capital University of Science & Technology, Islamabad, Pakistan
[*] Corresponding Author: nayyer@cust.edu.pk
[2] Department of Computer Science, Virtual University Rawalpindi, Pakistan

every university website shows its faculty information generally including faculty member's name, his/her designation, higher degree and research interests etc. Many of the university websites show this data in the form of web tables due to ease in creating, visualizing and understanding of tables. If schema extraction is applied to multiple universities' web sites to fetch the schema and data and store them in database tables; further schema integration can be on these database tables to get a canonical schema. This will give us a single schema/table containing data of faculty members belonging to different universities. We can then apply advanced/efficient queries to extract required information about faculty members of different universities.

Structure of this paper is as follows: section 2 presents the review of the related literature encompassing both schema extraction and schema integration. Section 3 presents proposed approach comprising three phases. Finally, section 4 concludes the paper.

2. Literature Review

The review of related literature falls into two major categories: schema extraction and schema integration. Both of them are discussed in the following:

2.1. Schema Extraction

Several approaches are available for schema extraction that can broadly be categorized as manual, wrapper induction and automatic extraction. In wrapper induction [2], [3], firstly the pages/data records are labeled manually and a set of extraction rules is deducted. These rules are then used to extract data from similar pages. This technique still involves manual effort. Automatic method [4], [6] finds patterns or grammars from similar pages containing similar data records and uses these patterns to extract data from new pages. The pages to extract the patterns are provided manually or by some other system.

The approach of Zhai & Liu [6] provides an automated system to extract data from web and put it in a database. Web pages are constructed using HTML tags. The <table> tag is used to represent table on web. The <tr> tag is used to insert the rows and <td> tag inserts the data in a particular cell of that row.

The proposed approach firstly identifies the data records. For this purpose, visual information is used to construct a tag tree which is constructed by following the nested structure of HTML code. Second task is to align and extract data from the identified data records using partial tree alignment technique. Tree edit distance is used to identify the data records. Trees are matched with each other node by node. Trees are aligned by gradually growing a seed tree Ts. The tree with the maximum records is chosen as the starting seed tree. Then for each node n_i in T*i* a matching node n_s is found in T_s. When a matching node is found in T_s, a link from n_i to n_s is created. If no match is found for n_i then seed tree is expanded by inserting this node into it. Data item nodes are not used during the matching. The data item in the matched nodes children is inserted into one column in database table.

Adelfio & Semat [4] proposed a conditional random field (CRF) based classification technique with the combination of logarithmic binning. Each web table contains different types of rows, like caption, heading, empty and data rows etc. Each row has been classified based upon the row features which include formatting, layout, style and values of the cell and then all these features are combined using binning to construct row features. In next step, logarithmic binning method is applied in which individual cell attributes are used collectively to encode row features. For each possible row feature a bin is formed and each bin is assigned a value which represents its feature. After row features extraction, row labels are assigned to each row based on CRF. CRF is trained with human classified rows. After training the CRF is used to label huge volume of data. The output of the CRF is a sequence of row labels like "TNNHGDDDAGDDDABN". This output helps in extracting schema of the relational

table. Column names are decided based upon the header row(s), data type is determined by the type frequency within the data rows of each column, additional attributes can be determined by the group header rows and data rows are determined by the data records.

George, David and Sharad [7] introduced a technique to covert the web tables to relational tables. This is the first end to end approach which produces an access compatible canonical table. The HTML table is converted into an excel table and from excel table its CSV file is generated. Table is segmented based upon the indexing property rather on appearance features. To segment the table minimum indexing point (MIP) and four critical cells CC1, CC2, CC3 and CC4 are calculated. CC1 and CC2 determine the stub headers; CC3 and CC4 indicate the data regions. MIP (CC2) is determined by searching from the cell A1 for unique columns and row header rows. The categories can be extracted by comparing the number of unique elements in the cross-product of a pair or header rows with the length of the header. From the category extraction output, canonical table is generated. This table can be used to query the data.

The technique proposed by Purnamasari, Wicaksana, Harmanto and Banowosari in [8] first finds the area of the table and then extracts data from it. First of all table is detected and then property (title) portion of the table is detected before extracting data from it. The technique is divided into three steps and algorithm for each step is formulated. In first step, number of rows and columns are calculated by counting the <tr>... </tr> tags in <table> tag and the <td>…</td> tags in each <tr> tag. The algorithm also checks the colspan attribute in the <td> tag. It adds the value of colsapn in the column count. In second algorithm the property of the table is detected. Generally the first row of the table contains the headings of the columns. The algorithm checks for the row span attribute in each <td> tag in <tr> tag of table to calculate the length of the property of the table. Third algorithm actually extracts the data from the table. It takes the value of the rowspan returned from the second algorithm to extract the heading of the columns. While reading the data in <td> tag of <tr> tag, it checks the value of colspan. If its value is greater than 1, it concatenates the content in this cell with the columns below it. After reading the header rows, it reads the cells row by row.

2.2. Schema Integration

It is the process that takes two or more schemas as input (called source or member schemas) and merges them into a single/canonical one. Other terms used for schema integration are database integration, data integration, database interoperability, etc. The most critical step in schema integration is schema matching in which two schemas are compared with each other to identify the elements modeling same or similar concepts. Once identified, the similar elements are merged into a common one in the schema merging phase, which is a relatively straightforward task. The main problem in schema matching is identification and resolution of semantic heterogeneities. A semantic heterogeneity reflects a situation when same or similar concept is modeled differently in two schemas. These differences arise mainly due to differences in the context of organization, popularity of using acronyms in defining schemas, idiosyncratic abbreviations and models of the same domain [11]. The schema integration approaches can be broadly categorized into two; schema based and instance based.

Schema based integration approaches exploit the information in the schema, like, name, data type and different types of constraints, to identify semantically similar schema elements. These techniques have been further classified as element-level and structure-level in [13]. Element-level schema matching approaches compare the entity types from different schemas in isolation without considering their links/relationships with other entity types. These approaches mainly include string-based techniques [14], [15] that use matchers like prefix, suffix, edit-distance

and N-gram; NLP-based techniques [16, 17] that apply natural language processing techniques, like tokenization, lemmatization and elimination on the names of the entity types and then apply some matching technique; constraint-based techniques [18] where constraints from schema are used to compare the schema elements, like data type, integrity constraints etc. Structure-level approaches mainly cover graph-based [19], taxonomy-based [20] and model-based [21]. A hybrid approach has been adopted in COMA++ [22], where a combination of matchers is applied on input schemas and the results are combined to establish final similarity between elements.

Table I. Comparison of Different SE Approaches

S #	Paper Reference	Schema Extraction	Fully Automated	File Format	Techniques	Data Domain	Multiple Sources
1	4	Yes	Yes	HTML, Table, Spreadsheet	Supervised	Different domain	No
2	6	No	Yes	HTML table	Tree based	Shopping data	No
3	7	No	Yes	HTML, table, Spreadsheet	Index Based	Statistical data	No
4	8	No	Yes	HTML tables	Programming	Not mentioned	No
5	9	No	Yes	HTML tables	Tree based	Different domains	No
6	10	No	Yes	HTML tables	Tree based	Different domains	No

The literature review of schema merging approaches reveals that most of the approaches strive to maximize the automated part of the process as performing SI completely manually is very time consuming and laborious task. Moreover, most critical part of SI process is handling semantic heterogeneities which exist across multiple schemas due to the fact that these schemas are built independently in certain contexts that are entirely different from each other even if they belong to same domain.

3. Proposed Approach

This article presents the novel idea of applying schema integration (SI) process on the schemas that have been extracted through schema extraction (SE) process from web tables of multiple web sites belonging to the same domain. To prove the concept, it is planned to test proposed approach on the domain of faculty members of computer science departments of different universities. However, the idea can be applied in any domain. The basic idea behind this approach is to access those websites where the data of faculty members have been placed in the form of tables, as shown in Fig. 1 below. Then, using the SE approach presented in [8], extract the schema and data from different websites. After that, different schema matching

approaches will be applied to identify semantically similar elements among the elements extracted from different universities websites.

The semantically similar elements are merged with each other and the data are finally stored in a single table for further queries. The proposed approach comprises three major phases; preprocessing, schema extraction and schema integration. In the following, these three phases have been explained.

3.1. Preprocessing

Basic objective of this step is to provide neat and clean web table source to SE step so that an accurate schema could be extracted out of it. Neat and clean web page source means removing all unnecessary or irrelevant code or tags from the web page source that means any source or tags other than that contains the web table including table headings and data.

- Faculty of CS & IT

Name	Designation	Qualification
Miss Marium Butt	HOD CSIT	MSCS UOL
Mr. Mohtishim Siddique	Lecturer	MSCS, MIT (MUL)
Mr. Saleem Akhtar	Lecturer	MSCS , M.Sc-IT(P.U), M.Sc. Mathematics(PU)
Mr. Sheraz Tariq	Lecturer	M.Phil CS Scholar MUL
Mr. Muhammad Hussain	Lecturer	MS Computer Sciences UOL
Mr. Muhammad Tahir Jan	Lecturer	M.Sc. CS UET
Mr. Irfan Shahzad	Lecturer	MSCS UET, MIT (MUL)
Dr. Muhammad Adeel Talib	Assistant Professor	Ph.D Information Engineering
Mr. Ghulam Yasin	Lecturer	M.Phil Scholar UOL

Figure 1: An Example Web Table of Faculty Data

It is a critical and difficult task, as there is too much difference the way web tables are defined on different web sites.

In the first step of preprocessing phase, web sites of universities will be found manually that store the faculty data in the web table form and store that URL in a database table along with the other basic information about the university and department. This is an ongoing process and the database of university pages will keep on increasing. Once we have that data, this table will be handed over to a crawler which picks the URL of websites one by one and downloads the source code of the web pages and stores it in a text file. In the next step, clipping is performed and additional or irrelevant code/tags are removed and only the part contained within <table> and <\table> tags are left that contains the web table. This is going to be a bit tricky, as a web page generally contains many tables (for example, for formatting purpose) and out of those tables the one that presumably contains required data will be picked. One possible strategy in this regard can be, to pick the table that contains multiple rows and within each row there are multiple columns; this is also a requirement of SE approach [8] that has been selected in proposed approach. As an example, parts of HTML code of two web tables (after clipping) have been shown in Fig. 2 below. Both of these pages present the faculty data in the form of a table, but the variation in the coding can still be seen as the code in the right column

contains a lot of formatting instructions whereas one on the left simply contains the data inside HTML tags. The SE approach that we have selected [8] assumes web table to be in a specific format (fig. 1). However, it is possible that a website does contain the <table>, <\table> tags but still is not in the required format. So one objective of preprocessing phase is to identify such pages and put them aside rather than passing those to next phase because the adopted approach will not be able to successfully extract schema out of such pages.

```
<table>
  <thead>
<th>Staff Name</th>
          <th>Staff
Designation</th>
          <th>Staff Image</th>
       </thead>
          <tr>
               <td>Dr. Muhammad
Anwar-ul-Rehman Pasha</td>
               <td>Professor</td>
               <td></td>
          </tr>
          <tr>
<td>Mr. Abid Rafique</td>
<td>Assistant Professor</td>
          <td></td>
          </tr>
          <tr>
          <td>Dr. Muhammad Din
Choudhry</td>
               <td>Assistant
Professor</td>
               <td></td>
          </tr>
          ……..
               <\table>
```

```
<table class="MsoTableGrid" border="1"
    cellspacing="0" cellpadding="0" width="100%"
    style="width: 100%; border-collapse: collapse;
    border: sube;">
    <tbody>
    <tr style="height: 18.4pt;">
 <td style="width: 10%; border: 1pt dotted
      windowtext; padding: 0in 5.4pt; height:
      18.4pt; background-color: #8db3e2;">
           <p class="MsoNoSpacing" style="text-
align:
                center;"><strong><span style="font-
size:
                9 pt; font-family: arial, sans-serif;">
 S.No<o:p></o:p></span></strong> </p></td>
  <td style="width: 56%; border-style: dotted dotted
   dotted none; border-top-color: windowtext;
   border-right-color:  windowtext; border-bottom-
   color: windowtext; border-top-width: 1pt; border-
   right-width: 1pt; border-bottom-width: 1pt;
   padding: 0in 5.4pt; height: 18.4pt; background-
   color: #8db3e2;">
   <p class="MsoNoSpacing" style="text-align:
      center;"><strong><span style="font-size: 9pt;
      font-family: arial, sans- serif;">
      Name<o:p></o:p></span></strong></p></td>
    ……
</table>
```

Figure 2: Sample Code for Two Different Web Tables

3.2. Schema Extraction

There are many approaches proposed for SE in literature, but we have selected one proposed in [8] because it is quite recent, simple to understand and implement, moreover, it performs comparatively well on the web pages that are in the specific format assumed by the approach. There are many extensions possible in this approach, but our main objective is to apply SE for integration purposes, so we are using approach of [8] as such rather than suggesting any enhancement.

The SE approach that we have selected assumes that the web table is contained inside <table>, <\table> tags, and there are multiple rows between <table> <\table> tags. The first row contains the headings of the columns and remaining rows contain the data. One special

feature of the approach is that it can manage the situations where header is spanned upon multiple rows. In this paper, we are not discussing SE approach of [8] in detail; interested readers can refer the actual paper. The preprocessing phase has already placed the clipped web table code in a text file. In this phase, SE will be applied to all web tables stored in text files and from each file the first row is separated as header row and remaining as data rows. Inside header rows there are different column names that are separately stored in an array. The n column names in the header row of the web table are stored in the first n elements of n array. Remaining rows of the web table are assumed to contain data. So data values from each row are stored in the next n elements of array. This process continues till all the rows of the web tables have been processed. At the end of this process, we have an array in which first n elements contain the names of the columns and each next set of n array components contain attribute values.

3.3. Schema Integration

Schema Integration (SI) is the third phase of our approach (SIWeT) where the extracted schemas in the previous phase will be merged to form a global schema. As mentioned earlier, semantic heterogeneities is the major problem faced in SI. This problem is further amplified when SI is implemented on extracted schemas where we do not have concretely defined schemas rather extracted from a web page by a semi-automated tool. Such schemas may have certain errors that may not exist in properly defined database schemas. Like, extracted schemas may have inappropriate or invalid data type assigned to an attribute because data types are assigned to attributes on the basis of evaluation of data, for example, a web table may contain date in number format like '021002' representing October 02, 2012, but SE approach may assign it numeric seeing all numeric data. There can be other such issues that make SI in extracted schema environment more complex as compared to a traditional database environment.

The SI approach in our SIWeT comprises applying multiple schema matchers on extracted schemas and then combining the output of these matchers. First of all, we define a global table for faculty members. This table is defined with maximum possible attributes that can be relevant to a faculty member. The SI task then becomes finding the most similar attribute in this global table for each of the attribute in every extracted schema. In order to find similarity between attributes, SIWeT builds taxonomy of similar terms using existing resources, like WordNet. In addition to this taxonomy, N-gram and edit-distance matchers are also applied. These matchers return the score of similarity between different terms. There scores are averaged and the pair having maximum similarity are considered as similar to each other. When corresponding attributes of extracted schemas have been found within the global table, then the data from web table will be inserted into the global table under the attributes found similar to the attributes of the extracted table, along with two additional attributes mentioning the university Id and department name. This process will be applied to all the extracted web tables from different universities and we will have a global table containing data containing data from many universities at a single place.

We plan to build a web application that lets the users query this table using simple keywords or SQL queries. This will be a great source of information for researchers and other interested users to find the relevant data.

4. Conclusion

In this paper, we have proposed application of schema integration approaches on the schemas extracted from web tables; so this work is basically a merger of two research areas, that is, SE and SI. This merger extends both areas as SE has been mostly applied on a single site in literature, whereas we are applying it on multiple sites. Our approach also extends SI research as the process has been mainly applied in database environment

where properly defined database schemas are available defined through DBMSs, whereas we are applying it on extracted schemas. This will help to establish new dimensions in SI.

In future, we plan to implement our approach in real environment by extracting data from large number of web tables and merging them into a single table. The SE approach that we have adopted works on web tables in a specific format, there are many other formats of the web tables on which this approach cannot be applied. The SE approach can be extended to handle other web table formats. There are many web tables that store multiple attributes in a single column; we need to evaluate the data extracted from one column to identify relevant attributes.

REFERENCES

[1] Ferrara, E., De Meo, P., Fiumara, G., Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-based systems*, 70, 301-323.

[2] Cohen, W. W., Hurst, M., Jensen, L. S. (2002). A flexible learning system for wrapping tables and lists in HTML documents. *In Proceedings of the 11th international conference on World Wide Web* (pp. 232-241).

[3] Kushmerick, N. (2000). Wrapper induction: Efficiency and expressiveness. *Artificial Intelligence*, 118(1-2), 15-68.

[4] Adelfio, M. D., Samet, H. (2013). Schema extraction for tabular data on the web. *In Proceedings of the VLDB Endowment*, 6(6), (pp 421-432).

[5] Zeeshanuddin, S. (2011) A library of schema matching algorithms for dataspace management systems.

[6] Zhai, Y., Liu, B. (2005). Web data extraction based on partial tree alignment. *In Proceedings of the 14th international conference on World Wide Web* (pp. 76-85).

[7] Nagy, G., Embley, D. W., Seth, S. (2014). End-to-end conversion of HTML tables for populating a relational database. *In 11th IEEE International Workshop on Document Analysis Systems (DAS)* (pp. 222-226).

[8] Purnamasari, D., Wicaksana, I. W. S., Harmanto, S., Banowosari, L. Y. (2015). HTML table wrapper based on table components. *International Journal of Computer Applications in Technology*, *52(4)*, 237-243.

[9] Lerman, K., Knoblock, C., Minton, S. (2001). Automatic data extraction from lists and tables in web sources. *In IJCAI-2001 Workshop on Adaptive Text Extraction and Mining* (Vol. 98).

[10] Gultom, R. A., Sari, R. F., Budiardjo, B. (2011). Proposing the new Algorithm and Technique Development for Integrating Web Table Extraction and Building a Mashup. *Journal of Computer Science*, *7(2)*, 129-136.

[11] Lukyanenko, R., Evermann, J. (2011). A Survey of Cognitive Theories to Support Data Integration. *In Proceedings of the Seventeenth Americas Conference on Information Systems*, All Submissions. Paper 30.

[12] Evermann, J. (2009) Theories of meaning in schema matching: An exploratory study. *Information Systems 34(1)*, 28-44.

[13] Shvaiko, P., Jérôme, E. (2005). A survey of schema-based matching approaches. *Journal on Data Semantics IV*, (pp. 146-171). Berlin, Heidelberg: Springer-Verlag.

[14] W. Cohen, P. Ravikumar, and S. Fienberg (2003). A comparison of string metrics for matching names and records. *In Proceedings of the workshop on Data Cleaning and Object Consolidation at the International Conference on Knowledge Discovery and Data Mining (KDD)* (Vol. 3, pp. 73-78).

[15] H. H. Do and E. Rahm (2001). COMA - a system for flexible combination of schema matching approaches. *In Proceedings of the Very Large Data Bases Conference (VLDB)* (pp. 610–621).

[16] F. Giunchiglia, P. Shvaiko, and M. Yatskevich (2004). S-Match: an algorithm and an implementation of semantic matching. In Proceedings of the European Semantic Web Symposium (ESWS), (pp 61–75).

[17] J. Madhavan, P. Bernstein, and E. Rahm (2001). Generic schema matching with Cupid. In Proceedings of the Very Large Data Bases Conference (VLDB), (pp 49–58).

[18] P. Valtchev and J. Euzenat (1997). Dissimilarity measure for collections of objects and values. Lecture Notes in Computer Science, 1280, (pp 259–272).

[19] D. Shasha, J. T. L. Wang, and R. Giugno (2002). Algorithmics and applications of tree and graph searching. In Proceedings of the Symposium on Principles of Database Systems (PODS), (pp 39–52).

[20] N. Noy and M. Musen (2001). Anchor-PROMPT: using non-local context for semantic matching. In Proceedings of the workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI), (pp 63–70).

[21] P. Bouquet, L. Serafini, and S. Zanobini (2003). Semantic coordination: A new approach and an application. In Proceedings of the International Semantic Web Conference (ISWC), (pp 130–145).

[22] D. Aum¨uller, H. H. Do, S. Massmann, and E. Rahm (205). Schema and ontology matching with COMA++. In Proceedings of the International Conference on Management of Data (SIGMOD), Software Demonstration.

Volume 1 | No. 2 | July – December 2017

Software Atom: An Approach towards Software Components Structuring to Improve Reusability

Muhammad Hussain Mughal[1], Zubair Ahmed Shaikh[2]

Abstract:

Diversity of application domain compelled to design sustainable classification scheme for significantly amassing software repository. The atomic reusable software components are articulated to improve the software component reusability in volatile industry. Numerous approaches of software classification have been proposed over past decades. Each approach has some limitations related to coupling and cohesion. In this paper, we proposed a novel approach by constituting the software based on radical functionalities to improve software reusability. We analyze the element's semantics in Periodic Table used in chemistry to design our classification approach, and present this approach using tree-based classification to curtail software repository search space complexity and further refined based on semantic search techniques. We developed a Global unique Identifier (GUID) for indexing the functions and related components. We have exploited the correlation between chemistry element and software elements to simulate one to one mapping between them. Our approach is inspired from sustainability chemical periodic table. We have proposed software periodic table (SPT) representing atomic software components extracted from real application software. Based on SPT classified repository tree parsing & extraction to enable the user to program their software by customizing the ingredients of software requirements. The classified repository of software ingredients assists user to exploit their requirements to software engineer and enables requirement engineer to develop a rapid large-scale prototype with great essence. Furthermore, we would predict the usability of the categorized repository based on feedback of users. The continuous evolution of that proposed repository will be fine-tuned based on utilization and SPT would be gradually optimized by ant colony optimization techniques. Succinctly would provoke automating the software development process.

Keywords**:** *Classification, Development, Prototyping, Extraction, Parsing, Re-usability, Software, Software Periodic Table (SPT), Softwares Repository.*

1. Introduction

Software industry is growing swiftly. Computing devices are interacting with human through software program. From personal assistant to business management and ubiquitous computation services, software gives solution for everything by automation that improves the efficiency and accuracy. In this emerging technologically evolving world significantly transformed the software development and diversity of software products, software organizations need to meet the market and their client requirements within short duration and with optimal quality. With increase in magnitude and complexity of the project the maintainability becomes difficult[1]. Software reuse does not only saves time and cost, but also give us reliable software product by integrating tested and reliable software components. We do not need to develop software from scratch, we extract the software components, which meet the

[1] Center for Research in Ubiquitous Computing, Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan.
[2] Center for Research in Ubiquitous Computing, Department of Computer Science, Muhammad Ali Jinnah University, Karachi, Pakistan.
*Corresponding author: muhammad.hussain@iba-suk.edu.pk

requirements, and the software is ready for use in short compelled by industry pressure.

Software product line [2, 3] a family based approach software contributed a lot for reusability. Software industry acceleration is not chaseable. There is huge collection of software component developed already. To reuse software artifacts, we need a classification scheme to classify our software repository from that enable developer easily search and retrieve software artifacts based on project requirements. We proposed classification scheme based on domains, attributes and, features and functions of software system. Classification of software [4] allows us to organize collections of software artifacts into a efficient searchable structure. In last decade, various techniques applied for software artifacts classification.

Our motivation behind this work is the "elements periodic table" in chemistry, where all the atoms exist in this world are classified. Everything that we see around us is either compound or mixture of these atoms. Similarly, through identification of the atomic functionalities of existing software, and their transition to other software by mutation of their internal functional composition we can develop a relation between software components. We will develop general classification scheme for software, and this approach will revolutionize the software reuse by identifying the pattern of basic functionalities in a software. We can pick these basic functionalities from our repository and create new software from existing software. Moreover, it will assist developer for RAD, prototyping, and even user-developed software.

The organization of remaining paper is as below: In section II, explains the related work and some classification schemes. Section III defines problem statement. We will explain our proposed approach in section IV. Section V, implementation, and VI reflect the integration of the semantic model approach along with function coding scheme. Section VII is conclusion of our work and its limitations.

2. Related Work

Over the years, many researchers have proposed different classification schemes for software reuse. Following are some crucial approaches, which we have covered in our literature review.

Rajender Nath and Harish Kumar[5] used the keyword base search for software reuse. Their approach has three parts. For storage, the software component is stored in the form of component files, and an index is maintained which has the keywords related to the component. The authors in [6] proposed an approach for efficient software component retrieval for reuse. They suggested software component retrieval system based on ontology, metadata, and faceted classification for storage and retrieval of components. [7] Proposed an approach to use automatic classification of software identifiers by using text based classification algorithms such as N-gram to split the identifiers and extract domain knowledge concepts. They have reviewed many text based classification algorithms and proposed their own algorithm called sword. The algorithm creates a set of candidate splitting terms and matches these terms with the list abbreviation found and analyzed in source code, and a list of words from dictionary. The terms, which are fully matched are retained from the list and are ranked using the score function from samurai algorithm [6]. Software reuse library is organized using faceted classification[8] scheme. Search and retrieval of different software artifacts and library functions is very effective in this system. It is very difficult to organize a reusable software artifact that is why they have used the faceted classification scheme. This scheme gives higher level of accuracy and flexibility in classification. The limiting factor of technology used is its manual classification nature. They have also put some limitations on their reuse infrastructure. This paper has important role of domain analysis. Integration

of reusable artifacts and their adaptation in the system is very difficult unless a high-level system is not proposed.

In [9], Lo, D. et al describes the technique depending on software reusable components are creation, management and extraction. Identifications for software function for reuse based on specification were used in[10]. Prieto-Diaz and Freeman [11] have proposed a software reuse library system, which they called Reuse Description Formalism (RDF), improves organization of software components. They have proposed two concepts forming the core of RDF's model: instance and classes. Instances include description of objects, which are reusable. Lo, Cheng [12], the step towards reliability of software using pattern mining techniques. They introduce classifier to generalize the failures and to identify the other unknown failures of it. Zina Houhamdi [13] defined the benefits of the software reuse, as it is a promising strategy for improvements in software quality, productivity and maintainability as it provides for cost effective, reliable, and accelerated. Software factories were developed extracting the pattern keeping in view critical axes of innovation from abstraction to specification[14].

3. Problem Statement

Software reuse enhances productivity and reliability of software product. It saves time and cost as there is no prior testing required for reusable software artifact. Our hypothesis is "To design a sustainable semantic software classification scheme that can incorporate the existing softwares designed without intent of reusability and support new software with semantic arrangement and efficient retrieval. From inspiration of "Element Periodic Table" in chemistry, we proposed semantic classification and retrieval.

3.1. Relation with Existing Approaches

In implementation of software classification, we have followed different research papers. Moreover, we found schemes relevant to our approaches given:

3.2. Faceted Classification:

Faceted classification scheme described by Gajala and Phanindra [4]presents solution to the problem many researchers face during classification. In faceted classification, classes and modules are assembled together and assigned predefined keywords from facets lists. It provides higher accuracy and flexibility in classification. Faceted classification scheme improves search and retrieval of reusable software artifacts and improves selection process of reusable artifacts.

In our approach, the periodic table and tree parsing are used for efficient organization of software components. Well-structured component improve accuracy of search and retrieval of artifacts and make flexible selection process of reusable artifacts.

3.3. Enumerative Scheme:

In this approach[4],all classes are predefined. It is mostly used in school libraries to arrange the books of different departments, like biology chemistry, computer etc. Librarian selects the books which best fit its location which illustration can be Dewey[13] Decimal system used to classify books.

In our project, we designed GUID for efficient search and storage. However, this scheme is one-dimensional with collision bucket support, means if we get more than massive similar reusable artifacts in with minor variation in one place, we save that item with collision number that would represent similar item with minor variation. Otherwise, it would not be scalable classification scheme. While in tree based, we evade this problem by determining the depth. Furthermore, we allocate the same position with structured metamorphosis of software component to improved scalability.

3.4. Attribute Value:

Gajala and Phanindra [4] uses set of attributes to classify an artifact. For example, different books in library have different attributes, like title, author, publisher, ISBN

number, date, and a classification code in Dewey decimal system.

In our proposed work, we used set of attributes like version number, domain, classes, projects, and functionalities for transitional threshold to classify software position in our proposed software periodic table.

3.5. Free Text

Classification: Free text approach states that search and retrieval is made using text in the documents of artifacts. It is the keyword-based search. However, there are disadvantages of this approach. One is its ambiguous nature, and second it may search many irrelevant objects.

We have used keyword based search approach in our implementation work to search the particular software artifact from the repository by generating a code of that particular keyword that accelerates the search efficiency.

4. Our Approach

We explored the dense literature related to our study; we came up with a novel approach of classifying software for reuse influenced by chemical element periodic table. It arranges all the known elements in an informative array. Elements are arranged left to right and top to bottom in order of increasing atomic number.

4.1. Mapping chemical elements to Software elements

In this section, we are developing relationship between software and software elements. We mapped software function (set of commands to compute) to chemical atom. Molecule to program for example, molecules of CO_2 and CO both contains same type of atoms, but due to difference in the number of atoms, they exhibit different behavior. It is in case of software where mutation single function would change the behavior of program interface. This simulate the variation same genre of software. A combination of basic functionalities of software with computability heuristic is valid user program compared to unstructured or invalid structure. We can embed these valid tested function and/or programs to develop software component. We can identify the valid programs from existing repository as well as upcoming software collections.

In Periodic table, the different rows of elements are called periods. The period number of an element signifies the highest energy level an electron in that element occupies and grouped based on semantic commonalities. We have developed mapping relation in chemical and periodic table characteristics as shown in table 1.

Software based on the functionality are categorized in groups such system software, application software, etc. People visualize elements from organization pattern. By examining an element's position on the periodic table, one can infer the electron configuration. Elements that lie in the same column on the periodic table (called a "group") have identical configurations and consequently behave in a similar fashion chemically. For instance, all the group 18 elements are inert gases. The periodic table contains an enormous amount of important information. People familiar with how the table are structured can quickly determine a significant amount of information about an element[15]. From the software table where software lie in groups based on the functionality exhibited by software that recursively from functions. .Therefore, in our software repository, we are classifying the software in groups based functionality and represented by ontology supporting high visibility. We designed coding scheme for efficient retrieval.

Table-I: Mapping of Chemical Elements and Software Attributes

CHEMICAL ATTRIBUTES	SOFTWARE ATTRIBUTES
Atomic Number: The atomic number indicates the number of protons within the core of an atom. The atomic number is an important concept of chemistry and quantum mechanics. An element and its place within the periodic table are derived from this concept [15].	We the version number of software has format Major, Minor, Build. It means release date or year, also it means some minor updating in software and major means some major changes in software, like altering the design of software.
Atomic Mass: the name indicated the mass of an atom, expressed in atomic mas units (amu). Most of the mass of an atom is concentrated in the protons and neutrons in the nucleus [15].	We can relate it to the granularity of the software. The usability of software according to functionalities and features it. It can vary every time the software updated.
Density: The density of an element indicated the number or units of mass of the element that are present in a certain volume of the medium.	In software terms, it is related to complexity of software a measure of resources expended by a system, interacting with a piece of software perform a given task.

5. Software Periodic Table (SPT) Approach

We have merged two different aspects for software classification into groups and assigning codes for their position.

5.1. Categorization

Categorization is essential aspect software periodic table. In which, we will classify the software into groups (based on the grouping semantic of periodic table). We utilize the concept of periodic table in order to categorize software's according to their level of complexities. As in periodic table, its elements are categorized with their respective to chemical properties, atomic number, and electron configuration. Therefore, in SPT the organization of the software's according to their category of respective functionality and type features they provide. We assign the ID to each category based on type of software. We establish grouping scheme for top-level hierarchy for reducing search space complexity. Complexity of search space directly referred to the repository size of a particular dataset. It means for the software enumeration problem, we should have structure which traverse the repository easily and at every level downward reduce complexity. We should have the knowledge of iteration for a particular search referring the dataset; more precisely enumerator can point directly our required search based on GUID coding scheme. Our proposed structure can be indexed at any level and based on segments characteristic. This will increase the search efficiency. Large set problem is broken down in subset of category in order to reduce search space. The detailed layout of the distribution of the bits and calculations shown in the Figure 1. We consider the realization of software search problem as:

- All Software as a universal set S,
- Software Types (Application Software and System Software,…) ST are subset of S
- Software Category (For each in ST) are Set SCA

- SCS which are again subset of each software type
- S ={ Set of All Softwares}
- ST={All Application Software, System Software}
- SCA ={All software Categories of Application Software }
- SCS ={All software Categories of System Software }
- O ={ All software of other category }
- SCA= { ERP Solution, Grid solution, realtime application, ...n}. Similarly, SCA={Operating Systems, disk utilities, device drivers.. n
- SCS ={All software Categories of System Software }
- O ={ All software of other category }

Since SCA, SCS, and O etc, are disjoints set. Therefore, search space is divided at each level up to more than 50% finite number the categories are explained as follows.

- SCA= { ERP Solution, Grid solution, real-time application ...n}. Similarly, SCA={Operating Systems, disk utilities, device drivers.. n}

Now if we consider set of operating system

- OS ={ windows , Linux, mac, ..} and disk utilities
- DU = {dr. disk, disk cleaner...} then both OS and DU are disjoint set results reduction in search space.

5.1.1. Software Node: This is top node. Initially controled search pointer will be here as shown in Fig. 2. Down to the hierarchy the complexity of the search space will be reduced gradually.

5.1.2. Software Type Node:

This will be the second node in the software enumerator. There major type nodes would be assigned ID of 32 bit for each type. This segment of 32 bits (From left to right bit 1 to bit 32) will contain the information for System Software as per given below detail SCA={ERP Solution etc } bits (From bit 9 to bit 16) will be length of function, module, package, class name as following classification.

- Software Project
- Package/release
- Software Module
- Software Sub-Module
- Software Class/Structure

5.2. GUID coding

Since the periodic table the elements are being classified according to their atomic number in the increasing order, we classify the software with their GUID in SPT on some particular location.

GUID for SPT is assigned on semantic of software functionality bases groups in form of classes as shown in figure 2. The evolved software component with additional feature and functionalities change its version number to a higher one. For example, if we have a software name A with have GUID, evolved version will be stored by almost the same ID but flipping a bit on collision bucket. We design an enumerator in java programming language that reads the name of the software system, software project, package, module, sub-module, class-name, and function then generate the code for that in and if there are more than one function with the same name then C bit is set 1 for second 10 for third and so on for counting collisions. The anomalies in this scheme are that the sum of ASCII of different function may result in the same code, but there would be very few collision comparative reducing to search complexity less than O(logN).

This coding scheme is performed for each of the following node. The binary format for each part of 32 bits separated by: and further divided in 2 parts that's separated by "." The first part is further divided by - and part left side of "-"is number of collision C and right side of the"-"and before "." is the length L of that software system, software project, package, module, sub-module, class-Struct-name, function, and second part after "." will be the number of the function. Software system, software project, package, module,

submodule, class-Struct-name, function. FN is number code of that Function assign that is generated base on summation of ASCII character of the Function name. For example, SUM =83+85+77= 245. Function/subroutine the enumerator converts the each category node (software project name, software package name, software module name, software sub-module name, software class-struct and function name) into equivalent binary code. Once the conversion is being completed then enumerator looks the repetition of each attribute in a file where the binary code value for each function detail is placed. If it finds the repetition then deals the particular attribute as collision and makes increment in the value by 1. Fig.1 represents the detail of collision of attribute and it also shows that how the binary value is represent in scheme for each of above five categorization nodes.

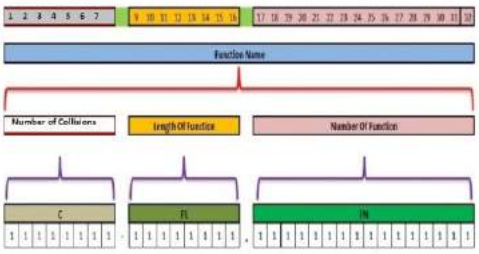

Figure 1: Layout of generating GUID for function name

The all bits one's represents absence associations in upper & lower hierarchy. Where the all bits zero represent absence of the particular nodes or skip codes. All other codes represent some node. The path root node to desired category is established and required source code file for the link is displayed to user. In the case of skip code or multiple version of same function or class multiple files are shared to user. Tree based representation. A tree is useful for exploring a large search space and reduces search complexity. Now in relation with our approach nodes in the tree are considered to be the domains, sub-domains, software, features, modules, classes etc. Branches of nodes are disjoints and leaves here would be the features, classes, modules, and functions of the software. We have used tree to show readers the flow of forthcoming software component and their auto-placement in the tree in their esteemed domain. The path from root node to leaf represents the semantics of the feature usability. Decision trees are very much helpful structures in building and interpreting because they are straightforward. The tree representation is efficient visualization and retrieval.

5.3. Software Quality

Quality of software is essential attribute in the Software development processes. The quality of the software has directed relation with customer satisfaction and organization's reputation. Developer appraisals, scheduling, and deadlines of release are affected by magnitudes of bugs in software products. The reusable software component enhance the productivity [16] and quality by via tested software component's reusability[17]. Software quality classification techniques described in [18]. The classification based modeling technique have proven to be better in achieving software quality facilitated the developer most relevant and minimum customize OTS component. The software quality evaluation techniques include CART (classification and regression tool)[19], S-PLUS[20], C4.5 Algorithm[21, 22], Treedisc algorithm[23], Sprint-sliq algorithm, logistic regression, case-based reasoning.

6. Implementation

Now according to our approach we have classified some of the real time software categories. We have extracted some basic functional components from GITHUB[24, 25] the one of most popular open source projects repository, collected from local industry and generated for experiments. The software component were stored in cloud based repository and structure is defined in ontology and data stored using xml format. The concept

was implemented and evaluated by designing a system as shown in figure 3.

6.1. Functionalities based Classification technique

In this technique, we have classified the software functionalities present in our software repository into different concerned domains and sub-domains. It behaves like a software tree in which at the first level the core domain of that functionality/feature is present and then at the next level its sub-domains are discovered and further sublevel are identified untill there are no more labels exists. Then on selecting a particular sub-domain, a user will then select their required functionalities from a list of those present in software repository. These nodes of functionalities behave like leaf as in n-ary tree with n number of disjoint categorization.

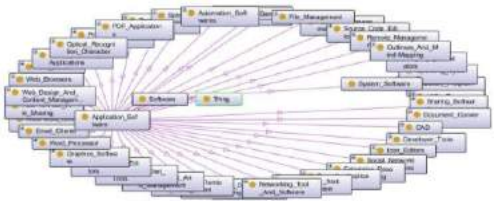

Figure 2: Classification of Software Components in Successive Hierarchy

6.2. Keyword-based technique

In this technique, we have considered a number of different keyword-based search techniques in order to select required functionalities. We applied keyword based search on each node for exploring the number similar function in the same domain. Further, it supports user to exclude the restriction of domain and exploring the repository on keyword based on parent node hierarchy. For optimization of search time the name of child node are assigned a unique key from their names ASCII sum. The collision are handled and assigned codes and position adjacent to colliding nodes.

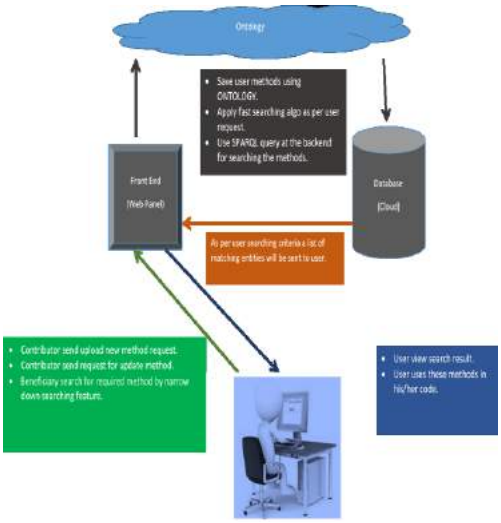

Figure 3: Experimental System Description

6.3. Hybrid technique:

This technique is the combination of the classification-based and keyword-based technique. In this approach, we have divided the softwares in different domains and sub-domains in the same way we have done in the classification-based technique. The element of keyword-based technique comes into play when user selects their required domain and sub-domain and then at that point they specify their required functionalities for that software domain. As all of the above techniques require us to maintain a repository for different functionalities of softwares from different domains, we have also considered in our system using an online software repository such as github[24] for accessing different software functionalities to evaluate the owr coding scheme. In this study, user simply specifies the programming language (such as java, C#) from which to obtain required software functionalities and then names those functionalities, and our system will fetch those software codes from the website based search narrow down by user. The distribution codes in our repository in respect of programming languages are shown in Table 2.

6.4. Semantic modeling approach

Semantic representation and manipulation of huge repository of the software component play essential role in retrieving right information with support of contextual flow to that particular components. We grouped the elements based on semantic resemblance characteristics as shown in figure 4.We developed a software repository to evaluate semantic storage and retrieval of different context software function, but with the same name such as withdraw () is different functionality in banking and university management system. The keyword based search will retrieve all functions with this keyword but by using semantic approach the right function will be reflected to software developed depending on the domain knowledge s/he working on the module, package, project context. The available functionality is provided to use and not existing functionality will be appended to that repository for later use, same or other software organizations.

Table- II: Distribution Programming Language Codes in Our Project

Rank	Languages	%Projects	Rank	Languages	%Projects
1	C++	30%	6	ASP. NET	5%
2	C	20%	7	CSS	5%
3	C#	10%	8	JavaScript	5%
4	PhP	5%	9	Java	10%
5	HTML	5%	10	Html 5	5%

6.5. Unique identification and representation.

Even domain, sub-domain and module distribution leads towards the constant time complexity. To improve more searching efficiency and reducing computation complexity we store the coding in binary.

Furthermore, we will be plugin that will parse the tree further reduce the number collision. To resolve the issues of the same computation but different naming would be resolve using name aliasing feature of the ontology development. Keyword based scheme would be embedded to facilitate exploration functionality of the system.

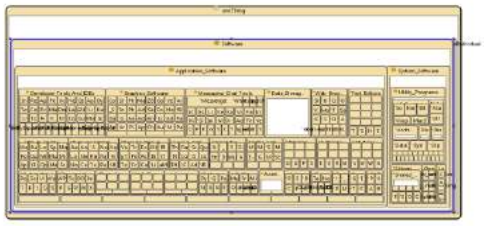

Figure 4: Semantic Periodic Table

7. Evaluation and Results:

We evaluated our software ontology to validate the software structuring hierarchy and information retrievals. The SPARQL[26, 27] is used for querying and semantic consistency validation. We configured apache jena fuseki [28] server on our system to query using SPARQL. We explored the software structure using SPARQL queries and results are shown in Fig.5. We hosted the software ontology and extracted the results from ontology from anywhere using, URL and adding prefix for names spaces for ontology. We can query and node, searching sub-tree form any particular node. We can traverse top down from root node to function leaves and vice versa. First few results of query are shown in Fig.5.

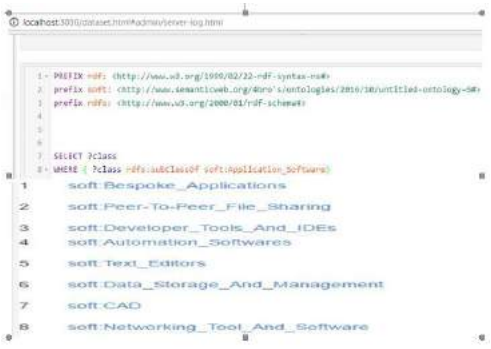

Figure 5: SPARQL query language for ontology

8. Conclusion

In our research, we have focused on the methodologies that focus more on the reuse of existing software components rather than developing a system right from the scratch that utilize a reasonable amount of efforts and resources. We have proposed a novel approach to classify software based on their basic functionalities.

In our proposed scheme, the software tree is designed in a way that helps in organizing software components in a hierarchical way, which in turn facilitates an efficient reuse of software components. The software tree is designed in such a way that inputting new software to this tree will prompt the process of traversing that software through the hierarchy of domains and sub-domains and finally assigning it to an appropriate node. The aspect of software periodic table in our methodology is that the software in different domains are so divided within the table that on adding some additional functionality/feature within them will promote them to become a new different version of software within a domain or a sub-domain.

9. Future work

The limitations regarding the proposed methodologies involve the collection of large amount of software for the software reuse repository. They feature for searching through different software available online because of their availability in compressed folder/files.

Our research work was an initiative towards the classification of software like periodic table. We will extend this research to design a classification model where all the software can be classified with sustainable structure semantic of SPT. We will use feedback mechanism to stabilize the position of software atom in logically justified position in SPT. Different genetic algorithm can be applied for location optimization of software components [29, 30]. Ant colony optimization [31] would be used to assure the quality of software component. We design and develop programming IDE Add-In that crawls our repository, facilitate developed available feature for reusability of existing component and update repository for custom build function from users

ACKNOWLEDGEMENT

We acknowledge the support of Sukkur IBA University & Muhammad Ali Jinnah University, Karachi for this research. We pay thanks to the CRUC (Center for research in Ubiquitous computing) team that provided us the environment promoting such research activities. We are thankful for github and, industry for sharing the software samples.

REFERENCES:

[1] Hu, J., et al. (2015). Modeling the evolution of development topics using Dynamic Topic Models. in Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on. 2015. IEEE.

[2] Clements, P. and L. Northrop, (2002). Software product lines: practices and patterns.

[3] Linden, F.J., K. Schmid, and E. Rommes, (2007). Software product lines in action: the best industrial practice in product line engineering. Springer Science & Business Media.

[4] Gajala, G. and M. Phanindra. A Firm Retrieval of Software Reusable Component Based On Component Classification.

[5] Rajender Nath, H.K.,(2009). Building Software Reuse Library with Efficient Keyword based Search, International Journal of Computing Science and Communication Technologies, VOL. 2, NO. 1.

[6] Suresh Chand Gupta1, P.A.K.,(2013). Reusable Software Component Retrieval System, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 1.

[7] P. Warintarawej, M.H., M. Lafourcade, A. Laurent , P. Pompidor,(2014). Software Understanding: Automatic Classification of Software Identifiers, Intelligent Data Analysis (IDA Journal) 18(6) (2014) in press.

[8] Prieto-Diaz, R.,(1991). Implementing Faceted Classification for Software Reuse, Software Production Consortium, Herndon, VA.: New York, USA. p. 88-97.

[9] Shireesha P., S.S.V.N.S.,(2010). Building Reusable Software Component For Optimization Check in ABAP Coding, International Journal of Software Engineering & Applications 1.3.

[10] Cimitile, A., A. De Lucia, and M. Munro. (1995). Identifying reusable functions using specification driven program slicing: a case study. in Software Maintenance, 1995. Proceedings., International Conference on. 1995. IEEE.

[11] Prieto-Diaz, R. and P. Freeman, (1987). Classifying software for reusability. IEEE software, 4(1): p. 6.

[12] Lo, D., et al. (2009). Classification of software behaviors for failure detection: a discriminative pattern mining approach. in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009. ACM.

[13] Zina Houhamdi, S.,(2001). Classifying Software for Reusability, Courrier du Savoir: Algeria.

[14] Greenfield, J. and K. Short. (2003). Software factories: assembling applications with patterns, models, frameworks and tools. in Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications. 2003. ACM.

[15] Greenwood, N.N. and A. Earnshaw, (2012). Chemistry of the Elements. Elsevier.

[16] Case, A.F., (1985). Computer-aided software engineering (CASE): technology for improving software development productivity. ACM SIGMIS Database, 17(1): p. 35-43.

[17] Tahir, M., et al., (2016). Framework for Better Reusability in Component Based Software Engineering. the Journal of Applied Environmental and Biological Sciences (JAEBS), 6: p. 77-81.

[18] Khoshgoftaar, T.M. and N. Seliya, (2004). Comparative assessment of software quality classification techniques: An empirical case study. Empirical Software Engineering, 9(3): p. 229-257.

[19] Steinberg, D. and P. Colla, (2009). CART: classification and regression trees. The top ten algorithms in data mining, 9: p. 179.

[20] Khoshgoftaar, T.M., E.B. Allen, and J. Deng, (2002). Using regression trees to classify fault-prone software modules. Reliability, IEEE Transactions on, 51(4): p. 455-462.

[21] Khoshgoftaar, T.M. and N. Seliya, (2003). Software quality classification modeling using the SPRINT decision tree algorithm. International Journal on Artificial Intelligence Tools, 12(03): p. 207-225.

[22] Quinlan, J.R., (2014). C4. 5: programs for machine learning. Elsevier.

[23] Khoshgoftaar, T.M. and E.B. Allen, (2001). Controlling overfitting in classification-tree models of software quality. Empirical Software Engineering, 6(1): p. 59-79.

[24] Dabbish, L., et al. (2012). Social coding in GitHub: transparency and collaboration in an open software repository. in Proceedings of the

ACM 2012 conference on Computer Supported Cooperative Work. 2012. ACM.

[25] Vasilescu, B., V. Filkov, and A. Serebrenik, (2015). Perceptions of diversity on GitHub: A user survey. CHASE. IEEE.

[26] Harris, S., A. Seaborne, and E. Prud'hommeaux, (2013). SPARQL 1.1 query language. W3C recommendation, 21(10).

[27] Prud, E. and A. Seaborne, (2006). SPARQL query language for RDF.

[28] Bansal, R. and S. Chawla, (2014). An Approach for Semantic Information Retrieval from Ontology in Computer Science Domain. International Journal of Engineering and Advanced Technology (IJEAT), 4(2).

[29] Bouktif, S., H. Sahraoui, and G. Antoniol. (2006). Simulated annealing for improving software quality prediction. in Proceedings of the 8th annual conference on Genetic and evolutionary computation. 2006. ACM.

[30] Washizaki, H. and Y. Fukazawa, (2005). A technique for automatic component extraction from object-oriented programs by refactoring. Science of Computer programming, 56(1): p. 99-116.

[31] Srivastava, P.R. and T.-h. Kim, (2009). Application of genetic algorithm in software testing. International Journal of software Engineering and its Applications, 3(4): p. 87-96.

Volume 1 | No. 2 | July – December 2017

Theoretical Insights into Coverage Analysis of Cellular Networks

Murk Marvi[1], Muhammad Khurram[1]

Abstract:

Recently tools from stochastic geometry have gained much attention for modeling and analysis of dense cellular networks. Although extensive studies are available in literature in this respect, the approaches are generalized and often lack significance towards practical scenarios where network conditions vary dynamically. The main objective of this research is to provide some insights into the stochastic geometry based analysis of cellular networks through suitable modifications so that it can be used to estimate parameters of interest i.e., intensity and coverage probability for different tiers and scenarios under consideration. The main definition for probability of coverage has been re-defined and complete closed form expression is derived. The intensity for different tiers have been estimated with the help of proposed approach, beyond which no more gain in coverage probability can be achieved. Monte-Carlo simulations have been performed for validation of proposed approach and results are also compared with state-of-the-art approach.

Keywords: *Stochastic geometry, Coverage analysis, Cellular network*

1. Introduction

The rapid development in handheld digital devices i.e., smartphones, tablets, laptops, and machine type communication have resulted in exponential growth for capacity requirements. Apart from this, explosion of emerging resource hungry services and applications are playing a dominating role in overall increase in traffic demands. The traditional cellular networks were designed to provide simple voice communication (1G and 2G) after that enhanced to provide data communication as well (beyond 2G) to its customers. Initially, the network users were only humans; however, in current scenario, the network has to serve machine type applications along with conventional human-to-human applications. Therefore, existing cellular network is not capable of fulfilling the growing demands with extreme variability in service requirements like delay, throughput, packet loss and other related metrics of measure. Thus, the concept of dense heterogeneous i.e., multi-tier networks [1]-[3] has been exploited in literature for fulfilling capacity requirements of customers. In traditional network. Since the demands were low, the focus was around coverage analysis only. However, in existing networks the focus has been shifted to capacity analysis as well in order to meet the requirements of customers and that's the main reason behind coined concept of dense networks. As ergodic capacity of a network is limited due to interference experienced by the customers which is random; hence, it is important to analyze the distribution of this random variable across the network.

1.1. Related Work

Traditionally, Wyner [4] and grid [5] models have been employed for estimating coverage and capacity bounds in cellular networks. Wyner model provides good insight into high interference regime [6], however in general it is considered to be highly inaccurate. On the other hand, grid model considers regular deployment of base stations (BSs) which is, practically, not the case especially for dense networks. Therefore, grid model provides upper bound to interference

[1] Computer & Information Systems Department, NED University of Engineering & Technology, Pakistan
Corresponding email:* marvi@neduet.edu.pk

distribution. Recently stochastic geometry has gained much attention for coverage and rate analysis of dense cellular networks due to its tractability for special cases and simple numerical integrations for general cases. In [7], a tractable approach has been presented for coverage and rate analysis of single-tier cellular network by modeling the location of BSs through a Poisson point process (PPP). The results provided in [7] are pessimistic due to presence of strong interfering nodes nearby; they are extremely generalized and tractable as well. In [8], coverage and rate analysis for multi-tier network have been presented by exploiting max SINR BS association rule. On the other hand in [9], same analysis has been carried out as in [8], but nearest BS association rule is used. The interesting finding in [8], [9] is that, for interference limited network with Rayleigh fading and path loss exponent of 4, the probability of coverage for multi-tier cellular network becomes equivalent to that of single-tier case presented in [7]. The authors have justified this result by making an argument that addition of low power tiers into existing dense Macro tier will not affect the coverage in any case. Since any increase in interference will be counterbalanced by increase in received signal strength, additional BSs can be deployed for achieving linear increase in capacity of the network without affecting the coverage. Due to pessimistic results provided by PPP model, in [10] detrimental point process (DPP) is used where correlation or repulsive effect has been taken into account. Therefore, authors in [10] have claimed that DPP provides better coverage and rate analysis as compared to PPP. Similarly in [11], Poisson cluster process (PCP) has been exploited by considering the fact that certain regions are denser as compared to the others. The authors in [11] have compared to the results of PCP with PPP and concluded that in PPP capacity increases linearly with increase in number of tiers; whereas in PCP the coverage and capacity degrades as number of tiers grow beyond certain threshold. In [12], spatial dependence for BSs deployment have been considered for two tier network i.e., Macro and Pico, where the locations of Macro BSs have been drawn from a PPP and for Pico BSs Poisson hole process (PHP) has been used. The authors have verified results through simulation and claimed that proposed numerical results are closely fitting to the simulation results. The approaches proposed by [7]–[9] are pessimistic; however, they are tractable under especial cases. On the other hand, approaches proposed in [10]–[12] involve numerical integration but provide better approximation to actual scenarios.

Recently, in [13] the researchers have discussed fundamental limits on wireless network densification. Over the past few years, the concept of densification has been exploited for enhancing data rates. The researchers in [13] have argued that at some point further densification will no longer be able to provide exponentially increasing data rates like Moore's law. Traditionally, simple path loss model is used for coverage analysis. However, recently researchers have exploited multi-slope path loss models for better coverage analysis of dense cellular networks [14]. Authors have suggested that, due to densification, the distance between customer and BS is less than certain critical distance $\boldsymbol{R_c \approx {4h_t h_r}/{\lambda_c'}}$ where path loss exponent $\boldsymbol{\alpha \approx 2}$. In general, for distance greater than $\boldsymbol{R_c}$ the path loss exponent $\boldsymbol{\alpha \approx 4}$. Thus, with the help of multi-slope path loss model coverage probability can be more accurately estimated. In [14], the authors have used dual slope path loss model and showed that as network density grows and $\boldsymbol{\alpha < 1}$, potential throughput approaches to zero. On the other hand, for $\boldsymbol{\alpha > 1}$ potential throughput grows with denser deployment. However, growth may be sublinear depending on the path loss exponent.

1.2. Motivation and Contribution

Even though extensive studies are available in literature for coverage and rate

analysis of cellular networks by exploiting tools from stochastic geometry; such approaches are extremely generalized and cannot be used for analyzing specific network scenarios, which are more useful for industry practitioners. For example, in [7]–[9] tractable models have been presented for coverage analysis of dense single and multi-tier cellular network, which are independent of power and intensity of BSs under special conditions; however, few questions mentioned here are really important to answer from practical perspective, 1) What is the limit of $\boldsymbol{\lambda}$, for particular tier, beyond which it can be considered as dense? 2) How one can analyze the coverage and rate for less dense cellular networks? Because, when it comes to practical situations not all regions need to be densely deployed. Since the requirements vary from region to region, for example, the traffic patterns in residential regions would definitely be different from commercial ones. 3) How much intensity of BSs belonging to certain tier is required to provide minimum coverage and rate in a given region? Thus, the prime motivation of this research is to explore answers of the above mentioned questions by incorporating some factors into existing stochastic geometry model [7] for single-tier cellular network. The main contributions of this work are listed as under.

- A modified model for coverage analysis of single-tier cellular network by exploiting tools from stochastic geometry. The presented model is not independent of intensity or transmitted power of tiers; hence, it is equally applicable to both dense and sparse networks.
- The intensity of BSs for Pico and Femto tiers has been approximated through simulations, such that they can at least meet the coverage bounds given in [7].
- Monte-Carlo (MC) simulations have been performed for validation of proposed approach and results are compared at the end in Section IV.

2. System Model

In this work we assume a single-tier cellular network, where the location of base stations (BSs) has been drawn from a Poisson point process (PPP) $\boldsymbol{\Phi}$ of intensity $\boldsymbol{\lambda}$. The mobile stations (MSs) i.e., customers, are assumed to be distributed according to another independent PPP $\boldsymbol{\Phi_u}$ with intensity $\boldsymbol{\lambda_u}$. Nearest BS association rule and downlink channel have been considered under this study. Thus, a customer is said to be under coverage if received signal to interference noise ratio (SINR) exceeds some pre-defined threshold θ. Where, SINR can be defined as,

$$\boldsymbol{SINR}(\boldsymbol{x_o}) = \frac{\boldsymbol{P_r}(\boldsymbol{x_o})}{\sum_{\boldsymbol{x}\in\boldsymbol{\Phi_c}\setminus\boldsymbol{x_o}} \boldsymbol{P_r}(\boldsymbol{x}) + \boldsymbol{\sigma^2}} \qquad (1)$$

here, $\boldsymbol{P_r}$ in numerator denotes the amount of power received from tagged BS and in denominator, the sum over it, represents the amount of received power from all interfering BSs. "$\boldsymbol{\sigma}$" denotes noise power, which is assumed to be additive and constant. The amount of power received $\boldsymbol{P_r}(\boldsymbol{x})$ at a randomly located mobile user from a BS *2* Φ, is actually function of three factors as given in (2).

$$\boldsymbol{P_r}(\boldsymbol{x}) = \boldsymbol{P_t}.\boldsymbol{h_x}.\boldsymbol{l}(\boldsymbol{x}) \qquad (2)$$

1) The amount of power transmitted $\boldsymbol{P_t}$, 2) Random fading component $\boldsymbol{h_x}\sim\mathbf{exp}(\boldsymbol{\mu})$, that is assumed to be exponentially distributed with mean 1 over the link between MS and tagged BS. For Interfering links, it follows a general distribution "*g"* however, closed form expressions can be obtained when fading over interfering links also follow exponential distribution. 3) Distance dependent path loss component $\boldsymbol{l}(\boldsymbol{x})$, which has been defined by considering free space path loss (FSPL) model with reference distance of 1m and given as $\boldsymbol{l}(\boldsymbol{x}) = \left(\frac{\boldsymbol{\lambda_w}.\boldsymbol{l}}{\boldsymbol{4\pi}}\right)^{\mathbf{2}} \boldsymbol{x}^{-\boldsymbol{\alpha}}$. Here, $\boldsymbol{\lambda_w}$ is wavelength of operating channel, $\boldsymbol{\alpha} > \mathbf{2}$ is the path loss exponent.

3. Proposed Approach

According to [7]–[9], an MS is said to be under coverage if it receives a target **SINR**, $\boldsymbol{\theta}$. According to [15], in modern cellular networks such as LTE, the threshold for received signal power ($\boldsymbol{P_r}$) is low enough to be comfortably exceeded by any modern receiver within a fairly large distance from transmitting BS. Therefore, for the purpose of analysis $\boldsymbol{P_r}$ can be dropped and without loss of generality only SINR can be considered key indicator of coverage, as given in eq. (3). Thus, an MS is said to be under coverage, if communication link from the BS serving that MS is sufficiently good; where good means $\boldsymbol{SINR > \theta}$. This definition for coverage analysis is very suitable when considering dense networks. However, this is not always the case; as some regions are less densely deployed as compared to the others due to variation in traffic patterns and intensity of customers i.e., $\boldsymbol{\lambda_u}$.

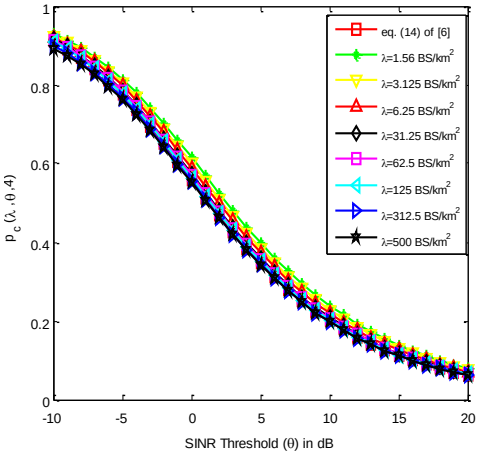

Figure 1: MC simulation results as a function of $\boldsymbol{\lambda}$, [7]

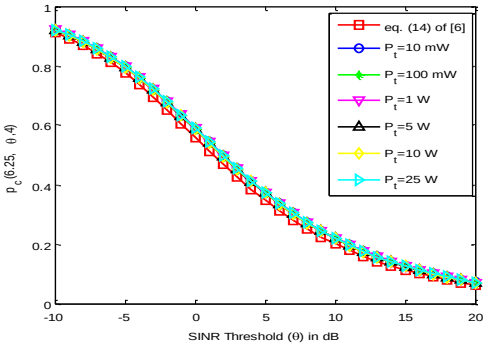

Figure 2: MC simulation results as a function of $\boldsymbol{P_t}$, [7]

Therefore, while considering less dense networks for certain regions; as distance between MS and tagged BS exceeds certain threshold $x > R$, the received signal strength $\boldsymbol{P_r}$ drops below defined threshold δ. Hence, MS can no more be considered under coverage. Since SINR is a ratio, it will remain the same either for dense or sparse networks. That's the reason, it does not get affected by intensity of BSs deployed, as discussed in [7]. Thus, for better insight into different parameters of interest from coverage and capacity point of view, it is important to consider the effect of $\boldsymbol{P_r}$ along with SINR as defined in eq. (4).

$$\boldsymbol{p_c(\lambda, \theta, \alpha) = P[SINR > \theta]} \quad (3)$$

$$\boldsymbol{p_c(\lambda, \theta, \delta, P_t) = P[SINR > \theta | x < R]} \quad (4)$$

Without loss of generality, we consider that MS is located at origin and the nearest BS is at a distance $\boldsymbol{x}$ from it; such that $\boldsymbol{x < R}$ which makes sure that $\boldsymbol{P_r > \delta}$. Hence, MS is said to be under coverage if it receives SINR greater than θ.

$$\boldsymbol{SINR(x_o) = \frac{P_r(x_o)}{I_x + \sigma^2}}$$

where, $\boldsymbol{I_x = \sum_{x \in \Phi_c \setminus x_o} P_r(x)}$ is the cumulative interference from all BSs *2* Φ other than tagged one. Since nearest BS association rule has been assumed, the probability density function (PDF) for distance *x* between MS and

tagged BS can be given as $\boldsymbol{f_x(x) = 2\pi\lambda x e^{-\pi\lambda x^2}}$. The maximum hard core distance *R* beyond which an MS cannot be considered under coverage has been obtained from eq. (2) as,

$$\boldsymbol{R = \left(\frac{P_t \cdot \mu \cdot K}{\delta}\right)^{1/\alpha}} \quad (5)$$

Where, $\boldsymbol{K = \left(\frac{\lambda_w \cdot l}{4\pi}\right)^2}$ and μ is mean of random fading variable, $\boldsymbol{h_x}$, which is exponentially distributed.

3.1. General Case

The general case results for coverage analysis of single-tier network have been presented in this section.

Theorem 1: The probability that an MS will be under coverage of cellular network with parameters described in Section II is,

$$\boldsymbol{p_c(\lambda, \theta, \alpha, \delta, P_t) = 2\pi\lambda \int_0^R e^{-\pi\lambda x^2 \beta(\theta,\alpha) - \mu\theta\sigma^2 x^{\alpha/2}} x dx}$$

where,

$$\boldsymbol{\beta(\theta,\alpha) = \frac{2(\mu\theta)^{2/\alpha}}{\alpha} \mathrm{E}\left[g^{2/\alpha}\left\{\Gamma\left(-\frac{2}{\alpha}, \mu\theta g\right) - \Gamma\left(-\frac{2}{\alpha}\right)\right\}\right]}$$

here, expectation is with respect to, *g*, channel fading distribution of interfering BSs. $\boldsymbol{\Gamma(a,x) = \int_x^\infty z^{a-1} e^{-z} dz}$ denotes incomplete gamma function, and $\boldsymbol{\Gamma(x) = \int_0^\infty z^{x-1} e^{-z} dz}$ is the standard gamma function.

Proof: By conditioning on the fact that, nearest BS is located at *x* distance from a randomly located MS such that $\boldsymbol{x < R}$, probability of coverage averaged over the plane can be given as,

$$\boldsymbol{p_c(\lambda, \theta, \alpha, \delta, P_t) = E_x[P\{SINR > \theta | x < R\}]}$$

$$\boldsymbol{= \int_{x>0} P[SINR > \theta | x < R] . f_x(x)\, dx}$$

$$\boldsymbol{= \int_0^R P\left[\frac{h_x x^{-\alpha}}{I_x + \sigma^2}\right] . f_x(x) dx}$$

$$\boldsymbol{= \int_0^R P[h_x > \theta x^\alpha (I_x + \sigma^2)] . f_x(x) dx}$$

Since probability of coverage for $\boldsymbol{x > R}$ is zero, the limit of integral has been extended from 0 to *R* instead of 0 to *1*, which is generally considered for dense cellular networks. Since it has been assumed that $\boldsymbol{h_x \sim exp(\mu)}$, the probability that random fading component $\boldsymbol{h_x}$ has value greater than required can be obtained as,

$$\boldsymbol{P[h_x > \theta x^\alpha (I_x + \sigma^2) | x] =}$$

$$\boldsymbol{E_{I_x}[P\{h > \theta x^\alpha (I_x + \sigma^2) | x, I_x\}]}$$

$$\boldsymbol{= E_{I_x}[exp\{-\mu\theta x^\alpha (I_x + \sigma^2)\}]}$$

$$\boldsymbol{= e^{-s\sigma^2} \mathcal{L}_{I_x}(s)}$$

where, $\boldsymbol{s = \mu\theta x^\alpha}$ and $\boldsymbol{\mathcal{L}_{I_x}(s)}$ is Laplace transform of random variable$\boldsymbol{I_x}$*x* evaluated at *s*, by conditioning on the fact that, MS
located at origin associates with nearest BS. Thus probability of coverage can be obtained as,

$$\boldsymbol{p_c(\lambda, \theta, \alpha, \delta, P_t) = 2\pi\lambda \int_0^R e^{-s\sigma^2} \mathcal{L}_{I_x}(s) e^{-\pi\lambda x^2} x dx} \quad (6)$$

For general fading distribution, the Laplace transform of interference for single-tier cellular network with BS locations drawn from PPP has been derived in [6] as,

$$\boldsymbol{\mathcal{L}_{I_x}(s) = e^{-\pi\lambda x^2 \{\beta(\theta,\alpha) - 1\}}} \quad (7)$$

Thus substituting eq. (7) in eq. (6) proves the theorem.

3.2. Special Case: General Fading, Interference Limited Network

The assumption for interference limited regime is frequently considered in literature [7]–[9], due to the fact that, in modern cellular networks thermal noise power is negligible as compared to desired signal power; while

considering interior of cells. On the other hand, at cell edges the interference is typically much larger than thermal noise. Hence for simpler expressions it is preferred to consider interference limited regime i.e., $\sigma 2 = 0$. With this assumption, Theorem 1 will be simplified as,

$$\boldsymbol{p_c(\lambda, \theta, \alpha, \delta, P_t)} = \frac{\mathbf{1} - \boldsymbol{e^{-\pi\lambda R^2 \beta(\theta,\alpha)}}}{\boldsymbol{\beta(\theta,\alpha)}} \quad (8)$$

It must be noted here that, as $\boldsymbol{\lambda}$ or $\boldsymbol{P_t}$ approaches *1*, the exponential term in eq. (8) approaches zero. The resulting expression becomes independent of $\boldsymbol{\lambda}$ or $\boldsymbol{P_t}$ of chosen tier and becomes equivalent to one given in eq.(8) of [7], for dense interference limited cellular network. Thus, with the help of expressions in eq. (8) we can estimate the limit of $\boldsymbol{\lambda}$ for particular tier with transmitted power $\boldsymbol{P_t}$, beyond which it can be considered as dense and no further improvements can be achieved in $\boldsymbol{pc}$.

3.3. Special Case: Rayleigh Fading, Interference Limited Network

Simplified expression can be obtained when fading on interfering links is also assumed to be exponentially distributed. In Theorem 2 of [7], the Laplace transform of interference while considering Rayleigh fading has been derived and resulting expression is given as,

$$\boldsymbol{\mathcal{L}_{I_x}(s)} = \boldsymbol{e^{-\pi\lambda x^2 \rho(\theta,\alpha)}} \quad (9)$$

where,

$$\boldsymbol{\rho(\theta,\alpha)} = \boldsymbol{\theta^{2/\alpha}} \int_{\boldsymbol{\theta^{2/\alpha}}}^{\infty} \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{u^{\alpha/2}}} \boldsymbol{du}$$

Thus, plugging eq. (9) into eq. (6), results in expression of coverage for a randomly located MS experiencing exponential fading on all links as,

$$\boldsymbol{p_c(\lambda, \theta, \alpha, \delta, P_t)} = \boldsymbol{2\pi\lambda} \int_{\mathbf{0}}^{\boldsymbol{R}} \boldsymbol{e^{-\pi\lambda x^2\{1+\rho(\theta,\alpha)\} - \mu\theta\sigma^2 x^{\alpha/2}}}\, \boldsymbol{x dx} \quad (10)$$

For no noise case i.e., $\boldsymbol{\sigma^2 = 0}$, the resulting expression will be further simplified as,

$$\boldsymbol{p_c(\lambda, \theta, \alpha, \delta, P_t)} = \frac{\mathbf{1} - \boldsymbol{e^{-\pi\lambda R^2\{1+\rho(\theta,\alpha)\}}}}{\mathbf{1} + \boldsymbol{\rho(\theta,\alpha)}}$$

where, $\boldsymbol{\rho(\theta,\alpha)}$ is much easier and faster to compute as compared to $\boldsymbol{\beta(\theta,\alpha)}$. By further assuming $\boldsymbol{\alpha = 4}$, a complete closed form expression for $\boldsymbol{\rho(\theta, 4)}$ can be obtained as,

$$\boldsymbol{\rho(\theta, 4)} = \sqrt{\boldsymbol{\theta}} \left\{ \boldsymbol{\pi/2} - \mathbf{tan^{-1}} \left(\frac{\mathbf{1}}{\sqrt{\boldsymbol{\theta}}} \right) \right\} \quad (12)$$

Thus, substituting eq. (12) into eq. (11) results in final expression for probability of coverage as,

$$\boldsymbol{p_c(\lambda, \theta, \alpha, \delta, P_t)} = \frac{\mathbf{1} - \boldsymbol{e}^{-\pi\lambda R^2 \sqrt{\theta}\left\{\pi/2 - \tan^{-1}\left(\frac{1}{\sqrt{\theta}}\right)\right\}}}{\mathbf{1} + \sqrt{\boldsymbol{\theta}}\left\{\boldsymbol{\pi/2} - \mathbf{tan^{-1}}\left(\frac{\mathbf{1}}{\sqrt{\boldsymbol{\theta}}}\right)\right\}} \quad (13)$$

The obtained expression in eq. (13) is very simple and approaches to one, given in eq. (14) of [7], as $\boldsymbol{\lambda}$ or $\boldsymbol{P_t}$ approaches to *1*.

4. Results and Discussions

For tiers like Femto or Pico, the assumption of PPP is directly valid but the deployment of Macro BSs involve planning; hence, cannot be considered as completely random. That's why in [7], [8], it has been justified that PPP though exhibits the property of complete spatial randomness; however, provides pessimistic bounds for coverage and rate analysis of dense Macro-tier. In this research authors suggest that, it is also equally valid to consider PPP for sparse Macro-tier. With significant decrease in $\boldsymbol{\lambda}$, the probability that, two BSs are close to each other approaches to zero. Thus, resulting expressions becomes free of high interference issues from nearby BSs.

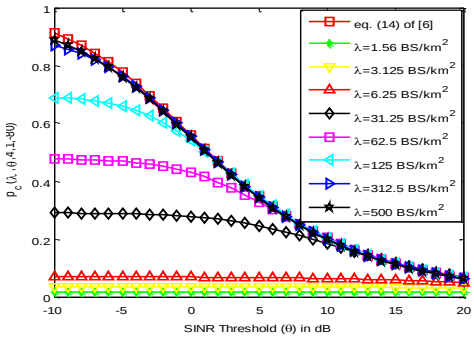

Figure 3: MC simulation results as a function of $\boldsymbol{\lambda}$, proposed approach

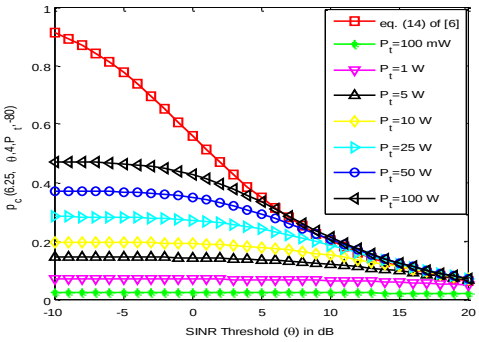

Figure 4. MC simulation results as a function of P_t, proposed approach

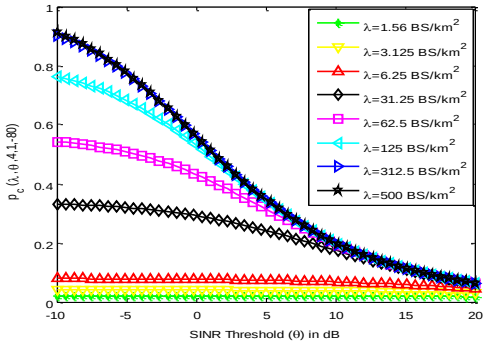

Figure 6. Analytical results as a function of λ, proposed approach

Similarly, while considering practical situations, though the deployment of Macro BSs is planned; however, a constant distance as supposed in grid models [5] cannot be maintained due to various issues like availability of space, terrain conditions, traffic requirements and others. Even though, there is not complete spatial randomness as supposed by PPP, but to certain extent it does involve randomness. Thus, the location of BSs for sparse networks can also be approximated by PPP as in the case of dense networks. In [7], the authors have already compared the results of stochastic geometry based approach with grid model. Therefore, in this research, we have performed Monte-Carlo (MC) simulations. The results are compared with analytical expressions obtained in this work and one presented in [7].

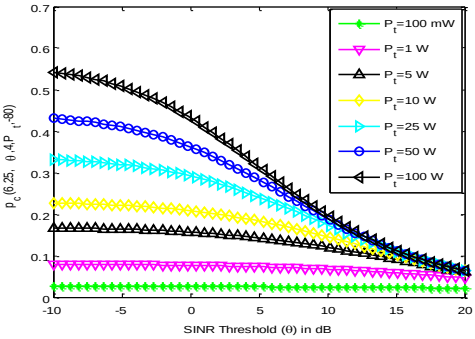

Figure 5. Analytical results as a function of P_t proposed approach

4.1. Monte-Carlo Simulations

For comparison of results, we realized the network described in Section II through Monte-Carlo (MC) simulations. An area of $\boldsymbol{4 * 4\ km^2}$ was considered and BSs of intensity $\boldsymbol{\lambda}$, with identical power of transmission $\boldsymbol{P_t}$, were randomly placed in considered region using uniform distribution. The distance between an MS and all BSs were measured by using expression, $\boldsymbol{xn = \sqrt{(x_n - x_o)^2 + (y_n - y_o)^2}}$ where, $\boldsymbol{x_o}$ & $\boldsymbol{y_o}$ represent coordinates of MS location and $\boldsymbol{x_n}$ & $\boldsymbol{y_n}$ denote coordinates for $\boldsymbol{n^{th}}$ BS location. For every MS location, the BS with $\boldsymbol{min(x_n)}$ was selected as tagged one and rest were treated as interfering BSs. Thus, SINR at every MS location was calculated by using eq. (1) with $\boldsymbol{\sigma^2 = 0}$ and results were averaged down at the end for obtaining probability of

coverage. Each simulated average result has been obtained by repeating 1000 iterations. It must be noted that all MC simulations have been performed for interference limited network with $\boldsymbol{\alpha = 4}$, $\boldsymbol{fc = 1800\ MHz}$, $\boldsymbol{\lambda}$ in BS/km2, $\boldsymbol{\delta}$ in dBm, $\mu = \mathbf{1}$, and $\boldsymbol{P_t}$ in watts, if specifically not mentioned.

4.2. Results for MC Simulations

Initially for a better insight into results, MC simulations have been performed. Coverage definition given in eq. (3) has been considered, which is extensively used in literature [7]–[9], [15]. It must be noted from MC results given in Fig. 1 and 2 that, no matter how sparse or dense network we chose, the resulting SINR distribution is similar to that proposed in [7]. Since SINR is a ratio, thus while considering sparse or dense networks the received signal strength from tagged and interfering BSs are affected by almost same amount. Therefore, in each case MC simulation produces same results. That's the reason that, under especial cases, $\boldsymbol{p_c}$ for multi-tier proposed in [8], [9] is the same as that for single-tier given in [7].

MC simulations have also been performed by redefining the *pc*, as proposed in this research and given in eq. (4), under the same scenario. In Fig. 3, the probability of coverage i.e., eq. (4), has been plotted as a function of BS density. It must be interesting to note that, by varying $\boldsymbol{\lambda}$, $\boldsymbol{p_c}$ is changing and after increasing $\boldsymbol{\lambda}$ to certain level, $\boldsymbol{p_c}$ does not vary and approaches to limit given in [7]. Similarly, in Fig. 4 probability of coverage has been plotted against BSs operating at different power levels. It must be clear that, BSs operating at higher power level provides better coverage as compared to ones operating at lower power levels. It is obvious; Macro BSs provide better coverage as compared to Femto or Pico BSs.

4.3. Analytical Results of Proposed Approach

Here we discuss main results of the proposed approach. As already clear from results of MC simulations that by following eq. (4), we can have better insight into coverage analysis of cellular networks. In Fig. 5 and 6, the closed form expressions obtained in eq. (13) have been plotted. As we increase the density of BS the coverage probability improves. It must be clear that proposed framework provides almost the same results as given in Fig. 3 for MC simulations. Similarly, in Fig. 6, the coverage probability has been plotted against transmitted power of BS by exploiting proposed analytic expression given in eq. (13). It must be clear that, coverage probability improves in proportion with transmitted power of BSs and is comparable with results given in Fig. 4. Thus, this verifies the validity of proposed approach. However, the proposed analytical expressions provide a bit over estimation for $\boldsymbol{p_c}$ as compared to results obtained through MC simulations. For example, in Fig. 3 and 5, for $\boldsymbol{\lambda = 125\ BS/km^2}$ compare two curves. In Fig. 3, at $\boldsymbol{SINR = -10dB}$, $\boldsymbol{p_c \approx 0.7}$. However, in Fig. 4, at $\boldsymbol{SINR = -10dB}$, $\boldsymbol{p_c \approx 0.75}$, but it is as per expectation. Since, we assumed hard core distance *R* for received signal strength which does not accounts for irregular boundary effects.

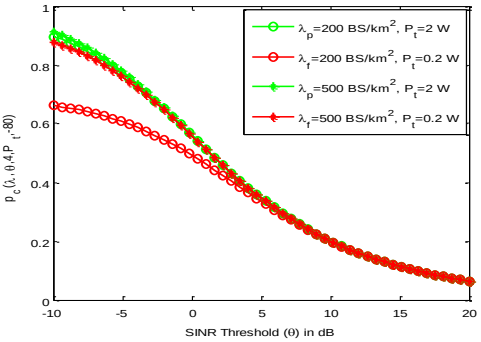

Figure 7. Estimating intensity of tiers i.e., $\lambda\ for\ \delta = -80dBm$

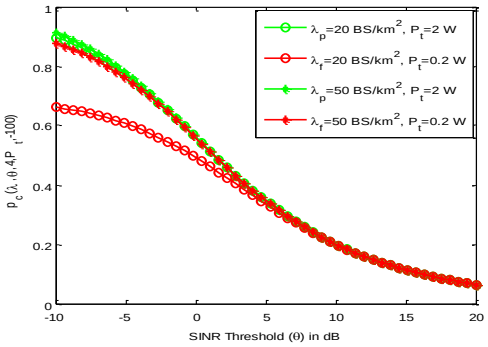

Figure 8. Estimating intensity of tiers i.e., $\lambda \; for \; \delta = -100 dBm$

On the other hand in Fig. 8, the results are obtained by assuming $\boldsymbol{\delta = -100 dBm}$, which is fair $\boldsymbol{P_r}$ in cellular network. As hard core distance *R* is increased in the case of $\boldsymbol{\delta = -100 dBm}$, therefore $\boldsymbol{\lambda}$ for each tier has been decreased. Another, important fact to note here is that, $\boldsymbol{\lambda_p < \lambda_f}$ since $\boldsymbol{P_t}$ for Pico tier is higher as compared to Femto. After increasing $\boldsymbol{\lambda}$ behind certain limit (i.e., in Fig. 8, $\boldsymbol{\lambda_p > 20}$) no further improvements in $\boldsymbol{p_c}$ has been achieved and it never exceeds the bounds defined in [7]. Thus, the proposed approach is equally valid for sparse as well as for dense networks. It is more important for industry practitioners, since they can tune the parameters of interest as per intensity of customers and traffic patterns in certain region under consideration.

5. Conclusion

A modified stochastic geometry based tractable approach has been presented in this research, which takes into account the effect of received signal strength in addition to SINR. With the help of simple modification into existing definition for probability of coverage, parameters of interest i.e., $\boldsymbol{\lambda}$ and $\boldsymbol{p_c}$ can be estimated for different tiers of cellular network under different scenarios. Various results have been reported where the effect of BS density and power of transmission on coverage probability has been analyzed in detail. Apart from that, the density of different tiers has also been estimated under different received signal strength thresholds. Although the work presented provides just an initial insight into the effects of considering a modified definition for $\boldsymbol{p_c}$, by limiting the coverage radius for single-tier cellular network to hard core distance *R*. However, more accurate models can be obtained by considering joint probabilistic definition for $\boldsymbol{p_c}$ or by using other point processes, which consider the effect of minimum distance or repulsion into account. For validation of proposed approach and its comparison with existing results, Monte-Carlo simulations have been performed. It has been concluded that, after certain limit for $\boldsymbol{\lambda}$ or $\boldsymbol{P_t}$, the results of proposed approach becomes equivalent to one presented in [7] for single-tier dense cellular network.

REFERENCES

[1] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. A. Thomas, J. G. Andrews, P. Xia, H. S. Jo et al., "Heterogeneous cellular networks: From theory to practice," IEEE Communications Magazine, vol. 50, no. 6, pp. 54–64, 2012.

[2] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5g wireless networks," IEEE Wireless Communications, vol. 21, no. 2, pp. 106–112, 2014.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?," IEEE Journal on selected areas in communications, vol. 32, no. 6, pp.1065-1082, 2014.

[4] S. Shamai and A. D. Wyner, "Information-theoretic considerations for symmetric, cellular, multiple-access fading channels. i & ii," IEEE Transactions on Information Theory, vol. 43, no. 6, pp. 1877–1894,1997.

[5] T. X. Brown, "Cellular performance bounds via shotgun cellular systems," IEEE Journal on Selected Areas in

Communications, vol. 18, no. 11, pp. 2443–2455, 2000.

[6] J. Xu, J. Zhang, and J. G. Andrews, "When does the wyner model accurately describe an uplink cellular network?" in 2010 IEEE Global Telecommunications Conference GLOBECOM 2010, Dec 2010, pp. 1–5.

[7] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," IEEE Transactions on Communications, vol. 59, no. 11, pp. 3122–3134, 2011.

[8] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, "Modeling and analysis of k-tier downlink heterogeneous cellular networks," IEEE Journal on Selected Areas in Communications, vol. 30, No. 3, pp. 550–560, April 2012.

[9] H.-S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink SINR analysis," IEEE Transactions on Wireless Communications, vol. 11, no. 10, pp. 3484–3495, 2012.

[10] Y. Li, F. Baccelli, H. S. Dhillon, and J. G. Andrews, "Statistical modeling and probabilistic analysis of cellular networks with determinantal point processes," IEEE Transactions on Communications, vol. 63, no. 9, pp. 3405–3422, 2015.

[11] Y. J. Chun, M. O. Hasna, and A. Ghrayeb, "Modeling heterogeneous cellular networks interference using poisson cluster processes," IEEE Journal on Selected Areas in Communications, vol. 33, no. 10, pp. 2182–2195, Oct 2015.

[12] N. Deng, W. Zhou, and M. Haenggi, "Heterogeneous cellular network models with dependence," IEEE Journal on Selected Areas in Communications, vol. 33, no. 10, pp. 2167–2181, 2015.

[13] J. G. Andrews, X. Zhang, G. D. Durgin, and A. K. Gupta, "Are we approaching the fundamental limits of wireless network densification?," IEEE Communications Magazine, vol. 54 no. 10, pp.184-190, 2016

[14] X. Zhang, and J. G. Andrews, "Downlink cellular network analysis with multi-slope path loss models," IEEE Transactions on Communications, vol. 63, no. 5, pp.1881-1894, 2015

[15] S. Mukherjee, Analytical Modeling of Heterogeneous Cellular Networks. Cambridge University Press, 2014.

Volume 1 | No. 2 | July – December 2017

Survey of Applications of Complex Event Processing (CEP) in Health Domain

Dr. Nadeem Mahmood[1], Madiha Khurram Pasha[1], Khurram Ahmed Pasha[2]

Abstract:

It is always difficult to manipulate the production of huge amount of data which comes from multiple sources and to extract meaningful information to make appropriate decisions. When data comes from various input resources, to get required streams of events form this complex input network, the one of the strong functionality of Business Intelligence (BI) the Complex Event Processing (CEP) is the appropriate solution for the abovementioned problems. Real time processing, pattern matching, stream processing, big data management, sensor data processing and many more are the application areas of CEP. Health domain itself is a multi-dimension domain such as hospital supply chain, OPD management, disease diagnostic, In-patient, out-patient management, and emergency care etc. In this paper, the main focus is to discuss the application areas of Complex Event Processing (CEP) in health domain by using sensor device, such that how CEP manipulate health data set events coming from sensor devices such as blood pressure, heartbeat, fall detection, sugar level, temperature or any other vital signs and how these systems respond to these events as quickly as possible. Different existing models and application using CEP are discussed and summarized according to different characteristics.

Keywords: *Business Intelligence (BI); Complex Event Processing (CEP)*

1. Introduction

Various examples of domains exist in the world such that product manufacturing system, fraud detection, anomaly management, cloud computing and many more. But in our perspective health issue is the major concern for all age groups of male and female at every stage of life from birth to death.

Health is the major concern for all age groups and for all genders as Health Information System [19] are on the demanding research field. Majority of patients who lie between the ages of 40 to 70 years having critical diseases which may be caused due to age factor, inherited family disease, and improper routine checkup. Most of the severe-ill, aged patients cannot travel at regularly basis due to critical health situation, they need proper monitoring and immediate treatment. To avoid the travelling time for checkups, for proper monitoring and self-assessment, and to get exactly the correct and trustworthy result, to facilitate both doctor and patients, different wearable body sensors are available which can sense different vital signs such as blood pressure, body temperature, ECG respiration rate etc. These body sensors can facilitate both patients and doctors as the results are accurate and true and also helpful for prevention and prediction of disease at the right time.

Business Intelligence (BI) [15], [16] and [17] describes business analysis of any field in which how extraction of required information is possible from different sources of raw data. To design new tactics and strategies for business having long-term solidity is the aim of BI. Many technologies of BI included data mining, text mining, predictive analysis, Complex Event Processing (CEP) and etc.

Change of state of any actor, system, machine or object when any method or function has triggered is an event. There are

[1] Department of Computer Science (UBIT), University of Karachi, Karachi, Pakistan
[2] IT Department (BSS), Beacon house School System, Karachi, Pakistan
Corresponding email: nmahmood@uok.edu.pk

two types of events: Event Stream Processing (ESP) and Complex Event Processing (CEP). Generally, CEP is a subpart and a technique of ESP. The CEP is appropriate for complex system in which the composition and correlation of atomic events constructs complex event or complex patterns which are very essential and beneficial for the system. CEP contracts with multiple events from different sources and produces significant events. The CEP is suitable for large real-time domains where production and growth of event occur at each level or layer of these domains.

Complex Event Processing (CEP) is a technique of BI, which creates actionable events or patterns from data of multiple sources. These events are helpful at any layers of organization for current development and future prediction. CEP can be used in many areas: weather prediction, traffic controlling, social media posts, health domain, finance, RFID management, Supply chain management etc. The CEP can be classified into two categories: Aggregation-oriented CEP executing on-line algorithms while, Detection-oriented CEP detecting events patterns.

2. Literature Review

A detail framework of BI&A (Business Intelligence and Analytics) which provides history, various applications and recent trends of BI&A [1]. The history of BI&A included BI&A 1.0, BI&A 2.0 and BI&A 3.0 have also been discussed in detail. Applications included E-Commerce and Marketing Intelligence, E-Government and Politics 2.0 and more. Also multiple recent and future trends of BI&A were discussed which included big data analysis.

"Designing and Developing Complex Event Processing Application" [2] is a white paper given a detail guideline of CEP, its techniques and developer consideration. Furthermore, how pattern matching technique with correlation and aggregation is used for CEP. Also how CEP deals with big data has been discussed in depth. A very detailed assessment of all aspects of CEP [3] and [4] which included CEP specification, methods, experiments and a very helpful guideline for tools selection at both commercial and academic levels.

The CEP models can be analyzed by using different logics such as first order logic (FOL) and fuzzy logic. The authors of [18] have given a detailed review of all logics.

2.1. Complex Event Processing (CEP) In Health Domain

In [5] those patients who need full-time continuous monitoring and alone at home, the web enabling body sensor are very helpful for proper monitoring of the patients. The wireless body sensor used to record patient's vital sign and by using web enabling sensors, doctor can access data anytime through web for accurate diagnosis.

Distributed and centralized are two different approaches to process continuous stream of data generated from multiple sources. This system used the centralized approach. The SOA used to create a gateway to combine multiple sensors network, web services sent integrated data to doctor via web and last complex event processing (CEP) generated multiple meaningful events from the cloud of events. The CEP, analyzing raw sensor data and identify abnormal health conditions.

The proposed system focused on two perspectives: healthcare monitoring of alone patients and fall detection of old patients for these two perspective environmental, body, physiological and motion sensor have been used. For rules generation Strawman EPL has been used. See table 1.

The [6] Emergency Medical Assistance (EMA) is the biggest example of Complex Event Processing (CEP) in healthcare, which quickly facilitates the sudden diseases. The reason to build EMA is to reduce the wait time to call an ambulance. For this assistance full real-time information about the current location of ambulance crew is needed.

To achieve this information built-in tablet sensors have used e.g. tablet GPS identify the ambulance location, the further

built-in sensors emitted different data, to filter out the required data and to improve ambulance availability (response time) CEP is used with EMA-DSS.

The event processing rules are generated by Event Processing Language (EPL) having two parts condition and action. The EMA-DSS is a 2-layer distributed architecture model. Layer1: *EMA Coordination Centre* located into "central control center" to locate the actual location of each ambulance to patients by using GPS. The GPS used to locate the nearest available ambulance's location. While, the layer 2: *Ambulance Vehicle (CEP)* is located into "ambulance vehicle". Every vehicle has tablets with built-in sensors. Built-in GPS used to locate ambulance position. This location received at "central control center".

In [7] the RFID plays verity of role in healthcare domain such as patient monitoring and healthcare unit, drug and medicine compliance etc. The RFID generates large volume of data, to extract medical information from this data stream, the CEP framework used with the RFID-mounted hospital.

The RFID have capability that it can detect and sense data sets from other mounted sensors in hospitals such as physiological and environmental sensors. This generates large volume of data. The CEP is the most suitable solution to get required data sets form raw data sets in real-time, do continuous processing and detect critical situation.

In this paper, the main focus is "Surgical Procedure" in the RFID-enabled hospital. Many issues of surgical procedure have been mentioned and handled in real-time by using CEP. The CEP rules generation has been defined through Event-Condition-Action (ECA) like expression of rule. While, CEP architecture designed by using an open source software Drools 5.0 including expert and Drools fusion.

The wearable sensors are efficient to senses human activities such as walking, running, cycling, eating and chewing etc. In [8] a detailed survey to recognize human activities by using wearable sensors.

Table 1: Result Table

Ref. No	***DOMIAN***	***TECHNIQUE***	***SOFTWARE SPECIFICATION***	***SENSOR***
[5]	Healthcare	Design "Patient Health Monitoring System" by using **CEP**	The rule generation by Strawman EPL.	Physiological, motion, environmental sensors
[6]	Healthcare	Develop an Emergency Medical Assistance Decision Support System (EMA-DSS) by using **CEP**	Event Processing Language (EPL) based on event processing rule (cond. Part, action. Part) Platform: Esper on an Android device	Built-in tablet's sensor and GPS
[7]	Healthcare	Develop surgical procedure system in RFID enabled hospital by using **CEP**	CEP rules generation by Event-condition-action (ECA) and CEP engine implemented by using Drools 5.0 including Drools expert and Drools Fusion.	RFID

[8]	A survey paper to recognize human activities by using wearable sensors.	**Machine Learning** (ML) application used for feature extraction.	Supervised and semi supervised algorithm such as decision tree, Bayesian methods etc.	Wearable body sensors
[9]	Healthcare	Proposed Remote Health Monitoring System by using **CEP** (CRHMS)	The CEP engine was implemented through Drools Fusion 5.4 and CRHMS system development by Java7. For IDE software development Netbeans7.0 and for backend database MySQL 5.5 is used.	Zephyr Bio-Harness 3 sensor and GPS
[10]	Healthcare	Develop a proactive Remote Patient Monitoring (RPM) with **CEP**	The implementation of this system is shown by using MyCardioAdvisor (MCA) is a mobile app	Zephyr HxM BT

[11]	Product Manufacturing	Design a Product Manufacturing System using **CEP**	Complex Event Processing Language (CEPL) used as EPL. Siddhi CEP engine by using WSO2 used Complex Event Processor.	RFID
[12]	Traffic condition monitoring	Traffic condition monitoring by using **CEP**	CoReMo simulation platform has been used. PetriNets used to model complex agents behavior. Open source "EsperTech" used as CEP engine.	GPS

This survey focuses on machine learning applications for features extraction such as supervised and semi-supervised learning algorithms. Recognition of human behavior and activity by using wearable sensors is the on-demand topic of research as it is applicable in many areas, for example medical, tactical scenarios, etc.

There are two different ways to recognition of humans' activities and behaviors. The first one is by using external sensor and the second by using wearable sensors. External sensors installed at any predefined points (camera) and wearable sensor attached to human's body.

ML application can be categorized into supervised or semi-supervised and unsupervised. Many supervised techniques have been in [8] such as decision tree, Bayesian methods, Instance based learning and neural networks, etc.

The traditional batch processing approach is slow to detect abnormalities form Remote Health Monitoring System (RHMS), while event driven approach of Complex Event Processing (CEP) using sensor is more appropriate solution for RHMS. Because CEP

correlates sensor data to generate complex events which are helpful for proper monitoring and abnormalities detection. The abnormalities detection is possible when vital signs are out of range from their normal range and that can be captured by body sensor. The proposed CEP based Remote Health Monitoring System (CRHMS) [9] is useful home alone old patients.

Several patients wear sensor, the sensor's data stream collects form patients smartphone, which also specifies patient's location. Now, the collected streams of raw data are send to CEP, the CEP system detects abnormalities in vital sign of patients and generates alert, These alerts notification are now sent to caretaker and doctor for immediate solution.

In this scenario, Zephyr Bio-Harness 3 sensor is used that can sense heart rate, ECG and respiration rate, etc. The patient's location is sensed by GPS activation in cell phone. While, CEP engine was implemented through Drools Fusion 5.4 and CRHMS system development by Java7. For IDE software development Netbeans7.0 and for backend database MySQL 5.5 is used.

In [10] Mobile-based monitoring through sensors is a very common approach in the field of medical and healthcare. In this, the person who requires remote monitoring wears a wearable sensor. The data collected form sensor are sent to smartphone for storage, where mobile CEP engine was implemented that generated complex event form raw data stream. Now, these events send to server CEP through Wi-Fi. The server CEP provides real-time feedback.

2.2. Further Application of Complex Event Processing (CEP)

Another real-time application of CEP in product manufacturing domain has been proposed in [11] by using RFID. By using RFID manufacturing system can easily detect any complexity. For real-time product's monitoring, RFID tags have been attached with products while, RFID reader mounted at different locations in the factory. Complex Event Processing Language (CEPL) is a query based language like SQL which is used to process event stream. CEPL having capability to process RFID data stream, data filtering and real-time analysis. The main part of this system is probabilistic CEP, used to detect complex event. Siddhi CEP engine by using WSO2 used Complex Event Processor.

[12] To monitor the traffic condition is another application of CEP. In this article the huge amount of traffic related data produces complex events in many scenarios and it can be handled through CEP. To implement traffic management system using CEP, CoReMo simulation platform has been used. PetriNets used to model complex agents behavior. Open source "EsperTech" used as CEP engine.

One of the common applications of CEP is Anomaly Management [13], which can be applicable in many systems such as in fraud detection to provide safety to user in case on stolen credit card or mobile, in health care to monitor different vital signs such as heartbeat, blood pressure, sugar level. of the severe ill patients, Just-in-line (JIT)-logistics, Stock market Analysis and many more areas. In [13] the anomaly management using CEP have been developed for the novel security monitoring tool for computer system.

The authors of [14] have proposed another application of CEP with Graphical Processing Unit (GPU). In this several CEP operators (event window, event filter and stream join) has been redesigned and implemented on GPUs by using NIVIDIA CUDA to increase the performance of parallel processing or parallel computing.

2.3 How Complex Event Processing (CEP) Systems respond to incoming events

The CEP is a "sense and respond" system. First CEP system 'SENSE' meaningful events from input event streams and secondly quickly 'RESPOND' them. (See figure: 1).

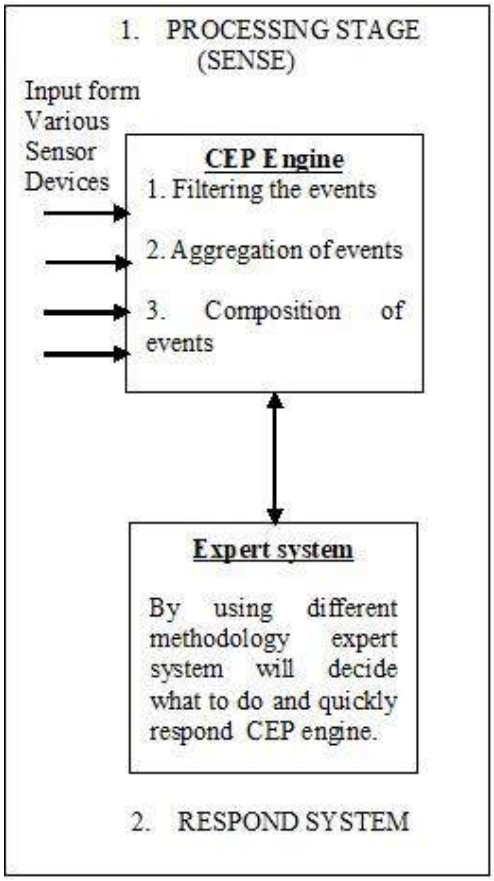

Figure 1: The CEP "Sense and Respond" Model

In this figure, a CEP sense and respond model are shown in which sense system based on CEP engine which takes input from multiple sensors. The CEP engine having following responsibilities filtering, aggregation and composition of events to generate meaningful events. While the respond system is an Expert system that quickly responds to these events by using different methodology such as machine learning algorithm, logic-based driven approach.

3. Result

The results/comparative study of the above mentioned survey is shown in table 1 which clearly indicates the techniques used and the use of wearable body sensor devices. Two of the approaches [11] and [12] are not from health domain but it can be applied to health domain.

4. Conclusion

The importance of the application of CEP in health domain is evident from the literature review. Health related data is very essential for short and long term decisions making and diagnosis of unusual and unexpected events in human body which may result in a human loss if not identified and cured.

In the light of given survey we conclude that Complex Event Processing (CEP) is the appropriate solution for many critical situation in health domains where speedy responses from expert system is the major concern. To manipulate sensor data and generate event stream form sensor data CEP is a very helpful tool.

The CEP approach will facilitate the individuals, patients and doctors to analyze and respond to unexpected events for early prediction, but also very helpful for specific patients to reduce their routine activities and traveling time. It is very beneficial for people for self-assessment and monitoring and to take precautionary measures in case of unwanted event or to immediately consult a physician.

5. Future Work

In this survey, we have shown the usefulness of Complex Event Processing (CEP) in health domain and also in other areas. In future, we will model Complex Event Processing (CEP) in health care for managing/monitoring patient health using body area sensor network.

REFERENCES

[1] Hsinchun chen, Roger H.L. Chiang, Veda C. Storey, "Business Intelligence and Analytics: Form Big Data to Big Impact", Business Intelligence Research, Vol. 36 No.4, pp. 1165- 1188/ December 2012.

[2] Mohd. Saboor, Rajesh Rengasamy, "Desinging and Developing Complex Event Processing Application", Sapient Global Markets, August 2013.

[3] Lajos Jeno Fulop, Gabriella Toth, Robert Racz, Janos Panczel, Tamas Gergely, Arpad Beszedes, "Survey on Complex

Event Processing and Predictive Analytics", Nokia Siemens Networks, July 13, 2010.

[4] Michael Eckert, Francois Bry, "Complex Event Processing (CEP)", German Language in Informatik-Spekturm, Springer 2009.

[5] Dr. V. Vaidehi, Bhargavi R, Kirupa Ganapathly, C. Sweetlin Hemalatha, "Multi-Sensor Based In-Home Health Monitoring using Complex Event Processing", IEEE, 2011

[6] Holger Billhardt, Marin Lujak, Ralf Burns, Jurgen Dunkel, "Intelligent Event Processing for Emergency Medical Assistance", SAC'14 March 24-28, 2014.

[7] Wen Yao, Chao-Hsien Chu, Zang Li, "Leveraging Complex Event Processing (CEP) for smart hospitals using RFID", Journal of Network and Computer Applications, 2011.

[8] Oscar D. Lara and Miguel A. Labrador, "A Survey on Human Activity Recognition using Wearable Sensor", IEEE Communications Surveys & Tutorials, 2013.

[9] Ravi Pathak, Vaidehi. V, "Complex Event Processing Based Remote Health Monitoring System", IEEE Computer Society", 2014.

[10] Nenad Stojanovic, Yongchum Xu, Aleksandar Stojadinovic, Ljiljana Stojanovic, "Using Mobile-based Complex Event Processing to Realize Collaborative Remote Person Monitoring", DEBS' 14, May 26-29 2014.

[11] V.Govindaramy, Dr. P. Thambidurai, "RFID Probabilistic Complex Event Processing in Real-Time Product Manufacturing System", International Journal of Engineering and Innovative Technology (IJEIT), Vol. 2, April 2013.

[12] Bogdan Tarnanuea, Dan Puiu, Dragos Damian, Vasile Comnac, "Traffic Condition Monitoring using Complex Event Processing", IEEE International Conference on System Science and Engineering, July 4-6, 2013.

[13] Bastian HoBbach, Bernhard Seeger, "Anomaly Management using Complex Event Processing", Extending Data Base Technology Paper, 2013, March 18-22, 2013.

[14] Prabodha Srimal Rodrigo, H.M.N Dilum Bandara, Srinath Perera, "Accelerating Complex Event Processing through GPUs", IEEE 22nd International Conference on High Performance Computing, 2015.

[15] David Loshin, "Business Intelligence The Savvy Manager's Guide", Newnes, 2012.

[16] Vicki L.Sauter, "Decision Support System for Business Intelligence", John Wiley & Sons, 2010.

[17] RAE. Andrade, RB. Perez, AC. Ortega, JM. Gomez, "Soft Computing for Business Intelligence", 2014.

[18] Aqil Burney and Nadeem Mahmood (2006), "A Brief History of Mathematical Logic and Applications of Logic in CS/IT" in Karachi University Journal of Science Vol. 34(1) pp. 61-75.

[19] Nadeem Mahmood, Aqil Burney, Zain Abbas, Kashif Rizwan (2012), "Data and Knowledge Management in Designing Healthcare Information Systems", International Journal of Computer Applications (IJCA) Vol. 50, No. 2, pp. 34-39.

Volume 1 | No. 2 | July – December 2017

Video Copyright Detection Using High Level Objects in Video Clip

Abdul Rasheed Balouch[1],Ubaidullah alias Kashif[1], Kashif Gul Chachar[2], Maqsood Ali Solangi[1]

Abstract:

Latest advancements in online video databases have caused a huge violation of copyright material misuse. Usually a video clip having a proper copyright is available in online video databases like YouTube without permission of the owner. It remains available until the owner takes a notice and requests to the website manager to remove copyright material. The problem with this approach is that usually the copyright material is downloaded and watched illegally during the period of upload and subsequent removal on request of the owner. This study aims at presenting an automatic content based system to detect any copyright violation in online video clips. In this technique, a video clip is needed from original video that is used to query from online video to find out shot similarity based on high level objects like shapes.

Keywords: *Digital Image Processing; Video Copyright Detection; Video Processing; Edge and Object Detection.*

1. Introduction

Increasing number of videos day by day on web servers are violating the copyright rules. They are uploading pirated videos on their web server. It is extremely difficult to keep a watch on a huge number of videos which are being uploaded on different web servers on daily basis. It is practically impossible for video producers to check each and every video manually to find out any copyright violation. In few cases, these pirated videos are uploaded even before the original version released by legitimate producers. Therefore, many systems are proposed for detecting pirated video. Like watermark and copyright detection by content based retrieval system. In watermark approach, some extra information is added to detect the video like signature [1]. But still there is no robust approach of watermark [1, 2].Content Based Copy Detection(CBCD) is an alternative approach of watermark. In Content Base Copy Detection (CBCD) Frames are taken as image to extract the features from image, motion of frames, spatial temporal features of video, events of video like fighting scene and love scene [3]. But extracting of video features and then converting these features in histogram, or calculation of features entropy are also time cost tasks specially when there are huge low level features calculations in a query. In many cases, videos are not copied from original version; they are copied from a copy. It means that when a second copy is compared with the original, simple histogram matching may not produce accurate results [1].

The approach followed in this study exploits high level features, which include objects. A user poses a query to video database, then it will extract key frame from query clip and apply image analysis method to highlight high level objects, which are highlighted by Canny Edge detection algorithm shown in figure.2. In this research authors calculate position and size of each object for comparing it with video database objects.

[1] *Department of Computer Science, Sukkur Institute of Business Administration University Sukkur, Pakistan*
[2] *Department of Computer Science, IQRA University Karachi, Pakistan*
Corresponding email: abdulrasheed.ms@iba-suk.edu.pk

Figure 1: Image taken from a Video

- Man
- Books
- Flowers
- Telephone Set
- Shelf
- Papers
- Table

For accuracy of objects retrieved by our system, authors will use Precision recall methods. These methods are used in object retrieval to measure accuracy of retrieval system[4].

Our system is extension of Canny Edge Detection in which system will extract objects highlighted by canny algorithm. Figure 1 is taken from video shows many object in one frame. Figure.2 shows the edge detection result of Canny Edge Detection algorithm which is clearly depicting the edge of each object in frame. Figure.1 is taken from a video, in this figure there are so many objects, authors have enlisted these objects beside the picture. Through Canny Edge Detection, authors have highlighted all objects in figure 2.

Figure 2: Objects detection Using Canny algorithm

Authors calculate the position and size of each object individually. It is rare to have two scenes in a video where same objects appear at the same location and size; thus, proposal can be used to differentiate scenes and extract a query scene.

2. Related Work

Content based video copy detection by motion vector is proposed by Kasm Tademir in [1]. It calculates the histogram of motion vector. The author has computed mean value of magnitude of motion and mean value of phase motion vectors of each macro blocks. An advantage of this study is that it does not consider the features of video; it just calculates motion of video. So, it reduces computation cost [1]. Another approach of copyright detection is Watermark approach. It inserts some extra bits in a video. By extracting this extra information, authors could detect pirated video [3]. However, watermark approach is not still robust [1][2][3]. Watermark approach has two significant limitations. First, since

watermarks must be introduced into the original content before copies/duplicates are made, it cannot be applied to content which is already in circulation. Second, the degree of robustness is not adequate for some of the attacks that authors encounter frequently. In CBCD approach, media itself is watermark approach in which features are extracted from image are key frame of video to match video database [3] Another drawback of CBCD is that when you compute the video similarities from a copied video, the results may not be the same as they may be when compared with original video. Another approach used by Idnyk in[5] exploits temporal fingerprints based on shot boundaries of video sequence. The drawback of this technique is that it does not work well with the short video [5]. Oostveen presents hash function for identification of video in [6]. B. Coskun et al present two methods for video copyright detection both based on discrete cosine transform for video copy detection [6]. Hampapur and Bolle have enhanced the work of Bhat and Nyar in which they compare the video on the basis of motion, color and spatial temporal distribution of intensity [7]. Y. Li ET also proposed a method to detect video copy. The author used a binary signature involving color histogram. The main advantage was that the system was robust regarding to inserting signature which frequently cropped in TV production [8]. T.N. Shanmugham and Priya Rajendran presented a good approach of content based retrieval based on query clip in [4]. The approach could be used to detect pirated video in which authors have proposed that spilt video in a sequence of elementary shots and represents each shot as elementary frame, and calculates the frame description including motion, color, edge, feature and, texture. Mani Malekesmaeili has proposed an approach namely video copy detection using temporally informative representative images [8]. Basically, author enhanced another work of video image hashing technique which is applied on either each frame or selected key frame of a video sequence. But approach never used temporal information in a video [4]. Maini has enhanced this approach and adds temporal as well as spatial information. Performance measured by a simple image hashing function on video database [5]. Earlier video matching method was reducing video sequence in small key frames [1][2][3]. Image sequence method used to match the key frame. In this approach when a shot is missed, the entire process of matching is also missed. They also ignore the temporal information [9]. A fast method of color histogram was discovered by many researchers. Yeh and Cheng have discovered a fast method that is 18 times faster than other sequence matching algorithm. They used and extended color method namely HSV color histogram [5]. Jolly, Frelicot and Buisson proposed a statistical similarity based approach, in which some interesting points are detected and then compared with Nearest Neighbor method. Interested points are extracted by Harris detector [6]. Law et al. proposed a video indexing method by using temporal contextual information which is extracted from interested point by voting function [7]. Another approach proposed by Deepak CR et al Query by video clip. They have proposed an algorithm which retrieves video from database by a query clip. Their results show that retrieval has high precision and comparatively low search time cost. They have used clustered approach for key frame matching [9]. Lienhart et al. used a mechanism to characterize key frame of the video by using color vector [8]. Another approach proposed by Deepak CR et al Query by video clip [9]. They have proposed an algorithm which retrieves video from database using query by clip. Their result shows that retrieval has high precision and searching query has low search time cost. They have used clustered approach for key frame matching [9]. Another efficient method is proposed by Kimiaki Shirahama and Kuniaki Uehara namely Query by shot. In this approach, they have classified features in high level objects and low level objects. In this study, they mainly focus on low level feature

color, region, and motion, high level object is just identified by them but not used further [10].

3. Methodlogy

The presented system extracts objects from the image. As shown in figure.3, each object in a frame is indicated by red circle. Here system extracts these objects from image. Figure.4 shows the result of extracted objects from image.

Figure 3: Objected detected in video

Figure 4: Detected objects and extracted objects from a video

Figure 5 depicts all the steps involved in our methodology. Rest of the paper defines each building block step by step.

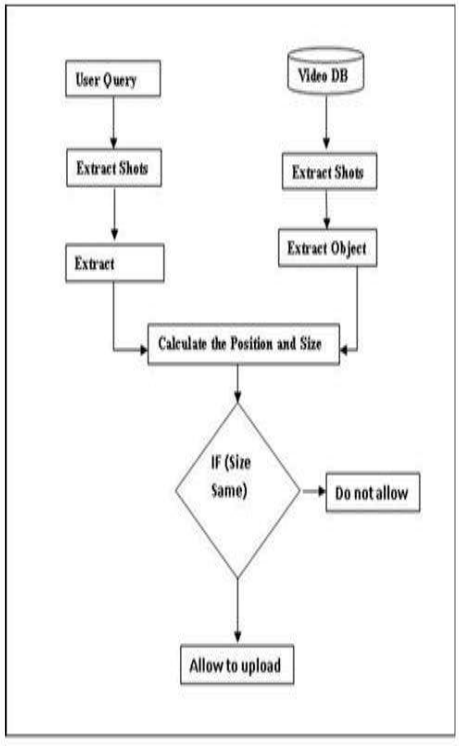

Figure 5: Showing all steps of Methodology

3.1. User Query

Authors define online video database V as,

$$V= \{v_1, v_2, v_3 \ldots\ldots v_n\}$$

Where each v_i represents a video file in an online database V.

Authors further define a set of query Q as,

$$Q= \{q_1, q_2, q_3, q_4 \ldots\ldots\ldots q_n\}$$

Where each qi represents a clip to be checked in online video database V.

In our approach, the set V represents all online video databases whereas each q in Q is submitted by video producer who want to check its video for piracy. Authors use following algorithm to determine if a video is illegally uploaded by a non-legitimated user.

Algorithm:

1. A set S containing all the shots in query clip q
2. A set V of online video database
3. For each vi in set V

4. For each si in S
    a. Determine shots in vi
    b. From each shot of si and vi, extract objects size and position
    c. if size and position of objects are found same in both clips
        i. Report video to be pirated
5. Continue to check next video

3.2. Extract Shot

Shot detection is primary step to analyze video [10].

Figure 6: Shots extracted from a video

Different algorithms are used to detect shots from video based on visual discontinuity along the time domain between two scenes of video [11]. Hard cut is most advanced algorithm to detect the shots from sequence of video frames [11, 12] .

In this research work authors have also used an existing approach of hard cut, an abrupt video shot detection based on color histogram used in [11] for finding the color histogram difference in frames authors used the expression to find the shot boundary of frames.

$$HistDif[i] = \sum_{j=1}^{M} |h_i(j) - h_{i-1}(j)|$$

Where h_i color histogram with M bins of frames *i* [11]. Figure 7 shows the step of shot detection. It compares current frame color histogram with the previous frame color histogram. If it finds a difference between these frames, it will detect as shot.

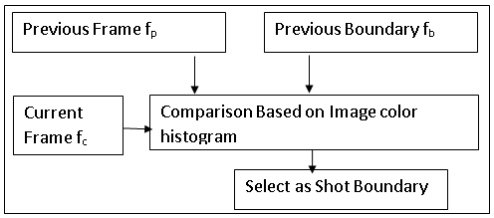

Figure 7: Frame comparison diagram

3.3. Extract Objects

Edge based object detection is a collection of homogenous edge which shows the continuity of image in same region when edge dissimilarity occurred in region shows that object edge is complete here and new object starts. For highlighting each object in image, authors have used existing approach of Canny Edge detection algorithm. This detects objects based on edge. Through canny expression of finding gradient for pixel which shows the color intensity of object [14], authors find the position of object. For the calculation of edge based intensity authors have applied a method 'Crop' which crops the segment of image highlighted by canny edge detection.

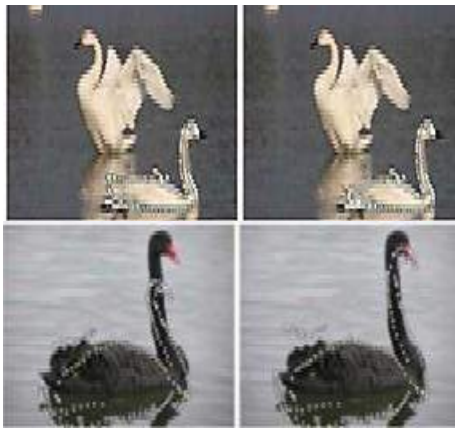

Figure 8: Objects highlighted by our System

$$|G| = \sum (Gx + Gy) \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Calculation of the total size of the object through following expression

$$S = X * Y \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where x and y show the horizontal and vertical pixel of highlighted objects. The Same process will apply on test database to test the position and size of objects.

4. Comparision

At the stage of comparison, the system compares the size and position of each object against each object of video database. If the size and position are found to be equal, the system will not allow the uploading of the video on video database server. If the size and position are not matched from both sides the system will allow to upload the video on database server.

5. Conclusion and Future Work

In this paper, authors presented a robust way to detect the pirated videos based on Content Based Copy Detection (CPCD). The presented system extracts objects from image and calculates the position and size of objects and apply same technique on video database server. Compare the size and position of objects to test the video either it is pirated or not. In future work, authors aim to fully implement the system on a large scale video database server and perform further analysis on the system.

REFERENCES

[1] K. Tasdemir, "Content based video copy detection using motion vectors," bilkent university, 2009.

[2] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, "Video copy detection: a comparative study," in Proceedings of the 6th ACM international conference on Image and video retrieval, 2007, pp. 371-378.

[3] D. Zhang and S.-F. Chang, "Event detection in baseball video using superimposed caption recognition," in Proceedings of the tenth ACM international conference on Multimedia, 2002, pp. 315-318.

[4] B. Taneva, M. Kacimi, and G. Weikum, "Gathering and ranking photos of named entities with high precision, high recall, and diversity," in Proceedings of the third ACM international conference on Web search and data mining, 2010, pp. 431-440.

[5] P. Indyk, G. Iyengar, and N. Shivakumar, "Finding pirated video sequences on the internet," Tecnical Report, Stanford University, 1999.

[6] J. C. Oostveen, T. Kalker, and J. Haitsma, "Visual hashing of digital video: applications and techniques," in International Symposium on Optical Science and Technology, 2001, pp. 121-131.

[7] A. Hampapur and R. Bolle, "Feature based indexing for media tracking," in Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on, 2000, pp. 1709-1712.

[8] S. Poullot, O. Buisson, and M. Crucianu, "Z-grid-based probabilistic retrieval for scaling up content-based copy detection," in Proceedings of the 6th ACM international conference on Image and video retrieval, 2007, pp. 348-355.

[9] L. Teodosio and W. Bender, "Salient video stills: Content and context

preserved," in Proceedings of the first ACM international conference on Multimedia, 1993, pp. 39-46.

[10] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of TRECVid activity," Computer Vision and Image Understanding, vol. 114, pp. 411-418, 2010.

[11] J. Mas and G. Fernandez, "Video shot boundary detection based on color histogram," Notebook Papers TRECVID2003, Gaithersburg, Maryland, NIST, 2003.

[12] S. M. Doudpota, S. Guha, and J. Baber, "Shot-Based Genre Identification in Musicals," in Wireless Networks and Computational Intelligence, ed: Springer, 2012, pp. 129-138.

[13] J. Canny, "A computational approach to edge detection," Pattern Analysis and Machine Intelligence, IEEE Transactions on, pp. 679-698, 1986.

.

Volume 1 | No. 2 | July – December 2017

Comparative Study of Load Testing Tools: Apache JMeter, HP LoadRunner, Microsoft Visual Studio (TFS), Siege

Rabiya Abbas[1], Zainab Sultan[1*], Shahid Nazir Bhatti[1], Farrukh Latif Butt

Abstract:

Software testing is the process of verifying and validating the user's requirements. Testing is ongoing process during whole software development. Software testing is characterized into three main types. That is, in Black box testing, user doesn't know domestic knowledge, internal logics and design of system. In white box testing, Tester knows the domestic logic of code. In Grey box testing, Tester has little bit knowledge about the internal structure and working of the system. It is commonly used in case of Integration testing.Load testing helps us to analyze the performance of the system under heavy load or under Zero load. This is achieved with the help of a Load Testing Tool. The intention for writing this research is to carry out a comparison of four load testing tools i.e. Apache JMeter, LoadRunner, Microsoft Visual Studio (TFS), Siege based on certain criteria i.e. test scripts generation , result reports, application support, plug-in supports, and cost . The main focus is to study these load testing tools and identify which tool is better and more efficient. We assume this comparison can help in selecting the most appropriate tool and motivates the use of open source load testing tools.

Keywords: *Testing, manual testing, automated testing, testing tools, load testing, stress test.*

1. Introduction

The objective of software testing is to find defects, errors and bugs in a software, system or product. Software testing is characterized into manual testing and automation testing.

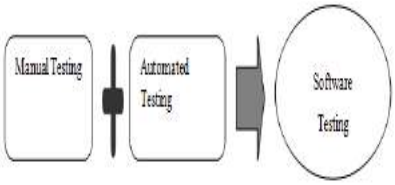

Figure 1: Software testing

Manual testing is executed by the testers. First of all a written test plan is created, and followed by testers that provides a guideline through different steps. But there are a lot of problems faced by testers like it is very time taking and consuming, no reusability, has no scripting feature, much human effort required, and still major or minor bugs remain unexplored. Therefore to cover all types of errors and bugs automation testing has introduced that explores all the issues exist in manual testing [11]. All Automation testing tools test software in less time, produce more reliable, repeatable, and reusable final product.

Load testing is activated when we steadily raise the load upon a system until it reaches a target load. Usually this is the maximum load, average or even zero load. The goal of a load testing is to discover functional and performance issues of a system under load. Load testing is appropriate for testing the performance of web site, and its framework[18].

The intention for writing this research is to carry out a comparison of four load testing tools i.e. Apache JMeter, LoadRunner, Microsoft Visual Studio (TFS), Siege based on some parameters. This research paper is divided into different sections. Section 1 consists of introduction.

[1] Department of Software Engineering, Bahria University Islamabad, Pakistan
[*] Corresponding Author: zaini.1984@gmail.com

Literature review is discussed in Section 2. Research methodology is presented in section 3 .In section 4, we present evaluation study. In section 5, and comparison table of automated testing tools is presented. In section 6, results and analysis of study is presented and in section 7, on the basis of research, conclusion is presented.

2. Literature review

This section presents the literature review of research topic. Farmeena Khan Et.al in their paper describes and compares three main testing techniques: "White box, Black box and Grey box testing". Authors' presents the future changes in software testing industry due to new technologies like SOA (Service Oriented Architecture) & mobile technologies etc[3]. Niranjanamurthy et .al in their paper discusses testing terminologies used in testing, levels of testing, analysis of automated and manual techniques, comparison of Selenium and QTP Tools, comparison of "White box, Black box and Grey box testing techniques" [2]. Taraq Hussain et al in their research mentions that testing can never completely diagnoses all errors of a software but it provides evaluation techniques which helps the tester to diagnose a problem. After comparison they show that the White Box Testing is best suitable for software reliability[4]. "The Growth of Software Testing" is written by "David Gelperin and Bill Hetzel". In this paper, authors describes the evolution of different software testing process models , their merits and demerits due to which some of these are failed . From 1956 to present, different software testing models are discussed; Changes in these models are evaluated[14].

Harpareet Kaur and Dr.Gagan Gupta in their paper discussed the two ways of testing: Manual testing and Automation testing. In this research paper, they discuss the parameters of "Selenium 2.0.0, Quick Test Professional 10.0, and TestComplete 9.0". These three tools comparison is based on different specification and parameters. After analysis, researchers concluded that anyone can choose the testing tool on the basis of budget and nature of software that has to be tested. Researchers found that QTP is more suitable among all that three tools[6].

Neha Bhatia in her paper discussed manual testing and automation testing. If the requirements are continuously changing and regression testing is needed to perform repeatedly, then automated testing is more suitable in that environment. In this paper, researcher discussed different automation tools[7]. Neha Dubey et al in their paper compare and study concepts and characteristics of two "software automated tools Ranorex and the Automated QA TestComplete" that are based on some criteria. After comparison they concluded that "Ranorex" is the best tool for web based applications [10].Monika Sharma, Vaishnavi S. Iyer, Sugandhi Subramanian, and Abhinandhan Shetty in their paper focuses on comparing load testing tools- Apache JMeter, HP LoadRunner, WebLOAD, and The Grinder on the basis of different parameters[11].

3. Related Work

Comparison between load testing tools has been done by many authors. Vandana E. [13] [20], have done comparative study of testing tools which are jmeter and load runner. They described advantages and disadvantages of both tools and recommended that Jmeter is much better than Load Runner because it has clean UI that offers much simplicity. Bhatti Et.al [19], described number of load testing tools for test web applications. The testing tools they discussed are Load Runner, NeoLoad, WAPT, Soasta Cloud Test, LoadStorm, Loadster, Apache, JMeter, HTTPERF, LoadUI, and LoadImpact. They analyzed that among all tools to test a web application, NEOLOAD is best for load testing due to its visual programming and its script less design. Rina [21] analyzed the NeoLoad, WAPT and Loadster tools on different browsers and compared the results of their performance. One site has tested on above three tools for performance. The comparison they done provides a better understanding for selecting the best tool according to requirements and possibilities, however they concluded that It is difficult to compare tools because many parameters values are not considers in all tools. Upadhyay [22]compared some specific performance testing tools for their usability and effectiveness. WAPT and RANOXEX performance testing tools inferences, implications and results have been presented and discussed. Different attributes, their ability to compare the

results, test cases documentation ability and regression testing performance ability have been compared. Dart Et.al compared software web tools in terms of their dynamic test generation ability [18]. A survey has been presented on static and dynamic testing analysis. Sufiani [23] compared different performance testing tools response time and justified these differences include architecture and simulation mechanism.

4. Research Methodology

Testing is an important and critical part of the SDLC. In recent times different automated software testing tools are available in market. Several studies are available in which comparisons of different testing tools are done. According to our observations, there is no comparative analysis on the load testing tools, such as "JMeter, Siege, LoadRunner, and Microsoft Visual Studio (TFS)". In this paper, we compare these load testing tools on the basis of different parameters.

4.1 Automated software testing tools

A brief explanation and comprehensive account of automated software testing tools is taken here in this section.

4.1.1 Apache JMeter

Apache JMeter is an "open-source testing tool" developed by "Apache Software Foundation (ASF)". JMeter's main function is to load test client/server. Moreover, JMeter is used in regression testing by generating test scripts [12]. JMeter provides offline reporting of test results. JMeter test reports are shown in fig 2.

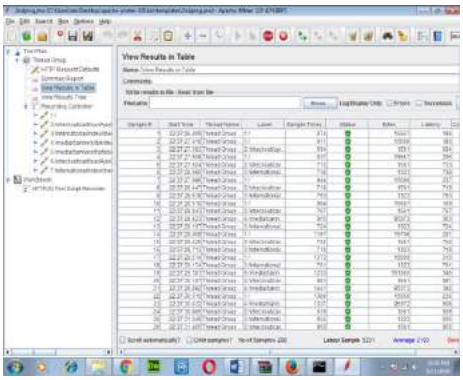

Figure 2: JMeter test reports

4.1.2 LoadRunner

HPE LoadRunner is an "automation testing tool" from "Hewlett Packard" enterprise [15]. HP LoadRunner software testing tool helps you to detect and prevent from software performance issues by identifying bottlenecks [16]. HP LoadRunner Scripting and test report summary is shown in Fig 3.

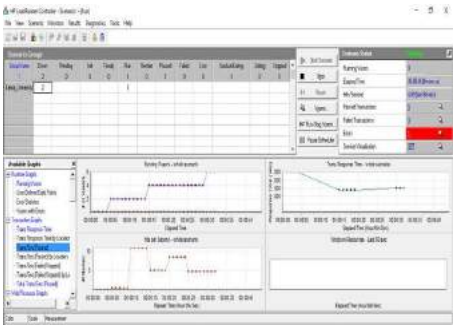

Figure 3: HP LoadRunner scripting & test summary reports

4.1.3 Siege

"Siege" was developed and implemented by "Jeffrey Fulmer" as Webmaster for Armstrong World Industries. It is a software load testing tool which is very productive in detecting the performance of system when load exists [13]. Siege executing commands and Test summary reports is shown in fig 4.

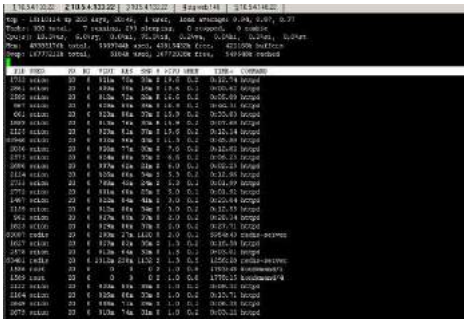

Figure 4: Siege executing commands &test report summary

4.1.4 Microsoft visual studio (TFS)

"Team Foundation Server (TFS)" is a load testing tool which facilitate with source code management, Project management, Requirement management, reporting, testing capabilities, automated builds, lab management, [17] TFS test Summary reports are shown in fig 5.

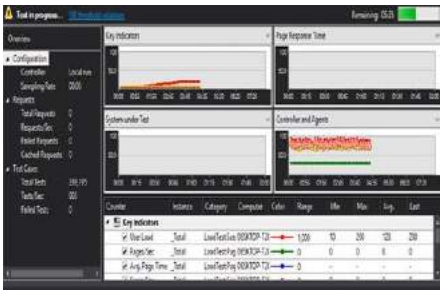

Figure 5: TFS test summary report

5. Evaluation Study

Now a day's different open source and profitable Load testing tools are available in the market. For this comparative study, we are using the latest and running versions of "Apache JMeter, LoadRunner, Microsoft Visual Studio (TFS), and Siege". Through these tools we test the Bahria University Islamabad campusWebsite (http://www.buic.edu.pk/) and for Siege, we test telecommunication company website (www.telenor.pk). Comparison between these four tools is made on the basis of list evaluation parameters with the explanation.

6. Result and Analysis of Study

For assessment of the parameters, we use 3-point scale in a graph i.e. 3, 2, 1 as Best, Average, and Worst respectively. Different value for different parameters with selected automated tools is verified. The calculated value of parameters is used for conclusion and investigation of this comparative study. The overall comparison based graph for these four automated load testing tools is shown in Fig 6.

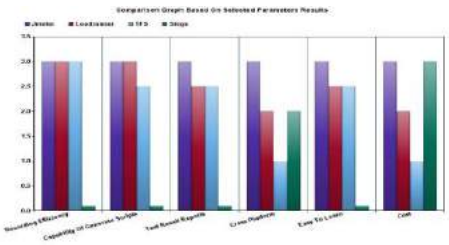

Figure 6: Comparison graph based on selected parameter results

7. Results and Findings

After comparison and analysis, we conclude that anyone can choose the testing tool on the basis of budget and nature of software that has to be tested. Apache JMeter, LoadRunner, Microsoft Visual Studio (TFS), Siege all four are good tools for test automation. We take two different websites Bahria University Islamabad campus Website (http://www.buic.edu.pk/) and for Siege, we test telecommunication company website (www.telenor.pk) because Siege cannot test live sites. Each tool has its own benefits and drawbacks too; a detail analysis in this context is in Table I below. It is to be noted for Telenor System we did have the access code available but for the Bahria University we did not have access to the code.

Siege can reduce the cost as it is open source but it has limited options to be used as it is command line tool and it sometimes generate inaccurate result.

HP LoadRunner is best for performance checking when load found. It can handle multiple users at same time butit has some configuration or installation problems across firewalls and its licensing cost is high. Microsoft Visual Studio (TFS) is user friendly .It has built-in testing capabilities whether there are 100 parallel users or 1000s, it is easy to test according to user requirements but it can only supports Windows OS and it has high licensing cost. Apache JMeter is best option as it is free of cost (see fig 7 in this). It takes more time on one time installation but it has broad set of options for result analysis and it is good for different tests to be run simultaneously .It has several plugins which raise its testing capabilities.

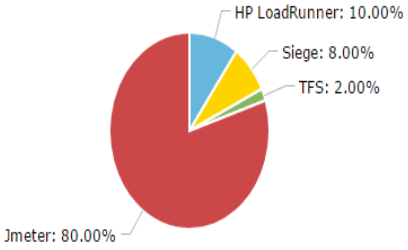

***Figure 7**: Pie chart showing JMeter as highest use*

TABLE-I: EVALUATION PARAMETERS WITH EXPLANATION DETAILS	
Appropriate Parameters	**Explanation**
Recording Efficiency	For handling the application which is to be tested.
Generating scripts capability	Generating the interrelated scripts.
Test Result Reports generation	Efficient investigation of test scripts
Cross platform	Operating systems on which tool operates on
Easy to learn	GUI(GRAPHICAL USER INTERFACE)
Application support	Applications that are supported by testing tool
Scripting languages	Languages that are used for scripting
Plugin support	Testing tools either support plugins or not
Licensing Cost	Reasonable or low cost

8. Comparison of Automation Testing Tools

In this section, we compare the automation testing tools. This comparison is beneficial for the testers/researchers (technical stake holders) to choose the more appropriate load test tool as per requirements. Table II presents comparison of automated testing tools i.e. Apache JMeter, LoadRunner, Microsoft Visual Studio (TFS), and Siege on the basis of different parameters (characteristics).

TABLE-II: COMPARISON OF WEB SERVICE TESTING TOOLS				
FEATURES	**JMeter**	**Siege**	**LoadRunner**	**Microsoft Visual Studio(TFS)**
Cost	It is free or no renovation cost for this tool.	It is free of cost.	License is expensive	About $2K licensing cost
Application support	Static and dynamic resources, Web services and databases.	Only used for testing websites.	Websites and other applications.	Website & Other applications.
Scripting language	Javascript ,BeanShell	Skrit	Citrix, ANSI C, .Net and Java	PowerShell ,Perl
Cross Platforms	Supports Windows PC/MAC/UNIX Platforms.	Supports UNIX, AIX, BSD, Solaris.	Supports Microsoft Windows and LINUX OS.	Supports Windows 7, Windows Vista, Windows Server 2008 or later operating systems.
Plugin support	It has several plugins which raise its testing capabilities.	Siege has small no. of plugins.	It has several plugins which raise its testing capabilities.	It has several plugins which raise its testing capabilities.
Interface	GUI	Command Line Interface	GUI	GUI
Benefits	It provides GUI and has many features that can be used while testing.	It has faster setup. It is good for quick results.	It is best for performance checking where there is actual load.	It is simple to use. It has inherent testing capabilities.

	It has vast set of options for result analysis. It is good for different tests to be run simultaneously. It gives accurate results.		It can handle large no. of users at the same time. It can also checks network and server resources for improving performance. It automatically trace client/server performance while testing	It uses graphical illustrations in reports. Whether there are 100 parallel users or 1000s, it is easy to test as per requirements.
Drawbacks	JMeter takes more time to setup as it involves many steps.	It has limited options to be used as it is command line tool. It sometimes generate inaccurate result.	It has some configuration or installation issues across firewalls.	It only supports Windows OS. It has high licensing cost.
Report Generation	JMeter supports dashboard report generation to get graphical illustrations.	Reports total no.of transactions ,server response etc.	It allow user to convert performance report into word, excel, pdf etc.	In this, reports are generated in SQL Server Reporting Services.

9. Conclusion

Improving software quality and performance has become a priority for almost every organization that relies on the software development. Thus the quality issue related to the software's industry becomes more important, apparent and more technical also considering the user's requirements in this aspect. Software systems have to ensure consistent and bug free execution at a rapid pace every time they are used especially in web based development.

In this work we have performed a thorough and comprehensive comparison and analysis using different tools/ technologies available for testing (load testing as case scenario). After a through analytical review of these different tools mentioned in sections IV and V for Load testing, we summarize here that anyone can choose the testing tool but on the basis of budget, time and nature of software system under consideration that has to be tested. Besides Each tool have its own benefits and drawbacks, and have to keep in queue when performing anyone of the mentioned testing strategies (or any other). Apache JMeter, LoadRunner, Microsoft Visual Studio (TFS), Siege all four are good tools for test automation. But we have shown that JMeter provides better results than any other tested tools (techniques) as it is a ratio scale methodology, and also includes a consistency check.

In future work, with the access to code (for web projects) the applications and values (attributes) of these tools can be estimated, especially in case of stress testing while performing Load testing. As stress testing evaluate the system when stressed to its limits over a short period of time and that following testing is especially important for systems that usually operate below maximum capacity but are severely stressed at certain times of peak demand.

REFERENCES

[1] Rapinder Singh ,Manprit Kaur ,"A review of software testing techniques",(IJEEE), ISSN 0974-2174, Volume 7, pp. 463, 2014.

[2] Nitesh S N, Niranjanamurthy M, Balaji Sriraman ,Nagesh S N, "Comparative Study of Software Testing Techniques ", IJCSMC, Vol. 3, Issue. 10, October 2014, pg.151 – 158.

[3] Farmeena Khan, Mohammad Ehmar Khan ,"A comparative study of white box , black box , grey

box testing techniques" , (IJACSA) , Vol. 3, No.6, 2012.

[4] Taraq Hussain, Dr.Satyaveer Singh, "A Comparative Study of Software Testing Techniques Viz. White Box Testing Black Box Testing and Grey Box Testing", (IJAPRR) , ISSN 2350-1294.

[5] Kamna Solanki, Jyoti, , "A Comparative Study of Five Regression Testing Techniques: A Survey ", International Journal of Scientific & Technology Research (IJSTR), Volume 3, Issue 8, August 2014.

[6] Neha Bhateja , "A Study on Various Software Automation Testing Tools" ,(IJACSA), Volume 5, Issue 6, June 2015 .

[7] Richa Rattan, "Comparative Study Of Automation Testing Tools: Quick Test Professional & Selenium", IJCSIT, Vol. 3, No. 6 June 2013.

[8] Raj Kumar, Manjit Kaur, "Comparative Study of Automated Testing Tools: Test Complete and Quick Testpro", Intl. Journal of Computer Applications (0975-8887), Volume 24, No. 1, June 2011.

[9] Monika Sharma, , Abhinandhan Shetty, Sugandhi Subramanian, Vaishnavi S. Iyer,"A Comparative Study on Load Testing Tools" , Int. Journal of Innovative Research in Computer and Communication Engineering ,Vol. 4, Issue 2, February 2016

[10] Neha Dubey, Mrs. Savita Shiwani , "Studying and Comparing Automated Testing Tools; Ranorex and TestComplete" , (IJECS) , Volume 3, Issue 5, Pp. 5916-5923

[11] Sanjay Tyagi , Pooja Ahlawat , "A Comparative Analysis of Load Testing Tools Using Optimal Response Rate" , (IJARCSSE), Volume 3, Issue 5, May 2013.

[12] Dr.K..V..K .K Prasad , Software Testing Tools: Covering WinRunner, SilkTest, LoadRunner, JMeter, TestDirector and QTP with Paperback , 2007 .

[13] Ibrahima Kalil Toure, Abdoulaye Diop, Shariq Hussain and Zhaoshun Wang, "Web Service Testing Tools: A Comparative Study", IJCSI Int. Journal of Computer Science Issues, Vol. 10, 2013.

[14] Daniel A. Menasce, "Load Testing of Websites", http:/computer.org/internet, IEEE Internet Computing, PP.70- 74, 2002.

[15] Li Xiao-jie, Zhang Hui-li and Zhang Shu."Research of Load Testing and Result Application Based on LoadRunner", National Conference on Information Technology and Computer Science, 2012.

[16] Sneha Khoria, Pragati Upadhyay, "Performance Evaluation and Comparison of Testing Tools", VSRD Int. Journal of Compt. Science & IT, Vol. 2, 2012.

[17] Sinha M, and Arora A., "Web Application Testing: A Review on Techniques, Tools and State of Art", (IJSER), Volume 3, Issue 2, 2012.

[18] Vandana Chandel, Shilpa Patial Sonal Guleria, "ComparativeStudy of Testing Tools: Apache JMeter and Load Runner",IJCCR, VOLUME 3 ISSUE 3 May 2013.

[19] Sandeep Bhatti, Raj Kumari, "Comparative Study of LoadTesting Tools", ijircce, Vol. 3, Issue 3, March 2015.

[20] Rina, Sanjay Tyagi, "A Comparative Study of PerformanceTesting Tools", Volume 3, Issue 5, May 2013.

[21] Dr. S. M. Afroz, N. Elezabeth Rani and N. Indira Priyadarshini, "Web Application– A Study on ComparingSoftware Testing Tools", International Journal of ComputerScience and Telecommunications, Volume 2, Issue 3, June2011.

[22] Muhammad Dhiauddin Mohamed Suffiani, Fairul RizalFahrurazi, "Performance Testing: Analyzing Differences ofResponse Time between Performance Testing Tools", inproceeding of International Conference on Computer &Inf. Science (ICCIS) 2012.

Volume 1 | No. 2 | July – December 2017

Utilization of Electronic Learning System in Swat Rural Areas

Nazir Ahmed Sangi[1*], Habib ur Rahman[1]

Abstract:

As developments in electronic technologies i.e. personal computers, laptops, tablets, mobiles and wearable devices, the way of learning is also changing. Therefore, utilization of Information and Communication Technology (ICT) has great important role in schools and colleges. ICT is using by students, teachers and societies in District Swat, KP, Pakistan in the form of mobiles internet (for social contact and chat), computers internet (for knowledge exploration and entertainment) and multimedia (for teaching and learning). One of the difficulties involved in rural areas' students of District Swat is that they cannot join class rooms due to their poor livelihood condition and far away from schools and colleges. Especially most of the females of rural areas of Swat do not come to schools and colleges for their family tradition and culture. Various questions were examined in every aspect of educational technologies in this study. We surveyed 50 responded randomly at District Swat from different schools and colleges and discovered that the responded were generally positive and have great interest about e-learning in Swat. The use of proposed electronic system for the learning, the literacy rate will increase in rural areas and students will achieve their individual goals.

Keywords: *ICT, educational technology, electronic learning system.*

1. Introduction

E-learning means to use electronic technologies and methods to study or learn. According to Beal web-based learning, computer-based learning, virtual classrooms and digital collaboration [19] are processes and applications of e-learning. In Swat Valley the home usage of personal computers, laptops and internet is relatively very moderate. Due to recent growth in ICT, Internet and ICT in the form of educational technologies are having great roles in the improvement of education which are providing more opportunities and flexibility to the teachers and students [1]. Dekel defined e-learning or electronic learning as online learning, internet-based learning, web-based learning, distributed-learning, or computer-assisted instruction [2]. Due to electronically support education and training pedagogy for student hub and collective learning has turned into familiar. E-learning is a totally new learning platform for students and teachers, therefore, computer skills are involving for its implementation. Due to evolutions in information communication technology, students study or learn without schools' places, therefore teachers' responsibilities and students' learning practices are also changing [3]. There are several definitions of e-learning offered such as using of ICT to help and facilitate students to learn or study [3]. A system which has abilities of teaching and learning methods [4]. The rest of the paper is organized as follows. In session 2 covers related literature review, session 3 covers survey objectives, session 4 shows the research questions, session 5 shows the survey sampling and population, session 6 describes survey results, session 7 is about the research conclusion and future work, session 8 discuss lesson learned and finally shows the related references.

2. Literature Review

According to E. Ross e-learning materials are delivered by means of the internet, intranet, audio and video clips i.e. it can be

[1] Department of Computer Science, Allama Iqbal Open University, Islamabad Pakistan
[*] Corresponding Author: nazir.sangi@gmail.com

self-paced or tutor guided [5] and includes different sources in the form of text, image, computer graphics, video and audio. Education experts have the same opinion that to organize students for the e-learning, our traditional educational system requires use of ICT. Therefore, E. Ross described that it is essential for teachers to accept e-learning i.e. global learning environment to use information communication technology with their old-style of education [5]. Knowledge and skills are sources of success and needed in current education system (primary schools, high schools and higher secondary schools) all over the world. R. Sims stated in his paper that e-learning model is new practice of learning and activity which improved and transformed traditional style of learning in the form of well-organized, effective and attractive new technology models [6]. Learners can get opportunities and enhance their knowledge from blended learning which combines face-to-face and online mode of instruction [7]. Here, we will discuss ICT utilization effort in education and e-learning implementation and status in different countries and Pakistan…

2.1. United State America

The R. Abel observed 89% in The National Center for Education Statistics that United State public institutions offered both degree and certificate programs in online mode of instruction [8]. Acellus Academy has started e-learning program which is benefited more than one million students in every state of U.S. [9]. Rebecca extracted a report from Sloan Consortium Survey on Online Education in the U.S. in year 2011 and stated that online enrolments evaluation is ten times higher than traditional mode [10].

2.2. China

Y. Zhang detected that China introduced e-learning in 1998 and China government is assisting e-learning in the form of funding and policy, therefore Ministry of Education (MoE) permitted 68 online colleges in 2004, during this year 190,000 students have awarded graduate degrees from these online colleges in different courses like administration, medicine, arts, social sciences, physical sciences, engineering and technologies etc. [11].

2.3. Thailand

Thailand Cyber University (TCU) initiated a web-based learning management system since 2004. TCU web based learning system registered 41 universities for their online curricula that's why now i.e. year 2016; 17 curricula, 635 lessons, 4040 lecturers and 107068 students registered online with TCU project [12].

2.4. Pakistan

Allama Iqbal Open University (AIOU) is the first Open University in Pakistan which started distance education from secondary school to postgraduate level courses and online learning system for the learning of bachelor level courses. Similarly, NUST, COMSAT and FAST universities are some of the prominent national level electronic programs to disseminate consciousness and education of ICTs. Now, there are many public and private universities which making contribution for the virtual learning facilities but the Virtual University (VU) started online learning support first time in Pakistan [16].

- Punjab province has initiated e-learning which only digitizes textbooks and adds some videos in between but e-learning awareness and understanding in students and teachers and government policy for the e-learning are still remaining issues in Punjab. Punjab government, along with PITB (Punjab Information Technology Board) are started combine efforts for the enhancement in education system using ICT specially in online textbooks, online syllabus, online tutorials and online activities [13].
- Taaleem Foundation [18] started virtual class rooms i.e. e-learning education in many districts of Sindh and Baluchistan provinces which utilize digital whiteboards, ICT components and online tutorial applications for the online instruction mode.
- The Khyber Pakhtunkhwa government initiated IT Labs [17] in province and laptop schemes, the KP pilot project

distributed 2800 tablets to teachers but e-learning concept, understanding and policy for the learning in KP are still remaining problems. Government only just focused on hardware rather than utilization of ICT for education.

- There is no such initiative like e-learning in schools and colleges of district Swat, KP Pakistan still started or ICT policy made for e-learning. Only traditional education system i.e. education or learning within class rooms exist in district Swat and it is great hurdle to increase literacy rate especially in females. Therefore, we arranged a survey and found a solution i.e. e-learning.

Naveed Ahmad asked from Government of Pakistan [14], it is clear that e-learning will be important for standard education but now the question is, why is the government not pursuing this 21st century movement seriously? Why doesn't the government integrate internet and computer-based technologies in its schools, colleges and universities?

E-learning gives equal and excellent education opportunities to urban and rural areas in the form of equal quality online learning contents, designs and pedagogy methods i.e. same education in everywhere [15].

After different literatures reviewed, we arranged a pilot survey in district Swat to find out the students' interest in e-learning because this is the modern education need and solution to increase literacy rate especially of females in rural areas.

3. Survey Objectives

The objectives of this pilot survey for the Swat, Pakistan students as under:

- To find out the usages of computers and internet for the learning.
- To find out problems faced by students from traditional learning system in Swat, Pakistan.
- To survey the District Swat, KP Pakistan students' willingness to adopt e-learning.
- In e-learning education, females can be capable to acquire benefits of these prospects and increase educational levels and increase literacy rate in rural areas of Swat.
- E-learning is learning anytime, anywhere i.e. personalize approach and highly attractive learning experience in rural areas of Swat.

We propose to make an e-learning system according to modern needs, government syllabus, and students' interest and to increase literacy rate especially in females. This model will help for the whole KP province, Pakistan country and for the whole universal.

4. Research Questions

A list of questions was used as the main tool for the research survey. The analysis included an assessment of the questions responses from which data were collected. This data from the quantitative questionnaires analyzed in Microsoft Excel statistically. The following some questions were proposed:

- Do you like to learn or study independently (self-study)?
- If you want to study or learn, then what kind of learning mode do you like to learn from home?
- Would you please identify the present learning system problems faced by you?
- What are learning needs that you are feeling flexible and easy for learning access?
- What kind of instruction mode do you like for learning?

5. Sampling and Population

A total of 66 questionnaires in hard copies were distributed across 14 education centres which were included 8 high schools (4 males and 4 females) and 6 higher secondary schools (4 males and 2 females).

The area of Swat is 5,337 km^2 and population is 2.2 million (rural 86% and urban 14%) (Source: Wikipedia 2015). Swat has a total of 1,530 public sector all levels schools i.e. from primary to and higher secondary level schools. The number of male high school is

16% i.e. 242 (government 77 and private 165) and female high school is 3% i.e. 52 (government 28 and private 24), male higher secondary school is 4% i.e. 64 (government 14 and private 50) and female higher secondary school is 1% i.e. 18 (government 8 and private 10) (Source: DEO Swat 2016). The literacy rate of Swat is 68% (53% males and 15% females) (UNICEF 2014).

6. Survey Result

The pilot survey was deployed during a period of two months (from April to May of 2016) in District Swat hilly areas. As far as the pilot survey sample was concerned, 76% (50 out of 66) respondents out of 60% (30 respondents) were males and 40% (20 respondents) were females responded to this question (see fig. 1). As fig.2. shows, the ratio of the respondents i.e. teachers 20% (10 out of 50) and students 80% (40 out of 50).

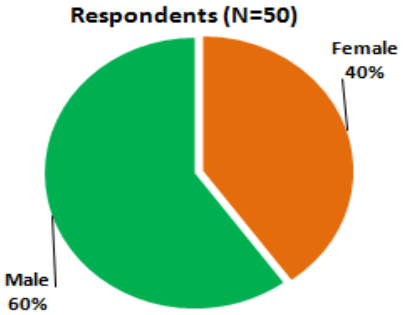

Figure 1: Gender base ratio

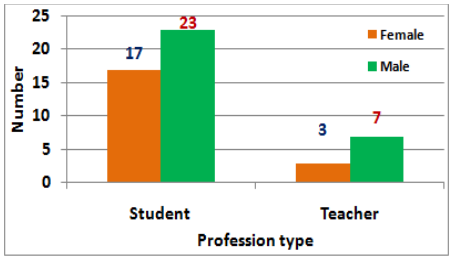

Figure 2: Profession type segregation

From the students' point of view, it is very important to know whether students at the time of taking survey, how many years they have been using computer. See the figure 3, the major respondents i.e. 74% have been used computers since 1 to 5 years. Similarly, 12% used since 6 to 10 years' group, 2% used since 11 to 15 years' group and 12% was not using computer group.

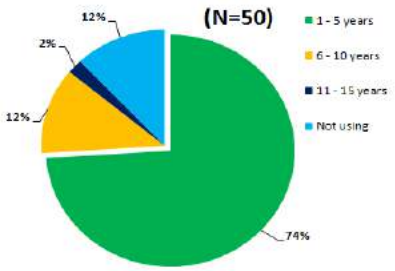

Figure 3: Usage of computer

In the survey respondents were asked some related valid questions about the utilization of e-learning in District Swat, KP Pakistan and found their response results.

6.1. Respondents like learn independently (self-study)

As fig.4. demonstrates, 27 respondents (54%) were frequently like to learn by self-study. Moreover, 14 respondents (28%) were occasionally like to learn independently, 6 respondents (12%) were not like and only 3 (6%) had no know about self-study (independently).

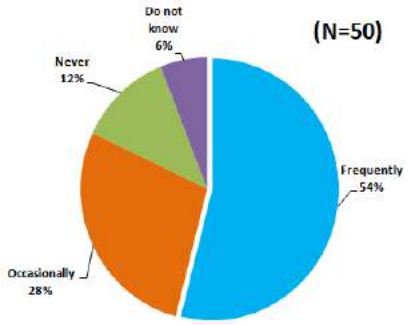

Figure 4: Like to learn independently

6.2. Respondents like learn mode from home

When answering this question of multiple choice question type, respondents could tick more than one option. Therefore, 10 respondents (20%) reported that they like having the study in online form while the 10

respondents (20%) said that they would prefer to self-study in net. 9 respondents (18%) responded that they would desire to learn in the form of DVD or USB or Audio or Video materials. Similarly, 9 respondents (18%) through distance education, 8 respondents (15%) through web courses, and fewer respondents do not know learning mode from home (see fig.5).

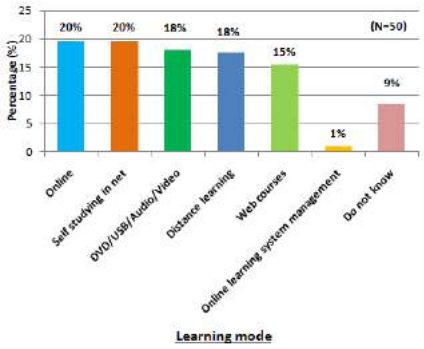

Figure 5: Respondents like learning mode from home

6.3. Present learning system (traditional system) problems

The survey respondents responded according to the present learning system i.e. traditional system problems faced by the different areas' respondents of Swat (see fig. 6). The majority respondents faced problem by distance i.e. 26% (13 respondents) followed by cost problem i.e. 24% (12 respondents). Similarly, instruction materials 18% (9 respondents), instruction mediums 8% (4 respondents) and 24% (12 respondents) have no time to learn in a class.

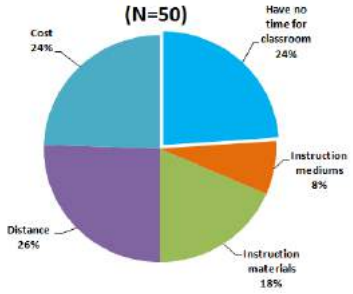

Figure 6: Present learning system's problems

6.4. Learning needs that respondents are feeling flexible and easy for learning access

The majority of respondents i.e. 10 (20%) having great interest that Online or Internet need is a flexible and easy for learning access. Similarly, respondents also like Video or Audio clips for easy learning access i.e. 9 respondents (19%) followed by 7 respondents (14%) shown interest for learning applications need, 8 respondents (17%) for handouts, 4 respondents (9%) for simulation tutorials, 4 respondents (8%) for distance education, 2 respondents (4%) for Blog/Wikis and 4 respondents (8%) do not know about learning need (see fig.7).

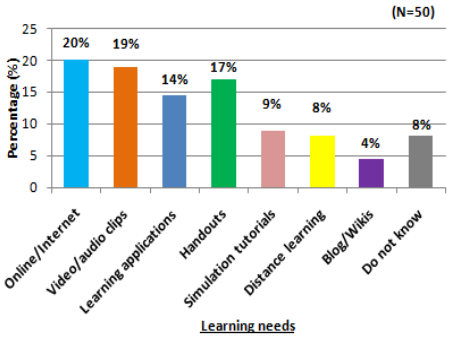

Figure 7: Learning needs for flexible and easy learning access

6.5. Respondents like instruction mode for learning

As fig.8. shows that the majority of respondents like Online/Internet instruction mode i.e. 21% (10 respondents), followed by Video/Audio clips instruction mode i.e. 19% (9 respondents), Learning application 17% (9 respondents), Simulation tutorials 12% (6 respondents) and Handouts instruction mode liked by respondents 12% (6 respondents). Similarly, Distance learning and Blog/Wiki instruction mode liked by respondents i.e. 6% (3 respondents) each respectively. 8% (4 respondents) do not know about instruction mode.

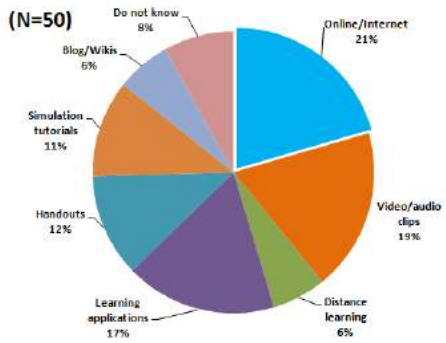

Figure 8: Respondents like instruction mode of learning

7. Conclusion and Future work

As this research of pilot survey showed, a majority of respondents welcomed and great interested a possibility of having their study or learn in e-learning mode i.e. 28 respondents: 57% (Online/Internet 10 respondents: 21%, Video/audio clips 9 respondents: 19% and Learning applications 9 respondents: 17%).

The most frequents reasons for the Swat rural area students' satisfaction with the e-learning study was as follows:

- Easy accessible anytime and anywhere.
- It saves time and cost.
- E-learning may be introduced in schools and colleges of district Swat to make the education sector much more efficient.
- E-learning is future and easy way of learning to increase literacy rate in rural areas and especially for females' education.

We will work on general formulation of interesting model for e-learning utilization in district Swat rural areas. This type will help for the whole KP province and Pakistan country and for the whole universal purpose.

8. LessonLearned

This section describes about some lesson learned which we noticed during pilot survey for e-learning utilization.

Policy: Need strong and consistent support with a clear policy of e-learning acceptance. Policy will execute e-learning in schools and colleges for the education improvement.

Technology: There should be advance, friendly and good technological support for the online mode of instruction.

Finance: For the long term plan of education quality, improvement and education for all need a strong finance support i.e. e-learning system needs ICT utilization.

Human Resources: Human resources are important role in e-learning implementation because teachers should know about computer using, internet surfing and ICT using. Therefore, qualified staff will need for the e-learning utilization.

ACKNOWLEDGMENT

This research was partially supported by Earn-to-Learn (E-to-L) scheme of Allama Iqbal Open University, Islambad Pakistan.

REFERENCES

[1] Oh E and Russell RDR, "Pre-service Teachers Perceptions of an Introductory Instructional Technology Course," Electronic Journal Integration of Technology Education, Vol. 3(1), pp.10, 2007.

[2] Wikipedia contributors, "Educational technology," Wikipedia, The Free Encyclopedia, https://en.wikipedia.org/w/index.php?title=Educational_technology&oldid=755651857 (accessed December 19, 2016).

[3] JISC, "E-Learning Pedagogy Programme," e-learning pedogogy JISC, https://www.jisc.ac.uk/rd/projects/e-learning-pedagogy. (accessed June 20, 2016).

[4] DfES Publication, "Towards a Unified E-Learning Strategy, July 2003," Department for Education and Skills,http://dera.ioe.ac.uk/id/eprint/4748. (accessed June 15, 2016).

[5] E. Ross, "The praise and perils of e-learning," Theory and Research in Social Education, Vol. 28(4), pp. 482-492, 2000.

[6] R. Sims, "Rethinking (e) learning: A manifesto for connected generations," Distance Education, Vol. 39(92), pp. 153-164, August 2008.

[7] L. G. Muradkhanli, "Blended learning: The integration of traditional learning and eLearning," *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, Baku Azerbaijan, 2011, pp. 1-4.

[8] A. Rob, "Implementing Best Practices in Online Learning," EDUCAUSE Review January 2005, http://er.educause.edu/articles/2005/1/implementing-best-practices-in-online-learning. (accessed June 29, 2016).

[9] Acellus Academy, "A Program of the International Academic Science," About Acellus Academy,https://www.acellus.com/about.php. (access June 30, 2016).

[10] Rebecca A. Clay, "What you should know about online education," American Psychological Association, Vol. 43(6), pp. 42, June 2012. (accessed Jun 30, 2016). http://www.apa.org/monitor/2012/06/online-education.aspx.

[11] Y. Zhang, "Unavoidable Challenge: Investigation of and Thoughts on Distance Higher Education in China," Education Newspaper of China, November 15, 2004.

[12] Thailand Cyber University, "Thailand Cyber University Project," 2004. http://www.thaicyberu.go.th/. (accessed June 27, 2016).

[13] MORE DESK, "PITB's E-learning can become future of education in Pakistan," January 2015.http://www.moremag.pk/2015/01/07/pitbs-e-learning-education-pakistan/. (accessed June 30, 2016).

[14] N. A. Wassan, "E-learning and Pakistan," May 2015. http://tribune.com.pk/story/891608/e-learning-and-pakistan/. (accessed 28, 2016).

[15] Jozef Hvorecky, "Can E-learning break the Digital Divide?," *European Journal of Open, Distance and E-Learning*, 2004. http://eurodl.org/?article=143.

[16] HEC, "Distance & E-learning Universities," Higher Education Commission. http://www.hec.gov.pk/. (accessed June 30, 2016).

[17] KPESE, "IT Projects, Elementary & Secondary Schools," KPESE,http://www.kpese.gov.pk/. (accessed June 30, 2016).

[18] Taleem Foundastion, "Education through Technology," Taaleem Foundation, http://educast.co.uk/taaleem-foundation/. (accessed July 01, 2016).

[19] Beal, Vangie, "e-learning," 2016.http://www.webopedia.com/TERM/E/e_learning.html. (accessed June 28, 2016).

0.7” Margin Top

Paper Formatting Guidelines

Title: 14pts Bold

Author Name: 11pts Bold

Affiliation: 11pts Italic

Abstract: Single Paragraph (Min: 150 – Max: 250 words)

Paper Length: Formatted as guided (Min: 4,000 – Max: 8,000 words)

Font: Times New Roman 10pts

Font Color: Black

Line Spacing: 1.0 (single line space throughout the paper)

Indentation: Justify

References & In-text Citations: Follow the IEEE & Use EndNote X7 for In-text citations and Bibliography.

Headings: 12pts Bold

1. **12pts Bold**
   - **1.1. 11pts Bold**
     - ***1.1.1. 11pts Bold Italic***

Margin Left **0.8”**

Margin Right **0.5”**

6.5”x10”

Page Size: 6.5” (Width) x 10” (Height)

Submission: Formatted paper as guided can be submitted through our online submission system at http://sjcms.iba-suk.edu.pk

1.9” Margin Bottom

Sukkur IBA Journal of Computing and Mathematical Sciences

SUKKUR IBA UNIVERSITY
Merit - Quality - Excellence

SUKKUR IBA UNIVERSITY
Airport Road, Sukkur -65200
Sindh, Pakistan
Tel: +92-71-5644233
Fax: +9271-5804419
Email: sjcms@iba-suk.edu.pk
URL: sjcms.iba-suk.edu.pk