Design of Survey Tools for Obstacle Height Monitoring on the Line of Sight Communication Path

Mochammad Taufik
Department of Electrical Engineering
State Polytechnic of Malang
Malang Indonesia

Hudiono
Department of Electrical Engineering
State Polytechnic of Malang
Malang, Indonesia

Ridho Hendra Yoga P.
Department of Electrical Engineering
State Polytechnic of Malang
Malang, Indonesia

Koesmarijanto
Department of Electrical Engineering
State Polytechnic of Malang
Malang, Indonesia

Abstract: The quadcopter, which is basically used as the top viewer picture taker, in this research will be used as a device to carry the laser rangefinder to measure the obstacle height on the communication path between the location points of near end to far end. The obstacle height reading results were transmitted using a 5.8 GHz wireless transceiver to the monitoring center in real time. The data received were then processed to be displayed in graphical form, which shows obstacle height as a function of line of sight communication distance. The test results show that this device is very helpful in line of sight survey work, especially for monitoring obstacle height on the communication pathway with an accuracy of more than 90%. This device is very efficient because without using a 3-dimensional map the highest obstacle can be detected so that it can be used as a basis for determining the high position of a microwave communication antenna with clearence status.

Keywords: Line of sight, Laser Rangefinder, Obstacle height, 5.8 GHz wireless transceiver, Near end, Far end

1. INTRODUCTION

Radio communication system requires two antennas installed within communication paths; one antenna is installed in the transmitter to radiate the signal to the air while the other is installed in the receiver to receive the signal in the air. Both of the antennas should be placed above high objects in the line of sight [1][2].

The standard operational procedure of radio transmission survey is carried out manually by determining the location coordinates “near end” and “far end”, followed by scanning the path by measuring the height of objects along the paths that are suspected to be the obstacles of communication system. The result of object height scanning is then analysed using pathloss application [6] as the basis for determining the optimum height of communication antenna[5], by ensuring there is no obstacles within the path of line of sight. [7].

The scanning of object height itself, however, cannot be done if the communication path lies on areas that are not accessible by transportation system in general. Therefore, the determination of the height of the antenna can only be based on estimation value of object height suspected as obstacles in the path. The estimation method is quite susceptible to errors thus the equipment installation recommendation can possibly be wrong or inaccurate.

This research aims at designing and creating a supporting equipment for survey line of sight in the micro-wave radio communication system. The equipment will do a monitoring of object height that is suspected to be the obstacles within the path of line of sight by utilizing proximity/ altitude sensor laser rangefinder. Laser rangefinder is used to find out the proximity of a particular object [8][9][10] surface by measuring the round trip travel time of light pulses.

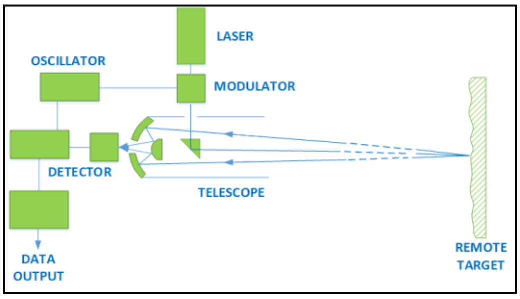


Figure 1. Object Proximity/ Altitude Sensor Using Laser

A light pulse is used to measure the distance from its travel time [11], starting from radiating the pulse until reaching the object, then the bounce of the pulse is received back by a detector. Since light propagation (c) and the travel time from being radiated to received back by the sensor (Δt) have been found out, the proximity/ height (d) can be measured using the equation below:

$$d = 2\ x\ c\ x\ \Delta t \qquad (1\text{-}1)$$

where :

d : distance from the equipment to the object (meter)
c : wave propagation (meter/second)
Δt : travel time (second)

Rangefinder Laser measures only the distance in the direction of view with a high level of accuracy [12][13]. The

Rangefinder laser brought by quadcopter autopilot to move between two point locations of “near end” and “far end”. Quadcopter is equipped by Rangefinder laser as the sensor to determine the height of the object under communication path of line of sight [7]. The result will be transmitted to the monitoring location using wireless transceiver 5.8 GHz. The data received in the location is then processed to obtain the real time display of object height graphics in the communication path. The graphics are useful for easing the analysis to directly determine the position and also height of the objects that possibly become the obstacles in the path.

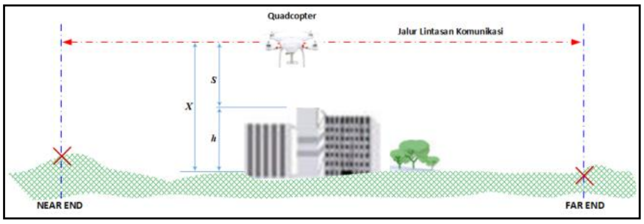


Figure 2. Robot Monitoring of Obstacle Height

Quadcopter with rangefinder laser constantly moves above the highest object along the communication path. Rangefinder detects the height of straight objects beneath, *h* meter (AGL), which is measured using the equation:

$$h = X - S \qquad (1\text{-}2)$$

X is the height of quadcopter and S is the distance of upper tip of the object towards the quadcopter position. By adjusting horizontal quadcopter speed and the setting of measurement periods of object distance using rangefinder, the object height graphics of the intended measurement will be as Figure 3. The graphics directly show the height and position of the highest object suspected as the obstacles, which in this case, the highest objects are d1 and d2 respectively towards “near end” and “far end”, with the height of h meter (AGL).

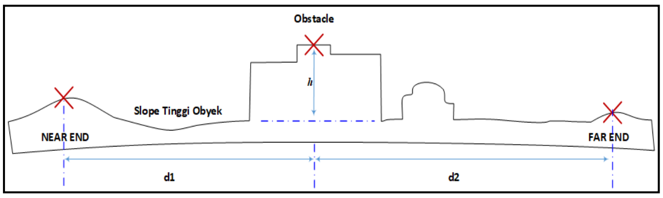


Figure 3. The Graphics of Object Height on the Signal Communication Lane

The last step of the survey is setting the optimum height of antenna position based on the highest object in the communication path, thus the line of sight communication is free from obstacles. It is continued by planning and selecting the equipment to install based on the calculation of link budget so the signal power value received can be as expected. These two steps are completed by using Pathloss application software.

2. METHOD OF RESEARCH

This research aims to design a survey tool for micro-wave radio communication system related to object height monitoring that is suspected to be obstacles on line of sight communication path using a quadcopter. The result of this research is expected to help people finding out the position and highest height of object quickly and easily that can be used as the basis to determine the optimum height position of an antenna so the communication path is clearance. [14][19][20].

Figure 4 shows the block diagram of survey equipment to monitor the height of object.

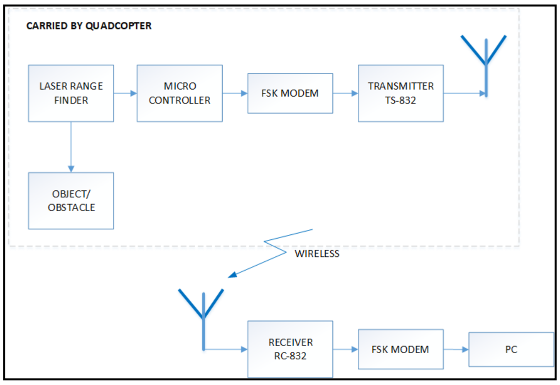


Figure 4. Tools Survey for Object Height Monitoring

2.1 Equipment Specifications

The survey equipment to measure the height of the obstacles in the communication path of line of sight consists of quadcopter to bring rangefinder laser as an altitude sensor, microcontroller, FSK modem, wireless transceiver 5,8 GHz to transmit the result of object height detection in the communication path of line of sight to the center of monitoring location. The data of object height received in the center of monitoring location will be processed to get the graphics, as a function of communication path distance between ”near end” and ”far end”.

The components of the equipment have the specifications as follows:

1) Quadcopter: The quadcopter must be able to move horizontaly straight from the location of near end to far end with constant height as planned that is above the highest object under the communication path of line of sight. It also must be controlable to move automatically and have adequate coverage to accomodate the distance of surveyed communication link. The specifications of the quadcopter are [21]:

TABEL 2.1. SPESIFIKASI QUADCOPTER

No.	Description	Specification
1.	Max. Ascent Speed	5 mps
2.	Max. Descent Speed	3 mps
3.	Max. Speed	16 mps (no wind)
4.	Max. Altitude above sea level	6000 m
5.	Wifi frequency	2.400 GHz – 2.483 GHz
6.	Max. Transmission Distance	FCC 1000 m, CE 500 m
7.	Transmitter Power (EIRP)	FCC 27 dBm, CE 20 dBm

2) Rangefinder laser: Rangefinder laser is used as a sensor to measure the height of the object. The measurement is controlled by a microcontroller that works as measurement control. The specifications of the rangefinder laser are: [22]:

TABEL 2.2. SPESIFIKASI LASER RANGEFINDER

No.	Description	Data
1.	Distance	5 - 600 meters
2.	Wavelength	905 nm
3.	Accuracy	+/-0.5 m
4.	Operating Voltage	5V to 6 V

3) Wireless Transmitter 5,8 GHz: the wireless transmitter works to transmit the rangefinder laser sensor readings in real time to the central location of monitoring. This research utilizes transmitter type AV (audio video) TS-832 that works in the frequency of 5,8 GHz. The detailed specifications of this AV wireless transmitter [23] is seen in the table below:

TABEL 2.3. TRANSMITTER SPECIFICATION TS-832

No.	Description	Data
1.	Item Name	TS-832 Transmitter
2.	Antenna gain	2 dBi
3.	Frequency	5.8 GHz
4.	Transmitting Power	600 mW
5.	Power Input	7.4 – 16V
6.	Video Format	NTSC/PAL Auto
7	Audio Bandwidth	6.5 MHz
8	Video Bandwidth	8 Mhz
9	Connector	RP-SMA jack

4) Wireless Receiver 5,8 GHz: the wireless receiver works to receive transmission data performed by the transmitter module. The technical specifications of the wireless receiver 5,8 GHz type AV RC-832 is seen in the table IV below:

TABEL 2.4. RECEIVER SPECIFICATION RC-832

No.	Description	Data
1.	Number of channels	32 CH
2.	Antenna gain	2 dBi
3.	Frequency	5.8 GHz
4.	Rx Sensitivity	- 90 dBm
5.	Video Output Level	75 Ohm
6.	Video Output Level	10 Kohm
7.	Video Format	NTSC/PAL Auto
8.	Video Bandwidth	8 Mhz
9.	Connector	RP-SMA jack

5) User interface: A user interface is created to connect the systems with users through a particular application program which can process the data from altitude sensor received in the central location of monitoring to be displayed in the form of object height graphics as a distance path function. It is also beneficial to set the storing process of object height logger data.

2.2 System Design

The system design to monitor the height of an object in a communication path of line of sight is performed based on block diagram in Figure 3. The detailed design of each part, especially for transmitting and receiving process, is as follows:

1) Obstacle Altitude Sensor Control*:* The laser rangefinder altitude sensor was controlled by a microcontroller (Arduino Uno R3) [24], then the generated height data were fed to the FSK modem to be transmitted to the monitoring location.

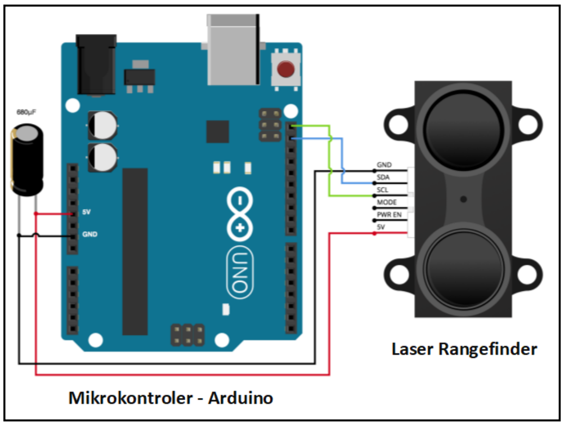


Figure 5. Obstacle Altitude Sensor Control Circuit

2) Obstacle Height Transmission Data: Obstacle height data from rangefinder laser sensor is transmitted to the central monitoring location using AV module (audio video) wireless transmitter TS-832. The obstacle height data is in the form of digital data so conversion to the analog one using modem is needed so that it can be transmitted using wireless transmitter TS-832. The modem used in this research is FSK type IC TCM-3105.

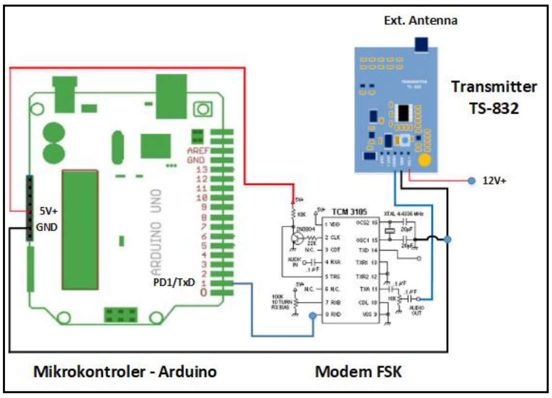


Figure 6. Obstacle Height Data Transmitting Circuit

3) Obstacle Height Data Receiver System: In the location of monitoring center, the obstacle height data is received by wireless receiver AV type RC-832. Since the data is analog, it needs to be converted first just like in the transmitter, that is turned into digital data using FSK modem type IC TCM 3105 [25].

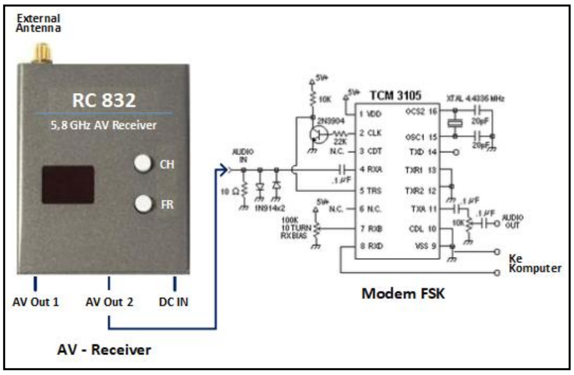


Figure 7. Obstacle Height Data Receiver System

The converted data from FSK modem is then processed by computer so that it can be displayed in graphics. In this way, the final result of obstacle height monitoring process is also in graphics of obstacle height above communication path of line of sight.

3. RESULTS AND DISCUSSION

The result of this research is an artificial model of obstacle height monitoring system device in the micro-wave *line of sight* communication link that consists of two parts:

1) The transmitter: the device is brought by a quadcopter to move above the communication path of line of sight which has been planned, to determine the height of object suspected to be the communication barrier.

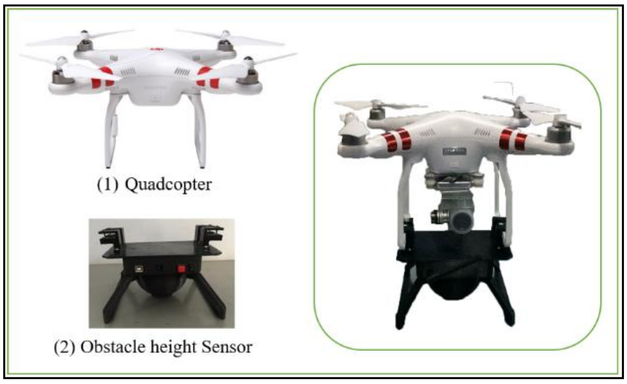


Figure 8. The Transmitter of Obstacle Height Monitoring System

2) The receiver: this device is placed in monitoring location, working to receive the signal of obstacle high sensor transmission data in the transmitter. This part plays the role to interpret the data based on the computer input using USB.

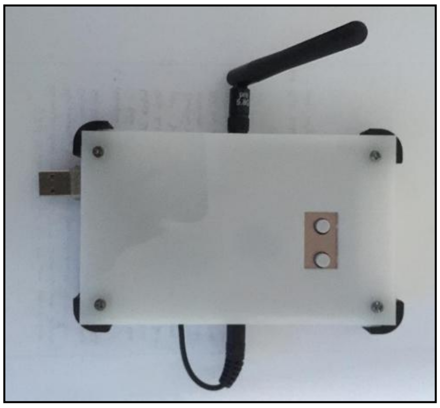


Figure 9. The Receiver of Obstacle Height Monitoring System

The measurement of the height of obstacle was performed using rangefinder laser, while the quadcopter was functioned as rangefinder laser sensor carrier that moves straight horizontal from the point of *near end* to the point of *far end*, following the communication path of line of sight which has been set previously with the height more than the highest obstacle in the path.

The data of rangefinder laser sensor interpretation was sent real time to the location of monitoring centre of obstacle height through wireless receiver 5.8 GHz. After that, the received data was processed to get the graphics version of it thus the users can easily determine the highest object within the communication path of line of sight suspected as the obstacle.

3.1 The Result of Obstacle Height Measurement

The device prototype of obstacle height monitoring system that did the measurement of object height suspected as obstacle in the communication path of line of sight has the distance of 170 meters with the path as seen in the Figure 9.

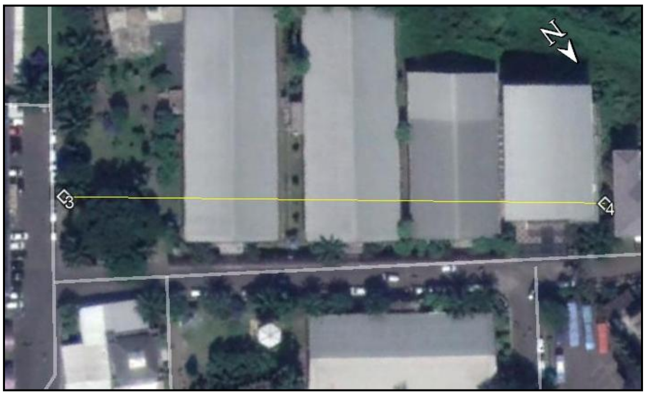


Figure 10. The Testing Path of The Rangefinder

The results of the obstacle height monitoring system are displayed in Figure 10.

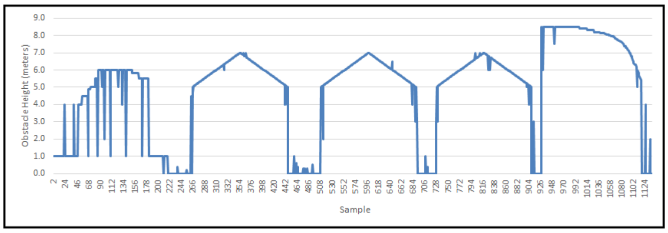
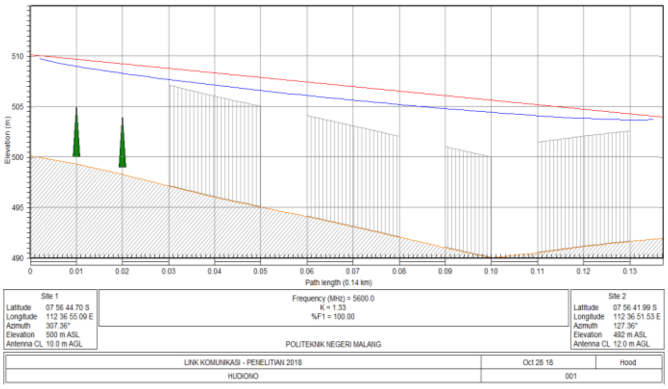


Figure 11. Communication link of the rangefinder testing path

When the laser rangefinder sensor hit leaves or a tree, the measurement results were not stable because some laser beams hit the object underneath or through the leaves. The highest obstacle between communication links was 8.5 meters which was close to the far end location. The Application of Obstacle Height Data is used as the basis to determine the height of the antenna. The obstacle height data which are generated by the monitoring system tool can be used as inputs of pathloss software, which is used to estimate the line of sight radio communication antenna's height [4].

Assuming the communication link had a WiFi frequency of 5.6 GHz, Fresnel of 100%, and the value of K = 1.33, then a communication link that was free of obstacle required the minimum antenna height at the near-end and far-end of 10 m (AGL) and 12 m (AGL), respectively. The results can be shown as in Figure 10 [4].



Site 1		Site 2	
Latitude	07 56 44.70 S	Latitude	07 56 41.99 S
Longitude	112 36 55.09 E	Longitude	112 36 51.53 E
Azimuth	307.36°	Azimuth	127.36°
Elevation	500 m ASL	Elevation	492 m ASL
Antenna CL	10.0 m AGL	Antenna CL	12.0 m AGL

Frequency (MHz) = 5600.0
K = 1.33
%F1 = 100.00

POLITEKNIK NEGERI MALANG

LINK KOMUNIKASI - PENELITIAN 2018	Oct 28 18	Hood
HUDIONO	001	

Figure 12. Determining The Antenna Height Using Pathloss-4 Software

4. CONCLUSION

The conclusions of this research are:

1. The laser rangefinder sensor has a maximum measurement specification of 600 meters and it can detect most types of objects, except for transparent objects or the materials that cannot reflect the laser beam.
2. The output of the height monitoring system is displayed in graphs; thus, it is easy to read without having to analyze it first.

3. Quadcopter that carried the monitoring system moved on the line of communication links as planned from the near-end location with coordinates of 07 56 44.70 South and 112 36 55.09 East to the far-end location with coordinates of 07 56 41.99 South and 112 36 51.53 East. The distance was 137 meters and the quadcopter flew above the height of the obstacle, which was 8.5 meters (AGL).

5. REFERENCES

[1] Balanis, C. A. (1992). Antenna theory: a review. Proceedings of the IEEE, 80(1), 7–23. doi:10.1109/5.119564

[2] Kiema, J. B. K., Siriba, D. N., Ndunda, R., Mutua, J., Musyoka, S. M., & Langat, B. (2011). Microwave Path Survey Using Differential GPS. Survey Review, 43(323), 451–461. doi:10.1179/003962611x131177488917

[3] De Floriani, L., Marzano, P., & Puppo, E. (1994). Line-of-sight communication on terrain models. International Journal of Geographical Information Systems, 8(4), 329–342. doi:10.1080/02693799408902004

[4] J. B. K. Kiema, D. N. Siriba, R. Ndunda, J. Mutua, S. M. Musyoka & B. Langat (2011) Microwave Path Survey Using Differential GPS, Survey Review, 43:323, 451-461, DOI: 10.1179/003962611X13117748891796

[5] Mohamed, Nazar & Nadir, Zia & Salam, M & Rao, Jegathese. (2005). Microwave attenuation studies due to rain for communication links operating in Malaysia. Georgian. 9-17

[6] Jeffrey Barney, M & M. Knapil, Jamie & L. Haupt, Randy. (2009). Determining an optimal antenna placement using a genetic algorithm. 1-4. 10.1109/APS.2009.5172055.

[7] Vucetic, B., & Du, J. (1992). Channel modeling and simulation in satellite mobile communication systems. IEEE Journal on Selected Areas in Communications, 10(8), 1209–1218. doi:10.1109/49.166746

[8] Markus-Christian Amann, Markus-Christian Amann, Thierry M. Bosch, Thierry M. Bosch, Marc Lescure, Marc Lescure, Risto A. Myllylae, Risto A. Myllylae, Marc Rioux, Marc Rioux, "Laser ranging: a critical review of unusual techniques for distance measurement," Optical Engineering 40(1), (1 January 2001). https://doi.org/ 10.1117/1.1330700

[9] Davis, Q. V. (1966). LASERS AND DISTANCE MEASUREMENT. Survey Review, 18(139), 194–207. doi:10.1179/003962666791277919

[10] Duchoň, F., Dekan, M., Jurišica, L., Vitko, A. (2012). Some Applications of Laser Rangefinder in Mobile, Journal of Control Engineering and Applied Informatics, Vol 14, No 2 pp. 50-57, ISSN 1454-8658

[11] Lee, Joohyung & Kim, Young-Jin & Lee, Keunwoo & Lee, Sanghyun & Kim, Seung-Woo. (2010). Time-of-flight measurement with femtosecond light pulses. Nature Photonics. 4. 10.1038/nphoton.2010.175.

[12] F. Panasenko, A. (1998). Estimating the accuracy with which the distance between spacecraft is measured with a laser rangefinder. Journal of Optical Technology - J OPT TECHNOLOGY-ENG TR. 65. 662-665.

[13] Hamilton, G.W. & Fowler, A.L.. (1966). The laser rangefinder. Electronics and Power. 12. 318-322. 10.1049/ep.1966.0245.

[14] Qiao, X & H. Lv, S & L. Li, L & J. Zhou, X & Y. Wang, H & Li, D & Y. Liu, J. (2016). APPLICATION OF DSM IN OBSTACLE CLEARANCE SURVEYING OF AERODROME. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLI-B2. 227-233. 10.5194/isprs-archives-XLI-B2-227-2016.

[15] Zhao, X., Vainikainen, P., & Kivinen, J. (1999). Diffraction Over Typical-Shaped Terrain Obstacles. Journal of Electromagnetic Waves and Applications, 13(12), 1691–1707. doi:10.1163/156939399x00169

[16] Fazi, Christian & Crowne, Frank & Ressler, Marc. (2012). Link budget calculations for nonlinear scattering. 1146-1150. 10.1109/EuCAP.2012.6206100.

[17] Nguyen, Tien & C Wang, Charles. (2001). On the Link Budget Calculation for CDMA Systems. Journal of Molecular Catalysis A-chemical - J MOL CATAL A-CHEM. 2. 10.1109/AERO.2001.931272.

[18] Merabtine, Nadjim. (2014). WiMax Link budget Calculations for Airport Surface Communications in the C Band. International Journal of Engineering and Technology. 4. 483-487.

[19] Sharma, V. T. V. (1974). Long Path Line of Sight Microwave Communication Links. IETE Journal of Research, 20(6), 282–286. doi:10.1080/03772063.1974. 11487396

[20] Khan, Mohammad A.R., et al., (2015). *Analysis for identify the problems of microwave installations and Solutions*. IJRCCT, Vol 4, Issue 1, pp. 1-8.

[21] Jalali, A. (2014). *Broadband access to mobile platforms using drone/UAV*. United States Patent Ser. No . : 14 /222,49. Filed Mar. 21, 2014.

[22] ShenZhen Rui Er Xing Electronic Co.,Ltd. (2011). *Laser Works 600m LRF laser Rangefinder RS232 interface distance sensor laser distance module*. Retrieved from https://www.laserexplore.en.alibaba.com/product-detail/ Laser-Works-600m-LRF-laser-rangefinder_60504979219.html

[23] Helipal, (2013). *TS832 and RC832 Instruction Manual*. Retrieved from http://www.helipal.com/boscam-5-8ghz-fpv-system-ts382-tx-rx832-rx-600mw.html

[24] Adafruit, (2010). *Arduino – ArduinoBoardUno (EAGLE files and schematic posted).* Retrieved from https://

blog.adafruit.com/2010/09/27/arduino-arduinoboarduno-eagle-files-and-schematic-posted

[25] Texas Instruments Incorporated, (1985), *TCM3105DWL, TCM3105JE, TCM3105JL TCM3105NE, TCM3105NL FSK MODEM*, Retrieved from http://tcm3105.com/ tcm3105.pdf, Revised May, 1994.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 04, 89-97, 2019, ISSN:-2319–8656

Optimal Placement of Capacitor Bank in Reorganized Distribution Networks Using Genetic Algorithm

Emma Bakker
Department of MBA or business management/ Leiden University, Netherlands

Vahid Deljou
Department of electrical Engineering/ Islamic Azad University, West Tehran Branch, Iran

Javad Rahmani
Department of digital electronics Engineering/ Islamic Azad University, Science and Research Branch, Iran

Abstract: Capacitor optimal placement is one of the most important designs and control issues of power systems in order to reduce network losses, improve the voltage profile, reduce the reactive load, and reducing the power factor. The distribution network operator, taking into account two major goals of reducing real power losses and maximizing the return on investment required for installation of capacitive banks for sale to the transmission system, obtains the position, number, and capacity of capacitive banks. In this paper, the optimization problem is formulated for different values of the parameter "reactive energy value". After evaluating the objective function and implementing an optimization algorithm for each value of this parameter, the arrangement and capacitance of the capacitors in the network load nodes are obtained. Meanwhile, using the objective function defined in this paper, you can obtain the threshold for the sale of reactive energy, and by selling it to the transmission network, the investment in installing capacitor banks will be profitable for the distribution network operator.

Keywords: Capacitor, Optimal Placement, Genetic Algorithm, Wind Energy Conversion Systems Reactive Power, Power Factor.

1. INTRODUCTION

Generally, in alternating current networks, the apparent power received from the generators is divided into two parts: real power (active) and reactive power. The way this division depends on the power factor of the consumers, that is, the more power factor closer to 1, the greater is the true power share and the less is imaginary power contribution. Due to the fact that many consumers in which the coil or inductor play a significant role, they are the resistive-inductive consumers, and because of the energy saving feature in the inductors, there is always an amount of power that moves between the network and inductor which cannot be used and is wasted on the path through the wires and cables [1, 2]. As a result, generators need to produce more power and increase the current flow, which, with the increasing of the current flow, the capacity of the transmission lines reduces for real power transmission. In fact, all the reactive power required for loads, lines, and transformers should be produced at the transmission level. Also, the power loss in the distribution networks is in the form of heat, voltage drop, and reduction in efficiency. The reactive power compensation means that the reactive power needed is generated besides the load instead of supplying through the generators of the power plant [3, 4]. This distribution can be done at the distribution and over-distribution level by parallel capacitors. Basically, the more capacitors are installed near the consumer centers, the higher the efficiency of the network will be. The use of parallel capacitors makes it possible to utilize the capacity of the transmission lines for more active power transmission [5, 6]. The power consumed by electricity subscribers varies, as a result of their power factor characteristics. By producing reactive power by capacitors, the effects of the reactive components reduce and the power factor increases, which will result in more favorable technical conditions for energy transfer [7-12].

In 1956, the first steps were taken to optimize the placement of capacitors in distribution networks, and this has continued so far. From the presented methods, there are many methods such as mathematical and analytic methods including two noncompressed methods (Kant Tuck theory, Hysin method, reduced gradient method, and quadratic programming) and compact (linear programming and nonlinear programming) [13-18]. [19-22] proposed a nonlinear programming model to find optimal locations of facilities throughout the network. In the past and the use of the birds breeding algorithm in 2007.

This paper presents a solution based on the genetic algorithm for optimal placement of capacitors.

2. GENETIC ALGORITHM

Genetic algorithms use Darwin's natural selection principles to find the optimal formula for predicting or matching patterns. According to Darwin's survival evolution theory, living organisms in the next generation are better than the previous generations [23, 24]. In general, these algorithms consist of four parts of the fitting, display, selection and modification function. It is briefly said that the genetic algorithm is a programming technique that uses genetic evolution as a problem-solving model. This is an evolutionary search algorithm to find an approximated optimal solution starting with a set of the initial solution [25-28].

The input of this program is a problem that needs to be solved and solutions are coded according to a template. Fit fitness evaluates candidate responses. First, a select number of inputs, x1, x2,…..,xn, which belong to the X space, are selected and represent them as a vector X=(x1, x2, ….,xn). This input vector is called the organism or the chromosome and the group of chromosomes is called colonies or populations [29, 30]. In each period, the colony grows and evolves in accordance with certain laws that indicate biological evolution. For each chromosome xi, there is a fitness function f(xi). Stronger elements or chromosomes that are closer to their current value, they are more likely to survive in other periods and re-produce, and the weaker ones will die. In other words, this algorithm keeps the inputs that are closer to the optimal answer and ignores the rest. Another important step in the algorithm is the birth or production of a child that occurs once in each period. Children can be generated through crossover and mutation operators at each step [31-34]. The contents of the two chromosomes that occur in the production combine to create a new chromosome called a child. Some of the genes are transmitted from the father and some others from mother to child, where the genes mutate from father to mother or vice versa, are called compound combinations 4. In addition, during a period, a series of chromosomes may find a gene mutation. A gene that does not exist in the parent is created in the child. As stated above, each entry X is located on a vector number X=(x1, x2,…, xn). For the implementation of the genetic algorithm, each entry must convert to one chromosome. In the zero step, a bunch of inputs X is randomly selected. Then, for each period, the fit value is calculated and the operators of production will change and select. When the fit value is obtained or the chromosomal matched around the constant value oscillates, because of the gene mutation operator the total matching of the chromosomes does not remain constant at all, the algorithm ends.

3. OPTIMAL REACTIVE POWER COMPENSATION IN A REDISTRIBUTED DISTRIBUTION GRID

3.1 Expression of Optimal Capacitor Positioning Approach with Attitude to Restructuring in Electricity Industry

Considering the compensation system, the compensation in the medium pressure distribution network and the MV / HV buses in the middle of the transmission and distribution, there is virtualization in a reactive power generator and there is no need to carry this power from the transmission system to the distribution system [35]. With the production of reactive power in HV buses, the demand for a distribution system or more can be provided. In fact, there is the advantage that there is no need for reactive power transmission and installation of energy systems. In this new scenario, investing in the installation of the compensation system for the distribution system operator has two advantages of reducing the loss of power and increasing the profit from the sale of reactive power to the network operator. The amount of economic allocation for the distribution system operator cannot be lower than a certain threshold value [36, 37].

In fact, if the system obtain the reactive power transfer at a cost above the threshold cost, it will be economically feasible to install capacitor banks and exploit them for the distribution network operator, and if this cost be lower than the threshold, according to the distribution system operator, there is no need to install a compensating system, more than what is needed to reduce the distribution network losses, while the production and transmission of reactive power is cheaper than buying it from Distribution Network [38-40]. The optimization model presented in this paper is examined from the point of view of the distribution network.

In a detailed description of the procedure, the program will be run and the system losses will be achieved before the compensation. Then, a target function and a suggested price for the sale of one kilo VAR of reactive energy to the transmission network are defined. In fact, the new target function, in relation to the target functions used in the previous methods, has an additional sentence that relates to the revenue from the sale of reactive power to the transmission network. By defining the objective function and using an optimization algorithm (genetic algorithm), the search begins to find the optimal response. By finding the optimal answer for a suggested price, the price is increased by one step, and again the load and the genetic algorithm is performed to find the optimal capacitor arrangement for the new price, and finally, the network losses are calculated after compensation. This process continues to the point where the increase in the price of reactive energy sales to the transmission network will not increase casualties compared to pre-compensation network losses. In fact, reducing the losses is more important than increasing the economic benefits.

3.2 Formulation of the Problem

In order to optimize the target function, the two current positions (Before compensation) and a new position (design response) are compared. The economic components of the current situation are as below:

- Variable cost related to power losses in distribution network lines
- Variable cost related to power losses in the transformer MV/HV

After exploiting the system in a new position and inserting capacitive banks in the MV / HV distribution network nodes, the economic components of the new position are:

- The variable cost is related to the new amount of line losses that is definitely less than the losses in the previous state.
- Variable cost related to the new value of MV / HV Transformer losses
- Revenues from sales of reactive power to the transmission system
- The total cost of installation of capacitive banks: All of these components are reviewed over a year. For this time period, load variations and capacitor banks are considered.

The condition (1) or the power loss per hour h, taking into account the daily load and changing the clock to the hourly load of each node of the medium network, is expressed as follow:

$$P_{loss}\text{ (h) = }\sum_{i=1}^{n_r}\frac{R_i}{V_i^2}\text{ [(}P_i\text{(h))}^2\text{+ (}Q_i\text{(h) - }Q_{ci}\text{(h))]}^2 \qquad (1)$$

In which, Q_{ci} is the capacity of capacitor banksat node i, P_i(h) and Q_i(h) are real and reactive powers of the ith branch (including the loads and losses under the i-th branch) and n_r is the total number of branches of the network.

In this case, because the current position is checked, no capacitance is installed, Q_{ei} (h) = 0). Energy losses in a year are as follows:

$$E_{loss}\text{=365}\sum_{h=1,24}P_{loss}\text{ (h)} \qquad (2)$$

The condition (3) is equivalent to the condition (1), in which (h) = 0). Condition (2) represents the losses in the MV / HV transformer station at hour h.

$$P_{lossTR}\text{(h)=}\frac{R_{TR}}{V^2}\text{[[}P(h)^2\text{]+[Q(h)-}Q_c\text{(h)]}^2\text{]} \qquad (3)$$

The R_{TR} is the transformer series resistor, V is the rated voltage of the intermediate voltage, and P(h) and Q (h) represent the active and reactive power requirements of the network (including loads and losses) that are in the medium voltage transformer. In the formula above, the expression Q_c(h) is zero and P_{lossTR} represents the loss of the transformer prior to the compensation. The loss of distribution transformers has been neglected due to the insignificance of their series resistance.
The energy losses per year for the transformer is as follows:

$$E_{TR}= 365 \sum_{h=1,24} P_{lossTR} (h) \tag{4}$$

The condition (4) is equivalent to the condition (2), which is opposite zero. In fact, this condition indicates the actual operating conditions of the capacitor banks connected to the MV / HV modulus of the transformer at any time of the day.

The economic benefits resulting from the loss of network losses is as follow:

$$R_{ET} = (C_{ETb} - C_{ETa}) = (E_{lossb} - E_{lossa}) C_{ET} \tag{5}$$

Which C_{ETb} and C_{ETa} represents the costs of energy losses before and after compensation, and E_{lossb} and E_{lossa} are the energy losses before and after compensation, and C_{ET} indicates the unit cost of energy in terms of (KWh / Rials), which according to the budget law of year 87 is equal to 773 Rials.

The resulting losses in MV / HV transformers are calculated as follows:

$$R_{ETR}=(C_{ETRb}-C_{ETRa})=(E_{TRb}-E_{TRa})C_{ET} \tag{6}$$

C_{ETRb} and C_{ETRa} are the costs of the transformer losses before and after the compensation, and E_{TRb} and E_{TRa} are the energy losses before and after the compensation.

If the phrases R_{ETR} and R_{ET} are lowered to zero, this means that for the distribution network operator not only economic gain is not achieved, but also incurred economic losses, while the cost of installation of capacitive banks is higher compared to the cost of reducing the losses.
In order to express the condition (5), that is, the proceeds from the sale of reactive power to the transmission system, the Rial value of the reactive production service unit (one kilo VAR hour) R, and the total value of the Rial of the reactive power of the capacitor banks assigned to the distribution network R_T.
In this case, we will have:

$$R_{\mathrm{T}} = 365\mathrm{R} \sum_{i=1}^{n} \sum_{h=1}^{24} Q_{ci}(h) \tag{7}$$

Where n I s the total number of network chains and $Q_{ci}(h)$ is the capacitance bank at h is in chin i.
The main problem is determining the optimal response obtained from the genetic algorithm for different values of the parameter R.

Unit R (kvarh / Rials) is the economic value of reactive energy.

Finally, the condition (6) is equivalent to the total cost of the purchase, installation, and maintenance of capacitor banks, which is obtained by multiplying the cost of one kilo VAR capacitive bank (C_{inst}) in the capacity of the entire installed capacitor banks (Q_{cinst}).

$$C_{instT} = C_{inst} \times Q_{cinst} \tag{8}$$

In general, an indicator should be defined to assess the answer to the problem, the index in this paper is called the return on capital. This indicator represents the difference in revenue and expenses over a year in the distribution network.

3.3 Solve the problem of capacitance in a redistributed distribution network with a genetic method

The steps to solve the problem are as follows:

First, enter the network information and power flow is done, then all the network load buses are considered as the candidate for the capacitor installation location. Initial capacity is then determined for the specified sites to begin the search using the genetic algorithm and the objective function. Input data of this program, active and reactive load data of the network, impedance and admittance lines, the maximum number of capacitors per bus, the rial equivalent of one kilowatt hour of energy, the purchase and installation cost per kilo VAR capacitor and the cost of sales one kilo of hours of reactive energy to the transmission network. The decision variables in this issue are the position and size of the nominal capacitive banks and their control variables, including the maximum capacitance banks size, the allowed range of buses' voltage and the maximum allowable amount of reactive energy sold to the transmission network, in such a way that the generator does not get unstable. The output data of this program are the capacitance and final position of the capacitors in the network, the active and reactive losses of the grid after compensation, the MV / HV transformer losses, and the amount of energy sold to the transmission network. Objective Function.

$$\frac{R_{Et}+ R_{ETR}+ R_T- C_{instT}}{C_{instT}} \tag{9}$$

Maximize = max objective function

The constraints of the objective function are:

$$V_{i\,min}<V_i<V_{i\,max},Q_{c(\mathrm{min})}\le Q_c\le Q_{c(\mathrm{max})} \tag{10}$$
$$0 \le n_i \le n_{i(\mathrm{max})}$$

Which V_i represents the value of the voltage in the bus-bar i, n_i is the number of capacitor banks in the bus-bar i and Q_c is the value of the reactive power that can be injected to or received from the transmission network. The cost of each economic component 1 to 6 is assessed in one year. If the deduction is greater than zero, the income will increase as the nominal capacity increases. The price threshold is the amount for which the return on investment index changes from negative or zero to a positive amount. In order to obtain the final configuration and capacitance of the capacitors in the network for each value of R, the program must be implemented so far as to allow the maximum number of repetitions to be allowed, or the sum of fitting the chromosomes in a generation to a constant value. Of course, in general, due to the presence of a gene mutation operator, a constant amount is not obtained, and the total fit of the chromosomes fluctuates around a constant value. In this case, the response is saturated and the optimal response is obtained from the genetic algorithm. Then R is increased one step and the algorithm is executed again.

The makeup and capacity resulting from the implementation of the algorithm for each value of R results in the highest revenue for that value for the distribution system operator. Considering R = 0, the term for selling revenues from the sale of power to the transmission network is eliminated. In this case, the goal is to find the makeup and optimal capacitance of the capacitors, regardless of the energy sales to the transmission network - and without considering restructuring in the electricity industry.

3.4 Solution Algorithm

According to the above, the problem-solving procedure can be summarized as follows (figure 1):

- Initial population formation (initial values of capacitance capacities in candidate positions for each chromosome from the initial population)
- Evaluation of superior values for each chromosome and evaluation of the objective function using:
  - ➢ Performing load flow based on the initial population of capacitors per chromosome
  - ➢ Calculate the losses of lines and transformers for this arrangement
  - ➢ Calculate the cost of installing capacitors for this arrangement
  - ➢ Calculation of R_T for different prices of R
  - ➢ Calculate the fit of each chromosome (the value of the objective function)
  - ➢ Determine the highest values of each population for reproduction
- Perform reproduction using combinator operators and gene mutations
- Repeat steps 2 and 3 to achieve the maximum reproduction specified or saturation of the fit function

Determine the capacitive capacity for each candidate position according to the best chromosome produced.

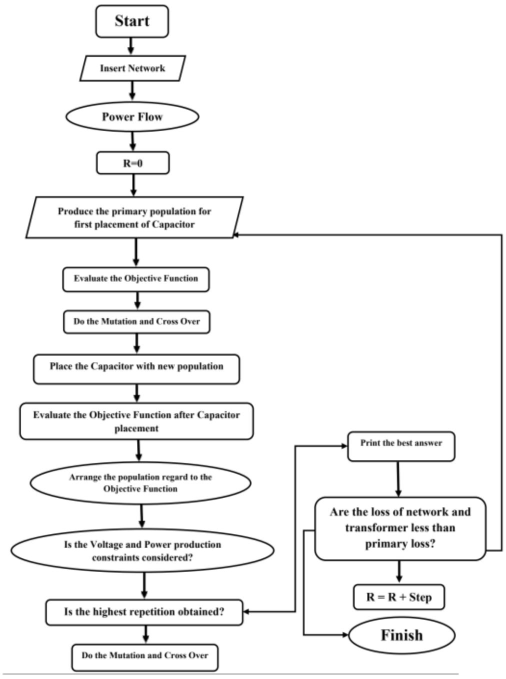

Start
Insert Network
Power Flow
R=0
Produce the primary population for first placement of Capacitor
Evaluate the Objective Function
Do the Mutation and Cross Over
Place the Capacitor with new population
Evaluate the Objective Function after Capacitor placement
Arrange the population regard to the Objective Function
Is the Voltage and Power production constraints considered?
Is the highest repetition obtained?
Do the Mutation and Cross Over
Print the best answer
Are the loss of network and transformer less than primary loss?
R = R + Step
Finish

Figure 1. The algorithm for solving the problem of capacitance in a distributed distribution network with a genetic method

4. SIMULATION AND RESULT

The simulation of this method is performed on an IEEE radial-circular grid. A computer program in the MATLAB software environment is written based on the Genetic Algorithm and Newton-Raphson load distribution for different loading modes in the distribution network and simulation is performed with this program.

4.1 Formulation of the problem in MATLAB software space: The modeling of the test network is done using two input data matrices and line data.

4.2 The input of each bus

Column 1: bus-bar number

Column 2: Bus-bar Code (Bus-bar PQ: 0, Bus-bar Reference: 1, Bus-bar PV: 2)

Column 3: Voltage value in P.U

Column 4: Phase angle in degrees

Column 5: Bus load in MW

Column 6: Bus load according to MVAR

Column 7 to 10: Production Megawatts, Mega Var Production, Minimum Mega Var Production Permitted

Column 11: Reactive Power (MVAR) injected by parallel capacitors

4.3 Data of transmission line

In this case, each line is marked with two nodes, with columns 1 and 2 including the number of nodes at the beginning and end of the line. In column 3 to 5 resistance, the reactance and half the total Susceptance of the line are expressed in terms of P.U based on the MVA. Due to the insignificance of the capacitance susceptibility of the line in distribution networks, its value is neglected. The last column of this matrix is used to adjust the transformer tap. For lines, number 1 should be entered in this column. The line information can be entered in any order, but if the input values are for the transformer, the side with the tap is considered as the left bus-bar. Once the network and line information is entered, there must be a connection between the load distribution program and the genetic algorithm. The output of the genetic program should be referred to as the input of column 11 of the data file. For this purpose, we have to consider the network load nodes as the initial population of the genetic algorithm and allocate a suitable initial capacity as inputs to them. Because of the discrete capacities of existing capacitor banks, chromosome bits are valued at the same level as the actual stacks of capacitor banks. In this program, the return on investment indicator is used to determine the fitness of the objective function.

4.4 Simulation on Roy-Billinton Test System (RBTS): Run the program assuming a variable daily load within a year

The proposed method on the Roy - Billinton (shown in Figure 2 of the single-mode model diagram of the sample network) with the assumption of changing the load over a 24-hour period of a day, is reviewed and the resulting responses are examined.

In the Billinton test network, bus 4 of the system has 71 chunks and 38 nodes. The capacitive capacities of the capacitor banks in the load cells of this network are 150 kilo Var. The maximum number of capacitor banks nominated at a node is 8 (maximum 1200 kilo Var) and the maximum number of 1,200 kV capacitive banks to be installed at the station is 230 kV (28 kV) (up to 33.6 megawatts) The 33 / 230kV transformer has a capacity of 160 MVs and a series resistance of 0/024 ohm. The losses before the capacitor are installed is 739kW. The average active and reactive loads connected to this network are respectively 24.578 MW and 15.232 MVAR, respectively.

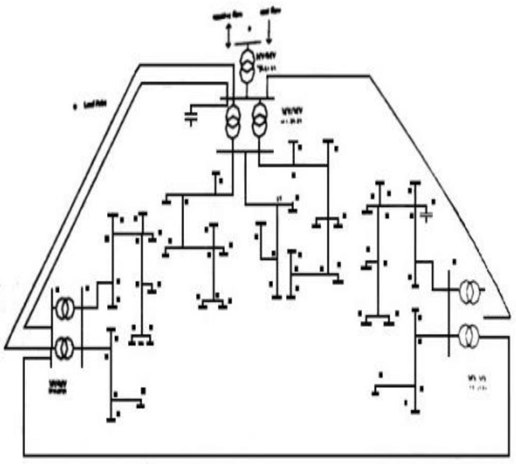


Figure 2. Single-line diagram of Bus-bar 4 in Roy – Billinton test system

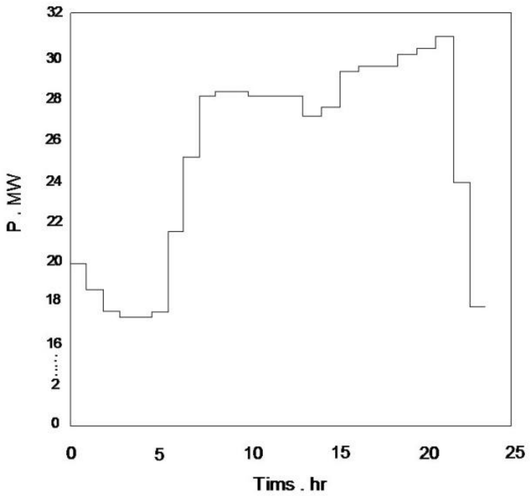


Figure 3. 24-hour load diagram of Bus-bar 4 in Roy – Billinton test system

By executing a program for a sample network with a 24-hour load diagram, for each value of R (the equivalent of a Rial per kilo Var per hour), and implementing the genetic algorithm at multiple times, the final energy dissipation in the network, the energy lost in the transformer 33/230 Within one year, the total capacitive capacity and capital return indicator are in accordance with Table (1).

It is seen that in this case, for the values of R near zero, the term capacitance of the nominal capacitor is often constant and the fit of the objective function is negative. As R increases, the return on capital increases gradually, meanwhile capacitive capacitance increases. This increase is slowing down to reach the threshold price, and after reaching the threshold price and positive fit, the objective function increases with a higher rate. This process is consistent with what was previously expected, installing a surplus capacitive bank on the distribution network needs a cost-effective economy if the operator purchases a reactive energy transfer network at a reasonable price from the distribution network. In this case, with a daily change in load, the threshold value of R is about 55 Rials. In fact, for the sale of one kilo Var of reactive energy per hour, the reactive power of 55 Rials and more to the transmission network operator, the investment in the installation of capacitor banks is more profitable for the distribution network operator than reducing the losses of the distribution network. The power loss diagrams in the grid lines and main station transformer, the installed capacitance and the return indicator of capital (Figures 4 to 7) for the different values of the parameter R represent this fact.

Table 1. Results from program execution for change 24-hour load parameter

(Rial/Kvarh) R	P_{loss} (MWH)	$P_{loss\ TR}$ (MWH)	Q_{sinst} (kvar)	ROI
0	2926/38	99/39	13800	-0/89735
5	2936/73	99/45	14550	-0/80974
15	2972/54	99/50	14550	-0/63655
25	2869/35	99/52	18900	-0/46279
35	3247/47	100/36	21150	-0/32181
45	3241/71	100/65	21150	-0/14872
55	3557/13	100/93	24900	0/00299
60	3728/24	100/99	27000	0/055728
65	3769/54	101/23	27600	0/16087
70	3993/10	101/44	27750	0/2555
75	4006/69	102/76	30900	0/32981
80	4232/14	104/95	31650	0/44149

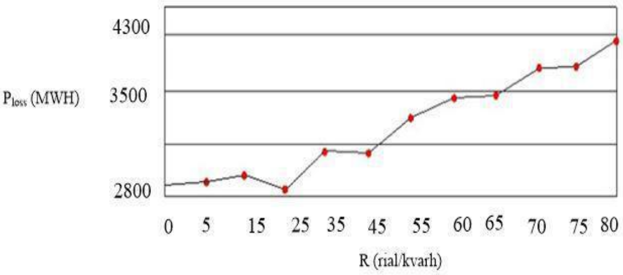


Figure 4. Energy lost in network grid for different values of the parameter R during a year

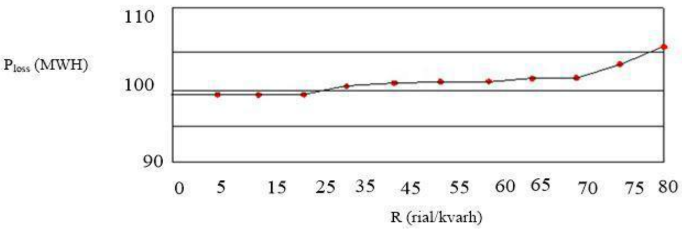


Figure 5. The lost energy diagram in the transformer 33/230 for different values of the parameter R during one year

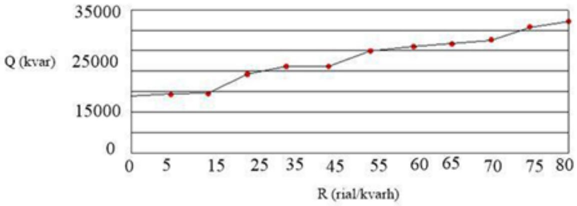


Figure 6. Capacitive capacitance diagram on the network for various values of R

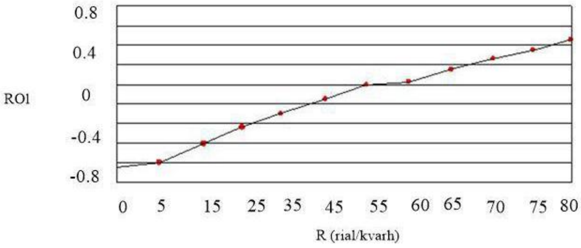


Figure 7. Return indicator of capital for various values of R

Also, in this program, the capacity of the capacitor bank is determined at load nodes at any time of the day. In fact, on the capacities obtained, it is easy to determine the switching state of the clock per hour of the capacitors.

In Figure 8, the switching state is provided in the state R = 0. Similarly, the switching status at other prices is also indicated in this program, due to the high volume of information, the total annual energy exchanged in R is presented. In Figure 9, reactive energy exchanged during one year for each parameter of R is presented.

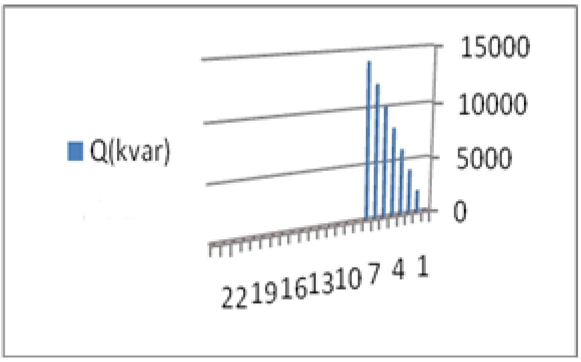


Figure 8. Change the nominal capacitance in length Overnight for R = 0

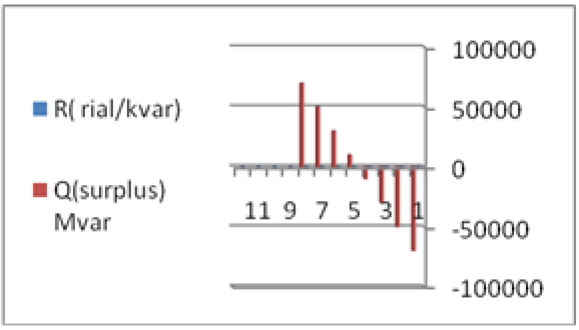


Figure 9. Reactive energy purchased from / sold to transmission network over a year

5. CONCLUSION

By formulating the problem of determining the position and capacitance of the capacitors in a redistributed distribution network and examining the results obtained from the implementation of the above method for the Roy - Billinton test network, the results are:

- Prior to the restructuring, a distribution network is required to continuously improve the voltage profile and reduce network losses, and this is possible through the installation of a reactive power compensation system in the distribution network.
- In the redistributed distribution networks, the capacity of the capacitor to reduce network losses depends on market conditions.
- In this new scenario, investing in the installation of a compensation system for the distribution system operator has two advantages in reducing the loss of power and increasing the profit from the sale of reactive power to the transmission network operator.
- The economic value assigned to the reactive power sold to the transmission system (R) should be reasonable. This economic value for a distribution system operator cannot be reduced to a certain threshold value.
- If the system gets the reactive power at a cost higher than the threshold cost, it is economically feasible to install capacitor banks and exploit them for the distribution network operator, and if this cost For the transmission system be less than threshold value, according to the distribution system operator, there is no need to install a compensation system - more than what is needed to reduce the distribution network losses - and this occurs when the generation and transmission of reactive power It's cheaper than buying it from the distribution network.
- For R values less than the threshold value, the nominal capacity of the capacitor and, consequently, the network losses are relatively constant, and changes in this parameter lead to similar arrangements of capacitor banks in the network bushings. In this case, the reactive power generated in the distribution network is equal to or less than the reactive power level of the reactors and losses, and considering the unit price of the reactive service and the cost of the purchase, installation and maintenance of the capacitor banks may be partially or entirely get the reactive energy you need from the transmission network.
- With reaching R to the threshold value and increase of this parameter, the phrase "installed nominal capacitance" increases along with the return on capital.
- At the same time, with the increase in the economic benefits of investing in the installation of capacitor banks more than the amount of network required, the network losses and transformers also increase to some extent. The reason for this is that the increase in reactive power in the network load cells is more than that which increases the losses.
- If the price of reactive energy is achieved regardless of constraint, it is possible that the resulting economic benefits will not only reduce the loss of the network but also increase it. Therefore, the price increase of the threshold value is also limited.
- Most of the reactive power sold to the transmission system - due to the limitations of the number of authorized capacitor banks in the network load nodes, the maximum Megavar injected to the transmission network and the voltage limit of each shaft - is due to the nominal capacity at the MV / HV station. Therefore, it can be said that the capacitors installed in the medium pressure network shins are used to reduce the losses of the distribution network and the capacity of the MV / HV capacitor banks is related to the investment and sale of energy to the transmission network operator.

The reason that the maximum number of capacitors in the MV / HV station is higher than the network load nodes is that the MV / HV station is closer to the transmission network and a lot of bus-bars are connected to the reactive load. Therefore, it is possible to move and sell energy in this bus-bar.

REFERENCES

[1] V. Miranda, J. Ranito, and L. M. Proenca, "Genetic algorithms in optimal multistage distribution network planning," *IEEE Transactions on Power Systems,* vol. 9, pp. 1927-1933, 1994.

[2] F. Rahmani, F. Razaghian, and A. Kashaninia, "Novel Approach to Design of a Class-EJ Power Amplifier Using High Power Technology," *World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering,* vol. 9, pp. 541-546, 2015.

[3] M. T. Andani, H. Pourgharibshahi, Z. Ramezani, and H. Zargarzadeh, "Controller design for voltage-source converter using LQG/LTR," in *Texas Power and Energy Conference (TPEC), 2018 IEEE,* 2018, pp. 1-6.

[4] M. Ketabdar, "Numerical and Empirical Studies on the Hydraulic Conditions of 90 degree converged Bend with Intake," *International Journal of Science and Engineering Applications,* vol. 5, pp. 441-444, 2016.

[5] M. T. Andani and Z. Ramezani, "Robust Control of a Spherical Mobile Robot," 2017.

[6] H. Pourgharibshahi, M. Abdolzadeh, and R. Fadaeinedjad, "Verification of computational optimum

tilt angles of a photovoltaic module using an experimental photovoltaic system," *Environmental Progress & Sustainable Energy,* vol. 34, pp. 1156-1165, 2015.

[7] M. Delfanti, G. P. Granelli, P. Marannino, and M. Montagna, "Optimal capacitor placement using deterministic and genetic algorithms," in *Proceedings of the 21st International Conference on Power Industry Computer Applications. Connecting Utilities. PICA 99. To the Millennium and Beyond (Cat. No. 99CH36351)*, 1999, pp. 331-336.

[8] R. A. Gallego, A. J. Monticelli, and R. Romero, "Optimal capacitor placement in radial distribution networks," *IEEE Transactions on Power Systems,* vol. 16, pp. 630-637, 2001.

[9] K. Yousefpour, "Placement of dispersed generation with the purpose of losses reduction and voltage profile improvement in distribution networks using particle swarm optimization algorithm," *Journal of World's Electrical Engineering and Technology,* vol. 3, pp. 118-122, 2014.

[10] M. Rostaghi-Chalaki, A. Shayegani-Akmal, and H. Mohseni, "Harmonic analysis of leakage current of silicon rubber insulators in clean-fog and salt-fog," in *18th International Symposium on High Voltage Engineering*, 2013, pp. 1684-1688.

[11] M. H. Imani, M. Y. Talouki, P. Niknejad, and K. Yousefpour, "Running direct load control demand response program in microgrid by considering optimal position of storage unit," in *2018 IEEE Texas Power and Energy Conference (TPEC)*, 2018, pp. 1-6.

[12] M. Rostaghi-Chalaki, A. Shayegani-Akmal, and H. Mohseni, "A study on the relation between leakage current and specific creepage distance," in *18th International Symposium on High Voltage Engineering (ISH 2013)*, 2013, pp. 1629-1623.

[13] J. Carlisle, A. El-Keib, D. Boyd, and K. Nolan, "A review of capacitor placement techniques on distribution feeders," in *Proceedings The Twenty-Ninth Southeastern Symposium on System Theory*, 1997, pp. 359-365.

[14] A. F. Bastani and D. Damircheli, "An adaptive algorithm for solving stochastic multi-point boundary value problems," *Numerical Algorithms,* vol. 74, pp. 1119-1143, 2017.

[15] P. M. Hogan, J. D. Rettkowski, and J. Bala, "Optimal capacitor placement using branch and bound," in *Proceedings of the 37th Annual North American Power Symposium, 2005.*, 2005, pp. 84-89.

[16] B. Rahimikelarijani, M. Saidi-Mehrabad, and F. Barzinpour, "A mathematical model for multiple-load AGVs in Tandem layout," *Journal of Optimization in Industrial Engineering,* 2018.

[17] M. AlHajri, M. AlRashidi, and M. El-Hawary, "A novel discrete particle swarm optimization algorithm for optimal capacitor placement and sizing," in *2007 Canadian Conference on Electrical and Computer Engineering*, 2007, pp. 1286-1289.

[18] R. Eini, "Flexible Beam Robust Loop Shaping Controller Design Using Particle Swarm Optimization," *Journal of Advances in Computer Research,* vol. 5, pp. 55-67, 2014.

[19] A. F. Bastani, Z. Ahmadi, and D. Damircheli, "A radial basis collocation method for pricing American options under regime-switching jump-diffusion models," *Applied Numerical Mathematics,* vol. 65, pp. 79-90, 2013.

[20] R. Bayindir, S. Sagiroglu, and I. Colak, "An intelligent power factor corrector for power system using artificial neural networks," *Electric Power Systems Research,* vol. 79, pp. 152-160, 2009.

[21] F. Rahmani, "Electric Vehicle Charger based on DC/DC Converter Topology," *International Journal of Engineering Science,* vol. 18879, 2018.

[22] M. Alizadeh, I. Mahdavi, S. Shiripour, and H. Asadi, "A nonlinear model for a capacitated location–allocation problem with Bernoulli demand using sub-sources," *Int J Eng,* vol. 26, pp. 1007-1016, 2013.

[23] M. Ketabdar, A. K. Moghaddam, S. A. Ahmadian, P. Hoseini, and M. Pishdadakhgari, "Experimental Survey of Energy Dissipation in Nappe Flow Regime in Stepped Spillway Equipped with Inclined Steps and Sill," *International Journal of Research and Engineering,* vol. 4, pp. 161-165, 2017.

[24] J. Rahmani, E. Sadeghian, and S. Dolatiary, "Comparison between ideal and estimated pv parameters using evolutionary algorithms to save the economic costs," 2018.

[25] Z. Gu and D. T. Rizy, "Neural networks for combined control of capacitor banks and voltage regulators in distribution systems," *IEEE transactions on power delivery,* vol. 11, pp. 1921-1928, 1996.

[26] B. Rahimikelarijani, A. Abedi, M. Hamidi, and J. Cho, "Simulation modeling of Houston Ship Channel vessel traffic for optimal closure scheduling," *Simulation Modelling Practice and Theory,* vol. 80, pp. 89-103, 2018.

[27] R. Eini and A. R. Noei, "Identification of Singular Systems under Strong Equivalency," *International Journal of Control Science and Engineering,* vol. 3, pp. 73-80, 2013.

[28] M. Alizadeh, I. Mahdavi, N. Mahdavi-Amiri, and S. Shiripour, "A capacitated location-allocation problem with stochastic demands using sub-sources: An empirical study," *Applied Soft Computing,* vol. 34, pp. 551-571, 2015.

[29] S. Dolatiary, J. Rahmani, and Z. Khalilzad, "Optimum Location of DG Units Considering Operation Conditions."

[30] M. Ketabdar and A. Hamedi, "Intake Angle Optimization in 90-degree Converged Bends in the Presence of Floating Wooden Debris: Experimental Development," *Florida Civ. Eng. J,* vol. 2, pp. 22-27.2016, 2016.

[31] J. R. Santos, A. G. Exposito, and J. M. Ramos, "A reduced-size genetic algorithm for optimal capacitor placement on distribution feeders," in *Proceedings of the 12th IEEE Mediterranean Electrotechnical Conference (IEEE Cat. No. 04CH37521)*, 2004, pp. 963-966.

[32] S. M. RakhtAla and R. Eini, "Nonlinear modeling of a PEM fuel cell system; a practical study with experimental validation," *International Journal of Mechatronics, Electrical and Computer Technology,* vol. 4, pp. 1272-1296, 2014.

[33] F. Rahmani, F. Razaghian, and A. Kashaninia, "High Power Two-Stage Class-AB/J Power Amplifier with High Gain and Efficiency," 2014.

[34] M. Alizadeh, N. Mahdavi-Amiri, and S. Shiripour, "Modeling and solving a capacitated stochastic location-allocation problem using sub-sources," *Soft Computing,* vol. 20, pp. 2261-2280, 2016.

[35] M. Taheri Andani, Z. Ramezani, S. Moazami, J. Cao, M. M. Arefi, and H. Zargarzadeh, "Observer-Based

Sliding Mode Control for Path Tracking of a Spherical Robot," *Complexity,* vol. 2018, 2018.

[36] A. Hamedi, M. Ketabdar, M. Fesharaki, and A. Mansoori, "Nappe Flow Regime Energy Loss in Stepped Chutes Equipped with Reverse Inclined Steps: Experimental Development," *Florida Civil Engineering Journal,* vol. 2, pp. 28-37, 2016.

[37] M. T. Andani, S. Shahmiri, H. Pourgharibshahi, K. Yousefpour, and M. H. Imani, "Fuzzy-Based Sliding Mode Control and Sliding Mode Control of a Spherical Robot," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*, 2018, pp. 2534-2539.

[38] A. Hamedi and M. Ketabdar, "Energy Loss Estimation and Flow Simulation in the skimming flow Regime of Stepped Spillways with Inclined Steps and End Sill: A Numerical Model," *International Journal of Science and Engineering Applications*, vol. 5, pp. 399-407, 2016.

[39] A. Rouholamini, H. Pourgharibshahi, R. Fadaeinedjad, and G. Moschopoulos, "Optimal tilt angle determination of photovoltaic panels and comparing of their mathematical model predictions to experimental data in Kerman," in *Electrical and Computer Engineering (CCECE), 2013 26th Annual IEEE Canadian Conference on*, 2013, pp. 1-4.

[40] Yousefpour, Kamran, et al. "Using a Z-Source Inverter as Rotor Side Converter in Doubly-Fed Induction Generator Wind Energy Conversion System." *2019 IEEE Texas Power and Energy Conference (TPEC)*. IEEE, 2019.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 04, 98-101, 2019, ISSN:-2319–8656

EYE BLINK CONTROLLED VIRTUAL KEYBOARD USING BRAIN COMPUTER INTERFACE

Palak M Jain
Student
SRM Institute of Science and Technology
Chennai, India
Palakrathod8423@gmail.com

Shipra Singh
Student
SRM Institute of Science and Technology
Chennai, India
Shipra65624@gmail.com

Nikita Singh
Student
SRM Institute of Science and Technology
Chennai, India
nikitasingh.cse2019@gmail.com

Akila V
Professor
SRM Institute of Science and Technology
Chennai, India
Akila.v@vdp.srmuniv.ac.in

Abstract – In our society there are more people suffered by paralytic diseases which causes them several disabilities like unable to talk and unable to move physically and unable to express their everyday basic needs, but they can still use their eyes and sometimes move their heads. This Project is working under the principle of Brain-Computer Interface (BCI). Our model helps them to type the letters using virtual keyboard, which is displayed in the monitor. Virtual keyboard contains alphabets, numbers and some sentences. Mouse pointer gets automatically shifted through every keys, characters can be chosen by making an Eye blink at particular position of mouse pointer at certain character.

Keywords – Brain-Computer Interface (BCI), EEG (Electroencephalography) waves, MATLAB, ThinkGear and Virtual Keyboard.

1. INTRODUCTION

BCI has been focusing on the development of communication between the human brain and the computer. An Individual's brain signals are used to access a system. The electrical activity of the brain is analyzed and measured by the electroencephalogram (EEG). For this process a Brain Sensor is placed on the scalp of human brain which consists of metal discs. These small metal disks are called electrodes. The electrodes carries the electrical waves (or) signals which passes through the skull of the human brain. These electrical signals extracted are categorized and recorded. The electrical signals are then translated into meaningful orders to make an application to perform its activity as per the order from the brain signal [1]. Electroencephalography (EEG) which monitors an Electric property of Brain along with the Scalp (Non Invasive). The Brain Sense measures intentionally directed EMG activity (Blink Strength). The goal of BCI is the movements, communication and environmental control for handicapped people [5]. The term virtual keyboard means it is a keyboard, which appears on the screen to get the input. The keyboard is designed in such a way that the cursor moves through each character in the keyboard. As the person blinks the character or alphabet is taken as input and the same process continues.

2. LITERATURE REVIEW

Rakesh Ranjan and Sasidharan have examined in this paper Utilization of EEG Signal as Virtual Keyboard for Physically Disabled about the Brain Computer Interface (BCI) and Electroencephalogram (EEG) that captures the signals from the brain using a BrainSense and a virtual Keyboard is set up in labview platform [1]. The Virtual Keyboard consists of three blocks (block 1, block 2 and block 3) with different set of alphabets. The Eye blinks are used as control signals or inputs in BCI, The user has to blink within a specific time interval, i.e. 5 seconds. The inputs captured are binary in nature. BCI detects the presence of an Eye blink within 5 seconds of time interval. A character selection is done in 20 seconds. In first 5 seconds the presence of single blink is detected. In the next 5 seconds the presence of two blinks are detected and in the last 5 seconds the presence of three blinks are detected. The final output is displayed in total time interval of 20 seconds. A character as an output is obtained as one character/minute.

The selection of a particular block is done in 20 seconds, the selection of a column in the block requires 20 seconds and the selection of a character from the selected column requires 20s and finally display the output in the screen.

Krzysztof Dobosz and Klaudiusz Stawski have examined in their paper Touchless Virtual Keyboard Controlled by Eye Blinking and EEG Signals about the disabled people suffering with tetra paresis and designed a system which consists of a Neuro keyboard, EMG sensors and EEG signals that are used as a support [7]. The algorithm implemented was Divide and Conquer. In this system The Virtual Keyboard consists of two modes Character selection and Predicted word selection. The row consisting of the character to be displayed is selected first and then the row gets divided into two parts and only one part consisting of the character is selected. Then the required character is then selected from the divided parts and the output is displayed.

3. PROPOSED SYSTEM

In this paper we have introduced a Brain Computer Interface (BCI). The main objective of our system is to help the physically disabled people (people suffering from paralytic diseases) to communicate easily with other people. This can operate through their blinking capability to recognize the letters in the Virtual Keyboard and selecting the required character. Virtual Keyboard contains alphabets, numbers and some punctuations. By using Brain Sense which has EEG (Electroencephalography) waves, the electrical signals are captured from the brain and the blinks are detected in an efficient way. EEG signals supports the controlling of the keyboard with the blinks and ThinkGear is the Novelty in this approach.

4. SYSTEM DESCRYPTION

4.1 BRAIN WAVE SENSOR

Brain wave sensor is a communication channel between the human brain and computer. It is composed of signals, pattern identification, processing, and control systems [1]. Brain wave sensor is a headset that transforms the computer into brain activity monitor. The headset estimates the brain wave signals. It observes the attention level of the user as they transmit the information to the virtual keyboard. More established EEGs require the use of the conductive gel between the sensors and the head. The systems also incorporates built-in electrical noise [6]. In this paper, we are using a brain wave sensor headset which has a electrode in the place of the forehead, which also has three sensor algorithm- attention sensor, meditation sensor and blink detection sensor algorithms. The measured electrical signals and determined interpretations are then yield as computerized messages to the computer.

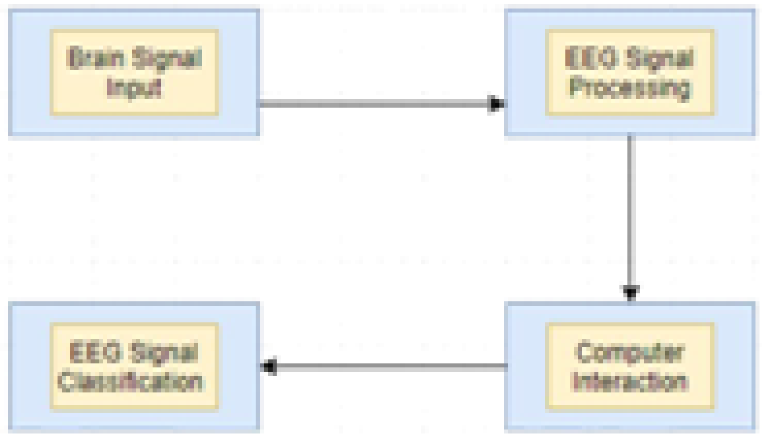


Fig. 4.1.1. BCI Flow diagram

4.2 MATLAB

MATLAB is an intuitive framework hose essential information component is an exhibit that does not require dimensioning. This enables you to solve many technical computing problems, particularly those with matrix and vector formulations. Matlab permits including the thinkgear library file with .dll extension. It would take a small amount of time to compose a program in a scalar non-intuitive language, for instance C or Fortran. It gathers the signs intermittently and it transfers the data to digital signal processing unit and maps the appropriate actions and gives instructions to the controller to operate the external unit [3].

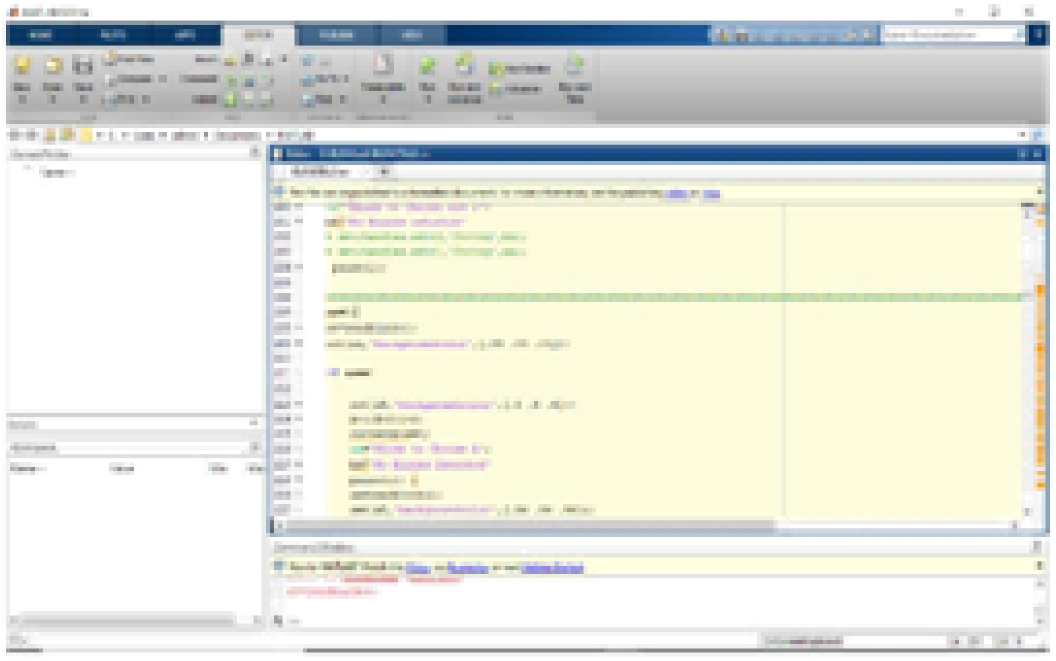


Fig.4.2.1 Screenshot of MATLAB

4.3 VIRTUAL KEYBOARD

The virtual Keyboard is developed using Matlab Gui. The entire alphabets (A - Z) are available in the Virtual Keyboard. The cursor moves through each alphabet and extra keys. The user has to focus on a key in the virtual keyboard and has to perform an eye blink when the cursor is on the key that has to be printed on the screen. [2] The key changes its color when the user blinks an eye to select the key.

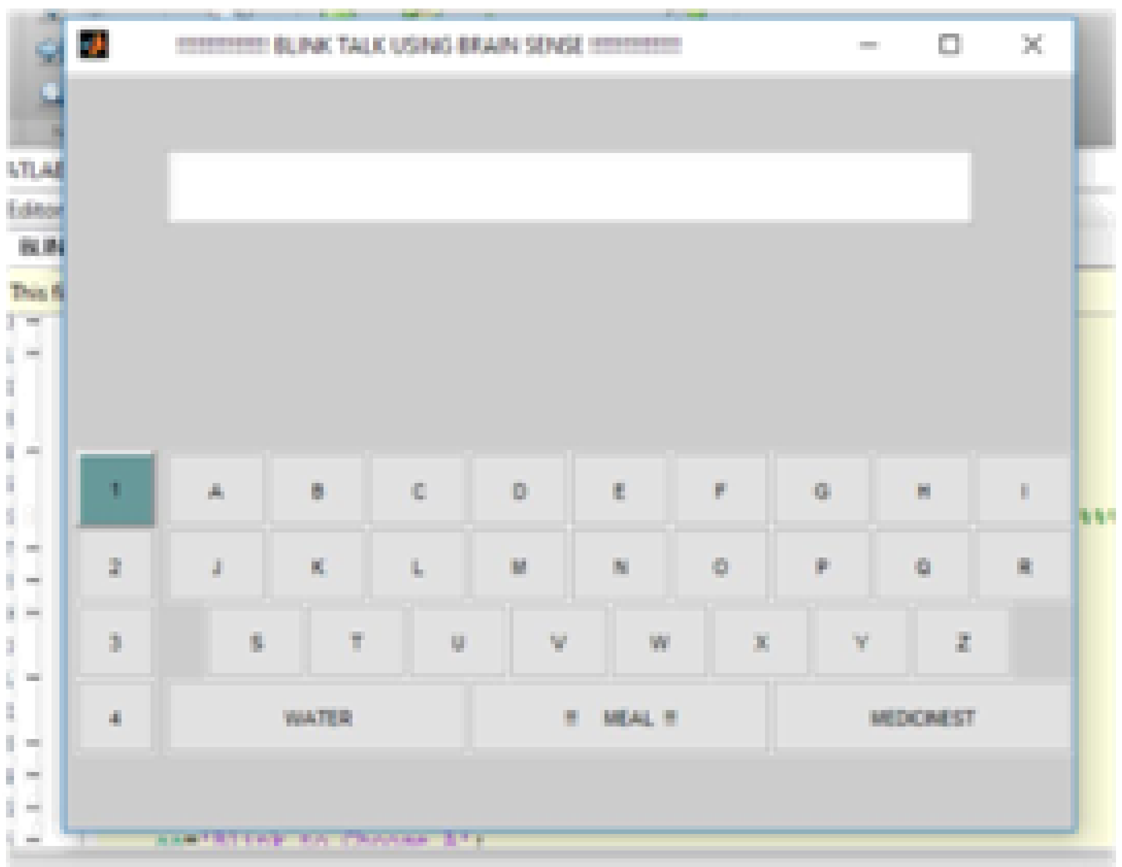


Fig. 4.3.1 Screenshot of virtual keyboard

5. METHODOLOGY

5.1 EEG BIOMETRIC ALGORITHMS

5.1.1 Attention Algorithm

This algorithm shows the intensity of mental focus or attention. The value ranges from 0 to 100. The attention level increments when a user focuses on a single thought or an external object and decrements when he/she is diverted. In gaming, attention has been utilized to make push command over virtual objects [4].

5.1.2 Meditation Algorithm

This algorithm calculates the dimension of calmness or relaxation. The value ranges from 0 to 100 and increments when the users loosens up or relaxes his/her mind and decrements when they are stressed or they feel uneasy. The Meditation Meter evaluates the capacity to find an inner state of mindfulness and thus helps users to beat the worries of regular day to day existence. This algorithm is also used in a variety of game design controls [4].

5.1.3 Blink Detection Algorithm

This algorithm calculates the signals from the user's blinks. A higher number shows a stronger blink and a lower number shows a weaker or fragile blink. The recurrence of the blinks is regularly correlated with fatigueness or nervousness [4]. For instance, in other applications, one blink means no and two means yes, giving people with special needs a basic and simple way to communicate.

5.1.4 ThinkGear Library file

ThinkGear is a library file which is installed in MATLAB to get signals from Brain wave sensor [3]. Brain wave headset transfers the signals using this library file to Matlab and the result is displayed on the screen [3]. ThinkGear is a dry sensor technology that lets the intensification, estimation and investigation of EEG signs and brain waves. This enables the headband to almost certainly measure the wearer's perspective and makes this information available to the application so that the application can react to psychological movement.

6. PERFOMANCES AND DISCUSSIONS

A character selection can be minimum at the rate of 2 or 3 characters/min and it is obtained with the current settings. Each selection by the user (Eye blink) is followed by giving a visual response to the user by changing the waves as peak and valleys in the brainwave visualizer.

Average typing rate (sec/letter)	**No. of typing errors**	**Accuracy**
5	5-6	75%

7. CONCLUSION & FUTURE SCOPES

This paper presented a development of the Brain Computer Interface application. The control signals used in this system is eye blinks. The BCI system can be used for the communication purposes, with eye blinks as control signals, especially for the patients suffering from Paralysis [2]. The Virtual Keyboard which is developed obtained a result at the rate of 2-3 characters/minute. This paper can be further extended with the voice-based output and including dictionary and produce expected result with more accuracy and this BCI system can also be used to send messages by the patients to their care taker. It can be used to transmit messages personally without the help of others.

REFERENCES

[1] Utilization of EEG Signal as Virtual Keyboard for Physically Disabled, International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), Volume 5, Issue 5, May 2016.

[2] EEG Based Brain Computer-A Virtual Keyboard Control Using Brainwave Sensor for Crippled With IAUI, Jour of Adv Research in Dynamical & Control Systems, 11-Special Issue, July 2017.

[3] Home Appliances Control Using Brain Wave Sensor by EEG, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 10, October-2015.

[4] http://neurosky.com/biosensors/eeg-sensor/algorithms/

[5] EEG based Brain computer interfaces: A novel Neurotechnology and computational intelligence method, IEEE Systems, Man, and Cybernetics Magazine, Volume 3, Issue 4, Oct. 2017.

[6] Wireless Control of Mobile Robot using Brain Computer Interface, International Journal of Scientific Engineering and Technology Research, Volume 4, Issue 45, November 2015.

[7] Touchless virtual keyboard controlled by eye blinking and EEG signals, Conference Paper · January 2018.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 04, 102-106, 2019, ISSN:-2319–8656

Research on Modulation recognition technology based on Machine learning

Biao Xu
School of Communication Engineering,
Chengdu University of Information Technology
Chengdu, China

Xiping Wen
School of Communication Engineering,
Chengdu University of Information Technology
Chengdu, China

Xi Wang
School of Communication Engineering,
Chengdu University of Information Technology
Chengdu, China

Abstract: This paper presents a method for modulation recognition of digital signals using machine learning. This method first extracted seven characteristic parameters according to the instantaneous parameters and high-order cumulant characteristics of the signal, then combined the decision tree classifier, neural network classifier and support vector machine classifier in machine learning with these characteristic parameters, and finally realized the modulation recognition of MASK, MFSK, MPSK and MQAM signals. This method not only has low computational complexity and more recognized signals, but also improves the recognition rate at low SNR.

Keywords: Machine learning; Modulation recognition; Decision tree; Neural network; Support vector machine.

1. INTRODUCTION

With the continuous innovation of communication technology, wireless communication network environment is becoming more and more complex. In the non-cooperative communication environment, modulation recognition technology can recognize and classify the modulation type of blind signal and play a key role in signal detection and demodulation. Modulation recognition technology is of great significance in both civilian and military fields. At present, there are two types of automatic modulation recognition technology: one is modulation recognition based on maximum likelihood theory; The other is modulation recognition based on statistical mode. The former requires many prior conditions and is not suitable for non-cooperative communication environment. The latter requires less prior conditions and can realize blind signal recognition, which is widely used in engineering practice. The principle of modulation recognition technology based on statistical mode is shown in figure 1, which mainly includes signal preprocessing, feature extraction and classifier recognition.

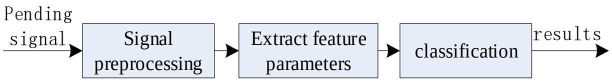


Figure 1. Modulation recognition based on statistical mode.

Signal preprocessing includes signal down conversion, carrier frequency estimation, bandwidth estimation, etc. Common methods for extracting feature parameters include instantaneous information[1], high-order cumulants[2], wavelet transform[3], spectral features[4], constellation map[5], etc. Common classification and recognition methods include decision tree[6], neural network[7], support vector machine[8], etc.

Modulation recognition based on statistical mode was first introduced in 1984. Liedtke proposed the concept of modulation recognition, and realized the recognition of digital modulation signals by using the parameters of signal amplitude histogram, frequency histogram, amplitude variance and frequency variance, etc [9]. With the development of machine learning, more and more people begin to use supervised learning to improve the efficiency of signal recognition. Recognition of digital modulation dignals using deep learning in document [10]; Cheol-Sun Park use support vector machine to improve the recognition effect [11]. Timothy J. O'Shea proposed using convolutional neural network to solve the problem of modulation recognition. Compared with the traditional classification algorithm in machine learning, it has been greatly improved [12].

In this paper, we use the machine learning classifier to realize the modulation recognition of signals. By analyzing the performance of different machine learning classifiers, we can provide guidance for engineering applications. At the same time, a new joint feature parameter is proposed, which can improve the accuracy of signal recognition.

The remainders of the paper are organized as follows: Section 2 discusses signal model and feature extraction of signal, which will provide a basis for using machine learning classifier. In Section 3, we will design the classifier and introduce the basic principles of the three classification algorithms. At Section 4, we introduce the experimental environment and carry out the experimental test, and analyze and compare the test results. Section 5 concludes the paper.

2. EXTRACTION FEATURE

The types of modulation signals to be identified in this paper are ASK, 4ASK, MSK, 2FSK, 4FSK, BPSK, QPSK, 8PSK, 16QAM and 64QAM. The unified mathematical expressions of these ten signals can be expressed as follows:

$$S(t) = a(t)\cos(w_c t + \phi(t)) \qquad (1)$$

In the formula, w_c is the angular frequency of the signal, $\phi(t)$ is the phase of the signal, $a(t)$ is the amplitude of the signal.

2.1 Instantaneous feature parameters

In this paper, instantaneous information parameters are selected from non-weak signal segments, and the zero center and normalization of the parameters are processed. The instantaneous feature parameters extracted in this paper include the following contents:

(1)Logarithm of second-order origin moment of instantaneous amplitude:

$$Ma_1 = \lg\left[\frac{1}{N_s}\sum_{n=0}^{N_s-1} A_{cn}^{\ 2}(n)\right] \tag{2}$$

In the formula, N_s is all sampled data points, and A_{cn} is the instantaneous amplitude with zero center and normalization. The instantaneous amplitude can be used to represent the envelope change of the signal.

(2)Logarithm of the origin moment of the absolute value of instantaneous amplitude:

$$Ma_2 = \lg\left[\frac{1}{N_s}\sum_{n=0}^{N_s-1} \left|A_{sn}(n)\right|\right] \tag{3}$$

A_{sn} is the normalized instantaneous amplitude after recursion. This parameter can distinguish 2ASK signal from 4ASK signal.

(3)Second-order amplitude moment of MQAM signal:

$$M_{asm} = M_{2,MQAM}^{2n} = \frac{M-1}{3}a^2 \tag{4}$$

The theoretical values of 16QAM and 64QAM are 5 and 21 respectively, so the two signals can be clearly distinguished by this parameter.

(4)Logarithm of origin moment of absolute value of instantaneous frequency:

$$Mf_1 = \lg\left[\frac{1}{N_s}\sum_{n=0}^{N_s-1} \left|f_{cn}(n)\right|\right] \tag{5}$$

f_{cn} is the normalized instantaneous frequency of zero center. Through this parameter, MPSK signal and MFSK signal can be distinguished.

2.2 Characteristic parameters of higher-order cumulants

The theoretical values of five kinds of higher-order cumulants for each signal are given in Table 1. According to the theoretical values, the following three characteristic parameters can be constructed：

Table 1. Recognition results of different modulated signals

signal	$\lvert C_{40}\rvert$	$\lvert C_{41}\rvert$	$\lvert C_{42}\rvert$	$\lvert C_{63}\rvert$	$\lvert C_{80}\rvert$
2ASK	$2E^2$	$2E^2$	$2E^2$	$13E^3$	$272E^4$
4ASK	$1.36E^2$	$1.36E^2$	$1.36E^2$	$8.32E^3$	$111.8E^4$
MSK	0	0	E^2	$4E^3$	0
2FSK	0	0	E^2	$4E^3$	0
4FSK	0	0	E^2	$4E^3$	0
BPSK	$2E^2$	$2E^2$	$2E^2$	$13E^3$	$272E^4$
QPSK	E	0	E^2	$4E^3$	$34E^4$
8PSK	0	0	E^2	$4E^3$	E^4
16QAM	$0.68E^2$	0	$0.68E^2$	$2.08E^3$	$13.98E^4$
64QAM	$0.62E^2$	0	$0.62E^2$	$1.80E^3$	$11.50E^4$

(1) Feature parameter Fx_1 :

$$Fx_1 = \frac{\left|C_{41}\right|}{\left|C_{42}\right|} \tag{6}$$

This parameter can divide the signal into {MASK, BPSK} and {MFSK, MPSK, MQAM} sets.

(2) Feature parameter Fx_2 :

$$Fx_2 = \frac{\left|C_{80}\right|}{\left|C_{42}\right|^2} \tag{7}$$

This parameter can divide the signal into {QPSK}, {16QAM, 64QAM} and {MFSK, 8PSK} sets.

(3) Feature parameter Fx_3 :

$$Fx_3 = \frac{\left|C_{63}\right|^2}{\left|C_{42}\right|^3} \tag{8}$$

This parameter can realize intra-class discrimination of MASK signals and intra-class discrimination of MFSK signals.

In this paper, instantaneous information and high-order cumulants are combined to form a 7-dimensional eigenvector, and then the signal is extracted according to these seven feature parameters.

3. CLASSIFIER DESIGN

3.1 Decision tree classifier

Decision tree classifier is the most commonly used classification algorithm in machine learning. A complete decision tree contains one root node, several internal nodes and several leaf nodes [13]. The leaf node corresponds to the decision result. The purpose of decision tree learning is to produce a decision tree with strong adaptability, which is trained as a recursive process. The most important step in the learning process is to select the optimal partition attribute, which can make the node more pure. Common partition methods include information gain, gain rate and Gini index. In this paper, the Gini index is used to measure the purity of each node of the decision tree. The smaller the Gini index is, the higher the purity of the sample will be.

Before using the decision tree classifier, it is necessary to set the relevant parameters of the decision tree, including the maximum number of categories, the depth of the decision tree, the minimum number of samples of nodes, etc. The specific parameters of decision tree classifier in this paper are shown in Table 2.

Table 2. Relevant parameters of decision tree classifier

Parameter name	**Value**
MaxCategories	4
MaxDepth	10

MinSampleCount	5
CVFolds	0
UseSurrogates	False
Use1SERule	False
TruncatePrunedTree	True
RegressionAccuracy	0

3.2 Neural network classifier

Neural network is a network formed by the interconnection of a large number of neurons. Its basic unit is neurons. In the process from input space to output space, the neural network constantly adjusts the weights and thresholds of the network to find the relationship between variables in order to achieve the best classification effect [14]. Because the feature parameters have been extracted before signal recognition, it belongs to modulation recognition based on shallow neural network. In order to be applied in engineering, this paper chooses BP neural network with stable and simple structure. The number of layers of neural network is chosen in three layers, namely input layer, hidden layer and output layer. Through a large number of experiments, the number of hidden layer neurons in this paper is 20. In this paper, the activation function of the neural network is Sigmoid function, and the learning algorithm is RPROP algorithm. In order to improve the accuracy, the maximum training times are set at 10000 times, and the training accuracy is 0.000001.

3.3 Support vector machine classifier

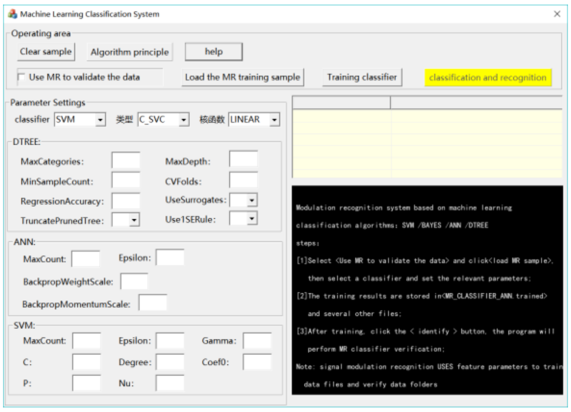
Support Vector Machine (SVM) is a linear binary classification model. Its basic idea is to transform the non-linear problem into a high-dimensional linear separable problem, and then find the optimal linear interface [15]. For linear separable problems, SVM can find the optimal classification hyperplane in the original space; for non-linear separable problems, SVM need to map low-dimensional spatial data into high-dimensional space, and then to find the optimal classification hyperplane. SVM can solve the problem of small sample and non-linearity better. In OpenCV, C_SVC and NU_SVC can realize the function of multi-classification, so they can only be selected when realizing modulation type recognition. C_SVC is called Class C Support Vector Machine (CSVM) classifier, which allows incomplete classification with outlier penalty factor C; NU_SVC is called Class C Support Vector Machine (CSVM). The commonly used kernels in support vector machines are linear kernels, Sigmoid kernels, RBF kernels and INTER kernels. This paper chooses INTER kernels to train data after many experiments.

4. SIMULATION AND ANALYSIS

This paper will use Matlab R2016b, OpenCV310 and Visual Studio 2010 to test the modulation recognition algorithm. The computer environment used in the experiment is Windows 1064-bit operating system, the CPU is Intel Core i5-6300HQ, and the computer memory is 12.0GB.

4.1 Establish classification model

We use MFC and OpenCV310 to construct a classification system of modulation recognition based on machine learning under Visual Studio 2010 platform. The interface of the classification system is shown in Figure 2. The system is divided into three modules: operation area, parameter setting and display area. The system can set the specific parameters of the classifier in detail, and realize the function of classifying and identifying the signals that have been extracted features, including the training of the classifier and the recognition of the signals.



classifier SVM 类型 C_SVC 核函数 LINEAR

```
Modulation recognition system based on machine learning
classification algorithms: SVM /BAYES /ANN /DTREE
steps:
[1]Select <Use MR to validate the data> and click<load MR sample>,
  then select a classifier and set the relevant parameters;
[2]The training results are stored in<MR_CLASSIFIER_ANN.trained>
  and several other files;
[3]After training, click the < identify > button, the program will
  perform MR classifier verification;
Note: signal modulation recognition USES feature parameters to train
  data files and verify data folders
```

Figure 2. Machine learning classification system.

4.2 Generate experimental data

For the training and validation data used in machine learning, this paper simulates the generation of training and validation data set under the platform of Matlab. The data set contains the following contents:

(1) There are ten kinds of digital modulation signals: 2ASK, 4ASK, MSK, 2FSK, 4FSK, BPSK, QPSK, 8PSK, 16QAM and 64QAM.

(2) The parameters of the simulation signal: carrier frequency is 4000Hz, sampling frequency is 2400Hz, symbol rate is 3000bit/s, symbol number is 1024, noise is Gauss white noise, SNR range of 0~29dB, step length of 1dB.

(3) The ten kinds of digital signals are labeled and numbered according to the sequence of signals in (1) starting from No. 1.

(4) According to the parameter set in the second section, the feature of the simulated signal is extracted, and the 7-dimensional feature vector is obtained by the feature. The set of characteristic parameters in this paper is $\{Ma_1, Ma_2, M_{asm}, Fx_1, Fx_2, Mf_1, Fx_3\}$.

(5) A total of 90,000 sets of feature data were generated by simulation, of which 60,000 were used as training classifiers and 30,000 were used to verify the classification effect.

4.3 Experiment and analysis

In this section, we will use three machine learning classifiers to analyze ten kinds of simulation signals.

4.3.1 *The results of using decision tree classifier*

Table 3 shows the recognition results using decision tree classifier. The table lists eight SNR recognition cases.

Table 3. Recognition results of different modulated signals

Signal	**Recognition rate(%)**							
	0dB	**3dB**	**5dB**	**7dB**	**10dB**	**15dB**	**20dB**	**25dB**
2ASK	100	100	100	100	100	100	100	100
4ASK	100	100	100	100	100	100	100	100
MSK	80	80	90	100	100	100	100	100
2FSK	80	100	100	100	100	100	100	100
4FSK	100	100	100	100	100	100	100	100
BPSK	100	100	100	100	100	100	100	100
QPSK	60	100	100	100	100	100	100	100
8PSK	80	80	95	100	100	100	100	100
16QAM	100	100	100	100	100	100	100	100
64QAM	100	100	100	100	100	100	100	100

From Table 3, it can be seen that the recognition rate of all signals obtained by using decision tree classifier can reach 100% under the SNR of more than 10 dB. Except MSK, 2FSK, QPSK and 8PSK, all the other digital signals can get 100% recognition rate at 0 dB SNR. The recognition rate of MSK, QPSK and 8PSK can still reach 100% under the SNR of more than 5dB. It can be proved that the decision tree classifier in machine learning can automatically find the decision threshold value of the classification boundary through the relationship between the data, which has a better classification effect.

4.3.2 The results of using neural network classifier

Table 4 shows the recognition results using neural network classifiers. The table lists eight SNR recognition cases.

Table 4. Recognition results of different modulated signals

Signal	**Recognition rate(%)**							
	0dB	**3dB**	**5dB**	**7dB**	**10dB**	**15dB**	**20dB**	**25dB**
2ASK	100	100	100	100	100	100	100	100
4ASK	100	100	100	100	100	100	100	100
MSK	50	70	90	100	100	100	100	100
2FSK	40	60	100	100	100	100	100	100
4FSK	80	90	100	100	100	100	100	100
BPSK	100	100	100	100	100	100	100	100
QPSK	60	100	100	100	100	100	100	100
8PSK	90	50	90	100	100	100	100	100
16QAM	90	100	100	100	100	100	100	100
64QAM	100	100	100	100	100	100	100	100

As can be seen from Table 4, the recognition rate of all signals increases with the increase of SNR. Except MSK, 2FSK, QPSK, 8PSK and 16QAM, the recognition rate of other signals can reach 100% at the SNR of 0dB.The recognition rate of MSK, 8PSK and QPSK is 100% when the SNR is more than 5dB. From the data in the table, it can be seen that the neural network classifier has better recognition performance and can improve the recognition rate of signals at low SNR.

4.3.3 The results of using support vector machine classifier

Table 5 shows the recognition results using support vector machine classifier. The table lists eight SNR recognition cases.

Table 5. Recognition results of different modulated signals

Signal	**Recognition rate(%)**							
	0dB	**3dB**	**5dB**	**7dB**	**10dB**	**15dB**	**20dB**	**25dB**
2ASK	100	100	100	100	100	100	100	100
4ASK	100	100	100	100	100	100	100	100
MSK	70	60	60	100	100	100	100	100
2FSK	20	60	60	80	90	100	100	100
4FSK	80	20	90	100	100	100	100	100
BPSK	100	100	100	100	100	100	100	100
QPSK	50	70	40	60	70	90	80	90
8PSK	50	70	50	50	90	60	80	80
16QAM	100	100	100	100	100	100	100	100
64QAM	100	100	100	100	100	100	100	100

From Table 5-8, it can be found that the recognition results obtained by using support vector machine are inferior to those obtained by decision tree classifier and neural network classifier. For 2ASK, 4ASK, BPSK, 16QAM and 64QAM, the recognition rate can still guarantee 100% at 0dB. The recognition rate of MSK and 4FSK signals can reach 90% above the SNR of 10 dB, but it is not stable enough, and there will be misjudgement under high SNR. The worst recognition results are 2FSK, 8PSK and QPSK signals. The recognition rate of 2FSK signals is very low at low SNR. The recognition rate of QPSK and 8PSK signals is also increasing with the increase of SNR, but can only be maintained at about 80%. It can be seen that when using support vector machine classifier, the extracted feature parameters are not clear enough for the classification boundaries of 8PSK and QPSK signals.

4.3.4 Performance comparison of different classifiers

In this experiment, three kinds of modulation recognition technologies are compared and analyzed. Table 5-10 shows the comparative data of different modulation recognition technologies in different aspects.

Table 6. Recognition results of different modulated signals

Index	Classifier types		
	decision tree	neural network	Support Vector Machine
Identification type	10	10	10
Recognition rate (5dB)	>95%	>95%	>85%
Recognition rate (10dB)	>99%	>99%	>90%
Recognition rate (15dB)	>99%	>99%	>95%
Total recognition rate	98.47%	95.90%	88.23%
Algorithmic complexity	Low	High	High
Simulation time	227.5ms	30882.3ms	17858.0ms

From Table 6, it can be seen that the types of signals recognized by the three modulation recognition methods are the same. Table 6 shows the statistical recognition rates of each modulation recognition mode under three SNR. It can be found that under the same SNR, the recognition rate using support vector machine is the lowest, and the recognition rate using neural network and decision tree classifier is higher. Especially, the recognition rate of modulation recognition technology based on support vector machine is unsatisfactory under various SNR. The main reason is that the distinguishing degree of feature parameters of 2FSK, QPSK and 8PSK is low. In addition, compared with the total recognition rate, algorithm complexity and simulation time, it can be seen that the application of decision tree classifier has the highest total recognition rate, low algorithm complexity, less simulation time, and the overall recognition effect is the best, which is suitable for engineering implementation.

5. CONCLUSIONS

This paper describes a modulation recognition method based on machine learning, which combines instantaneous information features with high-order cumulant features, and realizes modulation recognition of signals by using classification algorithm in machine learning. On the basis of a single feature, this paper combines the advantages of the two feature types to improve the signal discrimination under low SNR. In addition, the three different machine learning classification algorithms used in this paper can analyze the relationship between parameters adaptively, and better solve the problem of recognition under low SNR. Experimental results show that the method in this paper is simple, computational complexity is small, recognition rate is high, recognition types are many, and can be better applied in engineering.

6. REFERENCES

[1] Nandi, A. K. , and E. E. Azzouz . "Algorithms for automatic modulation recognition of communication signals." IEEE Transactions on Communications 46.4(1998):431-436.

[2] Wong, M. L. D. , and A. K. Nandi . "Automatic digital modulation recognition using artificial neural network and genetic algorithm." Signal Processing 84.2(2004): 351-365.

[3] Hassan, K , et al. "Automatic Modulation Recognition Using Wavelet Transform and Neural Networks in Wireless Systems." EURASIP Journal on Advances in Signal Processing 2010.1(2010):532898.

[4] Wang, Hong . "Spectral correlation function of basic analog and digital modulated signals." (2013).

[5] Mobasseri, Bijan G. "Digital modulation classification using constellation shape." Elsevier North-Holland, Inc. 2000.

[6] Dobre, O. A. , et al. "Survey of automatic modulation classification techniques: classical approaches and new trends." IET Communications 1.2(2007):137-0.

[7] Li, Jiachen , L. Qi , and Y. Lin . "Research on modulation identification of digital signals based on deep learning." 2016 IEEE International Conference on Electronic Information and Communication Technology (ICEICT) IEEE, 2016..

[8] Hsu, Chih Wei , and C. J. Lin . "A comparison of methods for multiclass support vector machines." IEEE Transactions on Neural Networks 13.2(2002):415-425.

[9] Nandi, A. K. , and E. E. Azzouz . "Algorithms for automatic modulation recognition of communication signals." IEEE Transactions on Communications 46.4(1998):431-436.

[10] Li, Jiachen , L. Qi , and Y. Lin . "Research on modulation identification of digital signals based on deep learning." 2016 IEEE International Conference on Electronic Information and Communication Technology (ICEICT) IEEE, 2016.

[11] Park, Cheol Sun , et al. "Automatic Modulation Recognition of Digital Signals using Wavelet Features and SVM." Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on IEEE, 2008.

[12] O'Shea, Timothy J , J. Corgan , and T. C. Clancy . "Convolutional Radio Modulation Recognition Networks." International Conference on Engineering Applications of Neural Networks Springer, Cham, 2016.

[13] Hazza, Alharbi , et al. "An overview of feature-based methods for digital modulation classification." International Conference on Communications IEEE, 2013.

[14] Pattanayak, S , P. Venkateswaran , and R. Nandi . "Artificial Neural Networks for Cognitive Radio: A Preliminary Survey." International Conference on Wireless Communications IEEE, 2012.

[15] Peng, Xinjun , et al. "L 1 -norm loss based twin support vector machine for data recognition." Information Sciences An International Journal 340-341.C(2016):86-103.

A Comparative Analysis of Standard and Ensemble Classifiers on Intrusion Detection System

Joseph Mbugua
Kabarak University
Kenya

Moses Thiga
Kabarak University
Kenya

Joseph Siror
Kabarak University
Kenya

Abstract
With the increased dependence on the Internet, Network Intrusion Detection system (NIDs) becomes an indispensable part of information security system. NIDs aims at distinguishing the network traffic as either normal or abmormal. Due to the variety of network behaviors and the rapid development of attack strategies, it is necessary to build an intelligent and effective intrusion detection system with high detection rates and low false-alarm rates. One of the major developments in machine learning in the past decade is the ensemble method that generates a set of accurate and diverse classifiers that combine their outputs such that the resultant classifier outperforms all the single classifiers. In this work a comparative analysis on performance of three different ensemble methods, bagging, boosting and stacking is performed in order to determine the algorithm with high detection accuracy and low false positive rate. Three different experiments on NSL KDD data set are conducted and their performance evaluated based on accuracy, false alarms and computation time. The overall performance of the different types of classifiers used proved that ensemble machine learning classifiers outperformed the single classifiers with high detection accuracy and low false rates.

Keywords Ensemble classifiers, intrusion detection, standard classifiers, false alarms

i. INTRODUCTION

Classification [1] algorithms takes instances from a dataset and assigns a class or category to each of them based on supervised learning techniques. The technique can be applied in intrusion detection system to classify the network data as normal or attack. Several researchers have developed models to evaluated machine classifiers to categorise intrusion data set such as Support Vector Machines (SVM), Bayesian belief networking, Artificial Neural Network (ANN). Despite the variety and number of proposed models the construction of a perfect classifier for any given task remains challenging. The main challenge related to machine learning techniques is that no single-classification technique is capable of detecting all classes of attacks within acceptable false alarm rates and detection accuracies [2]–[4].

The ensembles learning models [5]combines multiple and homogeneous, weak classifiers to solve advanced and complex problems and improve the classification accuracy of the final results. These models apply the same algorithm repeatedly through partitioning and weighting of a training data set and improves classification performance by the combined use of two effects i.e. reduction of errors due to bias and variance [6]. Adaptive hybrid systems have become essential in computational intelligence and soft computing, the main reason being the high complementary of its components. The integration of the basic technologies into hybrid machine learning solutions facilitates more intelligent search and reasoning methods that match various domain knowledge with empirical data to solve advanced and complex problems. Implementation of ensemble and combination of multiple predictions are mainly motivated by three aspects which characterize the intrusion detection domain [7]–[9]: (i) relevant information may be present at multiple abstraction levels, (ii) the information may be collected from multiple sources, and (iii) this information needs to be represented at the human level of understanding.

The remainder of this paper is organized as follows. First, some background on ensemble classifiers is given. Then, we present the proposed methodology, experiments and results Finally, we draw conclusions and future work.

ii. ENSEMBLE CLASSIFICATION SYSTEM

The strategy in ensemble classification systems is to create a set of accurate and diverse classifiers in order to combine their outputs such that the combination outperforms all the single classifiers. A classifier is accurate when its classification error is lower than that obtained when the classes are randomly assigned. Two classifiers are diverse if they make errors at different instances. classifier ensembles are built in two phases: generation and combination.

Generation phase: In the generation phase, the individual components of the ensemble, known as base classifiers, are generated. The techniques used to generate diverse classifiers are based on the idea that the hypothesis of a classifier depends on both the learning algorithm and the subset used to generate these classifiers. Three different approaches can be used to generate an ensemble of classifiers by varying the training set. (i) Resampling the training examples by bagging and boosting to construct the classifier ensemble, (ii) Manipulating the input features achieves diversity between classifiers by modifying the set of attributes used to describe the instances, and (iii) Manipulating the output target to generate a pool of diverse classifiers with each classifier solving a different classification problem. The category that solves multiclass problems by converting them into several binary subproblems falls in class no iii. Methods that vary the learning algorithm can be subdivided in two groups i.e. (i) approaches that use different versions of the same learning algorithm (homogeneous ensembles) and (ii) approaches where diversity is obtained using different learning algorithms (heterogeneous classifiers).

Combination Phase: In the combination phase, the decisions made by the members of the ensemble are combined to obtain one decision. There are two main strategies for combining classifiers i.e. selection and fusion.

(i) Classifier selection presupposes that each classifier is an expert in some local region of the space. Therefore, when an instance is submitted for classification, the ensemble decision coincides with the decision given by the classifier responsible for the region of the space to which the instance belongs.

(ii) In classifier fusion, the decisions from all members of the ensemble are combined in some manner to make the ensemble decision. Classifier fusion algorithms include combining rules, such as the average, majority vote, weighted majority vote, and the Borda Count, and more complex integration models, such as meta-classifiers. A meta-classifier is a second-level classifier generated from the outputs given by the base learners.

Methodology

Anomaly detection using Standard classifier

Support Vector Machine, (SVM) is a machine learning algorithms used for classification, regression and outlier detection. It is one of the most accurate and robust algorithms for classification and widely used in IDS as they provide high security and take less time to detect attacks [5], [10]. The major features of SVM according to [7] include:Deals with very large data sets efficiently, Multiclass classification can be done with any number of class labels, High dimensional data in both sparse and dense formats are supported, Expensive computing not required and can be applied in many applications like e-commerce, text classification, bioinformatics, banking and other areas. Even though SVMs are limited to making binary classifications, their superior properties of fast training, scalability and generalization capability give them an advantage in the intrusion detection application. Finding cost-efficient ways to speed up or parallelize the multiple runs of SVMs to make multi-class identification is also under investigation.

Random Forest is an ensemble learning technique for classification and predictive modeling. It is also an approach to data exploration and generates many trees by using recursive partitioning then aggregate the results [11]. Each of the trees is constructed separately by using a bootstrap sample of the data and the bagging technique[12] is applied to combine all results from each of the trees in the forest. The method used to combine the results can be as simple as predicting the class obtained from the highest number of trees.

J48 [11] is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 is a program that creates a decision tree based on a set of labeled input data. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. J48 classifier algorithms [13] are used to compare and build, using the information entropy process, a decision tree from a set of training dataset. These algorithms adopt a top down technique and inductively built the decision tree for classification. It's extremely efficient when handling large datasets. [14]. The extra features of J48 [15] includes accounting for missing values, decision trees pruning, continuous attribute value ranges and derivation of rules.To make actual decisions regarding which path of the tree to replace is based on the error rates used. The reserved portion can be used as test data for the decision tree to overcome potential overfitting problem (reduced-error pruning).

Instance based learners (IBL) are computationally simple and represent knowledge in the form of specific cases or experiences [16] . IBL rely on efficient matching methods to retrieve stored cases so they can be applied in novel situations. Instance based learners are also called lazy learners because learning is delayed until classification time, as most of the power resides in the matching scheme. IB1 [10] is an implementation of the k nearest neighbour based classifier where k is the number of neighbors. IB1 finds the stored instance closest according to a Euclidean distance metric to the instance to be classified and the new instance is assigned to the retrieved instance’s class.

Bayesian reasoning provides a probabilistic approach for inference and is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data [17]. A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, bayesian networks have several advantages for data analysis [18]. First, the Bayesian networks encode the interdependencies between variables and hence they can handle situations where data are missing. Secondly, the Bayesian networks have the ability to represent causal relationships. Therefore, they can be used to predict the consequences of an action. Lastly, the Bayesian networks have both causal and probabilistic relationships; they can be used to model problems where there is a need to combine prior knowledge with data.

iii. ENSEMBLES CLASSIFIERS

Bootstrap is a meta learning algorithm that improves classification and regression models in terms of stability and classification accuracy. The algorithm takes bootstraps samples of objects and the classifiers are trained on each sample. The classifier votes are then combined by majority voting. A bootstrap sample is a statistical sample taken uniformly and with replacement, this means that the result sample set will contain duplicates [19]. Given a training dataset of size N, Bagging creates M base models, each trained on a bootstrap sample of size N created by drawing random samples with replacement from the original training set.

Boosting [20] is a machine learning meta-algorithm that built ensemble classifier by incrementally adding and iteratively learning weak classifiers with respect to a dataset to a final strong classifier [21]. Bagging is better than boosting as boosting suffers from over fitting as it performs well only for the training data. While both can significantly improve accuracy in comparison to a single model, boosting tends to achieve greater accuracy[22]. However unlike bagging, boosting may also reduce the bias of the learning algorithm [9]

Stacking [11] or Stacked Generalization is a different technique of combining multiple classifiers. Unlike bagging and boosting that use a voting system, stacking is used to combine different learning algorithms e.g. decision tree, neural network, rule induction, naïve Bayes, logistic regression, etc. to generate the ensemble of classifiers. The Stacking produces an ensemble of classifiers in which, the base classifiers (level-0) are built using different training parameters. The outputs of each of the models are collected to create a new dataset which is related to the real value that it is supposed to predict. Then, the stacking model learner as (level-1) use the output from base classifier to provide the final output.

One of the issues in Stacking is obtaining the appropriate combination of base-level classifiers and the meta-classifier, especially in relation to each specific dataset. If only a small number of classifiers and algorithms will be used, this problem can be solved by a simple method, namely, exhaustive search, in a reasonable amount of time. However, it is difficult to determine the best Stacking configuration when the search space is large.

iv. Experiments and Results

Experiment Environment - Waikato Environment for Knowledge Analysis

Several data mining techniques which includes data cleaning and pre-processing, clustering, classification, regression, visualization and feature selection have been implemented in WEKA (Waikato Environment for Knowledge Analysis) [23]. Weka also offers some functionality that other tools do not, such as the ability to run up to six classifiers on all datasets, handling multi-class datasets which other tools continue to struggle with tools.

Experiment Data set - NSL-KDD dataset

The NSL-KDD dataset (Jain & Rana, 2016; Parsaei et al., 2016; Shahadat et al., 2017) which is derived from original KDD-99 and has eliminated some of its drawbacks is analysed. The simulated attacks in the NSL-KDD dataset fall in one of the following four categories Denial of service attack (Dos), Probe attacks, Remote-to-Local (R2L) attacks, and User-to-Root (U2R) attacks.

Figure 1 Flow Diagram of Main Steps in the Research Study

Experiment Setup

In the experiment, we apply full dataset as training set and 10-fold cross validation for the testing purposes. The available dataset is randomly subdivided into 10 equal disjoint subsets and one of them is used as the test set and the remaining sets are used for building the classifier. In this process, the test subset is used to calculate the output accuracy while the N_1 subset is used as a test subset and to find the accuracy for each subset. The process is repeated until each subset is used as test set once and to compute the output accuracy of each subset. The final accuracy of the system is computed based on the accuracy of the entire 10 disjoint subsets.

All experiments are performed using Windows platform with the following configuration Intel Core-i5 processor, 2.5GHz speed, and 8GB RAM.

Experiment 1

The experiment is conducted with two ensemble learning techniques, bagging and boosting and five classifier using 10-fold cross validation. The single classifier includes Bayes Net, IBK, Jrip and SVM and the results are illustrated in TABLE 1. The conducted experiments is evaluated according to four performance measures which are accuracy, false positive and execution time.

Experiment II

In this experiment, the researcher compaire eight different algorithms and SVM as a base learners and stacking as a multi classifier learner are used. We use various combinations of BayesNet, iBK, ANN, J48 and JRip. The classifications predicted by the base learners will be used as input variables into a stacking model learner. Each input classifier computes predicted classifications using cross validation from which overall performance characteristics can be computed. Then the stacking model learner will attempt to learn from the data how to combine the predictions from the different models to achieve maximum classification accuracy. The stacking algorithm experiment results are given in the Table 2

Experiment III

Intrusion detection performance using combination of four distinct classifiers based on stacking with SVM as a meta classifier and the results are illustrated in Table 3**.**

The conducted experiments will be evaluated according to four performance measures which are defined below:

i. The Classification Accuracy: is the percentage number of correctly classified instances (the number of correct predictions from all predictions made)

ii. Precision: is a measure of classifier exactness (used as a measure for the search effectiveness)

iii. Recall: is a measure of classifier completeness.

iv. F-Measure: also called F-Score, it conveys the balance between the precision and the recall.

v. Experiments' Results and Data Analysis

Experiment 1

Table 1: The performance of bagging and boosting with five classifier using 10-fold cross validation

Algorithm	Accuracy	False Positive	Execution time

	Single	Bagging	Boosting	Single	Bagging	Boosting	Bag	Boost
Bayes Net	95.7%	95.5%	99.3%	4.3%	4.5%	0.670%	6.8	6.5
IBK	99.2%	99.1%	99.3%	0.80%	0.90%	0.7%	0.25	6.5
Jrip	99.5%	99.5%	99.5%	0.5%	0.5%	0.5%	395	390
J48	99.5%	99.5%	99.6%	0.5%	0.5%	0.4%	29.7	6.47
SVM	95.4%	90.53%	90.6%	4.6%	9.4%	9.47%	1656.8	1643.8

Overall, all the algorithms achieved good results, with the highest accuracy being 99.6% and the lowest being 89.59%. Tables 3 show that Adaboost when implement with J48 as a weak classifier achieves the highest accuracy, which is 99.6%, with a false positive (FP) rate of 0.30%. On the other hand, the BayesNet Bagging algorithm achieves the highest FP rate of 4.5%. Unfortunately the computation time of the three ensemble classifiers are all very high; the slowest one is stacking followed in turn by boosting and bagging.

Table 1 show that the use of the bagging and boosting algorithms did not improve the accuracy significantly. Only the use of boosting on the BayesNet algorithm were able to improve the accuracy, by 3.6% respectively, while the others showed a less than 1% improvement. While the two ensemble algorithms failed to improve the accuracy, they succeed in reducing the false positive rates. Bagging was able to reduce the false positive rate by up to 0.1% and 0.02% when implemented with IBK and BayesNet respectively, boosting was able to reduce the false positive rate by up to 3.7% and 0.02% when implemented for Naïve Bayes and J48.

Many researchers compared the performance of bagging and boosting methods including some large-scale experiments [5], [22], [28]–[30]The overall agreement is that boosting reaches lower testing error have been crowned as the most accurate available off-the-shelf classifiers on a wide variety of datasets . Nonetheless, it is observed that boosting methods are sensitive to noise and outliers, especially for small datasets [31]. Bagging is effective with noisy data whereas boosting is quite sensitive to noise. Another benefit of bagging methods is that they are parallel in nature in both the training and classification phases, whereas the boosting method is sequential in nature.

Experiment II

Table 2: The performance of SVM as a base learners and stacking as a multi classifier learner with eight classifier using 10-fold cross validation

Stacking meta classifier	TP	FP	Precision	F measure	Execution time (sec)
SVM	96.4	0.029	96.1	96.1	762.33
SVM With Bayesian	98.9	0.7	98.9	98.9	21.8
SVM With RF	99.8	0.1	99.8	99.8	340.63
SVM With J48	99.8	0.1	99.8	99.8	40.1
SVM With boost	90.53	9.47	90.53	90.53	1643.8
SVM With bagging	90.6	9.4	90.6	90.6	1656.8
SVM With ANN	93.6	6.4	93.5	93.7	1057.8
SVM With IBK	95.7	4.3	95.7	95.7	2147.1
SVM With Jrip	97.1	2.79	97.1	97.1	985
SVM With oneR	91.77	8.23	91.77	91.77	876

Bayesian is a highly scalable classifier and performs well for classifying rough dataset like medical data. For NSL-KDD which is a preprocessed data set, Bayes Net is providing approximately the same results for accuracy, precision and recall of 98.9%. While its false positive rate i.e. in correctively classified instances is 0.7%. The time to build the model is moderate at 21.8 seconds.

Bagging and boosting Are ensemble machine learning algorithms used in integration with another classifier to improve its prediction power but when integrated with SVM its performance degrades. This is because SVM is a strong classifier, and as described by Khorshid et al., (2015) integrating SVM with Bagging or Boosting does not improve its performance. Similar results are found in our evaluation result as accuracy of both Ensemble with SVM is 90.53% and 90.6% respectively while the accuracy of individual SVM is 91.81%.

The JRip and OneR are both based on association rule mining. The JRip is a fast and ripper algorithm and OneR creates one rule for each attribute and then picks up rule with least error [32], [33]. Hence combining SVM with JRip gives high accuracy and sensitivity values i.e. 97.21% and 92.33% while with OneR, it provides accuracy and sensitivity values as 91.77% and 79.14%.

ANN [21] is a strong classifier and it is also a weak learner and it requires large data set to train classifier. Because of this stacking, both these algorithms do not give very good performance. It gives better performance than SVM as its accuracy, specificity, precision and recall is more than SVM.

The K - Nearest Neighbour (IBK) is a lazy trainer. This means it stores the training instances and do real work only at the time of classification [26]. and hence IBK gives strongly consistent results. However, equal weightage is given to each of the attributes. As a result, combining this algorithm with SVM gives moderately high rate of accuracy of. 95.79%. but slow in execution speed of 2147.1 sec.

The decision tree algorithms J48 and Random Forest provides high accuracy rate of 99.8 and a false positive rate of 0.1% and execution time of 40.1 seconds and 340.63 seconds respectively. Random Forest is giving the best performance in all evaluating parameters. NSL-kDD data set does not contain redundant records and it is easy for these classifiers to build their decision tree output and as a result combining them with SVM improves the overall performance of intrusion detection system.

Experiment III

Table 3. Intrusion detection performance using combination of four distinct classifiers based on stacking

Staking	TP	FP	PREC	RECAL	FMEASURE	ROC	TIME
	99.8	0.1	99.8	99.8	99.8	99.9	4757.87

CONCLUSION AND FUTURE WORK

Overall the application of bagging and boosting did not significantly improve the accuracy or reduce error rates. Only Stacking method was able to reduce the false positive rate by a moderately high amount. The key assumption that the errors due to the individual models are uncorrelated is unrealistic; in practice, the errors are typically highly correlated, so the reduction in overall error is generally small [28], [34]. Staking method took the longest execution time which is a drawback in its application in the intrusion detection field.

Although a lot of research on AI-based ensembles has been conducted, several research questions still remain unanswered, for example, how many base classifiers should be combined, how base classifiers should be combined, how to generate diverse set of base classifiers, how instances of training dataset should be partitioned to generate base classifiers, how feature space should be partitioned and in particular for ID quality training dataset among others.

REFERENCES

[1] S. Thaseen and C. A. Kumar, "Intrusion Detection Model using PCA and Ensemble of Classifiers," vol. 16, no. 2, pp. 15–38, 2016.

[2] M. Albayati and B. Issac, “Analysis of Intelligent Classifiers and Enhancing the Detection Accuracy for Intrusion Detection System,” *Int. J. Comput. Intell. Syst.*, vol. 8, no. 5, pp. 841–853, 2015.

[3] C. F. Tsai, Y. F. Hsu, C. Y. Lin, and W. Y. Lin, “Intrusion detection by machine learning: A review,” *Expert Syst. Appl.*, vol. 36, no. 10, pp. 11994–12000, 2009.

[4] M. Panda, A. Abraham, and M. R. Patra, “A hybrid intelligent approach for network intrusion detection,” *Procedia Eng.*, vol. 30, no. 2011, pp. 1–9, 2012.

[5] M. M. H. Khorshid, T. H. M. Abou-el-enien, and G. M. A. Soliman, “A Comparison among Support Vector Machine and other Machine Learning Classification Algorithms A Comparison among Support Vector Machine and other Machine Learning Classification,” no. August, 2015.

[6] M. Govindarajan, “Hybrid Intrusion Detection Using Ensemble of Classification Methods,” *Int. J. Comput. Netw. Inf. Secur.*, vol. 6, no. 2, pp. 45–53, 2014.

[7] S. Thaseen and A. Kumar, “Intrusion detection model using fusion of chi-square feature selection and multi class SVM,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 29, no. 4, pp. 462–472, 2017.

[8] I. Czarnowski and J. Piotr, “An Approach to Machine Classification Based on Stacked Generalization and Instance Selection,” 2016.

[9] G. Kumar and K. Kumar, “The Use of Multi-Objective Genetic Algorithm Based Approach to Create Ensemble of ANN for Intrusion Detection,” *Int. J. Intell. Sci.*, vol. 2, no. October, pp. 115–127, 2012.

[10] N. Chand, P. Mishra, C. R. Krishna, E. S. Pilli, and M. C. Govil, “A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection A Comparative Analysis of SVM and its Stacking with other Classification Algorithm for Intrusion Detection,” no. November, 2017.

[11] R. Pradhan, “Performance Assessment of Robust Ensemble Model for Intrusion Detection using Decision Tree Techniques,” vol. 3, no. 3, pp. 78–86, 2014.

[12] S. L. Pundir and Amrita, “Feature Selection Using Random Forest in Intrusion Detection,” *Int. J. Adv. Eng. Technol.*, vol. 6, no. 3, pp. 1319–1324, 2013.

[13] J. H. Assi and A. T. Sadiq, “NSL-KDD dataset Classification Using Five Classification Methods and Three Feature Selection Strategies,” vol. 7, no. 1, pp. 15–28, 2017.

[14] M. K. Gambo and A. Yasin, “Hybrid Approach for Intrusion Detection Model Using Combination of K-Means Clustering Algorithm and Random Forest Classification,” *Ijes*, vol. 6, no. 1, pp. 93–97, 2017.

[15] Dubb Shruti and Sood Yamini, “Feature Selection Approach for Intrusion Detection System,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 2, no. 5, pp. 47–53, 2013.

[16] S. V. Lakshmi, T. E. Prabakaran, and D. Ph, “Application of k-Nearest Neighbour Classification Method for Intrusion Detection in Network Data,” vol. 97, no. 7, pp. 34–37, 2014.

[17] A. A. Ramaki, M. Khosravi-Farmad, and A. G. Bafghi, “Real time alert correlation and prediction using Bayesian networks,” *2015 12th Int. Iran. Soc. Cryptol. Conf. Inf. Secur. Cryptol.*, vol. 978, pp. 98–103, 2015.

[18] R. Kaur and M. Sachdeva, “International Journal of Advanced Research in An Empirical Analysis of Classification Approaches for Feature Selection in Intrusion Detection,” vol. 6, no. 9, 2016.

[19] S. Kovac, “Suitability analysis of data mining tools and methods,” p. 53, 2012.

[20] S. S. Dongre and K. K. Wankhade, “Section V with concluding conclusion in Section VI. II. data mining for IDS,” vol. 2, no. 4, pp. 488–492, 2012.

[21] M. Amini, “Effective Intrusion Detection with a Neural Network Ensemble Using Fuzzy Clustering and Stacking Combination Method,” vol. 1, no. 4, pp. 293–305, 2014.

[22] I. Journal, F. Technological, A. Patel, and R. Tiwari, “Bagging Ensemble Technique for Intrusion Detection,” vol. 2, no. 4, pp. 256–259, 2014.

[23] M. Revathi, “Network Intrusion Detection System Using Reduced,” *J. Comput. Sci.*, vol. 2, no. 1, pp. 61–67, 2000.

[24] N. Shahadat, I. Hossain, A. Rohman, and N. Matin, “Experimental Analysis of Data Mining Application for Intrusion Detection with Feature reduction,” pp. 209–216, 2017.

[25] A. Thesis, “Using Support Vector Machines in Anomaly Intrusion Detection by,” 2015.

[26] A. Jain and J. L. Rana, “Classifier Selection Models for Intrusion Detection System (Ids),” *Informatics Eng. an Int. J.*, vol. 4, no. 1, pp. 1–11, 2016.

[27] M. R. Parsaei, S. M. Rostami, and R. Javidan, “A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset,” vol. 7, no. 6, pp. 20–25, 2016.

[28] M. P. Sesmero, A. I. Ledezma, and A. Sanchis, “Generating ensembles of heterogeneous classifiers using,” vol. 5, no. February, pp. 21–34, 2015.

[29] A. Shrivastava, M. Baghel, and H. Gupta, “A Review of Intrusion Detection Technique by Soft Computing and Data Mining Approach,” no. 3, pp. 224–228, 2013.

[30] C. Science, “A Hybrid Approach to improve the Anomaly Detection Rate Using Data Mining Techniques,” no. July, 2015.

[31] Y. Wahba, E. ElSalamouny, and G. ElTaweel, “Improving the Performance of Multi-class Intrusion Detection Systems using Feature Reduction,” *Ijcsi*, vol. 12, no. 3, pp. 255–262, 2015.

[32] J. Hussain and S. Lalmuanawma, “Feature Analysis, Evaluation and Comparisons of Classification Algorithms Based on Noisy Intrusion Dataset,” *Procedia Comput. Sci.*, v

[33] J. Song, “Feature Selection for Intrusion Detection System Jingping Song Declaration and Statement,” p. 132, 2016.

[34] M. Govindarajan, “Evaluation of Ensemble Classifiers for Intrusion Detection,” vol. 10, no. 6, pp. 1045–1053, 2016.

An Approach of Data Processing and Sales Prediction Model for E-commerce Platform

Guangpu Chen
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Zhan Wen
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Yahui Chen
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Yuwen Pan
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Xia Zu
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Wenzao Li
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan
School of Computing Science,
Simon Fraser University,
Burnaby BC, V5A 1S6, Canada

Abstract: For the e-commerce platform, obtaining sales status of the stores is important for formulating sale strategies, risk assessment plans and loan appraisal. The traditional way to obtain sales status is mainly based on the subjective judgment of relevant practitioners and the analysis of mathematical statistical models composed of historical data. These methods are inaccurate and too dependent on people’s judgment. Therefore, using data mining and machine learning technology to predict the sales amount came into being. In this paper, we propose a method to process a great deal of data from China's famous e-commerce platform called Jingdong. This method can make the messy data become uniform data sets which are more suitable for machine learning. Based on the uniform data sets, two sales prediction models are used to predict the sales amount of the stores in Jingdong. In experiment, 9-month historical sales and behavior data of 10,000 stores in Jingdong platform are processed by the proposed method. Furthermore, two prediction models including GBDT(Gradient Boosting Decision Tree)+DNN(Deep neural network) and GA(genetic algorithm) are used to predict the sales amount of stores in following 3 months. To verify the accuracy of the prediction, we import WMAE（Weighted Mean Average Error）score. In experimental results, the best WMAE is 0.39, which means accuracy is 61%. It shows the method of data processing and prediction models are effective compare with other models. This indicates the proposed method and model can be used for sales prediction in e-commerce platform.

Keywords: e-commerce; sales prediction; big data; data mining; machine learning

1. INTRODUCTION

Recent years, the rise of e-commerce platforms in China has led to online shopping becoming part of people's daily lives. The online shopping economy has become an inseparable part of China's economic development. Forecasting and analyzing plays an important role when the platforms try to make the right business decisions. Along with the explosion data, the traditional statistical analysis predict method has been unable to adapt to the ever-changing market. The sales forecast is also more favored for emerging Technologies such as big data and machine learning.

At present, the forecast of sales mainly has the following major categories, one is based on user research, the other is based on commodity research, and the third is hybrid research. These three types of research methods are based on the research of traditional store sales models. In the face of the rise of e-commerce platforms in the current environment, traditional forecasting methods are gradually unable to apply. There is still a lot of research space in the sales forecasting research of e-commerce platform in the new era of information.

Currently,Jingdong Financial Supply Chain has more than 100,000 companies and provided 250 billion yuan in loans. The majority of these enterprises are small, medium and small enterprises.Supply chain finance is a financial activity based on the industrial supply chain. Its purpose is to support financial activities through the industrial supply chain.In the past few years, based on the accumulation of Jingdong Big Data, Jingdong Finance has successively launched three core products of Jing Baobei, Jing Small Loan and Movable Finance, which greatly improved the financing difficulties and high financing costs of small and micro enterprises.In order to achieve accurate subsidies for each store, regular measurement and tracking of the operation status of each store has become an important evaluation standard for loans.Only accurate estimates of the future sales of the store can accurately assess their funding needs and set a reasonable loan amount. This paper will establish a forecasting model through the past sales records, product information, product evaluation, advertising costs and other information of each store on the platform, and predict the sales of each store in the next three months.

This study is based on the sales records, product information, product evaluation, and advertising costs of 10,000 stores on the Jingdong platform for the past 9 months. In order to predict sale amount in the following three months, we need to dig deeper into these data and build a machine learning prediction model. In the research, firstly, we will analyze and optimize the original data to make it more relevant to the model. Secondly, we will fit and predict the data through a machine learning model. Finally, we selected the best model' parameters that can accurately predict future sales by comparing the weighted mean absolute error (WMAE) of the predicted results.

This paper mainly has two contributions: First, we propose a method of data processing for e-commerce platform. This method can make these data become uniform data sets and it's more suitable for machine learning. Second, based on the processed data sets, a sale prediction model is used to predict the sale amount of the stores in e-commerce platform. It can provide a reference for e-commerce platform to provide loan for stores.

The rest of the paper is organized as follows: Sect.2 Introduces research on topics similar to this paper and detail the theoretical knowledge of the research methods. Section 3 presents the detailed process of the experiment. Sect.4 presents the analysis and conclusion of the result of the experiment. The five section presents the shortcomings of the experiment and propose some suggestions for the improvement.

2. RELATED WORKS

2.1 Research Status of Sales Forecast of E-commerce Platform

At present, the forecast of sales mainly has the following major categories, one is based on user research, the other is based on commodity research, and the third is hybrid research. These three types of research methods are based on the research of traditional store sales models. In the face of the rise of e-commerce platforms in the Internet environment, traditional forecasting methods are gradually unsuitable[1]. There is still a lot of research space in the sales forecasting research of e-commerce platform in the new era of information.

2.2 Feature Engineering

Feature engineering is an engineering activity that maximizes the extraction of features from raw data for use by algorithms and models[2]. Feature processing is the core part of feature engineering, including data preprocessing, feature selection, dimensionality reduction, and so on.In this project, we will use the sklearn library in Python to implement a series of operations on feature engineering.

2.3 Principal Component Analysis(PCA)

Principal component analysis(PCA) is a method for dimension reduction of features. It replaces the original features with a concise number of new features[3]. These new features are linear combinations of original features. It combinations maximize sample variance and try to make the new features uncorrelated. This makes it easy to study and analyze the influence of each feature on the model, and effectively reduce the complexity of models and increase the speed of the training model.

2.4 Gradient Boosting Decision Tree

GBDT(Gradient Boosting Decision Tree) is an iterative decision tree algorithm consisting of multiple decision trees[4]. All trees vote for the final result.It is recognized as one of algorithms with strong generalization.The idea of GBDT can be explained by a simple example. If a person is 30 years old, we first use 20 years old to fit and find that the loss is 10 years old. Then, we use 6 years old to fit the remaining losses and find the gap is 4 years old. In the third round ,we used 3 years old to fit the remaining gap and find the gap is only one year old. If the number of iterations is not yet complete, we can continue to iterate below. For each iteration, the age error of the fit will decrease.The specific steps of the algorithm are as follows.

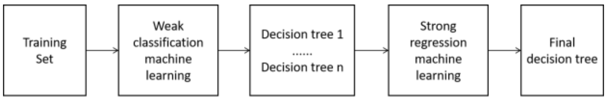

Training Set
Weak classification machine learning
Decision tree 1
......
Decision tree n
Strong regression machine learning
Final decision tree

Fig. 1:Process of Gradient Boosting Decision Tree

In this project, on the one hand, the algorithm is used to calculate the ranking of feature importance when crossing features, and on the other hand, the algorithm is used to transform features when building regression prediction models.

2.5 Deep Neural Network

DNN(Deep neural network) is a multi-layered neural network, also known as Multi-Layer perceptron(MLP).DNN is similar to the hierarchical structure of traditional neural networks. The system consists of a multi-layer network consisting of an input layer, a hidden layer (multilayer), and an output layer[5]. Only adjacent nodes have connections, the same layer, and cross-layer nodes. There is no connection between each other, and each layer can be regarded as a logistic regression model.Different from traditional neural networks, DNN adopts a forward-propagating training method.This is more suitable for processing high dimensional data.

In this project, we use DNN to process features transformed by GBDT and build prediction model.This can give full play to its advantages.

2.6 Genetic Algorithm

Genetic algorithm is a computational model that simulates the natural evolution of Darwin's biological evolution theory and the biological evolution process of genetic mechanism[6]. It is a method to search for optimal solutions by simulating natural evolutionary processes.In machine learning applications, the algorithm automatically selects the most appropriate model and parameters based on the input training data and tags, just like evolution.The steps of the algorithm are as follows.

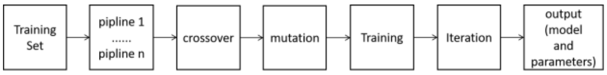


Fig. 2:Process of Genetic Algorithm

In this project, we use genetic algorithms to automatically select the algorithm that best fits this training set.We want to prove that model training can be finished and get good scores even without any machine learning foundation.We hope that we can promote the algorithm to make more effort on data analysis processing.

3. RESEARCH AND DESIGN OF SALES FORECAST MODEL FOR E-COMMERCE PLATFORM

3.1 Introduction to The Experimental Process

Based on the general process of machine learning training data, the experimental process of this paper is as follows.

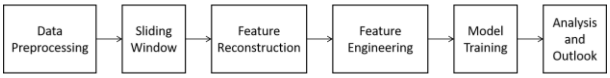


Fig. 3:Experimental Process

1)Data preprocessing:Perform preliminary inspection of origin data to handle missing, error, and abnormal data.

2)Sliding window:Divide origin data to fit model training and expand training set size to increase training coverage.

3)Feature reconstruction:Due to a mismatch between the original data and the predicted target,rebuild the training set based on the predicted target.

4)Feature engineering:Further optimize the training set generated by feature reconstruction and deeply mine more features that are beneficial to the training model.

5)Model training:The training set generated by the first few steps is input into a preset machine learning model, and the weighted mean absolute error (WMAE) of the output result is used as a measurement standard for selecting the optimal parameter.

6)Analysis and outlook:Analyze the results obtained from model training and draw conclusions.

3.2 Introduction of The Origin Data

This project uses the order quantity, sales, number of customers, evaluation number, advertising cost and other data of several stores within 270 days before 2017-04-30 provided by Jingdong Finance as training data, and sales of 90 days after the end of each month as labels.This project will use this data to build a predictive model to predict the total sales for the platform within 90 days after 2017-04-30. The original data structure used in this project is shown in Table 1 to Table 5.

Table 1. Orders data

Name	Type	Description	Example
shop_id	int	Store id	1631
Figures	Good	Similar	Very well
pid	int	Product id	41
ord_dt	date	Order date	2016/9/4
ord_cnt	int	Order count	1
sale_amt	float	Sales Amount	19.82
user_cnt	int	User count	1
rtn_amt	float	Return amount	0
rtn_cnt	int	Return count	0
offer_amt	float	Discount Amount	0
offer_cnt	int	Discount count	0

Table 2. Products data

Name	Type	Description	Example
shop_id	int	Store id	1014
pid	int	Product id	8141588
brand	int	Brand id	785
cate	int	Cate id	243
on_dt	date	put on shelves date	2017/2/9
off_dt	date	pull off shelves date	2017/5/1

Table 3. Comments data

Name	Type	Description	Example
shop_id	int	Store id	1494
good_num	int	Positive feedback	2
mid_num	int	Neutral feedback	0
bad_num	int	Negative feedback	0
dis_num	int	Number of share order	0
cmmt_num	int	Number of comments	2
create_dt	date	Create time	2016/8/7

Table 4. Advertisements data

Name	Type	Description	Example
shop_id	int	Store id	1036
charge	float	Advertising recharge	65298.6
consume	float	Advertising spending	25096.86
create_dt	date	Create time	2016/9/30

Table 5. Total sales data

Name	Type	Description	Example
shop_id	int	Store id	2143
sale_amt_3m	float	Sales within 90 days after the end of the month	72983.09
dt	date	Last day of the month	2016/12/31

As those tables shows,the data in Tables 1 to 4 will be used as training data,Table 5 is the target data.Among them, about 400W data in Table 1 account for 2G.,Table 2 has about 14W data, accounting for 0.8G,Table 3 is about 80W data, accounting for 0.5G,Table 4 is about 30W data, accounting for 0.3G.

3.3 Data Preprocessing

Data processing is mainly divided into three parts.The first is the processing of abnormal data.There are some obvious erroneous data in the original data used in this project,for example, the date on which the item is placed is later than the date of the removal, the number of return orders is more than the number of orders, etc.Since the overall proportion of such data is not high, this time it will be deleted directly.Secondly,there are still some missing data in the original data,mainly in Table 1, so we used the average of the last 6 days of missing data to fill in missing values.Finally, the purpose of this modeling is to predict sales for the three months after April 30, 2017, but the training data includes special festivals such as November 11, December 12(China Online Shopping Festival),for example, we plot the sales of the stores numbered 1630 in October and November as follows. Comparing the sales of these two months, it can be seen that there will be a significant abnormality in the sales of special event days (November 11).

To deal with this situation, the project adopted a smoothing treatment for these special sales days, using the average of the two days before and after the day instead of the day.

3.4 Sliding Window

The data used in this experiment belongs to time series data and contains a lot of time information.A time series is a sequence in which the values of the same statistical indicator are arranged in chronological order in which they occur.By processing such information through sliding window division, on the one hand, the training data set can be expanded, and on the other hand, the influence of time on the prediction effect of the model can be weakened as much as possible.The specific division method is as follows.

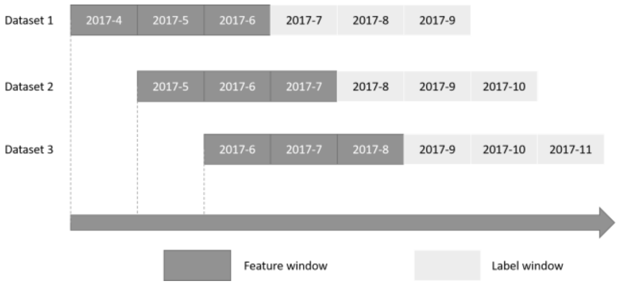


Fig. 4:Schematic Diagram of Sliding Window Division

3.5 Feature Reconstruction

As can be seen from the above tables, the original data of this experiment is a one-to-many situation, and the goal of our model is the sales of each store in the next 3 months. These multi-dimensional features cannot be used as input features of the prediction model. Therefore, it is necessary to perform a dimension reduction and reconstruction process on the original data through statistics, sampling, crossover, etc., so that the prediction model can be successfully constructed.

Financial data often accompany with multiple collinearity and financial leverage effects, discount, income, and cash flow are related to each other and have many combined features. Therefore, during the feature reconstruction, not only need comprehensive characteristics, but also need some sampling and crossing characteristics .The process of feature reconstruction is as follows.

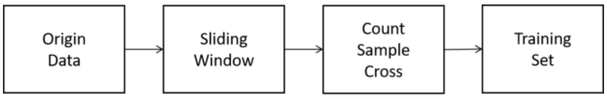


Fig. 5:Process of Feature Reconstruction

3.6 Feature Engineering

After feature reconstruction,we initially formed the training set required for the prediction model, and features that can use to build the model are 80 dimensions. But as in the field of data mining, the data and features determine the upper limit of machine learning, and the models and algorithms just approximate this upper limit.Therefore,we still need to mine more and better features,so the Feature engineering is proposed.

3.6.1 Data Cleaning

The first thing we need to do is to clean up the training set.The missing rate of training set is as follows.

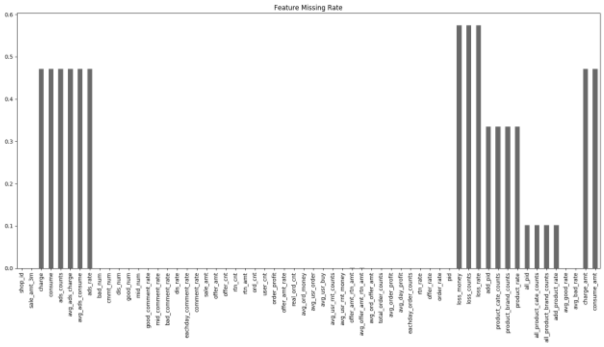


Fig. 6:Missing Rate of Training Set

As can be seen from the figure, about 20% of the features have missing data, including advertising features, order features, and product features.

For advertising features,the reasons for the missing are mainly the preference of the merchant and the size of the sliding window is smaller than the advertising investment period.In order to weaken the impact of this situation on the model, we set missing data as a special attribute value, which is different from any other attribute value and is filled with "-999" in the actual programming process.

Then about the order and product features,the main reason for the lack of data is due to the solution adopted by feature reconstruction,like there is no loss of orders or put on new goods in the sliding window period.Because the lack of these data is due to the absence of this situation,we fill all missing data with "0".This will make it work in the model.

3.6.2 Feature Crossing

Feature crossing is a mathematical combination of two or more category attributes into one feature.This project takes the Grid Search method on the feature crossing,make the reconstructed 60 features with no missing values are multiplied by each other to obtain a total of 870 features.Then, put the 870 features into the GBDT algorithm model as a training set for sales forecasting model.The parameters of GBDT are as follows.

After the training, use the method of ”get feature importance” to get the ranking of the importance of those features, and the top 100 are as follows.

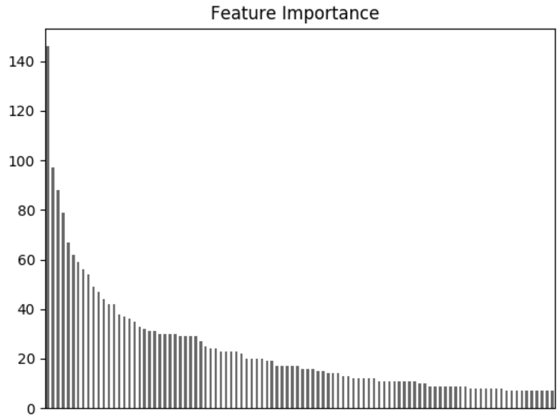


Fig. 7:Feature Importance of Cross-Features

3.6.3 Feature Selection

After the above processing, we have a total of about 200 features. Although it enhances the persuasiveness and prediction accuracy of the regression prediction model, it also increases the time for model training, and too many features make the model too complex and generalized, which is easy to cause dimensionality disaster. Therefore, in this project, we use PCA（principal component analysis）to select the most effective features from the original features to reduce the dimension of data set. The parameters of PCA are as follows.

After numerous experiments, the project got the highest score when reserve 90% features.So far, we have obtained a total of 161 features for model training.

3.7 Model Training

3.7.1 GBDT+DNN

The first model is constructed by GBDT combine DNN,the specific combination is as follows.

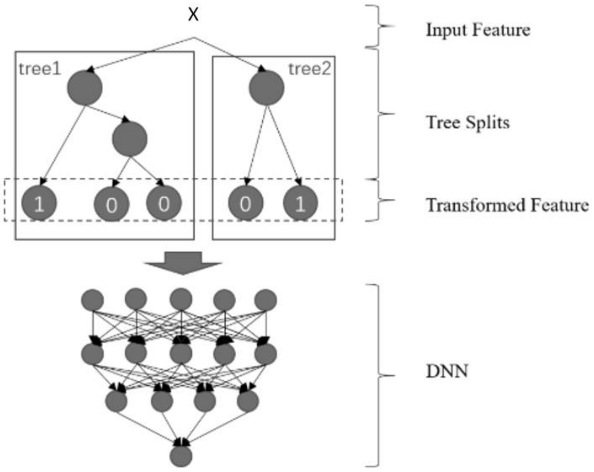


Fig. 8:Combination of GBDT and DNN

The purpose of this is to firstly exploit the advantages of GBDT algorithm mining feature combination. Second, DNN deep learning is more suitable for processing massive data and multi-dimensional input features. Through the combination of the two, the advantages of the respective algorithms can be fully utilized.The parameters are selected as follows.

3.7.2 Genetic Algorithm

The second model is built in two steps. First, the training data is input into the genetic algorithm to automatically select and iterate out the best regression prediction model suitable for the project. Secondly, the training set is input to the model of the first step output for training.The specific process is as follows.

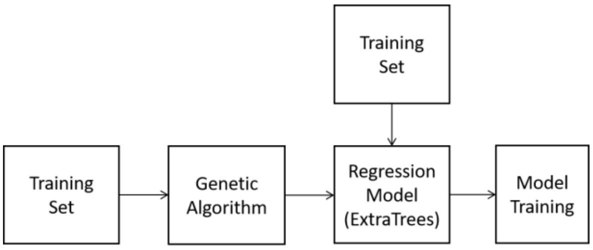


Fig. 9:Process of Genetic Algorithm

The purpose of this is to try genetic algorithm, a method of streamlined machine learning.This algorithm does not require an in-depth understanding of complex algorithm principles, and can automatically select the best predictive model based on input data.It will provide new ideas for establishing machine learning prediction models.The parameters of Genetic algorithm and the model output by it as follows.

4. ANALYSIS AND CONCLUSION

4.1 Result of Prediction

After a series of processing in Chapter 3, we generated the optimal training set used by the training model and selected the appropriate training model based on the characteristics of the training data.Through experiments,we found that feature engineering has greatly improved the model prediction results compared with different algorithms choices.We compared the scores and feature importance of the training sets that with or without feature engineering under each algorithm as follows.Also we used K-fold cross-validation in the experiment,and took WMAE as the evaluation criteria.The predicted scores for training sets used under different algorithms and with or without feature engineering are shown as follows.

Table 6. Scores of the training sets that with or without feature engineering under each algorithm

Models	WMAE (without feature engineering)	WMAE (with feature engineering)
LR	0.7853	0.7125
RF	0.5132	0.4589
SVR	0.4982	0.4464
GBDT	0.5044	0.4478
GBDT+DNN	0.4522	0.3982
GA	0.4500	0.3901

As can be seen from the table, firstly, the result has greatly improved under each algorithm when the training set is processed by feature engineering. Secondly, the two sets of algorithms finally adopted in this experiment have a certain improvement on the project compared with other regression prediction algorithms. These can prove the quality of the prediction.

4.2 Conclusion

As can be seen from the above table, both feature engineering and algorithm model selection have a certain improvement on regression prediction results.This proves that the project can effectively process data and predict future sales.After summarizing, this paper mainly has the following contributions:

1)Adding analysis of financial data in the process of data preprocessing and feature reconstruction.

2)Further mining of data through feature engineering and get a great improvement in scores.

3)After deep mining the characteristics of financial data, the method of inputting DNN after GBDT extended feature number was used and some improvements were successfully achieved.

4)Automatically select regression prediction models by genetic algorithm and obtain good scores, which provides new ideas for model selection.

5. PROSPECT

The desensitization data used in this study was provided by JD,therefore, we are unable to know the specific products sold in each store.I think it is very important for mining potential information.Second, we are unable to quantify the direct or indirect impact of unscheduled promotions and government policies on the platform, so these factors are not included in the training model, it may be benefit to the model.In addition, in terms of algorithms, the prediction effect can be further improved by means of model fusion.It will take a lot of time and resources to find better algorithms or fusion solutions.These outlooks are all issues that need to be considered in the future to explore such topics.

6. ACKONWLEDGEMENTS

The data used in this research is from JingDong Finance, and we are very grateful to the agency for supporting this experiment.

This research is supported by Science and Technology Department of Sichuan Province, Fund of Science and Technology Planning (No. 2018JY0290), Science and Technology Project of Sichuan Province(No. 2019YFS0493), Humanity and Social Science Research Project of Ministry of Education, Youth Foundation(17YJC190035) and Meteorological Information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes (No. QXXCSYS201606)

7. REFERENCES

[1] Hui-Chih Hung, Yu-Chih Chiu, Huang-Chen Huang, et al. An enhanced application of Lotka–Volterra model to forecast the sales of two competing retail formats[J]. Computers & Industrial Engineering, 2017, 109:325-334.

[2] Goodness C. Aye, Mehmet Balcilar, Rangan Gupta, et al. Forecasting aggregate retail sales: The case of South Africa[J]. International Journal of Production Economics, 2015, 160:66-79.

[3] Andrés Martínez, Claudia Schmuck, Sergiy Pereverzyev, et al. A machine learning framework for customer purchase prediction in the non-contractual setting[J]. European Journal of Operational Research, 2018.

[4] Giuseppe Piras, Fabrizio Pini, Davide Astiaso Garcia. Correlations of PM10 concentrations in urban areas with vehicle fleet development, rain precipitation and diesel fuel sales[J]. Atmospheric Pollution Research, 2019.

[5] Patrícia Ramos, Nicolau Santos, Rui Rebelo. Performance of state space and ARIMA models for consumer retail sales forecasting[J]. Robotics and Computer-Integrated Manufacturing, 2015, 34:151-163.

[6] Bin Zhang, Dongxia Duan, Yurui Ma. Multi-product expedited ordering with demand forecast updates[J]. International Journal of Production Economics, 2018, 206:196-208.

[7] Maobin Li, Shouwen Ji, Gang Liu, et al. Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA, Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model[J]. Mathematical Problems in Engineering, 2018, 2018.

[8] Eric W. K. See-To, Eric W. T. Ngai. Customer reviews for demand distribution and sales nowcasting: a big data approach[J]. Annals of Operations Research, 2018, 270(1-2):415-431.

[9] A.L.D. Loureiro, V.L. Miguéis, Lucas F.M. da Silva. Exploring the use of deep neural networks for sales forecasting in fashion retail[J]. Decision Support Systems, 2018.

[10] G. Dellino, T. Laudadio, R. Mari, et al. Microforecasting methods for fresh food supply chain management: A computational study[J]. Mathematics and Computers in Simulation, 2018, 147:100-120.

International Journal of Computer Applications Technology and Research
Volume 8–Issue 04, 122-134, 2019, ISSN:-2319–8656

Designing Framework for Data Warehousing of Patient Clinical Records using Data Visualization Technique of Nigeria Medical Records

Dr. Oye, N. D.
Department of Computer Science,
Modibbo Adama University of Technology,
Yola, Adamawa state, Nigeria

Emeje, G.D
Department of Computer Science
Modibbo Adama University of Technology,
Yola, Adamawa state, Nigeria

Abstract:

The availability of timely and accurate data is vital to make informed medical decisions. Today patients data required to make informed medical decisions are trapped within fragmented and disparate clinical and administrative systems that are not properly integrated or fully utilized. Therefore, there is a growing need in the healthcare sector to store and organize size-able clinical record of patients to assist the healthcare professionals in decision making processes. This research is about integrating and organizing disparate clinical record of patients into a Data Warehouse (DW) for data analysis and mining, which will enable evidence-based decision-making processes. This study uses SQL Server Integration Service (SSIS) for the extraction transformation and loading (ETL) of patient clinical records from fragmented administrative systems into the DW, Data visualization technique was used for data presentation while SQL Server Reporting Services and Business Intelligent (BI) tools for designing the output results. This research will assist medical experts and decision makers in the healthcare industry in planning for the future. This research also provides an architecture for designing a clinical DW that is not limited to single disease and will operate as a distributed system, Periodic update on a daily basis is possible because contemporary technologies have narrowed the gap between updates which enable organizations to have a "real time" DW which can be analyzed using an OLAP Server. In conclusion if this research is deployed it will aid medical decision-making process in the Nigerian medical sector.

Keywords: Clinical records; Data marts; Data Warehousing; Framework; Patient record

1. INTRODUCTION

1.1 Background of the study

Knowledgeable decision making in healthcare is vital to provide timely, precise, and appropriate advice to the right patient, to reduce the cost of healthcare and to improve the overall quality of healthcare services. Since medical decisions are very complex, making choices about medical decision-making processes, procedures and treatments can be overwhelming. (Demetriades, Kolodner, & Christopherson, 2005). One of the major challenges of Information Technology (IT) in Healthcare services is how to integrate several disparate, standalone clinical information repositories into a single logical repository to create a distinct version of fact for all users (Mann, 2005; Zheng et al., 2008; Goldstein et al., 2007; Shepherd, 2007).

A massive amount of health records, related documents and medical images generated by clinical diagnostic equipment are created daily (Zheng, Jin, Zhang, Liu, & Chu, 2008). Medical records are owned by different hospitals, departments, doctors, technicians, nurses, and patients. These valuable data are stored in various medical information systems such as HIS (Hospital Information System), RIS (Radiology Information System), PACS (Picture Archiving and Communications System) in various hospitals, departments and laboratories being primary locations (Zheng, Jin, Zhang, Liu, & Chu, 2008). These medical information systems are distributed and heterogeneous (utilizing various software and hardware platforms including several configurations). Such processes and data flows have been reported by Zheng, Jin, Zhang, Liu, & Chu (2008).All medical records are located in different hospitals or different departments of single hospital. Every unit may use different hardware platforms, different operating systems, different information management systems, or different network protocols. Medical data is also in various formats. There are not only a tremendous volume of imaging files (unstructured data), but also many medical information such as medical records, diagnosis reports and cases with different definitions and structures in information system (structured data), Zheng et al., (2008).

This causes Clinical Data Stores (CDS) with isolated information across various hospitals, departments, laboratories and related administrative processes, which are time consuming and demanding reliable integration (Sahama, & Croll, 2007). Data required to make informed medical decisions are trapped within fragmented, disparate, and heterogeneous clinical and administrative systems that are not properly integrated or fully utilized. Ultimately, healthcare begins to suffer because medical practitioners and healthcare providers are unable to access and use this information to perform activities such as diagnostics, prognostics and treatment optimization to improve patient care (Saliya, 2013).

1.2 Problem Statement

The availability of timely and accurate data is vital to make informed medical decisions. Every type of healthcare organization faces a common problem with the considerable amount of data they have in several systems. Such systems are unstructured and unorganized, demanding computational time for data and information integration (Saliya, 2013). Today

Patient's data required to make informed medical decisions are trapped within fragmented and disparate clinical and administrative systems that are not properly integrated or fully utilized. The process of synthesizing information from these multiple heterogeneous data sources is extremely difficult, time consuming and in some cases impossible. Due to the fast growing data in the healthcare sector there is need for health industries to be open towards adoption of extensive healthcare decision support systems, (Abubakar, Ahmed, Saifullahi, Bello, Abdulra'uf, Sharifai and Abubakar, 2014). There is a growing need in the healthcare scenario to store and organize sizeable clinical data, analyze the data, assist the healthcare professionals in decision making, and develop data mining methodologies to mine hidden patterns and discover new knowledge (Ramani, 2012). Data warehousing integrates fragmented electronic health records from independent and heterogeneous clinical data stores (Saliya, 2013) into a single repository. It is based on these concepts that this study plan to design a Data Warehousing and Mining Framework that will organize, extract, and integrate medical records of patients.

1.3 Aim and Objectives of the Study

The aim of this study is to design a Data Warehousing and Mining Framework for clinical records of patients and the objectives of the study are to:

i. Design a database for the Data Warehouse (DW) Prototype Model using Dimensional Modeling and Techniques.

ii. Simulate the data warehouse database in order to generate reports, uncover hidden patterns, and knowledge from the DW to aid decision making, using the SQL Server 2014, Business Intelligence (BI) Tools and Microsoft Reporting Services.

iii. Develop a web platform that will integrate the front-end, middle-end and the back-end using visual studio 2015 as the development platform. ASP.NET, bootstrap Cascading Style Sheet (CSS), HTML5 and JavaScript will be used for the frond end, C# as the middle end programming language and Microsoft SQL Server Management Studio for the back end development.

1.4 Significance of the Study

It is clear that advanced clinical data warehousing and mining information systems will be a driver for quality improvements of medical care. This ability to integrate data to have valuable information will result in a competitive advantage, enabling healthcare organizations to operate more efficiently. The discovered knowledge in the Human Leaning technique can be used for community diagnoses or prognosis. It will eliminate the use of file system and physical conveyance of files by messengers. It will provide a platform for data mining operations on patient's clinical data. This research will encourage and challenge many government and non-governmental healthcare providers to opt for data warehouse and mining investment in order to improve information access within their organization, bringing the user of their information system in touch with their data, and providing cross-function integration of operation systems within the organizations.

1.5 Definition of Terms

In this section, we have operationally defined some technical terms that were used in this research.

i. Decision Support System: This refers to the system that uses data (internal and external) and models, to provide a simple and easy to use interface, hence, allowing the decision maker to have control over the decision process.

ii. Prototype: Refers to the replica of the complete system ("DW prototype framework") working as if it were real.

iii. Architecture: Is the process that focuses on the formation of data stores within a DW system, along with the procedure of how the data flows from the source systems to the application use by the end users.

iv. Heterogeneous: consisting of or composed of dissimilar elements or ingredients.

v. Data Mining: The process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis.

vi. Data Visualization: Technique used in presenting results obtained in a graphical view for easier understanding and comprehension.

1.5.1 *Data warehouse definition*

Inuwa and Garba (2015) define data warehouse as a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management's decisions.

Subject-oriented: Classical operations systems are organized around the applications of the company. Each type of company has its own unique set of subjects.

Integrated: Data is fed from multiple disparate sources into the data warehouse. As the data is fed it is converted, reformatted, re-sequenced, summarized, and so forth. The result is that data once it resides in the data warehouse has a single physical corporate image.

Non-volatile: Data warehouse data is loaded and accessed, but it is not updated. Instead, when data in the data warehouse is loaded, it is loaded in a snapshot, static format. When subsequent changes occur, a new snapshot record is written. In doing so a history of data is kept in the data warehouse.

Time-variant: Every unit of data in the data warehouse is accurate at given moment in time. In some cases, a record is time stamped. In other cases, a record has a date of transaction. But in every case, there is some form of time marking to show the moment in time during which the record is accurate.'

According to Kimball and Ross as cited by Inuwa and Oye (2015) DW is the conglomerate of all Data Marts within the enterprise. Information is always stored in the dimensional model. Kimball view data warehousing as a constituency of Data Marts. Data Marts are focused on delivering business objectives for departments in the organization, and the DW is a conformed dimension of the Data Marts.

Over the last few years, organizations have increasingly turned to data warehousing to improve information flow and decision support. A DW can be a valuable asset in providing easy access to data for analysis and reporting. Unfortunately, building and maintaining an effective DW has several challenges (Güzin, 2007).

2. LITERATURE REVIEW

2.1.1 *Data warehouse modelling*

Ballard cited by Inuwa and Oye (2015) gave an assessment of the evolution of the concept of data warehousing, as it relates to data modeling for the DW, they defined database warehouse modeling as the process of building a model for the data in order to store in the DW. There are two data modeling techniques that are relevant in a data warehousing environment and they are:

i. Entity Relationship (ER) Modelling: ER modeling produces a data model of the specific area of interest, using two basic concepts: entities and the relationships between those entities. Detailed ER models also contain attributes, which can be properties of either the entities or the relationships. The ER model is an abstraction tool because it can be used to understand and simplify the ambiguous data relationships in the business world and complex systems. ER modeling uses the following concepts: entities, attributes and the relationships between entities. The ER model can be used to understand and simplify the ambiguous data relationships in the business world and complex systems environments.

ii. Dimensional Fact Modeling: Dimensional modeling uses three basic concepts: Measures, facts, and dimensions, Dimensional modeling is powerful in representing the requirements of the business user in the context of database tables. Measures are numeric values that can be added and calculated.

2.1.2 *Data warehouse modelling techniques*

Thomas and Carol cited by Inuwa and Oye (2015) derived the way a DW or a Data Mart structure in dimensional modelling can be achieved. Flat schema, Terraced Schema, Star Schema, Fact Constellation Schema, Galaxy Schema, Snowflake Schema, Star Cluster Schema, and Star flake Schema. However there are two basic models that are widely used in dimensional modeling: Star and Snowflake models.

i. Star Schema: The Star Schema (in Figure 2.1) is a relational database schema used to hold measures and dimensions in a Data Mart. The measures are stored in a fact table and the dimensions are stored in dimension tables. For each Data Mart, there is only one measure surrounded by the dimension tables, hence the name star schema. The centre of the star is formed by the fact table. The fact table has a column or the measure and the column for each dimension containing the foreign key for a member of that dimensions. The key for this table is formed by concatenate all of the foreign key fields. The primary key for the fact table is usually referred to as composite key. It contain the measures, hence the name “Fact”. The dimensions are stored in dimension tables. The dimension table has a column for the unique identifier of a member of the dimension, usually an integer of a short character value. It has another column for a description. (Inuwa&Oye, 2015).

ii. Snowflake Schema: Snowflake Schema model is derived from the star schema and, as can be seen, looks like a snow flake. The snowflake model is the result of decomposing one or more of the dimensions, which generally have hierarchies between themselves. Many-to-one relationships among members within a dimension table can be defined as a separate dimension table, forming a hierarchy as can be seen in Figure 2.2.

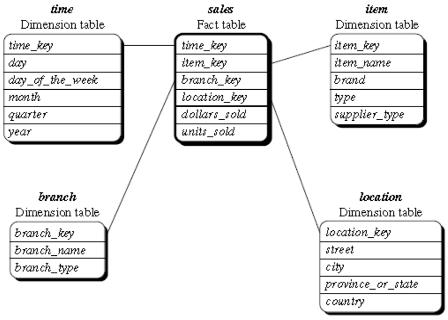


Figure 2. 1: Star Schema (Jiawei, 2012)

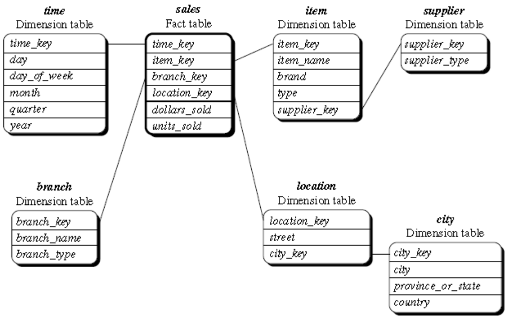


Figure 2. 2: Snowflake Schema (Jiawei, 2012)

2.1.3 **Data mart**

A Data Mart is a small DW built to satisfy the needs of a particular department or business area. The term Data Mart refers to a sub entity of Data Warehouses containing the data of the DW for a particular sector of the company (department, division, service, product line, etc.). The Data Mart is a subset of the DW that is usually oriented to a specific business line or team. Whereas a DW combines databases across an entire enterprise, Data Marts are usually smaller and focus on a particular subject or department. Some Data Marts are called Dependent Data Marts and are subsets of larger Data Warehouses. Gopinath, Damodar, Lenin, Rakesh& Sandeep, 2014, also summarized the types of Data Marts as follows:

2.1.3.1 Independent and dependent data marts

An independent Data Mart is created without the use of a central DW. This could be desirable for smaller groups within an organization. A dependent Data Mart allows you to unite your organization's data in one DW. This gives you the usual advantages of centralization as can be seen in Figure 2.3.

2.1.3.2 Hybrid Data Marts

A hybrid Data Mart allows you to combine input from sources other than a DW. This could be useful for many situations, especially when you need Ad-hoc integration, such as after a new group or product is added to the organization as can be seen in Figure 2.4:

a). A hybrid Data Mart transform data to combine input from sources other than a DW.

b). Extracting the data from hybrid Data Mart based on required conditions.

c). After extracting load into as a departmental Data Marts.

The Data Mart typically contains a subset of corporate data that is valuable to a specific business unit, department, or set of users. This subset consists of historical, summarized, and possibly detailed data captured from transaction processing systems (called independent Data Marts), or from an existing enterprise DW (called dependent Data Marts). It is important to realize that the functional scope of the Data Mart's users defines the Data Mart, not the size of the Data Mart database (Chuck, Daniel, Amit, Carlos and Stanislav, 2006).

2.1.4 Architecture of the DW

Data contained in a DW holds five types of data: data currency, existing data, data summarization (lightly and highly summarized data), and Metadata (Inuwa&Garba 2015). This traditional data warehousing architecture in Figure 2.5 encompasses the following components (Inuwa&Garba 2015):

i. Data sources as external systems and tools for extracting data from these sources.

ii. Tools for transforming, which is cleaning and integrating the data.

iii. Tools for loading the data into the DW.

iv. The DW as central, integrated data store.

v. Data Marts as extracted data subsets from the DW oriented to specific business lines, departments or analytical applications.

vi. A metadata repository for storing and managing metadata

vii. Tools to monitor and administer the DW and the extraction, transformation and loading process.

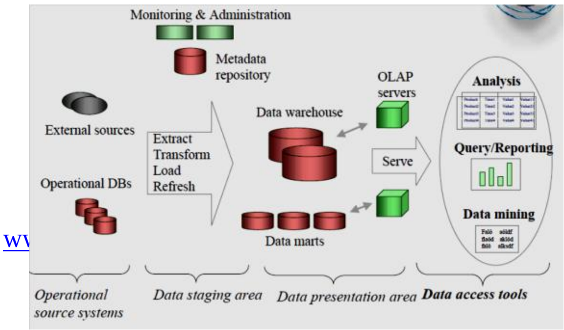
viii. An OLAP (online analytical processing) engine on top of the DW and Data Marts to present and serve multi-dimensional views of the data to analytical tools.

ix. Tools that use data from the DW for analytical applications and for presenting it to end-users.

This architecture exemplifies the basic idea of physically extracting and integrating mostly transactional data from different sources, storing it in a central repository while providing access to the data in a multi-dimensional structure optimized for analytical applications. However, the architecture is rather old and, while this basic idea is still intact, it is rather unclear and inaccurate about several facts:

Firstly, most modern data warehousing architectures use a staging or acquisition area between the data sources and the actual DW. This staging area is part of the Extract, Transform and Load Process (ETL process). It temporarily stores extracted data and allows transformations to be done within the staging area, so source systems are directly decoupled and no longer strained (Thilini& Hugh, 2010). Secondly, the interplay between DW and Data Marts in the storage area are not completely clear.

Figure 2. 3: Traditional Data Warehousing Architecture (Inuwa&Garba, 2015).



Actually, in practice this is one of the biggest discourses about data warehousing architecture with two architectural approaches proposed by Bill Inmon and Ralph Kimball (Inuwa&Garba, 2015). Inmon places his data warehousing architecture in a holistic modelling approach of all operational and analytical databases and information in an organization, the Corporate Information Factory (CIF). What he calls the atomic DW is a centralized repository with a normalized, still transactional and fine-granular data model containing cleaned and integrated data from several operational sources (Inuwa&Garba, 2015).Inmon's approach, also called enterprise DW architecture by Thilini and Hugh (2010) is often considered a top-down approach, as it starts with building the centralized, integrated, enterprise-wide repository and then deriving Data Marts from it to deliver for departmental analysis requirements.

However, it is possible to build an integrated repository and the derived Data Marts incrementally and in an iterative fashion. Kimball on the other hand proposes a bottom-up approach which starts with process and application requirements (Kimball, Reeves & Ross) as cited by Inuwa and Garba (2015). With this approach, first the Data Marts are designed based on the organization's business processes, where each Data Mart represents data concerning a specific process. The Data Marts are constructed and filled directly from the staging area while the transformation takes places between staging area and Data Marts.

The Data Marts are analysis-oriented and multi-dimensional. The DW is then just the combination of all Data Marts, where the single Data Marts are connected and integrated with each other via the data bus and so-called conformed dimensions that are Data Marts use, standardized or 'conformed' dimension tables (Inuwa&Garba, 2015).

When two Data Marts use the same dimension, they are connected and can be queried together via that identical dimension table. The data bus is then a net of Data Marts, which are connected via conformed dimensions. This architecture (also called Data Mart bus architecture with linked dimensional Data Marts by Thilini and Hugh (2010) therefore forgoes a normalized, enterprise-wide data model and repository.

In Figure 2.4, there are a number of options for architecting a Data Mart. For example:

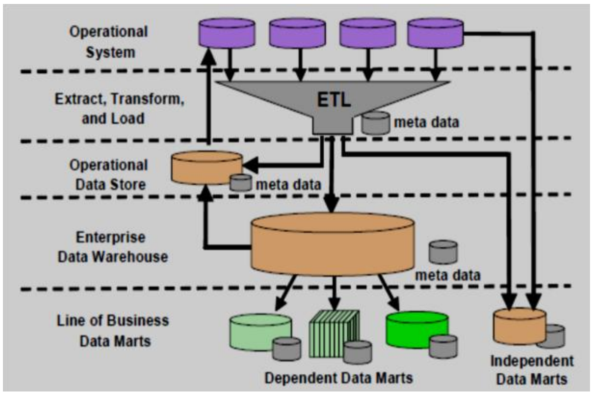


Figure 2. 4: DW Architecture (Inuwa&Garba, 2015)

i. Data can come directly from one or more of the databases in the operational systems, with few or no changes to the data in format or structure. This limits the types and scope of analysis that can be performed. For example, you can see that in this option, there may be no interaction with the DW Meta Data. This can result in data consistency issues.

ii. Data can be extracted from the operational systems and transformed to provide a cleansed and enhanced set of data to be loaded into the Data Mart by passing through an ETL process. Although the data is enhanced, it is not consistent with, or in sync with, data from the DW.

iii. Bypassing the DW leads to the creation of an independent Data Mart. It is not consistent, at any level, with the data in the DW. This is another issue impacting the credibility of reporting.

iv. Cleansed and transformed operational data flows into the DW. From there, dependent Data Marts can be created, or updated. It is a key that updates to the Data Marts are made during the update cycle of the DW to maintain consistency between them. This is also a major consideration and design point, as you move to a real-time environment. At that time, it is good to revisit the requirements for the Data Mart, to see if they are still valid.

However, there are also many other data structures that can be part of the data warehousing environment and used for data analysis, and they use differing implementation techniques. Although Data Marts can be of great value, there are also issues of currency and consistency. This has resulted in recent initiatives designed to minimize the number of Data Marts in a company. This is referred to as Data Mart Consolidation (DMC). Data Mart consolidation may sound simple at first, but there are many things to consider. A critical requirement, as with almost any project, is executive sponsorship, because you will be changing many existing systems on which people have come to rely, even though the systems may be inadequate or outmoded. To do this requires serious support from senior management. They will be able to focus on the bigger picture and bottom-line benefits, and exercise the authority that will enable making changes (Chuck et al., 2006).

2.1.5 Benefits of data warehousing

There are several benefits of data warehousing. The most important ones are listed as follows (Kimball & Ross, 2002):

i. DW improves access to administrative information for decision makers.

ii. It can get data quickly and easily perform analysis. One can work with better information, make decisions based on data. DW increases productivity of corporate decision-makers.

iii. Data extraction from its original data sources into the central area resolves the performance problem, which arises from performing complex analyses on operational data.

iv. Data in the warehouse is stored in specialized form, called a multidimensional database. This form makes data querying efficient and fast.

v. A huge amount of data is usually collected in the DW. Compared with relational databases that are still very popular today, data in the warehouse does not need to be in normalized form. In fact, it is usually de-normalized to support faster data retrieval.

2.1.6 Data warehousing in medical field

Health care organizations require data warehousing solutions in order to integrate the valuable patient and administrative data fragmented across multiple information systems within the organization. As stated by Kerkri cited by Saliya (2013), at a technical level, information sources are heterogeneous, autonomous, and have an independent life cycle. Therefore, cooperation between these systems needs specific solutions. These solutions must ensure the confidentiality of patient information. To achieve sufficient medical data share and integration, it is essential for the medical and health enterprises to develop an efficient medical information grid (Zheng et al., 2008).

A medical data warehouse is a repository where healthcare providers can gain access to medical data gathered in the patient care process. Extracting medical domain information to a data warehouse can facilitate efficient storage, enhances timely analysis and increases the quality of real time decision making processes. Currently medical data warehouses need to address the issues of data location, technical platforms, and data formats; organizational behaviors on processing the data and culture across the data management population. Today's healthcare organizations require not only the quality and effectiveness of their treatment, but also reduction of waste and unnecessary costs. By effectively leveraging enterprise wide data on labour expenditures, supply utilization, procedures, medications prescribed, and other costs associated with patient care, healthcare professionals can identify and correct wasteful practices and unnecessary expenditures (Sahama & Croll, 2007).

Medical domain has certain unique data requirements such as high volumes of unstructured data (e.g. digital image files, voice clips, radiology information, etc.) and data confidentiality. Data warehousing models should accommodate these unique needs. According to Pedersen and Jensen cited by Saliya (2013) the task of integrating data from several EHR systems is a hard one. This creates the need for a common standard for EHR data.

According to Kerkri cited by Saliya (2013), the advantages and disadvantages of data warehousing are given below.

Advantages:

1. Ability to allow existing legacy systems to continue in operation without any modification
2. Consolidating inconsistent data from various legacy systems into one coherent set
3. Improving quality of data
4. Allowing users to retrieve necessary data by themselves

Disadvantages:

1. Development cost and time constraints

2.1.7 Cancer data warehouse architecture

Cancer known medically as a malignant neoplasm, is a broad group of various diseases, all involving unregulated cell growth. In cancer, cells divide and grow uncontrollably, forming malignant tumours, and invade nearby parts of the body. The cancer may also spread to more distant parts of the body through the lymphatic system or bloodstream. Not all tumours are cancerous. Benign tumours do not grow uncontrollably, do not invade neighbouring tissues, and do not spread throughout the body. Determining what causes cancer is complex. Many things are known to increase the risk of cancer, including tobacco use, certain infections, radiation, lack of physical activity, poor diet and obesity, and environmental pollutants. These can directly damage genes or combine with existing genetic faults within cells to cause the disease. Approximately five to ten percent of cancers are entirely hereditary. People with suspected cancer are investigated with medical tests. These commonly include blood tests, X-rays, CT scans and endoscopy (Sheta & Ahmed, 2012).

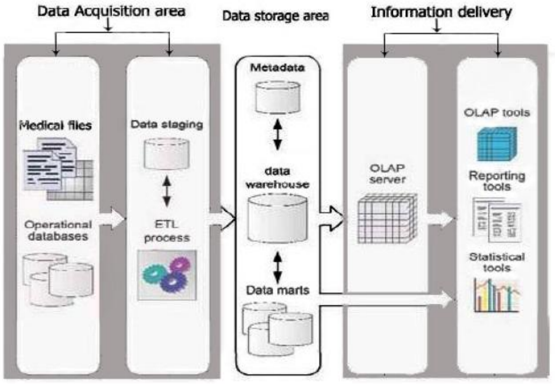


Figure 2. 5: Cancer DW Architecture (Sheta& Ahmed, 2012)

2.1.8 *Influenza disease data warehouse architecture*

Influenza, commonly known as the 'flu', is an infectious disease of birds and mammals caused by ribonucleic acid (RNA) viruses of the family Orthomyxoviridae, the influenza viruses, (Rajib, 2013). The most common symptoms are chills, fever, sore throat, muscle pains, headache (often severe), coughing, weakness/fatigue Irritated, watering eyes, Reddened eyes, skin (especially face), mouth, throat and nose, Petechial Rash and general discomfort. Influenza may produce nausea and vomiting, particularly in children. Typically, influenza is transmitted through the air by coughs or sneezes, creating aerosols containing the virus. Influenza can also be transmitted by direct contact with bird droppings or nasal secretions, or through contact with contaminated surfaces. Influenza spreads around the world in seasonal epidemics, resulting in about three to five million yearly cases of severe illness and about 250,000 to 500,000 yearly deaths (Rajib, 2013). People who suspected influenza are investigated with medical tests. These commonly include Blood test (white blood cell differential), Chest x-ray, Auscultation (to detect abnormal breath sounds), Nasopharyngeal culture.

Figure 2.6 shows the proposed architecture for the health care data warehouse specific to Influenza disease by Rajib in 2013. Architecture of Influenza specific health care data warehouse system builds with Source Data components in the left side where multiple data that comes from different data source and transform into the Data Staging area before integrating. The Data staging component present at the next building block. Those two blocks is under Data Acquisition Area. In the middle Data Storage component that manages the data warehouse data. This component also with Metadata, that also keep track of the data and also with Data Marts. Last component of this architecture is Information Delivery component that shows all the different ways of making the information from the data warehouse available to the user for further analysis (Rajib, 2013).

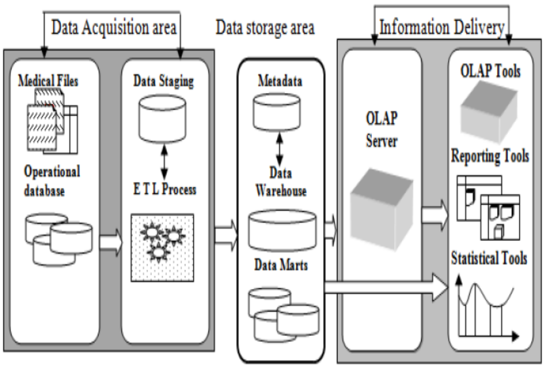


Figure 2. 6: Data Warehouse Architecture for Influenza Disease (Rajib, 2013)

2.1.9 *Cardiac surgery data warehousing model*

Cardiac Surgery clinical data can be distributed across various disparate and heterogeneous clinical and administrative information systems. This makes accessing data highly time consuming and error prone .

A data warehouse can be used to integrate the fragmented data sets. Once the data warehouse is created it should be populated with data through Extract, Transform and Load processes. Figure 2.12 shows a graphical overview of cardiac surgery data warehousing model (Saliya, 2013).

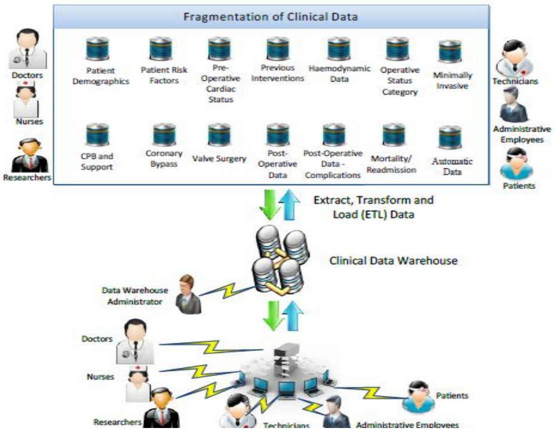


Figure 2. 7: Data Warehousing Model for Integrating Fragmented Electronic Health Records from Disparate and Heterogeneous Cardiac Surgery Clinical Data Stores (Saliya, 2013).

2.1.10 *Diabetic data warehousing model*

Diabetes is a defect in the body's ability to convert glucose (sugar) to energy. Glucose is the main source of fuel for our body. When food is digested it is changed into fats, protein, or carbohydrates. Foods that affect blood sugars are called carbohydrates. Carbohydrates, when digested, change to glucose. Examples of some carbohydrates are: bread, rice, pasta, potatoes, corn, fruit, and milk products. Individuals with diabetes should eat carbohydrates but must do so in moderation. Glucose is then transferred to the blood and is used by the cells for energy. In order for glucose to be transferred from the blood into the cells, the hormone - insulin is needed. Insulin is produced by the beta cells in the pancreas (the organ that produces insulin). In individuals with diabetes, this process is impaired. Diabetes develops when the pancreas fails to produce sufficient quantities of insulin, Type 1 diabetes or the insulin produced is defective and cannot move glucose into the cells and occurs most frequently in children and young adults, although it can occur at any age. Type 1 diabetes accounts for 5-10% of all diabetes in the United States. There does appear to be a genetic component to Type 1 diabetes, but the cause has yet to be identified. Type 2 diabetes. Either insulin is not produced in sufficient quantities or the insulin produced is defective and cannot move the glucose into the cells, is much more common and accounts for 90-95% of all diabetes. Type 2 diabetes primarily affects adults, however recently Type 2 has begun developing in children. There is a strong correlation between Type 2 diabetes, physical inactivity and obesity, (Abubakar et al., 2014).

Gap Analysis

The research carried out by other researchers, only focus on building a data warehouse for a particular disease. No single study exists which adequately covers the integration of clinical record of patients into a data warehouse for data analysis and mining that is not specific to a single disease. There is also no specific research that covers the integration of medical record of patients using the Nigerian medical records for analysis and mining using data visualization for decision making.Therefore this research is designing a framework for integrating patient clinical records into a single data warehouse for data analysis and mining using data visualization technique that is not specific to single disease using Nigeria medical record.

3. METHODOLOGY

The following research method and tools were used in achieving these research objectives:

System Analysis: UML was used in analysing all the physical and logical models. **OLAP:** Online Analytic processing (Server) was used for processing and analysing data from the server on a web browser. **ETL:** Extraction transformation and loading of data into the data warehouse for mining and visualization purpose **SSIS:** Integrating the operations data with the main data warehouse database.

Visualization Technique: Presentation of mine data from the data warehouse as visual effects rather than row and columns. Data are displayed in form of graphs and shapes for easy interpretation and understanding. Microsoft Visio 2015 as the UML tool, was used in in designing all the logical models of the proposed system. Aspx, HTML 5 and Bootstrap CSS were used in designing the GUI and SQL Server Reporting Services was used in designing the data visualization, clinical decision support and mining outputs of the system. The programming language of choice for the middle end was C# pronounced C-sharp. The database used in designing the physical models is Microsoft SQL Server 2014, while, SQL Server 2012 server tools were used for writing the ETL (Extraction, Transformation and Load) program.

4. RESEARCH RESULTS

System Prototype Development and Validation

The system prototype development is the actual application of the analysis and design that has been carried out. In this phase of the study, we have designed the DW (Fact and dimension tables) prototype, the ETL (Extract, Transform and Load), the front end (GUI) and the Middle end of the application for the purpose of this study. Validation process involves the confirmation by medical system administrators and personnel's that an information system (DW prototype) has been realized appropriately and it is in conformity with the User's needs and intended use. Figures 1-7 are available in the appendix.

FIGURE:1 Shows the architectural design of the Framework for Data Warehouse and Mining Clinical Records. The architecture has the followings abilities:

i. Integration of the independent and dependent Data Marts within the architecture.

ii. It has a multi-tier ETL which are simpler and in different stages.

iii. It has a standard access points for all medical and non-medical experts (Using a three tier server).

iv. It has the ability to Mine and analyse records using Data Visualization

The architecture has a back and frond end system in which so many activities are carried out. The back end systems comprise of the operational data source system, data staging area and the data presentation area. Data are first extracted from different operational data source systems using ETL and then stored at the data staging area where it is being processed as soon as it is captured. The activities of the ETL at the data staging area include data cleansing and validation, data integration, data fixing and data entry errors removal, transforming and refreshing data into a new normalized standard. As soon as data is cleaned, the transformed data are loaded and indexed into the data presentation area where the DW is located. The frontend systems in the other hand comprise of the main servers (OLAP) and data access tools. The OLAP server hold the copy of the cleansed clinical records. The data access tool is the interface where applications are stored, which allows for data Analysis, Reporting/Querying and Data mining activities.

FIGURE: 2, is about ***Creating and loading of the Clinical DW database***

The DW database was created on an MSSQL Server 2014 database, and the data loading was also done through the SSIS package. Fact and Dimension tables were pushed into the DW database, Figure 2 shows Physical database of the DW Model by Star Schema. The DW database is the one that integrates the Fact and Dimension tables of patient's clinical records into the MSSQL Server 2014 database. It was part of the MSSQL Server database that we used as the database repository. The DW database was populated with the correct data of good quality that we can make use of as the data repository. Data visualization is all about presenting data in graphical format rather than rows and columns thereby making data interpretation easy and is best use by decision makers to take decisions for organizations.

FIGURE 3, shows some of the Dimension and Fact tables that were extracted from the source system into the staging database area. The tables were cleansed and refreshed at this point and ready for transfer into their respective Data Marts. Having designed the Fact and Dimension tables and the extraction of data from the source system, then the researcher populated the Data Marts with the data that was extracted earlier from the staging databases. It is now from the Data Marts that another ETL was performed to transport the data into the DW database.

System Verification, Reports and Data Visualization

The best way to verifying the data in the DW is to prepare queries on these data. In this study, certain reports, data analysis and data visualization were presented and the purpose of these reports and analysis is to demonstrate the usefulness of the DW approach to data presentation and decision making. Even though these reports and analysis were based on some random requirements, this set of sample reports and analysis can be used as a basis for generating more comprehensive sets by applying complex queries on the data. We have classified the Analysis, Reports and Data Visualization that we generated from the DW based on all the analysis on patient diagnosis and details. **Figure 4**, illustrates the general system architecture for this study. It shows how the Source System database, Servers and the User/administrator system are connected for possible data visualization and decision making. **FIGURE: 5**, shows data visualization of Malaria patients and their LGA of residence within the year 2015, 2016 and 2017. **Figure 6**, shows data visualization of Tuberculosis patients and their LGA of residence within the year 2015, 2016 and 2017. Decision makers can use the visualized report to determine the trend of infection between the various patients LGA of residence. Proper decision can be taken on how to curb the rate of infection within the LGA's with high rate of illness by setting appropriate infrastructure and medical personnel in those LGA**. Figure 7**, shows a report of patients diagnosed with typhoid-fever based on their gender, month and the year of diagnosis.

5. CONCLUSION

This research has design and implemented a Clinical DW and Mining system using data visualization technique within the context of the healthcare service, to better incorporate patient records into single systems for simpler and improved data mining, analysis, reporting and querying. The Clinical DW we built contains only the data that is required for data mining, reporting and analysis for the purpose of this study and it can be updated periodically, such that all the data can be integrated from different source systems into the central DW. The system is design to operate as a distributed system. Periodic update on a daily basis is possible because contemporary technologies have narrowed the gap between updates. This can enable organizations to have a "real time" DW which can be analyzed using an OLAP Server.

The research focused on developing a Framework for DW and Mining of clinical records of patient, for improve data analysis and mining using data visualization. The study also considered the ideologies of data warehousing in the course of this research and demonstrated how data can be incorporated from diverse desperate heterogeneous clinical data stores into a single DW for mining and analysis purpose, to aid medical practitioners and decision makers in decision making. The researchers have also been able to develop a web based data mining, reporting and analysis tool (GUI) where users can interact with the system to get a speedy and timely information needed for the clinical decision making and community diagnosing.

The Framework for DW and Mining of clinical record of patients was design not to be specific to only a number of disease but to accept as many diseases as possible without any limitation. The ideologies that the researchers followed to develop this system makes it scalable and as such, it can be adopted for any form of disease, infections analysis and mining by medical institutions and healthcare providers in Nigeria. The designed framework can be used by industry professionals and researcher for implementing data warehousing system in the medical field, for long time data analysis and mining. The framework can also be used for community diagnosing in cases of outbreak of certain disease. Developing a Framework for DW and Mining of Clinical records is very essential, particularly for medical decision-makers, academic researchers, IT professionals and non-professional. Clinical data mining must not only support medical professionals and decision makers to understand the past, but also it strive professionals to work towards new prospects.

6. ACKNOWLEDGMENTS

We thank the experts who have helped to review this paper.

7. REFERENCES

[1] Abubakar, etal (2014). Building a diabetes data warehouse to support decision making in healthcare industry. *IOSR Journal of Computer Engineering (IOSR-JCE), 16*(2), e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. IX (Mar-Apr. 2014), PP 138-143 www. Retrieved from iosrjournal.org

[2] Chuck, B., Daniel, M. F., Amit, G. C., & Stanislav, V. (2006). *Dimensional Modeling: In a DSS Environment.* NY 10504-1785 U.S.A.: IBM Corporation, North Castle Drive Armonk

[3] Demetriades, J. E., Kolodner, R. M., & Christopherson, G. A. (2005). Person Centered Health Records. Towards Healthy People. . *Health Informatics Series*. USA: Springer. Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[4] Goldstein, D., Groen, P. J., Ponkshe, S., & Wine, M. (2007). *Medical informatics 20/20: quality and electronic health records through collaboration, open Solutions, and innovation.* Massachusetts, Sudbury, USA: Jones and Bartlett Publishers, Inc.

[5] Gopinath, T., Damodar, N. M., Lenin, Y., Rakesh, S., & Sandeep, M. (2014). Scattered Across Data Mart Troubles Triumph Over in the course of Data Acquisition through the Decision Support System. *International Journal of Computer Technology and Application, 2*(5), 411-419.

[6] Güzin, T. (2007). *Developing a Data Warehouse for a University Decision Support System.* Atilim University, The Graduate School of Natural and Applied Sciences.

[7] Ibrahim I., & Oye, N. (2015). Design of a Data Warehouse Model for a University Decision Support System. *Information and Knowledge Management, 5*.

[8] Inuwa, I., & Garba, E. (2015). An Improved Data Warehouse Architecture for SPGS, MAUTECH, Yola, Nigeria. *West African Journal of Industrial & Academic Research, 14*(1).

[9] Kimball, R. &. (2002). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling. In R. Kimball, & M. Ross, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modelling.* New York: John Wiley & Sons, Inc.

[10] Mann, L. (2005). From “silos” to seamless healthcare: bringing hospitals and GPs back together again. *Medical Journal of Australia, 1*(182), 34-37. Retrieved from http://www.mja.com.au/public/issues/182_01_030105/man10274_fm.html

[11] Rajib, D. (2013). *Health care data warehouse system Architecture for influenza (flu) Diseases.* Global Institute of Management & Technology, Krishnanagar., Department of Computer Science & Engineering, West Bengal, India.

[12] Sahama, T. R., & Croll, P. R. (2007). A data warehouse architecture for clinical data warehousing. *First Australasian Workshop on Health Knowledge Management and Discovery.* Retrieved from http://crpit.com/confpapers/CRPITV68Sahama.pdf

[13] Saliya, N. (2013). *Data warehousing model for integrating fragmented electronic health records from disparate and heterogeneous clinical data stores.* Queensland University of Technology, School of Electrical Engineering and Computer Science Faculty of Science and Engineering, Australia.

[14] Shepherd, M. (2007). Challenges in health informatics. *40th Hawaii International Conference on System Sciences.* doi:10.1109/HICSS.2007.123

[15] Sheta, O., &Ahmed, A. N. (2012). Building a Health Care Data Warehouse for Cancer Diseases. *International Journal of Database Management Systems (IJDMS), 4*.

[16] Thilini, A., & Hugh, W. (2010). Key organizational factors in Data Warehouse architecture selection. *Journal of Decision Support Systems, 49*(2), 200–212.

[17] Zheng, R., Jin, H., Zhang, H., Liu, Y., & Chu, P. (2008). Heterogeneous medical data Share and integration on grid. *International Conference on BioMedical Engineering and Informatics.* doi:10.1109/BMEI.2008.185

[18] Zhu, X., & Davidson, I. (2007). *Knowledge Discovery and Data Mining: Challenges and Realities.* New York: Hershey, New York.

APPENDIX

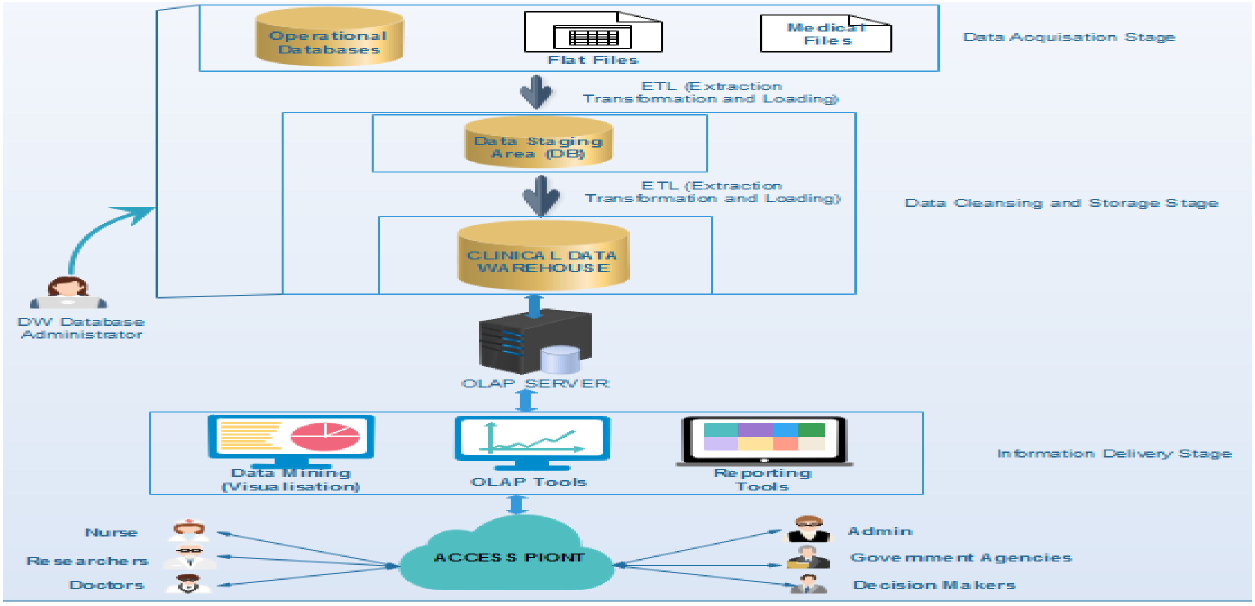


Figure 1: Framework for Data Warehousing and Mining Clinical Records of Patients

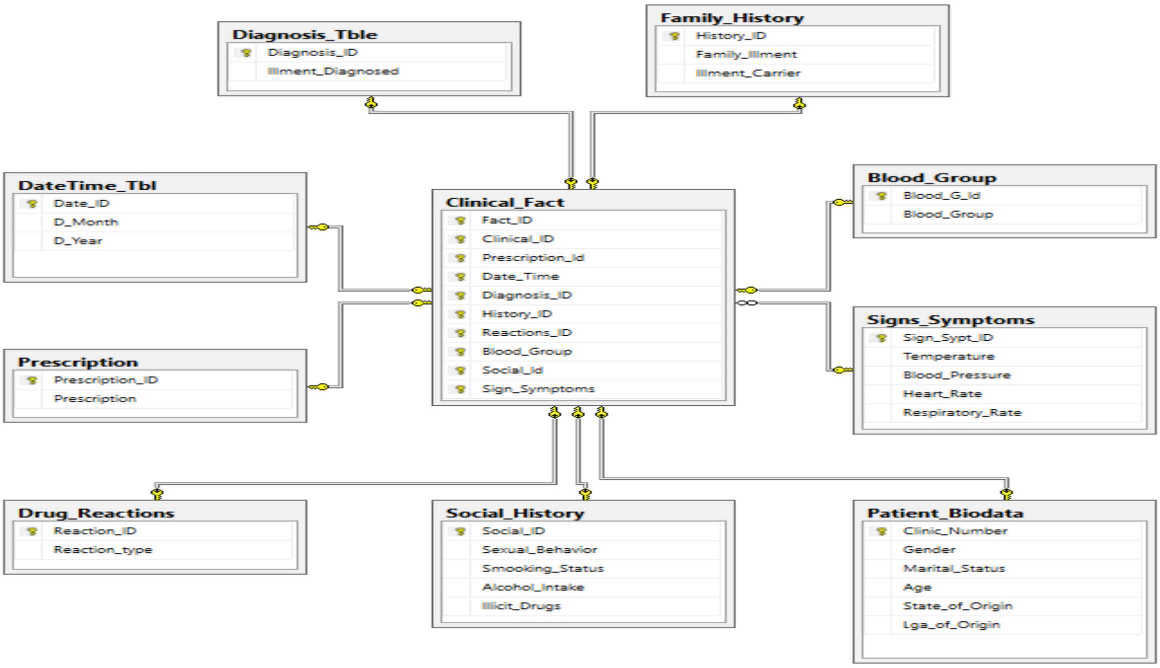


Figure 2: Star schema of the designed Framework for Data Warehousing and Mining Clinical Records of Patients

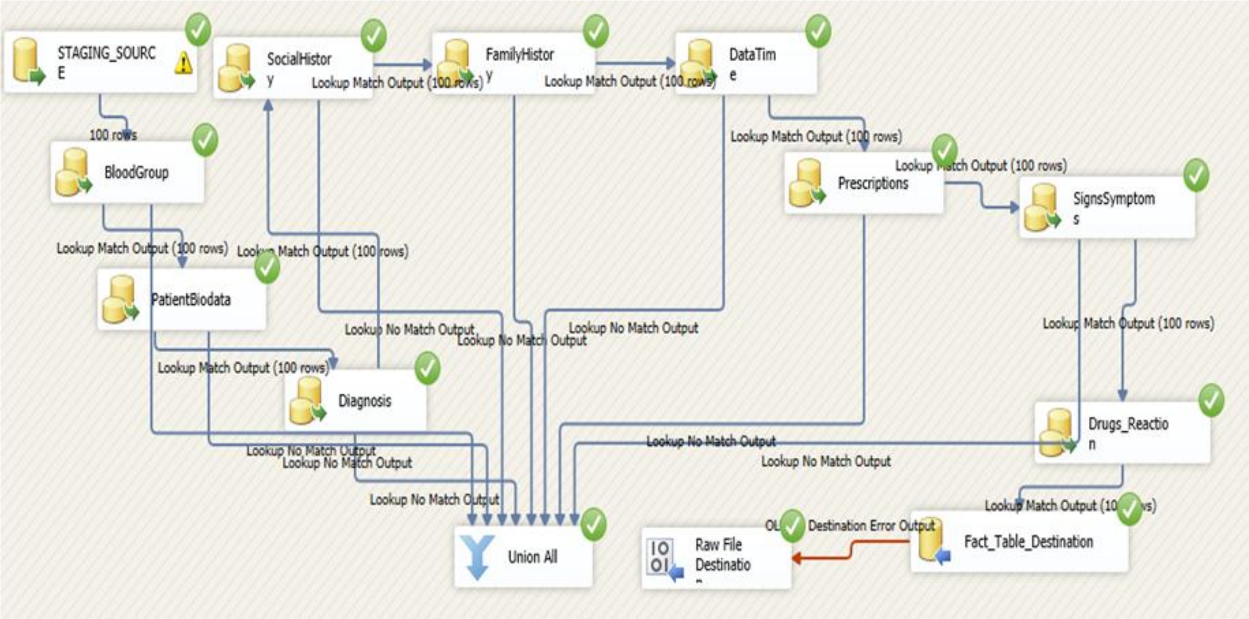


Figure 3: ETL for data cleansing and loading of the DW

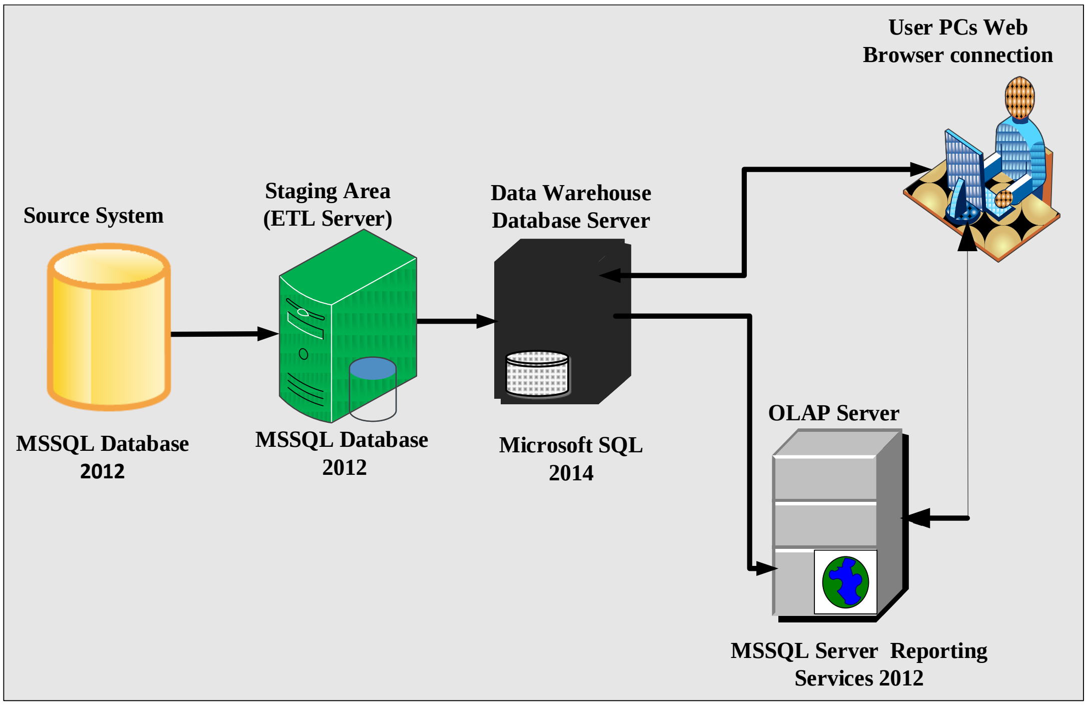

User PCs Web Browser connection
Source System
Staging Area (ETL Server)
Data Warehouse Database Server
MSSQL Database 2012
MSSQL Database 2012
Microsoft SQL 2014
OLAP Server
MSSQL Server Reporting Services 2012

Figure 4: Showing the structural architecture of the system

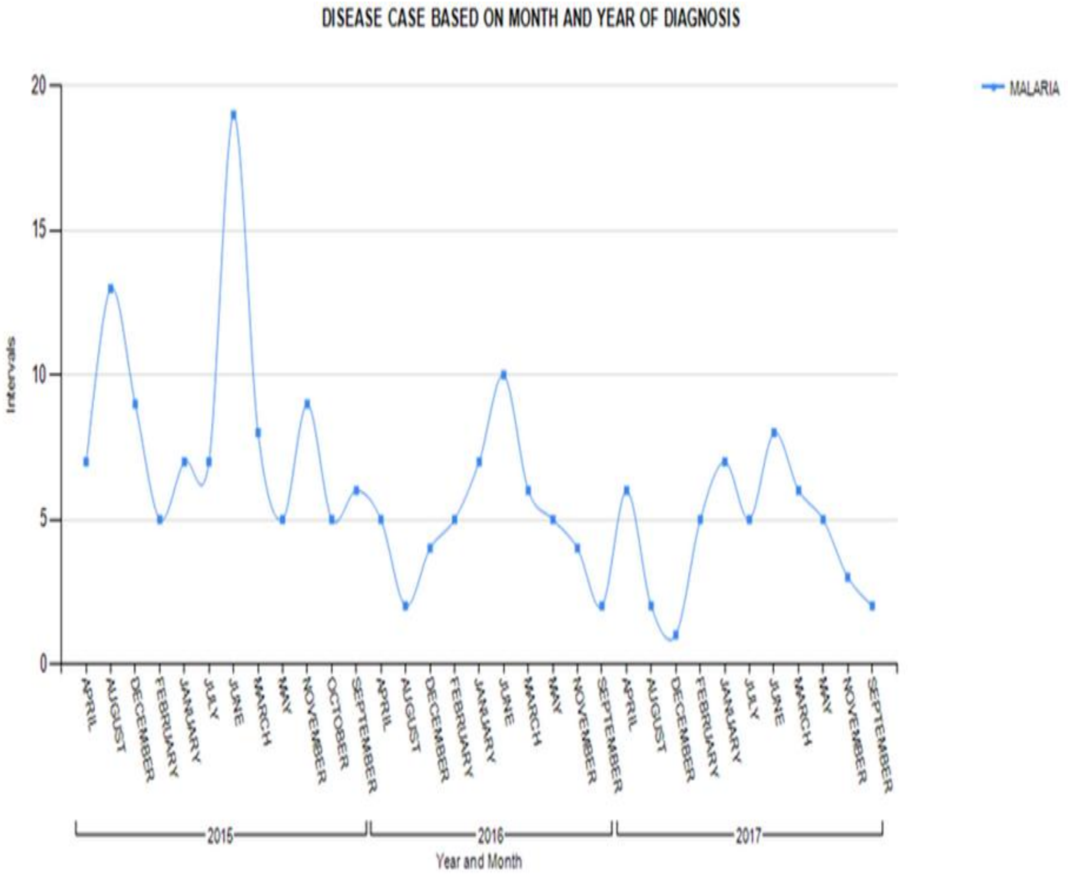


Figure 5: Malaria output result, based on month and year of diagnosis

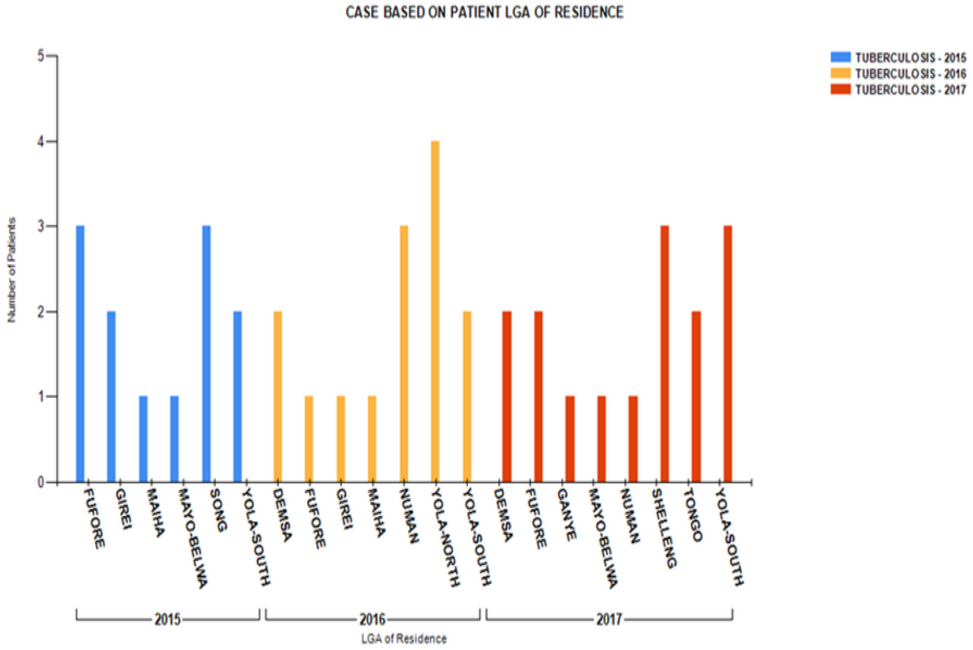


Figure 6: Tuberculosis report based on LGA of residence and year of diagnosis

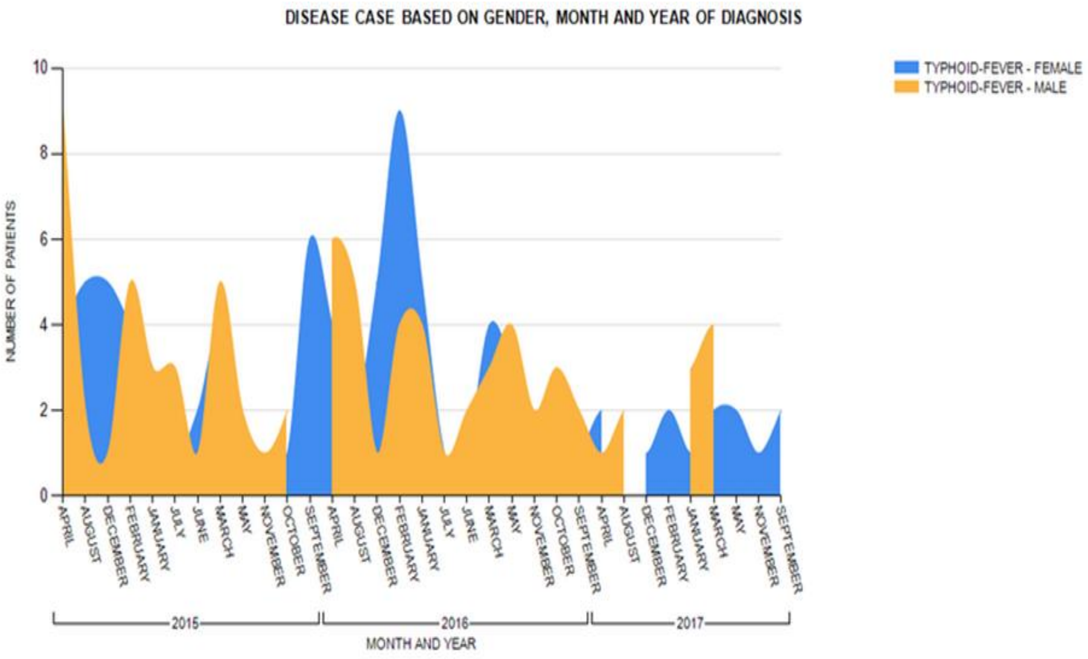


Figure 7: Typhoid-Fever report based on gender, month and year of diagnosis

International Journal of Computer Applications Technology and Research
Volume 8–Issue 04, 135-148, 2019, ISSN:-2319–8656

Fuzzy Logic Model for Analysis of Computer Network Quality of Experience

Walter B. Kihuya
Institute of Computer Science and Information Technology,
Jomo Kenyatta University of Agriculture and Technology,
Mombasa, Kenya.

Dr. Calvins Otieno
Institute of Computer Science and Information Technology,
Jomo Kenyatta University of Agriculture and Technology,
Nairobi, Kenya.

Dr.Richard Rimiru
Institute of Computer Science and Information Technology,
Jomo Kenyatta University of Agriculture and Technology,
Nairobi, Kenya.

Abstract: The estimation of the QoE provides valuable input in order to measure the user satisfaction of a particular service/application. Network QoE estimation is challenging as it tries to measure a subjective metric where the user experience depends on a number of factors that cannot easily be measured. All the Network analysis models can be divided into two major groups: qualitative and quantitative. In recent years many quantitative models have been developed in terms of quantitative measures i.e. use of scale of numbers between 1 to 5 to represent user perception of QoS. The challenge with this model is where user perception is subjective and not precise thus cannot be clearly measured using quantitative methods. On the other side qualitative models are in early stages of exploration. Little has been done on qualitative methods. Basing on previous studies, few models exists that measure qualitative analysis of computer network quality of experience. However none incorporated all the four parameters of integrity of service; throughput, delay, packet loss and jitter as parameters of network QoE. The study's objective is to address this gap by proposing a fuzzy logic model for analysis of computer network QoE. The tools used in the study are Linux MTR tool for data extraction, Ms. Excel for data cleaning and presentation, Visual paradigm for constructing of Unified Modeling language diagrams, mat lab software for plotting of functions/data, implementation of algorithms and creation of user interfaces. Experimental research design and sampling mechanisms is applied for 15 samples. The methodology in use is fuzzy logic. In order to deal with fuzziness associated with linguistic variables, inference rules are introduced. Five input linguistic terms are identified: Very High, High, Medium, Low and Very Low. Five output linguistic terms are defined to describe the opinion scores: Excellent, Good, Fair, Poor and Bad. Four variables are used: delay, jitter, packet loss and throughput. This results to a total of 625 rules (5^4). The rules are further condensed to 240 logical rules basing on expert knowledge. The collected data was used for simulation in matlab environment basing on the 240 rules. The results shows, analysis of Computer network QoE is subjective in nature rather than objective thus requires a resilient mechanism like fuzzy logic in order to capture clear-cut results to be used for decision making. The target population for this model is the ISPs' clients. This will enable ISPs to have the best responsive measures to deal with clients' QOE parameters so as to meet the QOS as per SLAs.

Keywords: fuzzy logic, ISPs (Internet Service Providers), quality of experience (QoE), Quality of service (QoS), SLAs (Service Level Agreement)

1. INTRODUCTION.

QoE in the context of telecommunications networks is defined as the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state. [1]

QoE is defined by International Telecommunication Union: *ITU* as the overall acceptability of an application or service, as perceived subjectively by the end-user. [2]

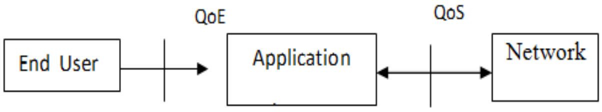


Figure2. 1: Relationship between QoS and QoE

Fuzzy logic is a natural, continuous logic patterned after the approximate reasoning of human beings [3]. As a theory mathematical discipline, fuzzy logic reacts to constantly changing variables [3].It challenges traditional logic by not being restricted to the conventional binary computer values of zero and one. Instead, fuzzy logic allows for partial truths and multivalued truths [3] .Fuzzy logic is especially advantageous for problems that cannot be easily represented by mathematical modeling because data is either unavailable or incomplete or the process is too complex [3].The real world language used in fuzzy control enables engineers to incorporate ambiguous, approximate human logic into computers using linguistic modeling, as opposed to mathematical modeling, greatly simplifies the design and modification of a fuzzy logic system [3].

Fuzzy set theory [4]was developed to address contexts in which decision Makers need to accurately analyze and process information that is imprecise in nature.

Fuzzy sets provide a conceptual framework, as well as an analytical tool to solve real World problems where there is a lack of specific facts and precision [4].

Human semantics are embedded in the meaning of fuzziness and comparison on the other hand; the usage of multi granularity linguistic information can eliminate the difference from evaluators [5].

Using fuzzy logic algorithms could enable machines to understand and respond to vague human concepts such as hot, cold, large, small, etc. It also could provide a relatively simple approach to reach definite conclusions from imprecise information [3].

The information technology (IT) and electronics industries apply the QoE model to businesses and services because QoE depends on customer experience; assessments are compiled from large user group polls. The most commonly used model is mean opinion score (MOS). The MOS is expressed as a single rational number, typically in the range 1–5, where 1 is lowest perceived quality and 5 is the highest perceived quality. Other MOS ranges are also possible, depending on the rating scale that has been used in the underlying test. This model is thus quantitative in nature while user perception is subjective and not precise thus cannot be clearly measured using quantitative methods. [2]

Fuzzy set theory was first introduced by Zadeh in 1965. Fuzzy logic is a problem solving methodology that provides a simple way of definite conclusions from vague and imprecise information. He was motivated by observing that human reasoning can utilize concepts and knowledge that don‘t has well [5].

In the case of Network analysis, all Network analysis models can be divided into two major groups: qualitative and quantitative. Qualitative metrics do not own quantitative values and cannot be measured numerically. Purposely, linguistic terms are used to evaluate performance of qualitative metrics [6]. Fuzzy logic controller is useful when the problem is too difficult to be solved with quantitative approaches [4].

2. RELATED WORK.

Several researches have been done on fuzzy logic in relation to quality performance though little has been done on fuzzy logic model for analysis of quality of experience.

The study in [7] proposed a Fuzzy logic aggregation of wireless sensor network data for smart traffic light control. This approach uses smart traffic control systems (STCS) to make traffic routing decisions. STCS use real time data and mimic human reasoning thus prove promising in vehicle traffic control. This presents a smart traffic light controller using fuzzy logic and wireless sensor network (WSN). The approach is designed for an isolated four way roundabout. It employed fuzzy logic to control the lights and determine how the green light will be assigned for each approach. The WSN collected the traffic data in real time. This data is aggregated and fed into a fuzzy logic controller (FLC) in form of two inputs – traffic quantity (TQ) and waiting time (WT) for each approach. Based on the inputs, the FLC then computes an output priority degree (PD) that controls green light assignment. Using the PD, an algorithm is formulated that assigns green light to the lane with highest PD. The cycle continues until all approaches get green.

In [8] a research study on a Fuzzy Logic System for Evaluating Quality of Experience of Haptic-based Applications was proposed. The proposed taxonomy was modeled with a fuzzy logic system and finally was tested by a Mamdani fuzzy inference system. In the mentioned study, by making some assumption like rule selection and membership function selection, the effect of different perception measures parameters such as rendering quality, physiological and psychological was studied. Here, fuzzy logic system was applied for objective measuring of QoE parameters.

The research work in [9] exhibited QoE estimation for web service selection using a Fuzzy-Rough hybrid expert system.

A methodology to estimate the quality of web services based on a fuzzy-rough hybrid algorithm is proposed. The estimated web QoE is used to select the most performing service among different web services. Fuzzy expert systems are good at making decision with imprecise information; however, they cannot automatically formulate rules that they require for making the decisions. Therefore, a fuzzy-rough hybrid expert system is proposed in this study where rough set theory is used to define the rules necessary for the fuzzy expert system. Three QoS parameters: reliability, execution time (in seconds), and availability (in seconds) are measured during the performance of the tests. Input linguistic terms are: Low, Medium and High. The output linguistic terms in use are: Bad, Poor, Fair, Good and Excellent.

The research work in [10] proposed analysis of Quality of Experience by applying Fuzzy logic: A study on response time. In this work, with a fuzzy perspective, the effect of response time variation in a network on the quality perceived by users is shown. Later, shows how by applying fuzzy techniques the linguistic terms and the users' perception can be translated into quantitative values. The main objective of this project was to analyze the fuzziness of QoE in order to provide more understandable user perception. This included proposing response time performance criteria that correlate well with QoE measurement result presented by fuzzy concepts. The proposed methodology provides a fuzzy relationship between QoE and Quality of Service (QoS) parameters. To identify this fuzzy relationship a new term called Fuzzi ed Opinion Score (FOS) representing a fuzzy quality scale is introduced. A fuzzy data mining method is applied to construct the required number of fuzzy sets. Then, the appropriate membership functions describing fuzzy sets are modeled and compared with each other. The proposed methodology intended to assist service providers for better decision-making and resource management [10] .

In [11] an efficient algorithm for transmitting packet for better quality of service in adhoc mobile network was proposed. In this study, Fuzzy Self Organizing Map (FSOM) provide very efficient algorithmic tools for transmitting packet in an efficient manner by taking the most efficient route and also the bandwidth, latency and range network parameters are considered to determine how good is the data delivered. The results indicated that fuzzy logic can guarantee QoS of every packet in the network. Incorporation of fuzziness in the input and output of the proposed model was seen to result in better performance. Input variables were only three properties: low, normal, and high. The output variables were poor, good and excellent.

In [12], a fuzzy logic based approach is in use for maintaining VoIP Quality in a network which is affected by many network factors (packet loss, packet delay, and jitter).In this case, Resource Reservation Protocol application was configured to control Token Bucket Algorithm and the simulation experiments were carried out with Opnet. In addition, comparison between Token Bucket with and without Quality of Service aimed at measuring network factors was performed. In this paper, building Fuzzy Token Bucket System consists of three variables (Bandwidth Rate, Buffer Size, and New Token) in order to improve Token Bucket Shaper output variable (New Token) by Fuzzy Stability model for Voice over IP quality maintaining. The linguistic values in use for each variable were: *Buffer Size* {VL, L, M, H, and VH}, Bandwidth Rate {VL, L, AL, BA, AV, AA, BH, H, and VH} *and* New Token {VL, L, BA, AV, AA, H, and VH}

The study in [9] revealed the analysis of the impact of different network QoS parameters on users perceived video QoE for VoD (Video-on-Demand) services. Network parameters in use included: Packet loss rate, Burst packet loss and Jitter. The input linguistic terms involved were Very annoying, slightly annoying, Imperceptible, Annoying and perceptible but not annoying. The output linguistic terms in use were Very annoying, slightly annoying, Imperceptible, Annoying and perceptible but not annoying. This study proposed a methodology based on a fuzzy expert system to objectively estimate the video QoE. To validate the methodology, the developed system was integrated as part of a monitoring tool in an industrial IPTV (Internet Protocol Television) test bed and compared its output with standard Video Quality Monitoring (VQM). The evaluation results show that the proposed video quality estimation method based on fuzzy expert system can effectively measure the network impact on the QoE.

3. MOTIVATION.

In recent years many network analysis models have been developed in terms of quantitative measures. This mechanism is quantitative in nature (use of scale of numbers between 1to5) to represent user perception of QoS. The challenge with this model is where user perception is subjective and not precise thus cannot be clearly measured using quantitative methods. The qualitative model is in early stages of exploration thus little have been done on this research.

Few models exists that measure qualitative network QoE but none incorporated all the four integrity of service parameters. Therefore, this study is inspired to address this gap by presenting an alternative approach of measuring network QoE parameters under integrity of service which measures underlying network QoS related parameters (throughput, packet loss, delay and jitter) by using fuzzy logic concept.

4. OBJECTIVES.

4.1 General Objectives

To design a Fuzzy logic model for Analysis of computer Networks Quality of Experience.

4.2 Specific Objectives

i) To analyze the fuzziness of Network QoE in order to provide more understandable user perception.

ii) To design a fuzzy logic framework for analysis of computer networks Quality of Experience through developing linguistic terms and variables, designing fuzzy membership function for different linguistic terms, designing fuzzy rules ,fuzzification and defuzzification

iii) To develop a fuzzy logic framework for analysis of computer networks Quality of Experience.

iv) To test the performance of network QoE and estimate the variation of user satisfactions level in function of network integrity of service parameters.

v) To implement the developed framework for use by end users.

5. RESEARCH QUESTIONS.

1) What are the tools and techniques to analyze the fuzziness of network QoE in order to provide more understandable user perception?

2) What are the requirements to design a fuzzy logic model for analysis of computer networks Quality of Experience?

3) What are the methodologies to device a method to estimate the variation of the user satisfaction level in function of the network QoS conditions?

4) Which relevant mechanisms are in place to test the performance of network QoE and estimate the variation of user satisfactions level in function of network integrity of service parameters?

5) Which techniques are used to implement the developed framework for use by end users?

6. RESEARCH METHODOLOGY.

This phase covers Research Design, Data Collection procedure involved in the study, Data analysis based on Research Methodology.

6.1 Research Design:

This study focuses on the fuzzy logic model for analysis of computer network quality of experience, a study on integrity of service.

The research approach is experimental Research Design.

This allows the researcher to have complete control over the extraneous variables & can predict confidently that the observed effect on the dependable variable is only due to the manipulation of the independent variable.

Experimental research is often used where there is time priority in a causal relationship (cause precedes effect), there is consistency in a causal relationship (a cause will always lead to the same effect), and the magnitude of the correlation is great.

The causal relationship in this case is between four variables. The variables are underlying network Qos-related parameters (Throughput, delay, packet loss and Jitter) which lay under integrity of service Network QoE parameters whose manipulation will affect the output.

6.2 Data Collection:

In a nutshell, activities involving data can be grouped into Data extraction, Data processing/ cleaning, data presentation and data analysis based on research methodology.

6.2.1 Data Extraction:

Data collection procedure involves an experiment for data extraction from a network setup. Data is extracted from 15 Autonomous Systems (AS)/ network connections for 19 services at an interval of 5minutes using Linux mtr tool. This data is acquired inform of TXT file (.txt) which can be viewed by notepad ++ tool.

There is a vast challenge when it comes to data extraction of network QoE data. These factors ranges from the type of tool to use, the kind of data to acquire, the method to use for data cleansing to make it relevant for use etc.

The command to use in Linux mtr tool to capture the required data for dataset is **mtr -rw -o "DRAM".** For instance,For the case of mtr command to access Gmail's Dropped packets, Received packets, Average RTT/Delay and Jitter Mean/Avg is accessed by below order of fields:

$ mtr -rw -o "DRAM" --aslookup www.gmail.com

Whereby the initials represent:

mtr –My Traceroute.

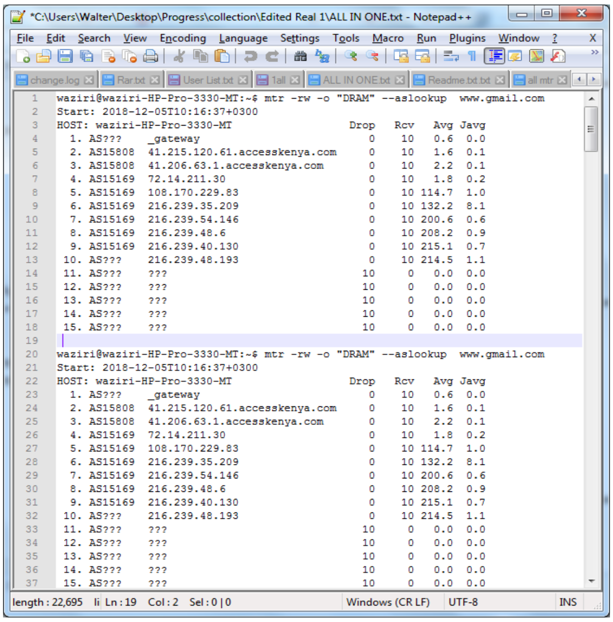
rw- Report wide mode. Without the **--report option**, mtr will run continuously in an interactive environment. The interactive mode reflects current round trip times to each host. In most cases, the --report mode provides sufficient data in a useful format.

O-In the order of e.g. In the order of DRAM.

DRAM- Dropped packets, Received packets, Average RTT/Delay and Jitter Mean/Avg respectively.

aslookup will be used to display Autonomous systems.

This will produce the following output:


*C:\Users\Walter\Desktop\Progress\collection\Edited Real 1\ALL IN ONE.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ? X
change.log Rar.txt User List.txt 1all ALL IN ONE.txt Readme.txt.txt all mtr
length : 22,695 li Ln : 19 Col : 2 Sel : 0 | 0 Windows (CR LF) UTF-8 INS

```
waziri@waziri-HP-Pro-3330-MT:~$ mtr -rw -o "DRAM" --aslookup  www.gmail.com
Start: 2018-12-05T10:16:37+0300
HOST: waziri-HP-Pro-3330-MT                   Drop   Rcv   Avg Javg
  1. AS???    _gateway                           0    10   0.6  0.0
  2. AS15808  41.215.120.61.accesskenya.com      0    10   1.6  0.1
  3. AS15808  41.206.63.1.accesskenya.com        0    10   2.2  0.1
  4. AS15169  72.14.211.30                       0    10   1.8  0.2
  5. AS15169  108.170.229.83                     0    10 114.7  1.0
  6. AS15169  216.239.35.209                     0    10 132.2  8.1
  7. AS15169  216.239.54.146                     0    10 200.6  0.6
  8. AS15169  216.239.48.6                       0    10 208.2  0.9
  9. AS15169  216.239.40.130                     0    10 215.1  0.7
 10. AS???    216.239.48.193                     0    10 214.5  1.1
 11. AS???    ???                               10     0   0.0  0.0
 12. AS???    ???                               10     0   0.0  0.0
 13. AS???    ???                               10     0   0.0  0.0
 14. AS???    ???                               10     0   0.0  0.0
 15. AS???    ???                               10     0   0.0  0.0

waziri@waziri-HP-Pro-3330-MT:~$ mtr -rw -o "DRAM" --aslookup  www.gmail.com
Start: 2018-12-05T10:16:37+0300
HOST: waziri-HP-Pro-3330-MT                   Drop   Rcv   Avg Javg
  1. AS???    _gateway                           0    10   0.6  0.0
  2. AS15808  41.215.120.61.accesskenya.com      0    10   1.6  0.1
  3. AS15808  41.206.63.1.accesskenya.com        0    10   2.2  0.1
  4. AS15169  72.14.211.30                       0    10   1.8  0.2
  5. AS15169  108.170.229.83                     0    10 114.7  1.0
  6. AS15169  216.239.35.209                     0    10 132.2  8.1
  7. AS15169  216.239.54.146                     0    10 200.6  0.6
  8. AS15169  216.239.48.6                       0    10 208.2  0.9
  9. AS15169  216.239.40.130                     0    10 215.1  0.7
 10. AS???    216.239.48.193                     0    10 214.5  1.1
 11. AS???    ???                               10     0   0.0  0.0
 12. AS???    ???                               10     0   0.0  0.0
 13. AS???    ???                               10     0   0.0  0.0
 14. AS???    ???                               10     0   0.0  0.0
 15. AS???    ???                               10     0   0.0  0.0
```

Figure 4.12: Mtr output in Ubuntu Linux platform.

Each numbered line in the report represents a hop. Hops are the Internet nodes that packets pass through to get to their destination. They are also referred to as Network Connections/Autonomous systems in the network initialized by **"AS".**

In the case where we have **"AS???"** in the Autonomous systems, it's an indication that:

The question marks appear when there is no additional route information.

Sometimes as a result of a poorly configured router will send packets in a loop.

6.2.2 Data Processing/Data Cleaning:

The acquired data from Linux Mtr tool is exported to Ms. Excel for cleaning/ processing. The processed data is obtainable in form of excel workbook format (.xlsx) each having crisp values for Delay, jitter, packet loss, throughput and users dataset.

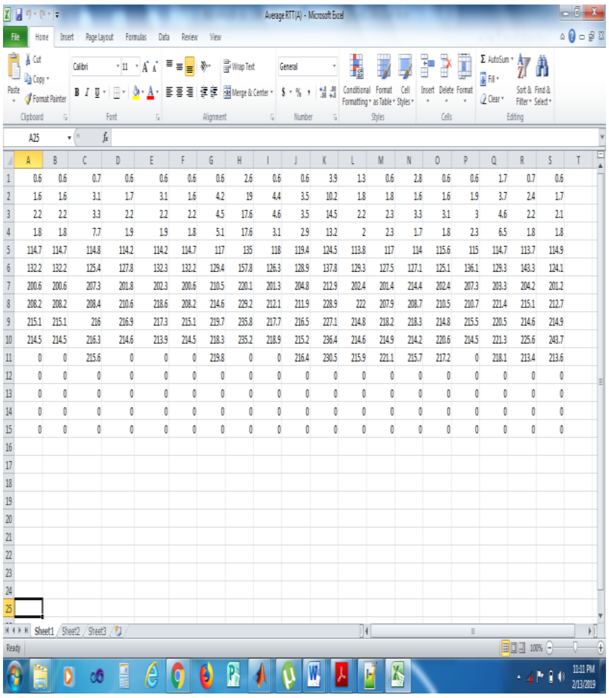
6.2.3 Data Presentation:

Each column field from mtr txt data file is given a different excel workbook for different dataset.

This results into four workbook in excel each having packet loss, Throughput, Delay, and jitter datasets as shown below. The fifth extra workbook will be used to display the User List i.e. Network Connections/Autonomous systems in the network.

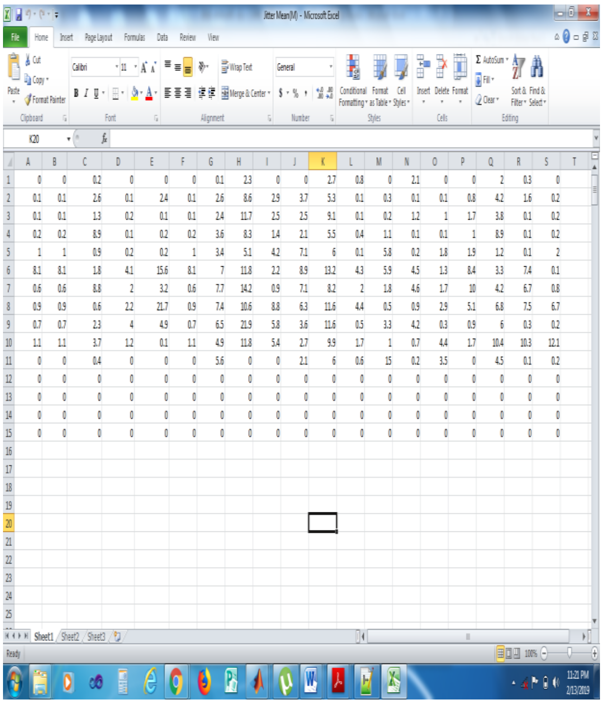
The obtained crisp values for Delay, jitter, packet loss and throughput are used as input dataset for developed framework.

The values are organized into tables each arranged in a matrix of 15 *19 as shown below to indicate the experiment was carried out on 15 autonomous systems/network connections, 19 times at an interval of 5 minutes each.



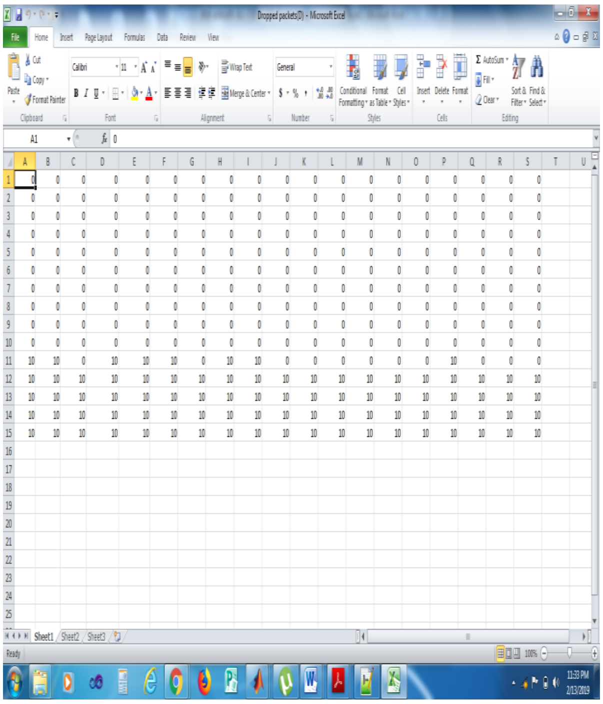
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	0.6	0.6	0.7	0.6	0.6	0.6	0.6	2.6	0.6	0.6	3.9	1.3	0.6	2.8	0.6	0.6	1.7	0.7	0.6
2	1.6	1.6	3.1	1.7	3.1	1.6	4.2	19	4.4	3.5	10.2	1.8	1.8	1.6	1.6	1.9	3.7	2.4	1.7
3	2.2	2.2	3.3	2.2	2.2	2.2	4.5	17.6	4.6	3.5	14.5	2.2	2.3	3.3	3.1	3	4.6	2.2	2.1
4	1.8	1.8	7.7	1.9	1.9	1.8	5.1	17.6	3.1	2.9	13.2	2	2.3	1.7	1.8	2.3	6.5	1.8	1.8
5	114.7	114.7	114.8	114.2	114.2	114.7	117	135	118	119.4	124.5	113.8	117	114	115.6	115	114.7	113.7	114.9
6	132.2	132.2	125.4	127.8	132.3	132.2	129.4	157.8	126.3	128.9	137.8	129.3	127.5	127.1	125.1	136.1	129.3	143.3	124.1
7	200.6	200.6	207.3	201.8	202.3	200.6	210.5	220.1	201.3	204.8	212.9	202.4	201.4	214.4	202.4	207.3	203.3	204.2	201.2
8	208.2	208.2	208.4	210.6	218.6	208.2	214.6	229.2	212.1	211.9	228.9	222	207.9	208.7	210.5	210.7	221.4	215.1	212.7
9	215.1	215.1	216	216.9	217.3	215.1	219.7	235.8	217.7	216.5	227.1	214.8	218.2	218.3	214.8	215.5	220.5	214.6	214.9
10	214.5	214.5	216.3	214.6	213.9	214.5	218.3	235.2	218.9	215.2	236.4	214.6	214.9	214.2	220.6	214.5	221.3	225.6	243.7
11	0	0	215.6	0	0	0	219.8	0	0	216.4	230.5	215.9	221.1	215.7	217.2	0	218.1	213.4	213.6
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.3: Processed data for Delay (A)



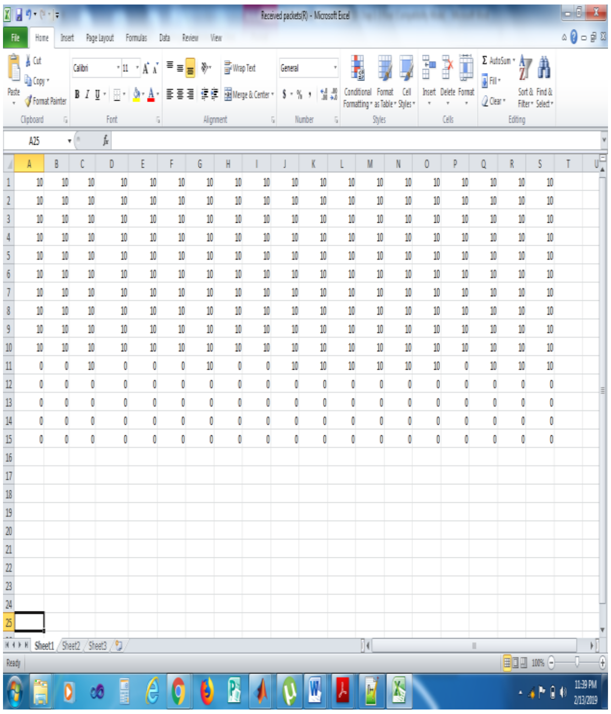
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	0	0	0.2	0	0	0	0.1	2.3	0	0	2.7	0.8	0	2.1	0	0	2	0.3	0
2	0.1	0.1	2.6	0.1	2.4	0.1	2.6	8.6	2.9	3.7	5.3	0.1	0.3	0.1	0.1	0.8	4.2	1.6	0.2
3	0.1	0.1	1.3	0.2	0.1	0.1	2.4	11.7	2.5	2.5	9.1	0.1	0.2	1.2	1	1.7	3.8	0.1	0.2
4	0.2	0.2	8.9	0.1	0.2	0.2	3.6	8.3	1.4	2.1	5.5	0.4	1.1	0.1	0.1	1	8.9	0.1	0.2
5	1	1	0.9	0.2	0.2	1	3.4	5.1	4.2	7.1	6	0.1	5.8	0.2	1.8	1.9	1.2	0.1	2
6	8.1	8.1	1.8	4.1	15.6	8.1	7	11.8	2.2	8.9	13.2	4.3	5.9	4.5	1.3	8.4	3.3	7.4	0.1
7	0.6	0.6	8.8	2	3.2	0.6	7.7	14.2	0.9	7.1	8.2	2	1.8	4.6	1.7	10	4.2	6.7	0.8
8	0.9	0.9	0.6	2.2	21.7	0.9	7.4	10.6	8.8	6.3	11.6	4.4	0.5	0.9	2.9	5.1	6.8	7.5	6.7
9	0.7	0.7	2.3	4	4.9	0.7	6.5	21.9	5.8	3.6	11.6	0.5	3.3	4.2	0.3	0.9	6	0.3	0.2
10	1.1	1.1	3.7	1.2	0.1	1.1	4.9	11.8	5.4	2.7	9.9	1.7	1	0.7	4.4	1.7	10.4	10.3	12.1
11	0	0	0.4	0	0	0	5.6	0	0	2.1	6	0.6	15	0.2	3.5	0	4.5	0.1	0.2
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.4: Processed data for Jitter Mean (M)



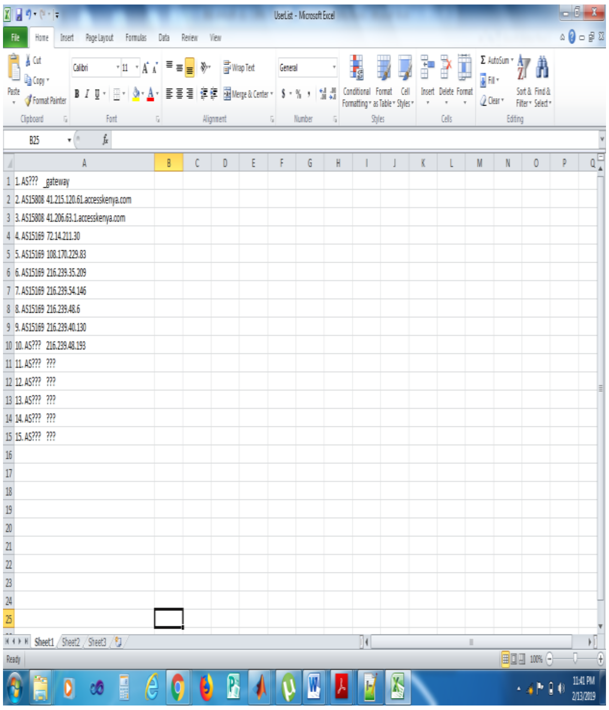
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	10	10	0	10	10	10	0	10	10	0	0	0	0	0	0	10	0	0	0
12	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
13	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
14	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
15	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10

Figure 3.5: Processed data for Packet Loss/Dropped packets (D)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
3	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
4	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
5	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
7	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
8	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
9	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
11	0	0	10	0	0	0	10	0	0	10	10	10	10	10	10	0	10	10	10
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.6: Processed data for Received Packets (R)



1. AS??? _gateway
2. AS15808 41.215.120.61.accesskenya.com
3. AS15808 41.206.63.1.accesskenya.com
4. AS15169 72.14.211.30
5. AS15169 108.170.229.83
6. AS15169 216.239.35.209
7. AS15169 216.239.54.146
8. AS15169 216.239.48.6
9. AS15169 216.239.40.130
10. AS??? 216.239.48.193
11. AS??? ???
12. AS??? ???
13. AS??? ???
14. AS??? ???
15. AS??? ???

Figure 3.7: User List (Autonomous systems/Network Connections)

6.3 Data Analysis Based on Research Methodology:

The acquired data from designed experiment is used to construct the proposed framework of fuzzy logic model for analysis of computer network quality of experience. The data is used as input variables data of the framework to make rational analysis based on the membership function and fuzzy rules.

Data analysis is guided by fuzzy logic Methodology framework design process/guidelines. The methodology is grouped into four phases: Initialization, Fuzzification, Inferencing and defuzzification.

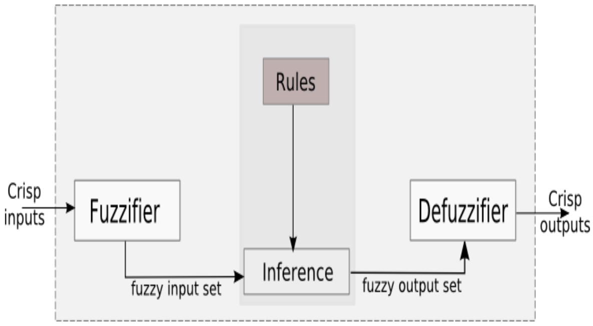


Figure 3.3: Fuzzy Logic System [13].

6.3.1. Initialization:

This phase comprises of defining the linguistic variables and terms, constructing the membership functions and constructing the rule base.

Five linguistic terms are identified for use i.e. “Excellent”, “Good”, “Fair”, “Poor”, and “bad”. Four variables for network integrity of service QoE parameters are identified i.e. Throughput, Delay, Delay variation/Jitter and packet loss. These are the primary factor for QoS quantification of any network according to [14].

Constructing the membership functions (MF) is done at this stage by determining a curve that outlines how each point in the input universe is plotted to a membership value (or degree of membership) between 0 and 1.A triangular membership function is used to acquire the notch of membership for each linguistic term because of its computational efficiency.

Moreover, the initialization phase involves constructing the rule base. The identified Five linguistic terms and the Four variables for network integrity of service QoE parameters results into 625 rules(5^4).The rules are further dropped to 240 rules basing on expert knowledge by discarding the illogical rules thus remaining with logical rules to make rational decisions.

The illogical is as a result whereby some conditions cannot exist at the same time for instance in rule 1 of the 625 rules indicates:

1. If delay is very low, jitter is very low, packet loss is very low and throughput is very low then User Satisfaction N/A.

This rule is N/A thus illogical since when delay, jitter and packet loss are very low then throughput is supposed to be high or very high in ideal network situation as these three variables which are supposed to make the throughput very low, their existence too are very low not to certain levels to affect the network throughput to match being very low.

6.3.2 Fuzzification:

A crisp set (subset elements of the set, definitely do belong to the set), of input unprocessed information is assembled and transformed to a fuzzy set (**sets** whose elements have degrees of membership) by using fuzzy linguistic variables, membership functions and fuzzy linguistic terms through **fuzzification** [3]. This is achieved by Fuzzifier component of the fuzzy Logic System.

6.3.3 *Inference:*

This stage involves evaluating the rules in the rule base. Each rule follows the order to fulfill certain condition. The logical 240 rules are interpreted one after the other. This is achieved by Fuzzy Inference system component of the Fuzzy Logic System. In this work, Mamdani fuzzy inference system is used to achieve the inferencing in the developed framework.

The Fuzzy set operator "AND" is used to aggregate the output of each rule. The results of each rule are combined at this phase. The matched fuzzy rules are then used in the defuzzification process.

The properties of logical "AND" and "OR" makes the result of a logical expression to be sometimes fully determined before evaluating all of the conditions.

The logical operator "AND" is selected for connecting the inputs in this experiment since the operator returns logical 0 (false) if even a single condition in the expression is false in an ideal situation. For instance in one of the 625 rules:

Example 1:

If delay is very low, jitter is very low, packet loss is very low and throughput is very high then User Satisfaction EXCELLENT:

In an ideal situation, when delay, jitter and packet loss are very low then throughput is very high as the network suffers no hitches thus resulting to user satisfaction being excellent.

6.3.4 *Defuzzification Of the Output:*

The linguistic variables & terms are matched, fuzzy rules generated and the output results obtained for each parameter are aggregated into one crispvalue through **defuzzification**.

Five linguistic terms considered at this phase include "Excellent", "Good", "Fair", "Poor", and bad. The linguistic terms are quantified on a numerical scale on a range of 5,4,3,2 and 1 respectively whereby the higher the value, the better the QoE and the lower the value the worse the QoE.

This process involves producing a quantifiable result in Crisp logic, given fuzzy sets and corresponding membership degrees.

Moreover this process maps a fuzzy set to a crisp set. It is typically needed in fuzzy control systems. These will have a number of rules that transform a number of variables into a fuzzy result, that is, the result is described in terms of membership in fuzzy sets [13]. This is the purpose of the defuzzifier component of a FLS.

In this work, weighted average method is the defuzzification technique to use because of its computational efficiency.

Graphical tools to build, edit and view fuzzy inference systems:

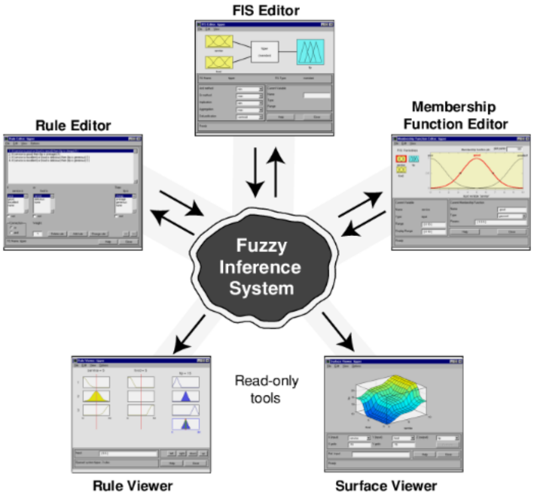


Figure 4.2: Graphical tools for fuzzy inference systems.

7. ANALYSIS/ DISCUSSIONS OF THE RESULTS.

Unit of analysis is based on the objectives achieved. This research achieved the following specific objectives:

7.1 To Analyze the Fuzziness Of Network QoE In Order To Provide More Understandable User Perception:

This objective was achieved by demonstrating the fact that user satisfaction is not precise rather its subjective in nature thus use of linguistic terms to categorize the fuzzy sets in terms of the level of user satisfaction : Very Low, Low, High, Medium, and Very High. The use of membership function to determine the degree of relationship between the linguistic terms and the linguistic variables demonstrates fuzziness i.e. all information in fuzzy set whether the elements in fuzzy sets are discrete or continuous.

Moreover there is no specific value to determine User satisfaction level thus the use of Triangular membership function was effective to manipulate the range of values for each record set for instance “GOOD” Network QoE Membership function has a range of values from [2.5 3.75 5] whereby 2.5 is the lowest value that someone can rate the QoE as Good, 3.75 is the mean value while 5 is the highest value for this membership.

Likewise, the use of “IF THEN” statement in fuzzy rules to join the different imprecise input parameters in order to obtaining a single output demonstrates the fuzziness in network QoE. The rules for defining fuzziness are fuzzy too. Furthermore, the need to use either AND, OR and NOT operators of Boolean logic that exists in fuzzy logic for the purpose of manipulation of different input values in order to obtain an output value demonstrates the fuzziness of network QoE. For instance in this work, AND operator is used to aggregate the fuzzy rule outputs.

7.2 To Design a Fuzzy Logic Framework For Analysis Of Computer Networks Quality Of Experience Through Developing Linguistic Terms & Variables, Designing Fuzzy Membership Function For Different Linguistic Terms, Designing Fuzzy Rules:

This activity achieved the following results: identification of five input linguistic terms: Very Low, Medium, Low, High and very High), Four input variables (Delay, Jitter, Throughput and packet Loss), five output linguistic term (Excellent, Good, Fair, Poor and Bad), designing Triangular membership function for different linguistic terms, designing 625 fuzzy rules (5^4) which were later reduced to 240 logical rules to be fired in the experiment.

Fuzzy rules operate using a sequence of if-then statements. For instance in this work: if delay is very low, jitter is very low, packet loss is very low and throughput is high then User Satisfaction GOOD.

The AND, OR, and NOT operators of Boolean logic exist in fuzzy logic, usually defined as the minimum, maximum, and complement; when they are defined this way, they are called the *Zadeh operators*, because they were first defined as such in Zadeh's original papers [4] In this work, AND operator is used to aggregate the fuzzy set values in order to acquire the output.

From the membership functions, we can calculate the truth value of each fuzzy proposition and of the fuzzy conjunction the minimum degree of membership is taken as the output when AND operator is used. In a nutshell, the design stage happened to design the following components:

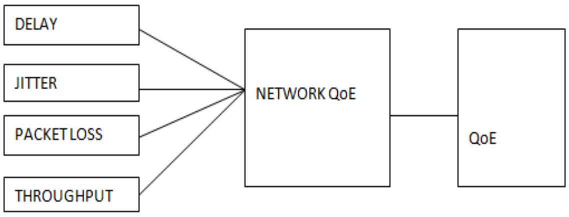


Figure 4.10: Designed Network QoE framework.

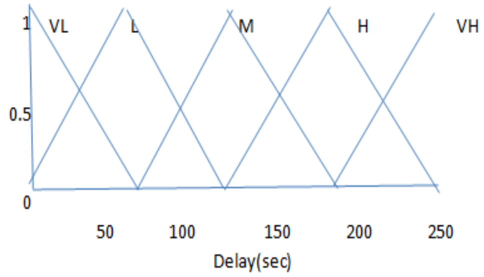


Figure 4.11: Designed Triangular membership function for Delay input linguistic term.

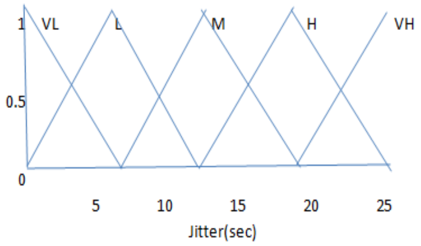


Figure 4.12: Designed Triangular membership function for Jitter input linguistic term.

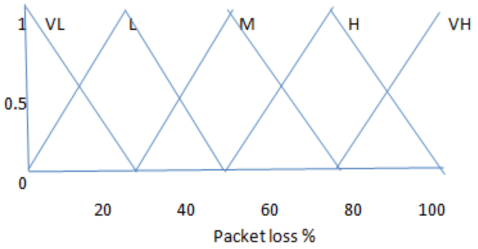


Figure 4.13: Designed Triangular membership function for Packet loss input linguistic term.

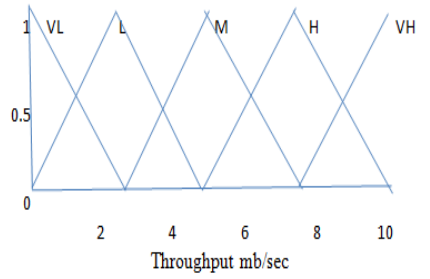


Figure 4.14: Designed Triangular membership function for Throughput input linguistic term.

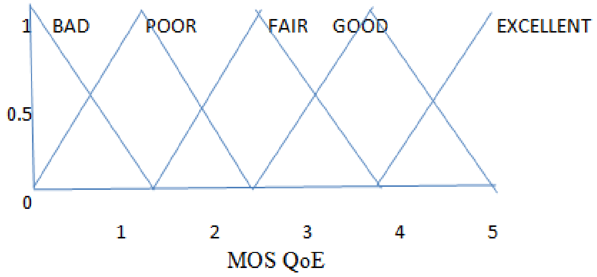


Figure 4.15: Designed Triangular membership function for different output linguistic terms.

7.3 To Develop a Fuzzy Logic Framework For Analysis Of Computer Networks Quality Of Experience:

The following was achieved in this activity: The designed framework was developed by use of fuzzy logic methodology. At this juncture, whatever was designed in the design phase is being developed in matlab environment.

In this scenario, the collected crisp values data are converted into fuzzy sets through fuzzification. The output fuzzy sets are further converted into a single crisp value through weighted average defuzzication technique. The acquired value is used for analysis of computer networks Quality of Experience.

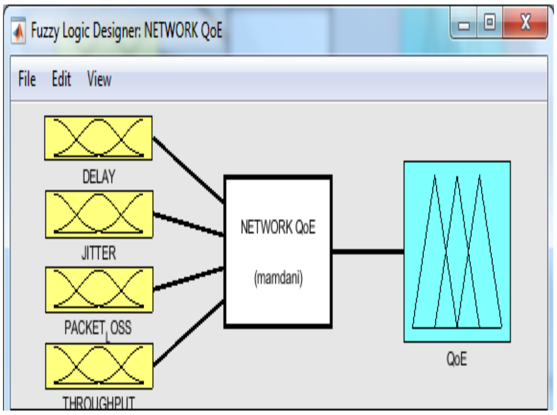


Figure 4.16: Developed Network QoE framework.

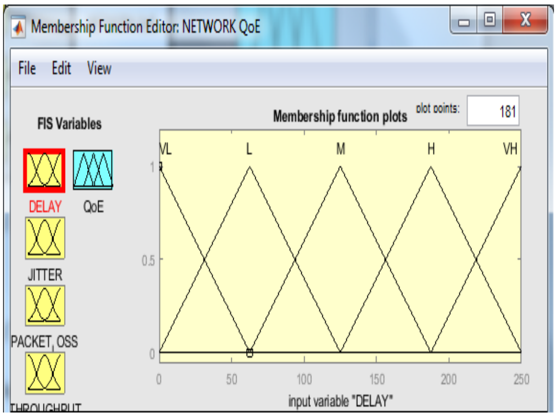


Figure 4.17: Developed membership function plots for Delay input linguistic term.

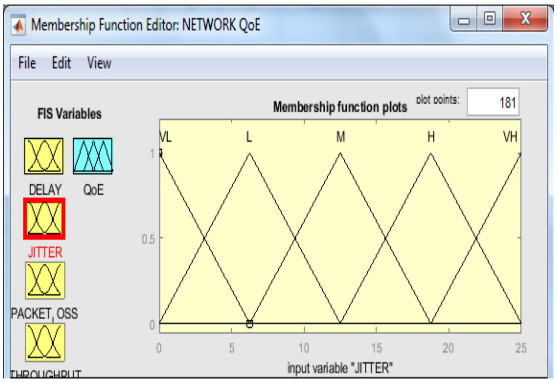
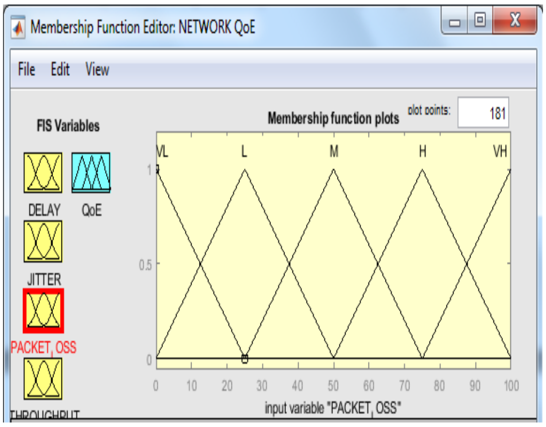


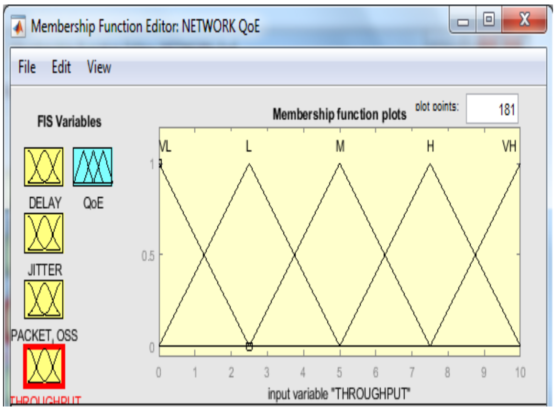
Figure 4.18: Developed membership function plots for Jitter input linguistic term.



plot points: 181

input variable "PACKET_OSS"

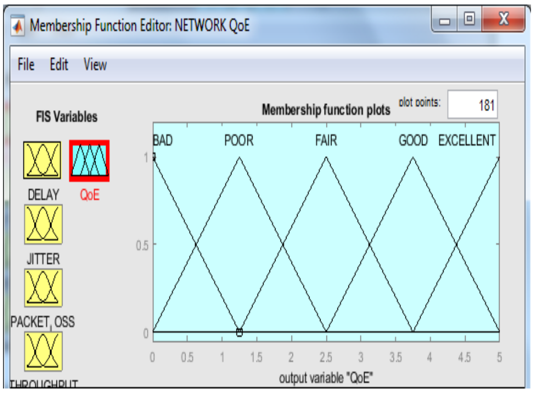
Figure 4.19: Developed membership function plots for Packet loss input linguistic term.



plot points: 181

input variable "THROUGHPUT"

Figure 4.20: Developed membership function plots for Throughput input linguistic term.



plot points: 181

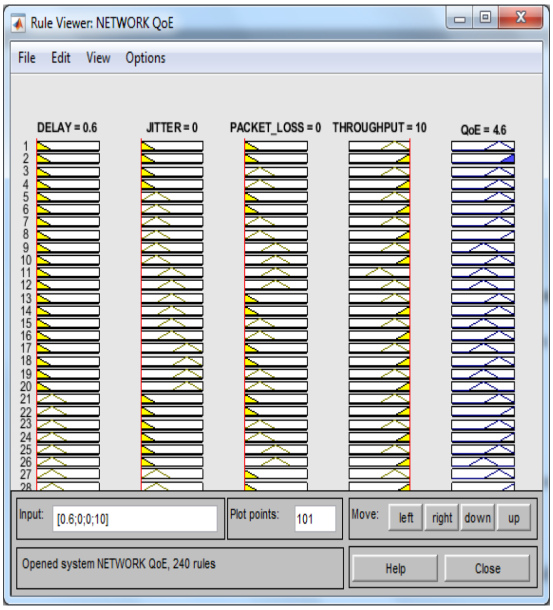
output variable "QoE"

Figure 4.21: Developed membership function plots for different output linguistic terms.

7.4 To Test the Performance Of Network QoE And Estimate the Variation Of User Satisfactions Level In Function Of Network Integrity Of Service Parameters:

This objective was achieved by testing the developed framework under different input crisp values of network integrity of service parameter values while observing the performance of the framework based on the obtained QoE values.

This activity was performed in the Matlab's Rule Viewer as shown below.



DELAY = 0.6
JITTER = 0
PACKET_LOSS = 0
THROUGHPUT = 10
QoE = 4.6

Input: [0.6;0;0;10]

Plot points: 101

Opened system NETWORK QoE, 240 rules

Figure 4.22: Developed Framework Testing environment in the rule viewer.

7.5 To Implement the Developed Framework For Use By End Users:

This activity was achieved by the help of "guide" command in matlab to develop a Graphical User Interface (GUI).Once the command is run, it prompts for an option to either create GUI or Opening an existing GUI. In this scenario, there was need to create a new GUI which when saved into the selected folder it resulted into two file formats i.e. .fig (to access the underlying objects in the figure) and .m file formats (To indicate MATLAB code is in files with extension .M)

This objective was achieved by developing a framework that is executable having the capability to be installed in computing devices and perform network QoE analysis based on the inputted data as shown below:

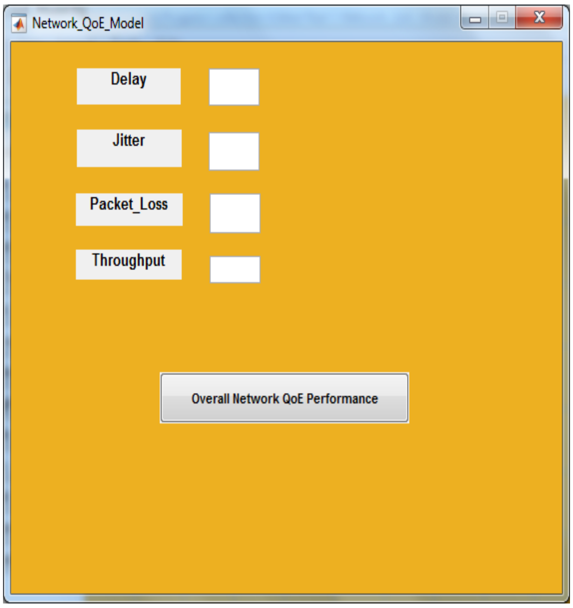


Figure 4.23: Developed User interface.

8. CONCLUSION:

Based on summary of work done in this research, in conclusion this research work successfully designed, developed, tested and implemented a computer networks QoE framework based on fuzzy logic methodology. The framework analyses the QoS provided by the service providers as perceived by the end users to be used for decision making in order to achieve QoS as per service level agreement. For this scenario, the model was intended for Internet Service Providers to analyze the products & services they deliver to their clients including internet services though the framework can be customized to suite other product/service provider industries. The framework allows users to use qualitative method (Fuzzy Logic concepts) instead of quantitative method (MOS) for analysis of computer networks QoE. This allows vagueness and subjectivity nature into the developed framework.

This work intensified in the underlying QoS-related parameters, which are linked to the integrity of service QoE parameters as the area of study. These are the primary factor for QoS quantification of any network [14].

9. RECOMMENDATIONS.

Based on the conclusion, it is greatly recommended to adopt Frameworks that allows users to use qualitative methods for instance Fuzzy Logic concepts that have capabilities to accept vague and subjective values for analysis and decision making based on certain concepts or methodology.

Quantitative methods for instance use of Mean Opinion Score for analysis of QoE has a drawback in that user satisfaction is not precise rather it’s subjective in nature hence difficult to be quantified using such methods.

10. KNOWLEDGE CONTRIBUTION TO THE FIELD OF STUDY.

Based on this research work, some of the accomplished tasks have contributed some knowledge in this field of study. This is clearly outlined whereby, basing on previous studies, few models exists that measure qualitative analysis of computer network quality of experience. However none incorporated all the four parameters of integrity of service; throughput, delay, packet loss and jitter as parameters of network QoE. The study’s objective is to address this gap by developing a fuzzy logic model for analysis of computer network QoE basing on all the four parameters for integrity of service to check for efficiency of the model.

Moreover, the experiment phase happened to gather relevant data using Linux mtr tool. This data was the hardest part to crack through, thus hindering many research works to be completed. Data to do with networks can be best acquired from service providers. Unluckily, this data can’t be rendered to the public for use or analysis due to security reasons and privacy issues. Therefore, having captured the live data through experiment, at this juncture I can avail it in public domain using various sites for instance Git-hub to assist other researchers to accomplish their work too in this field of study.

11. FUTURE WORK/FURTHER GAPS IN RELATED RESEARCH.

Future work in this work can be accomplished by evaluating the best Fuzzy inference system applicable to network QoE analysis for instance to compare the output results obtained by both Mamdani and sugeno fuzzy inference systems.

Based on the conclusion, this work intensified in the underlying QoS-related parameters, which are linked to the integrity of service QoE parameters as the area of study. These are the primary factors for QoS quantification of any network [14]. Since the Accesibility and retainability QoE parameters have not been tackled, future work can incorporate their respective underlying QoS-related parameters into the model to test for efficiency of the model.The respective underlying parameters are represented below;

TABLE 2.1: Mapping between QoE and QoS Related parameters [15]

QoE parameters	**Underlying QoS-related parameters**
Accesibility	Unavailability Security Activation Access Coverage Blocking Setup time
Retainability	Connection loss
Integrity of Service	Throughput Delay Delay variation/Jitter Packet loss

12. ACKNOWLEDGEMENT

Firstly, I would like to acknowledge God, the creator of heaven and earth for the gift of wisdom to think towards this research work. I would like to express my deep gratitude to my thesis supervisors; Dr. Calvin Otieno (PhD), for his patient guidance, enthusiastic encouragement and keeping my progress on schedule. Dr. Richard Rimiru (PhD), for his advice, assistance in keeping my progress on schedule and useful critiques of this research work.

My grateful thanks are also extended to Mr.George Ochieng, the Chair of IT department at JKUAT Mombasa CBD for ensuring we are exposed to several presentation seminars as one of the requirements before being awarded with master's degree, to Mr.Abdulrehman Ahmed who helped me in documenting my work and provision of technical advice and to Mrs. Amina Omar for her support in Matlab environment. My sincere appreciation also goes to all my lecturers who took me through this journey of academics, support staff including the Librarian for provision of access to research materials.

I would also like to extend my thanks to Technical University of Mombasa (TUM) laboratory technician, Mr.Aggrey Shitsukane for his inclusive support for provision of access to Laboratory environment at TUM in order to carry out my research work with precision.

Finally, I wish to thank my Family and Friends for their support and encouragement throughout my study. May God bless you all and guide you with these wise Quotes;

Start where you are. Use what you have. Do what you can. – Arthur Ashe. Moreover, there are no traffic jams on the extra mile. – Zig Ziglar.

13. REFERENCES

[1] K. Brunnstr *et al.*, "Qualinet White Paper on Definitions of Quality of Experience Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting , Novi Sad ," 2014.

[2] ITU, "P.800: Methods for subjective determination of transmission quality," *ITU-T Recomm.*, vol. 800, 1996.

[3] L. A. Zadeh, I. Introduction, and U. S. Navy, "Fuzzy Sets * -," vol. 353, pp. 338–353, 1965.

[4] A. S. Omar, M. Waweru, and R. Rimiru, "Application of Fuzzy Logic in Qualitative Performance Measurement of Supply Chain Management," vol. 5, no. 6, 2015.

[5] M. Madhoushi and A. N. Aliabadi, "Environmental Performance Evaluation Based on Fuzzy Logic," *Int. J. Appl. Sci. Technol.*, vol. 1, no. 5, pp. 432–436, 2011.

[6] E. Olugu, "Supply Chain Performance Evaluation : Trends and Challenges Supply Chain Performance Evaluation : Trends and Challenges," no. January 2009, 2016.

[7] R. Hawi, G. Okeyo, and M. Kimwele, "Techniques for Smart Traffic Control : An In-depth Review," vol. 4, no. 7, pp. 566–573, 2015.

[8] A. Hamam, M. Eid, A. El Saddik, and N. D. Georganas, "A Fuzzy Logic System for Evaluating Quality of Experience of Haptic-Based Applications," no. October 2016, 2008.

[9] J. Pokhrel, "Intelligent quality of experience (QoE)

analysis of network served multimedia and web contents Analyse intelligente de la qualité d ' expérience (QoE) dans les réseaux de diffusion de contenu Web et Multimédia," 2015.

[10] S. M. Ataeian and M. J. Darbandi, "Analysis of Quality of Experience by applying Fuzzy logic A study on response time," no. June, 2011.

[11] O. F. W. Onifade, "Better Quality of Service Management With Fuzzy Logic In Mobile Adhoc Network," vol. 6, no. 1, pp. 59–68, 2013.

[12] M. E. A. Ebrahim and H. A. Hefny, "Fuzzy Logic based Approach for VoIP Quality Maintaining," vol. 9, no. 1, pp. 537–542, 2018.

[13] P. Singhala, D. N. Shah, and B. Patel, "Temperature Control using Fuzzy Logic," vol. 4, no. 1, pp. 1–10, 2014.

[14] Y. Chen, T. Farley, and N. Ye, "QoS Requirements of Network Applications on the Internet," vol. 4, pp. 55–76, 2004.

[15] F. Farid, S. Shahrestani, and C. Ruan, "A Fuzzy Logic Approach for Quality of Service Quantification in Wireless and Mobile Networks," pp. 629–636, 2014.