nigeria computer society
ncs
www.ncs.org.ng

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

PREDICTION OF PEDIATRIC HIV/AIDS SURVIVAL IN NIGERIA USING NAÏVE BAYES' APPROACH

P.A. Idowu[1], T. A. Aladekomo[2], O. Agbelusi[3], and C. O. Akanbi[4]
[1,2] *Obafemi Awolowo University, Ile-Ife, Nigeria,*
[3] *Rufus Giwa Polytechnic, Owo, Nigeria*
[4] *Osun State University, Nigeria*
[1]paidowu1@yahoo.com , [4]*akanbico@uniosun.edu.ng*

ABSTRACT
Epidemic diseases have highly destructive effects around the world and these diseases have affected both developed and developing nations. Disease epidemics are common in developing nations especially in Sub Saharan Africa in which Human Immunodeficiency Virus /Acquired Immunodeficiency Disease Syndrome (HIV/AIDS) is the most serious of all. This paper presents a prediction of pediatric HIV/AIDS survival in Nigeria. Data are collected from 216 pediatric HIV/AIDS patients who were receiving antiretroviral drug treatment in Nigeria was used to develop a predictive model for HIV/AIDS survival based on identified variables. Interviews were conducted with the virologists and pediatricians to identify the variables predictive for HIV/AIDS survival in pediatric patients. 10-fold cross validation method was used in performing the stratification of the datasets collected into training and testing datasets following data preprocessing of the collected datasets. The model was formulated using the naïve Bayes' classifier – a supervised machine learning algorithm based on Bayes' theory of conditional probability and simulated on the Waikato Environment for Knowledge Analysis [WEKA] using the identified variables, namely: CD4 count, viral load, opportunistic infection and the nutrition status of the pediatric patients involved in the study. The results showed 81.02% accuracy in the performance of the naïve Bayes' classifier used in developing the predictive model for HIV/AIDS survival in pediatric patients. In addition, the area under the receiver operating characteristics [ROC] curve had a value of 0.933 which showed how well the developed predictive model was able to discriminate between survived and non-survived cases. Model validation was performed by comparing the model results with that of historical data from two (2) selected tertiary institutions in Nigeria.

Keywords: HIV/AIDS survival, naïve Bayes' classifier, predictive model, pediatrics, machine learning

1.0 **Introduction**

Epidemic diseases have highly destructive effects around the world and these diseases have affected both developed and developing nations. Disease epidemics are common in developing nations especially in Sub Saharan Africa in which Human Immunodeficiency Virus /Acquired Immunodeficiency Disease Syndrome (HIV/AIDS) is the most serious of all [1]. HIV is one of the world's most serious health and development challenges [2]. It is a type of virus called a retrovirus which infects humans when it comes in contact with tissues such as those that line the vagina, anal area,

mouth, eyes or through a break in the skin [3], while Acquired Immunodeficiency Syndrome (AIDS) is the advanced stage of the retroviral infection that swept through sub-Saharan Africa with venom [4-5].

Globally, HIV continues to be a very serious health issue facing the world [6]. About 34 million (31.4 million–35.9 million) people were living with HIV at the end of 2011 and an estimated0.8% of adults aged 15-49 years worldwide are living with the virus, although the burden of the epidemic continues to vary considerably between countries and regions [6]. In Sub-Saharan Africa, roughly 25 million people were living with HIV in 2012, accounting for nearly 70 percent of the global total. The epidemic has had widespread social and economic consequences, not only in the health sector but also in education, industry and the wider economy [8-9]. The epidemic has had a heavy impact on education, school attendance drops as children become sick or return home to look after affected family members[10]. Moreover, Sub-Saharan Africa remains the most severely affected, with nearly 1 in every 20 adults (4.9%) living with HIV; accounting for 69% of the people living with HIV worldwide. Although, the regional prevalence of HIV infection is nearly 25 times higher in sub-Saharan Africa than in Asia, almost 5 million people are living with HIV in South, South-East and East Asia combined and sub-Saharan Africa region is the most heavily affected region follow by the Caribbean, Eastern Europe and Central Asia, where 1.0% of adults were living with HIV as at 2011 [2].

Nigeria is the most populous nation in Africa with an estimated population of over 160 million people. Government reports claim that over 300,000 Nigerians die yearly of complications arising from AIDS. Nigeria has the highest HIV populations in Africa with 5.7 million infected people. It is estimated that over 200,000 people die yearly in Nigeria as a result of HIV/AIDS [11]. At present, there is no cure for HIV but it is being managed with antiretroviral drugs [ARV]. There is optimal combination of ARV which is known as Highly Active Antiretroviral drug [HAART] [12-13]. Antiretroviral therapy is the mechanism of treating retroviral infections with drugs. The drugs do not kill the virus but they slow down the growth of the virus [14]. HAART refers to the use of combinations of various antiretroviral drugs with different mechanisms of action to treat HIV.

The epidemic of HIV/AIDS affects two classes of people: the paediatric and the non-paediatric individual. The non-paediatric patients are patients above 15 years of age while the paediatric patients who form the main target of this research are patients whose age is less than 15 years [15]. There are four distinct stages of HIV infection which includes: the primary HIV infection stage or clinical stage 1 which involves asymptomatic and acute retroviral syndrome ; clinically asymptomatic stage or clinical stage 2 which involves moderate and unexplained weight loss [<10% of presumed or measured body weight; symptomatic stage or clinical stage 3 is also a conditions where a presumptive diagnosis could be made on the basis of clinical signs or simple investigations like unexplained chronic diarrhea for longer than one month, unexplained persistent fever [intermittent or constant for longer than one month] and progression or Clinical stage 4 that is a condition where a presumptive diagnosis can be made on the basis of clinical signs or simple investigations like HIV wasting syndrome, pneumocystis pneumonia, recurrent severe or radiological bacterial pneumonia etc. Patient at this stage is in a condition where confirmatory diagnostic testing is necessary [16-17].

There are different modes of transmission of this virus one of which is mother-to-child transmission. Here, about nine out of ten children exposed are infected with HIV during pregnancy, labour, delivery or while breastfeeding [18]. Without treatment, 15%-30% of babies born to HIV positive women are infected with the virus during pregnancy and delivery and a further 5%-20% are also infected through breastfeeding [19]. In high-income countries, preventive measures are undertaken to ensure that the transmission of HIV from mother-to-child is relatively rare and in cases where it

occurs, a range of treatment options are undertaken so that the child can survive into adulthood. Blood transfusion is another route in which HIV infection can occur in medical setting [20].

HIV epidemic in Nigeria is one of the incurable deadly diseases and it varies widely by region [21]. The impact of HIV/AIDS is pervasive and far-reaching, affecting individuals and communities not only psychologically but also economically and socially. Families lose their most productive members to this disease, leaving children and elderly people without means of support [22]. Despite the state of this deadly disease in Nigeria and most especially among children, there is no existing model in which survival of infected patient can be predicted. Therefore, this paper present the development of survival model among paediatric HIV/AIDS patients in South Western Nigeria and the objectives were to identify survival variables for HIV/AIDS paediatric patients in the South Western Nigeria and formulate survival predictive models based on variables identified [CD4 count, viral load, nutritional status and opportunistic infection] from the interview conducted with the virologist and paediatrician at the study area.

Machine learning is a branch of artificial intelligence that allows computers to learn from past examples of data records [23-24]. Machine learning does not rely on prior hypothesis unlike traditional explanatory statistical modeling techniques do [25]. Machine learning has found great importance in the area of predictive modeling in medical research especially in the area of risk assessment, risk survival and risk recurrence. Machine learning techniques can be broadly classified into: supervised and Unsupervised techniques; the earlier involves matching a set of input records to one out of two or more target classes while the latter is used to create clusters or attribute relationships from raw, unlabeled or unclassified datasets [26]. Supervised machine learning algorithms can be used in the development of classification or regression models. Classification model is a supervised approach aimed at allocating a set of input records to a discrete target class unlike regression which allocates a set of records to a real value. This research is focused at using classification models to classify pediatric HIV/AIDS patients' survival as either Survived or not survived.

2.0 Related Works

Researchers had worked on the prediction of HIV/AID prediction using different types of variables like CD4 count, CD8. Some of the researcher and the result of their finding are in the following paragraphs.

[27] developed a model to predict survival of HIV/AIDS using sequential and standard neural networks. The aim of the study was to produce a model of disease progression in AIDS using sequential neural network and compare the model's accuracy with that of a model constructed using only standard neural networks based on demographic and socioeconomic variables. The strength of the study was that sequential neural networks could discriminate patients who die and patients who survive more accurately than the standard neural networks. The weakness of the study is that only demographic and socioeconomic variables were used, which is not sufficient to predict survival. CD4 count, viral load, opportunistic infections and nutritional status would be enough to predict survival accurately in this paper.

[28] applied neural network in the prediction of HIV status of an individual based on demographic and socioeconomic characteristics. The aim of the study was to use supervised learning to train neural networks, to classify the HIV status of an individual given certain demographic factors. The strength of the study is that the neural networks used for prediction has high predictive capability. The weakness is that demographic and socioeconomic characteristics are not enough to accurately predict survival status.

[29] used highly active antiretroviral therapy to predict survival among HIV-infected children in Asian countries. The aim of this research is to conduct a general review of Paediatric ART effectiveness in Asian

countries using Kaplan- Meier survival analysis to estimate survival time probability, after the introduction of ART and Cox proportional hazard model was used for multivariate analysis. The strength of this research is that there were beneficial outcomes of first-line antiretroviral therapy for HIV infected children in Asian countries. The weaknesses were [1] limited information about the management of children who failed first line NHRT I regimen and [2] There was need to improve access to early diagnostic testing and treatment in infancy. CD4 count was the only predictive variable used in this study which is not enough to determine the survival of infected HIV patients CD4, viral load, opportunistic infection and nutritional status were added to the existing predictive factor in this thesis.

[30] used Classification and Regression Tree [CART] for the Prediction of Survival of Aids Patients receiving antiretroviral therapy in Malaysia. The aim is to investigate the use of CART as a tool for prediction of AIDS survival using CD4, CD8, Viral Load and Weight as predictor variables. The strength of the research is that the potential treatment methods and monitoring the progress of treatment of AIDS patients could be determined with the approach experimented and the results obtained. Fewer variables were considered for the prediction of survival of AIDS patients and data limitation is also a constraint in the study. Also opportunistic infections and nutritional status that are very important for HIV/AIDS prediction were not used in the research. CD4, CD8, viral load and weight are not enough factors to predict survival.

[31] carried out a prospective cohort on the Predictors of mortality in HIV-1 infected children on antiretroviral therapy in Kenya. The aim of this work was to carry out a study on early mortality following highly active antiretroviral therapy [HAART] using Cox proportional hazard model to determine the baseline characteristics associated with mortality and Kaplan-Meier method to estimate the probability of survival. The study shows that Low baseline haemoglobin was an independent risk factor for death. The weakness is that only haemoglobin was used as predictor variable in this research which is not enough to determine survival rate.

Haemoglobin is not enough to predict the survival of paediatric infected HIV patients but in this thesis, CD4, CD8 and viral load, opportunistic infection and nutritional status were used as predictive factors.

[32] developed a predictive model for AIDS Survival using Data Mining Approach. The aim of the research was to describe the feasibility of applying data mining technique to predict the survival of HIV/AIDS. An adaptive fuzzy regression technique, FuReA, was used to predict the length of survival of AIDS patients based on their CD4, CD8 and viral load counts. The strength of the research is that CD4, CD8 and viral load counts were used because the authors believed that predictors / markers are appropriate for predicting AIDS survival due to the high accuracies demonstrated by Fuzzy regression analysis [FUREA]. The weakness is that fuzzy neural network prediction results on AIDS survival could not be made possible because of data limitation. This is because Fuzzy neural network requires the use of large volume of data for prediction. Opportunistic infections and Nutritional status are important predictor variables together with CD4, CD8 and viral load counts that can be used to predict the survival of HIV/AIDS patients.

[33] carried out a retrospective cohort study on the survival status of HIV positive adult on antiretroviral treatment in Debre Markos Referral Hospital, Northwest Ethiopia. The aim of this study was to determine survival status and associated factors among HIV positive adult on antiretroviral treatment using Kaplan Meier to estimate survival and Cox regression for the analysis. Lost, drop out, transfer out and transfer in patients with high risk of death were excluded so as not to under estimate mortality of infected patients. Secondary data in which some important variables were not documented well were used and many opportunistic infections were presumed diagnosis. Pre-processing exercise was carried out in order to remove incomplete

data before classification and prediction of survival took place and more independent variables were also used in this study. CD4, CD8, viral load, opportunistic infection and nutritional status were in this thesis as predictive factors.

[34] applied logistic regression in modelling of survival chances of HIV-positive patients under highly active antiretroviral therapy [HAART] in Nyakach District, Kenya. The aim of this study was to outline the various social and economic factors affecting survival of HIV patients under highly active antiretroviral therapy [HAART]. The study was expected to provide suitable model for predicting the chances of survival among the HIV positives attending ART clinic in Nyakachi District and also provide information for policy makers on the factors affecting survival of HIV positive ARVs. The strength shows that the survival of infected patient under study can be improved if their access to socio-economic factors is considered. The outcome may only be obtained in services that have smaller numbers of patients. Socioeconomic factors are not enough to predict survival as CD4, CD8, viral load, opportunistic infections and nutritional status were added to the existing study in this paper as predictive factors.

3.0 Methods

Extensive review of literature on related areas in HIV/AIDS survival prediction was carried out in order to understand the body of knowledge surrounding the area and the extent of research alongside the gaps in knowledge. Following this, interview was conducted with virologist and Paediatrician in order to identify the required survival variables for HIV/AIDS survival among paediatric patients receiving treatment in Nigeria.

3.1 Data Identification and Description

The datasets used for this study to develop the predictive model for HIV/AIDS survival among paediatric patients was collected from two [2] tertiary hospitals in Nigeria, namely: Obafemi Awolowo University Teaching Hospital Complex (OAUTHC), Ile-Ife, Osun state and the Federal Medical Centre [FMC], Owo, Ondo state. The data collected contained the variables: age, sex, religion, HAART, weight, tribe, CD4 count, viral load, opportunistic infection, nutritional status and survival is shown in Table 1.

The datasets collected contained the following information; 97 paediatric patients who survived treatment and 119 paediatric patients who did not survive the treatment – all totalling to 216 records. 79 datasets were collected from OAUTHC while 137 datasets were collected from FMC. The final dataset contained 216 paediatric patients records consisting of four identified attribute values for the CD4 count, viral load, opportunistic infection and the nutritional status of the patients as the independent factors while the survival status of the HIV/AIDs paediatric patient was tagged as the output/dependent variable. A description of the variables is as follows:

a. CD4 Status: is a description of the number of white blood cells available in every mm^3 of blood sample collected from the pediatric AIDS patient; which fights infection in the body – the lesser this value the lesser the likelihood of surviving AIDS. It the value is less than 500 then it is said to be low but if greater it is said to be high – it is one of the variable of interest in determining the survival of pediatric AIDS patients;

b. Viral Load: is a description of the amount of HIV present in every mm3 of blood sample collected from the paediatric AIDS patient; it is a nominal value which takes a value of high when the CD4 status is low and vice versa – it is also one of the input variables which is used in determining paediatric AIDS survival;

c. Nutritional Status: is a description of the health and nutritional state of the pediatric HIV/AIDS patient determined by the BMI of each patient; which is a nominal value classified as high or low depending on the relationship between each patient's age and weight – it is also

one of the input variables used in predicting paediatric AIDS survival;

d. Opportunistic infection: is a description of the presence of other determining factors like tuberculosis or cancer – the presence of these diseases reduces the chances of survival of disease; which is a nominal value classified as yes or no – it is one of the input variables used in predicting the survival of paediatric AIDS survival; and

e. Survival: is a description of the survival status of each paediatric AIDS patient receiving treatment. It is an indication of whether a patient will survive treatment; which is a nominal value classified as either yes or no. It is the required output variable.

3.3 Model Formulation and Simulation

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

The predictive model for HIV/AIDS survival was formulated using the naïve Bayes' classifier – a supervised machine learning algorithm that is based on the naïve Bayes' statistical theory of conditional probability shown in equation [1].

$$P(Class|Attributes) = \frac{P(Attributes|Class) * P[Class]}{P[Attributes]} \quad [1]$$

Where:

i. P[Class]: Prior probability of class [survived/not survived]
ii. P[Attributes]: Prior probability of training data attribute values
iii. P[Attributes|Class] : Probability of attribute values given the class
iv. P[Attributes|Data] : Probability of Class given the attribute values

Equation [2] is derived from [1] and is used to estimate the probability of the record belonging to either of the two classes [survived and not survived] and the HIV/AIDS patient is allocated to the class with maximum probability as shown in equation [3]. Equations [2] and [3] are used by the naïve Bayes' classifier to formulate the prediction model for HIV/AIDS paediatric patients using the attributes, X_k = {CD4_count = "value", Viral_load="value", Nutritionl_status="value", Opportunistic_infection="value"} and allocate a patient to either of class, C_i = {Survived, not_survived}.

$$P(C_i|X_k) = \prod_{k=1}^{n} P(X_k|Ci) \cdot P(C_i) \quad [2]$$

$$survival\ class = MAX \begin{bmatrix} P(survived|X_k), \\ P(not_{survived}|X_k) \end{bmatrix} \quad [3]$$

The simulation of the predictive model for HIV/AIDS survival was performed using the Waikato Environment for Knowledge Analysis (WEKA) – a light weigh Java-based environment of machine learning tools. The model was evaluated using a performance evaluation tools called the confusion matrix-from which other measures of performance were evaluated. Every classification [prediction] problem can be evaluated using a confusion matrix. This is a diagram that displays the results of a prediction model figuratively. It allows a researcher to determine the amount of correctly classified cases and hence, evaluate the performance of a prediction model based on the results.

The output class is divided into positive and negative cases; all predicted

positive and negative cases are mapped to the actual cases as correct [true positives/negatives] and incorrect [false positives/negatives] classifications [see Figure 1]. Correctly classified cases are placed in the true cells [positive and negative] while incorrect classifications are placed in the false cells (positive and negative).

i. True positives are correctly classified positive cases;
ii. False positives are incorrectly classified positive cases;
iii. True negatives are correctly classified negative cases; and
iv. False negatives are incorrectly classified negative cases.

From a confusion matrix, different measures of the performance of a prediction model can be determined using the values of the true positive/negatives and false positives/negatives (Figure 2). The simulation was performed using the k-fold cross-validation technique; in the case of this study, the 10-fold cross validation method was used which has the following process (Figure 3):

i. The whole dataset is first divided into 10 parts;
ii. 9 parts (90%) are used for training while leaving 1 part (10%) out for testing;
iii. The process is repeated 10 times by keeping 1 part for training from the first part to the last part; and
iv. The predicted results of the 10 test parts are used to evaluate the performance of the prediction model developed for the study.

4.0 Results

The classification of each training data was performed via the implementation of the Naïve Bayes' classification algorithm which calculates the probability and manipulates them into the necessary results. A typical demonstration of how this is achieved is shown in the following paragraphs.

For HIV survival model, the variables used are: CD4 count, viral load, Opportunistic Infections and Nutritional Status and they are represented as X [input value] and the output class by C – survived and not survived. Consider the training data provided and the classification of the following data, X containing the values of each attributes for HIV/AIDS patients identified.

X=(CD4_status(X_1)="value",viral_load(X_2)="value",nutritional_status(X_3)="value",opportunistic_infection(X_4)="value")

a. First determine the probability of the output class being YES

$$P(YES|X_i) = [P(X_1|YES) * P(X_2|YES) * P(X_3|YES) * P(X_4|YES)] \cdot P[YES]$$

b. Second, determine the probability of the output class being NO

$$P(NO|X_i) = [P(X_1|NO) * P(X_2|NO) * P(X_3|NO) * P(X_4|NO)] \cdot P[NO]$$

c. Determine the maximum class probability

$$Survival_{class} = \boldsymbol{MAXIMUM}[P(YES|X_i), P(NO|X_i)]$$

Each probability $P[Attribute, X_i|Class, C_i]$ is calculated using the formula in equation [4] below.

$$P(X_i|C_j) = \frac{P[X_i \cap C_i]}{P[C_i]} \quad [4]$$

The results of the simulation of the predictive model for the survival of pediatric HIV/AIDS patients revealed that out of the 97 patients that survived the model was able to correctly classify 95 patents while 2 were misclassified as NO while out of the actual 119 patients that did not survive the disease, the model was able to correctly classify 80 patients while 39 were misclassified as YES.

Figure 4 shows a graphical plot of the correct and incorrect classifications performed by the naïve Bayes' classifier. The blue crosses identify a Survival (YES) and red crosses identify a no survival (NO) while

the boxes show misclassifications (i.e. YES classified as NO and vice versa). Figure 5 also shows the interpretation of the results using a confusion matrix from which the evaluation measures for the performance of the naïve Bayes' classifier was made.

Table 2 shows the results of the interpretation of the confusion matrix for the performance of the predictive model developed using the naïve Bayes' classifier. The accuracy of the predictive model was determined to be 81.02% owing for the 175 correct classifications out of the total 216 records. The sensitivity (true positive rate) was determined to be 97.9% - the percentage of Yes cases correctly classified while the specificity [true negative rate] was determined to be 67.2% - the percentage of No cases correctly classified. The area under the ROC curve showed a value of 0.993 – the ability of the model to discriminate between the patients that survived and did not survive was about 99.3%.

5.0 Model Validation Results

Following the development of the predictive model for HIV/AIDS survival prediction among pediatric patients, the model had to be validated using an external set of historical dataset of live clinical data about other HIV/AIDS pediatric patient which were not included in the training data.

Validation is the task of demonstrating that the model is a reasonable representation of the actual system. It reproduces system behaviour with enough fidelity to satisfy analysis objectives. Table 3 shows the result of the validation of the naïve Bayes' classifier using dataset of some patients and comparing the results of the naïve Bayes' with that of the actual results from the tertiary institutions considered in the study.

The dataset contained 24 records, from which naïve Bayes' classifier was able to correctly classify 17 records owing for an accuracy of about 71%. This further reaffirms the ability of the predictive model for HIV/AIDS survival to predict the survival of pediatric HIV/AIDS patients in Nigerian pediatrics.

6.0 Discussion

Following the results of the performance of the naïve Bayes' classifier on the datasets provided containing 216 pediatric HIV/AIDS survival data, a number of things need to be noted.

The naïve Bayes' showed the capacity to be able to clearly distinguish between pediatric HIV/AIDS patients survival with an accuracy of about 81% and the discrimination showed a value of about 99.3% owing to the value of the area under the ROC curve. It is also clear that the model is able to identify patients that will survive (Yes cases) more than those that will not survive (NO cases) the disease owing from the values of the TP and TN rates shown in Table 2.

The ability of the predictive model developed to misclassify a negative case as a positive case was observed to show a value of about 32.8% which is quite significant and can be reduced by providing more dataset about patients that did not survived the disease.

The result of the validation also further justifies the suitability of the developed predictive model for the survival of pediatric HIV/AIDS patients in Nigeria.

7.0 Conclusion

Child mortality is a factor that can be associated with the well-being of a population and taken as one of the development indicators of health and socioeconomic status in any country. HIV/AIDS epidemic has devastated many individuals, families and communities. Therefore, in order to reduce child mortality which is one of the important millennium goals, there is need to have effective and efficient model that can be used to forecast the survival of Paediatric HIV/AIDS patients in South Western Nigeria.

It will also help individuals, NGO and the government to make adjustments in areas where these affected children are suffering. The naive baye's predictive model serves as an effective model from the analysis above and will be recommended for use to predict Pediatric HIV/AIDS patients survival in order to justify results collected from the two health institutions [FMC, Qwo and

OAUTHC, Ile-Ife]. Pediatric HIV/AIDS patients' survival prediction depends on some major factors as shown in the result of this paper and the factors serves as contributing factors to the patient's status.

Acknowledgements

We acknowledge all the referees, and those that contributed to the success of this work.

References

[1] Idowu P et al. Spatial Predictive Model for Malaria in Nigeria. Journal of Health Informatics in Developing Countries, 2009; 3[2]:31-36.

[2] Henry K. The Global HIV/AIDS Epidemic. Available from www.kff.org/global-health-policy/fact.../the-global-hivaids-epidemic/, 2013. Accessed 2014 January 12.

[3] Eric J and Daria J. HIV-1 Antiretroviral Drug Therapy. Journal of Cold Spring Harb Perspect Medical, 2012; 2[4]:23-45

[4] Hoa M. ART Adherence among People Living With HIV/Aids in North Vietnam, Queensland University of Technology, Brisbane Australia, 2011.

[5] Idowu P. Development of A Web-Based Geo-Spatial Environmental Health Tracking System for South Western Nigeria. Unpublished PhD Thesis Submitted To Department Of Computer Science and Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria, 2012.

[6] Ojunga N *et al.* The Application of Logistic Regression in Modeling of Survival Chances of HIV-Positive Patients under Highly Active Antiretroviral Therapy [HAART]: A Case of Nyakach District, Kenya. Journal of Medicine and Clinical Sciences, 2014 3[3]:14-20.

[7] UNAIDS. Together We Will End AIDS. Available from www.unaids.org, 2012. Accessed 2013 June 12.

[8] WHO. Towards Universal access: Scaling up Priority HIV/AIDS interventions in the Health Sector. Progress Report 2010. Available from http://whqlibdoc.who.int, 2012. Accessed 2014 January 15.

[9] Shearer W. Evaluation of Immune Survival Factors in Pediatric HIV-1 Infection. Annual National Academic Journal, 2000; 91[8]:298-312.

[10] Picat M *et al.* Predicting Patterns of Long-Term CD4 Reconstitution in HIV-Infected Children Starting Antiretroviral Therapy in Sub-Saharan Africa: A Cohort-Based Modeling Study. Journal of Pediatric Medicine, 2013; 10[10]:45-49

[11] Nigerian Bulletin. World HIV/AIDS Day: 10 Facts about HIV/AIDS in Nigeria You Probably Didn't Know. Available from www.nigeriabulletin.com/threads/world-hiv- aids-day-10-fact-about-hiv-aids-in-nigeria-you-probably-didn't-know.24303/,2014. Accessed 2014 July, 28.

[12] Rosma M *et al.* The Prediction of AIDS Survival: A Data Mining Approach. 2nd WSEAS International, 2012. Conference. Malaysia.

[13] Kama K and Prem S. Utilization of Data Mining Techniques for Prediction and Diagnosis of Major Life Threatening Diseases Survivability-Review. International Journal of Scientific & Engineering Research, 2013; 4[6]:923-932.

[14] Ojunga N *et al.* The Application of Logistic Regression in Modeling of Survival Chances of HIV-Positive Patients under Highly Active Antiretroviral Therapy [HAART]: A Case of Nyakach District, Kenya. Journal of Medicine and Clinical Sciences, 2014; 3[3]:14-20.

[15] Sanjel S. An Application of Cox Model for Lifetimes of HIV Patients. Unpublished M.Sc thesis submitted to the department of Power Analysis, 2009. USA. MC Master University.

[16] Avert. Stages of HIV Infection. Available from www.averts.org/stages-hiv-infection.htm, 2014. Accessed 2014 July 28.

[17] Centre for Disease Control. HIV classification: CDC and WHO staging system. Available from www.hab.hrsa.gov/deliverhivaidscare/clinicalguide11/cg-205_hiv_classification.html, 2011. Accessed 2014 August 4.

[18] UNAIDS. Report on the Global AIDS Epidemic. Available From www.unaids.org/globalreport/global_report.htm, 2010. Accessed 2013 June 12.

[19] WHO. Preventing HIV/AIDS in Young People, A Systematic Review of the Evidence From Developing Countries, UNAIDS Inter-Agency Task Team on Young People. World Health Organization, Geneva. Available From: www.who.org, 2006. Accessed 2013 January 12.

[20] Mira J and Denis A. The Cost-Effectiveness of Preventing Mother-to-Child Transmission of HIV in Low- and Middle-Income Countries: Systematic Review. Journal of Bio Med Central, 2010; 9[3]:12-19.

[21] Petros I *et al*. High Survival and Treatment Success Sustained After Two and Three Years of First- Line ART for Children in Cambodia. Journal of International AIDS Society, 2010; 3[11]:21-29.

[22] UNAIDS. Global Report on HIV. Available from http://www.unaids.org/en/resources/documents/2013/name,85053,en.asp, 2013. Accessed 2014 August 2.

[23] Quinlan JR. Induction of Decision Trees. Machine Learning, 1986; 1: 81-106.

[24] Cruz JA and Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics, 2006; 2:59 – 75

[25] Waijee AK, Joyce JC and Wang SJ. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bone disease to thiopurines. Clin Gastroenterol Hepatol, 2010; 8:143 – 150.

[26] Mitchell T. Machine Learning, McGraw Hill, 1997, New York.

[27] Lucia O. Sequential Use of Neural Network for Survival Prediction in AIDS. Available from www.ncbi.nlm.nih.gov/pmc/articles/PMC2233186/, 1996. Accessed 2014 August 4

[28] Brain, *et al*. Prediction of HIV status from Demographic Data using Neural Networks. IEEE conference on Systems Man and Cybernetics, 2006. Taipei, Taiwan.

[29] Torsak A *et al*. The Effectiveness of Highly Active Antiretroviral Therapy among HIV-Infected Children in Asian Countries. Asian Biomedicine Journal, 2009; 3[1]:89-100.

[30] Sameem A. *et al*. Classification and Regression Tree in Prediction of Survival of AIDS Patients, Malaysian Journal of Computer Science, 2010; 23[3]:153-165.

[31] Dalton C *et al*. Predictors of Mortality in HIV-1Infected Children on Antiretroviral Therapy in Kenya: A Prospective Cohort. Journal of Pediatrics, 2010; 10[33]: 1471-2431.

[32] Rosma M *et al*. The Prediction of AIDS Survival: A Data Mining Approach. 2nd WSEAS International Conference, 2012. Malaysia.

[33] Nurilign A *et al*. Survival Status of HIV Positive Adults on Antiretroviral Treatment in Debre Markos Referral Hospital, Northwest Ethiopia: Retrospective Cohort Study. The Pan African Medical Journal, 2014; 17[88]:19-24.

[34] Ojunga N *et al.* The Application of Logistic Regression in Modeling of Survival Chances of HIV-Positive Patients under Highly Active Antiretroviral Therapy [HAART]: A Case of Nyakach District, Kenya. Journal of Medicine and Clinical Sciences, 2014; 3[3]:14-20.

Table 1. Variables identified from the data collected for HIV/AIDS survival

S/N	Field Name	Sample Values
1	Sex	Male, Female
2	Age	Numeric
3	Weight	Numeric
4	Tribe	Yoruba, Ibo, Idoma, Ibo, Kwale, Isoko, Ijaw, Ebira,
5	Religion	Christianity, Islam
6	Stage	1, 2, 3, 4, none
7	On HAART	Yes, No
8	CD4 Status	High [> 500], Low [< 500]
9	Viral Load	High [if CD4 status is Low], Low [if CD4 status is High]
10	Nutritional Status	High, Low
11	Opportunistic Infection [Presence of Tuberculosis or Cancer]	Yes, No
12	**Survival**	**Yes No**

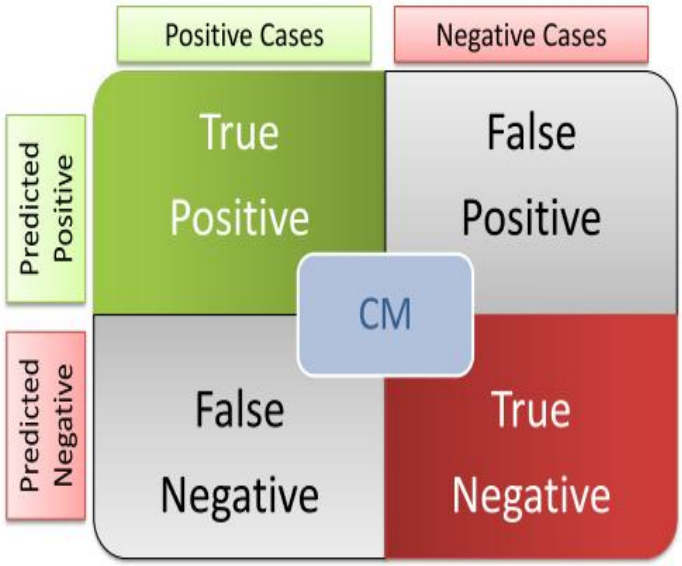


Figure 1. Confusion matrix showing the different components

		Target		
		Y	N	
Model	Y	*TP*	*FP*	**Positive Predictive Value** TP/(TP+FP)
	N	*FN*	*TN*	**Negative Predictive Value** TN/(TN+FN)
		Sensitivity TP/(TP+FN)	**Specificity** TN/(TN+FP)	**Accuracy** $\frac{TP+TN}{TP+TN+FP+FN}$

Figure 2. Confusion matrix showing the performance metrics

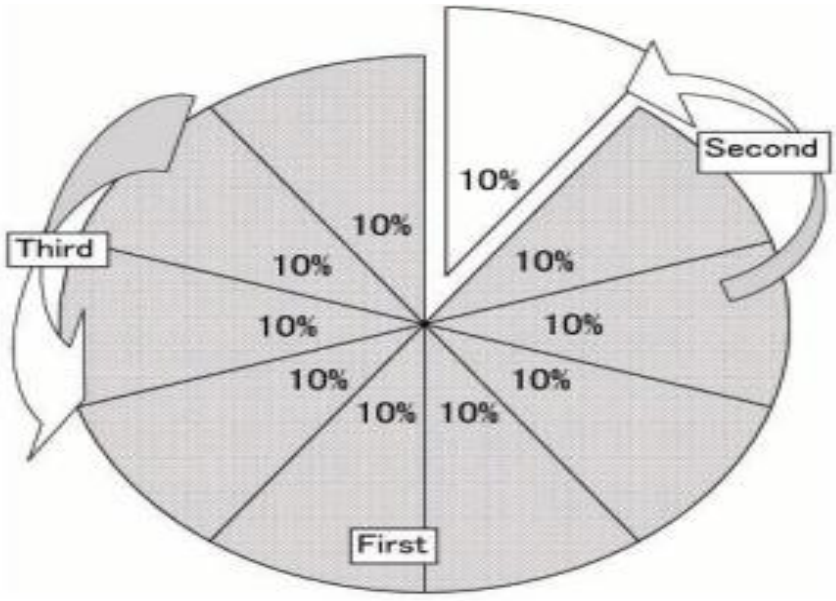


Figure 3. 10-Fold Cross Validation Process

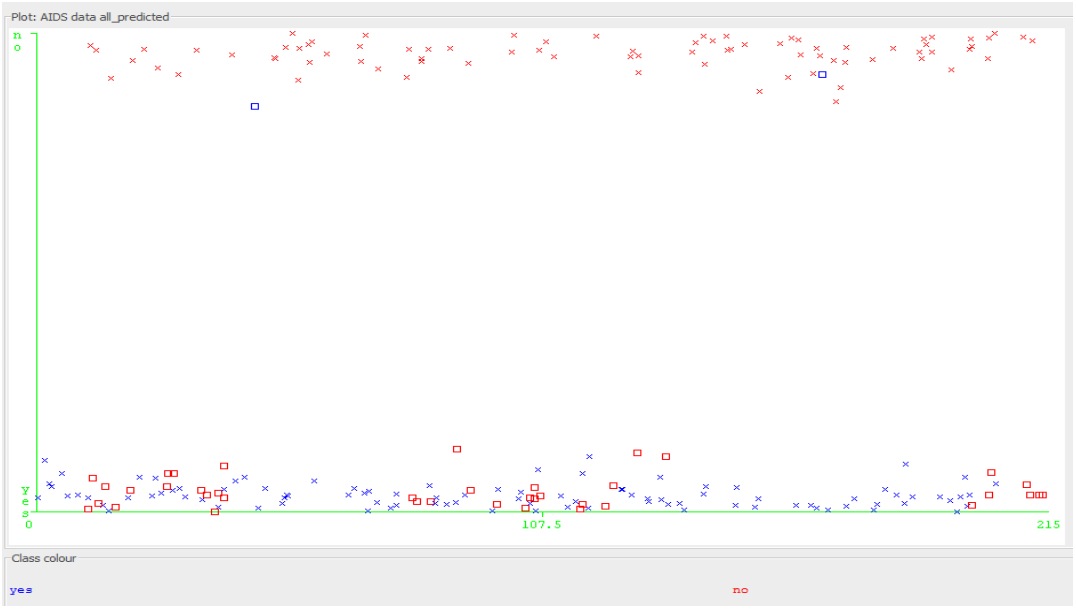


Figure 4. Results of the predictive model using naïve Bayes' classifier

Classified as	YES	NO
YES	95	2
NO	39	80

Figure 5. Confusion matrix evaluated from the simulation results

Table 2. Performance Evaluation Results of the Developed Model

Performance Metrics	Values
Accuracy	81.0185%
Correct Classification	175
Incorrect Classification	41
True Positive [TP] rate/Sensitivity	0.979
True Negative [TP] Rate/Specificity	0.672
False Positive [FP] rate/False Alarm	0.328
Precision	0.709
Area under ROC	0.993
MAE	0.2025
RMSE	0.292
RAE	40.9205%

Table 3. Validation Results of the Predictive Model

CD4	Viral Load	Nutritional Status	Opportunistic Infection	Survival	Decision Tree
High	Low	High	Yes	Yes	Yes
Low	High	High	Yes	No	No
High	Low	Low	Yes	**No**	**Yes**
High	Low	High	Yes	Yes	Yes
High	Low	Low	Yes	**No**	**Yes**
High	Low	High	Yes	Yes	Yes
High	Low	High	Yes	Yes	Yes
High	Low	Low	Yes	**No**	**Yes**
High	Low	High	Yes	Yes	Yes
Low	High	Low	Yes	No	No
Low	High	Low	No	No	No
High	Low	High	Yes	Yes	Yes
High	Low	High	Yes	Yes	Yes
High	Low	Low	Yes	Yes	Yes
High	Low	Low	Yes	**No**	**Yes**
High	Low	High	Yes	Yes	Yes
High	Low	High	Yes	Yes	Yes
High	Low	Low	Yes	No	No
High	Low	High	Yes	Yes	Yes

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

A NEW TWO-TIERED STRATEGY TO E-EXAMINATION SYSTEM

A.J. Ikuomola
Department of Mathematical Sciences,
Ondo State University of Science and Technology Okitipupa, Nigeria
deroniknng@yahoo.com

ABSTRACT
Examiners have used many techniques in coping with examination malpractice and still we do not have an effective approach to examination system. The aim of this work is to combine fingerprint biometric and N-types techniques in designing an effective e-examination system. The idea is that fingerprint biometric is used as a suitable solution for rapid authentication of users in accessing the questions in the exam via the use of biometric devices. Also, an N-type technique is used to generate multiple question types for a given subject and map these types to each candidate using the monotonic (1:1) mapping scheme such that candidates adjacent to each other do not have the same type even if they are taking the same subject in the examination. The system is made up of three components: the pre-registration phase, configuration phase, and examination phase. The new system design was tested with data and the result shows that the proposed system is highly efficient in verification of user/student fingerprint and allows for error free and faster process of conducting and writing examination.

Keywords – Biometric, e-Examination, Fingerprint, N-type, Traditional Examination.

1.0 INTRODUCTION

Examination is a tool or technique intended to measure students' expression of knowledge, skills and/or abilities. It is an official exercise designed to evaluate knowledge and skills, covering the contents of a course or program of studies. A test may be administered orally, on paper, on a computer, or in a confined area that requires a test taker to physically perform a set of skills [14]. Examination can be conducted manually or electronically. Manual/conventional/traditional paper-based exams are performed with the use of sheets of paper, biros and/or pencil.

In traditional paper-based examination system, eligible students are usually authenticated or checked-in manually by the examiners. Examiners are also saddled with the responsibility of ensuring that students comply with the rules of examination conduct. There is non-anonymity of teachers in traditional paper-based examination system i.e. students have knowledge of who will mark their answer sheets; hence they can bribe or threaten them in order to receive better grades.

As useful as examinations are, traditional paper-based method of conducting examination opens way for examination malpractices. Having the question-types pre-printed on the question papers as it is the case in some examinations such as Senior Secondary Certificate Examination (SSCE), Universities Matriculation Examination (UME) and Polytechnics and Colleges of Education Examination (PCE) has not solved the problem of cheating. Often time, candidates employ more hands from outside to assist in solving different question-types and with the aid of invigilators to distribute the answers in the exam halls.

The problems with traditional paper-based exams are absence of transparency and effective preparation of exams, lower grade of objectivity, no instant feedback on examination result, increased costs, increased teacher's workload, no elaborate feedback on teaching success, does not enhance audit quality management and bad security.

Electronic examinations (e-examinations) are now a viable alternative method of assessing student learning. They provide freedom of choice in terms of the location of the examination (whether ex-aminations are running synchronously or asynchronously) and can provide immediate feedback [3]. The Electronic examinations are those examinations performed through a computer where questions and answers are computer based rather than sheets of paper. [1] defines e-examination as a system that "involves the conduct of examinations through the web or the intranet". Electronic examination is of great interest from both educational and pedagogical points of view. It is aimed to resolve many questions and limitations in the conventional or traditional paper-based examination. It is flexible and handy to use with complete question types and excellent security strategy so as to make e-examination automatic and reduce the cost. It can be used for all kinds of different scaled e-examinations in different subjects in primary, secondary and tertiary institutions and in class examinations or exercises [4].

Basically, the electronic examination (e-examination for short) system involves the conduct of examinations through the web or the intranet. Its aim is to reduce the large proportion of workload on examination, grading and reviewing. The set of questions often used in the e-examination system are multiple choice objective tests and quizzes that can be formally and easily evaluated online. In essence, e-exam system provides the existence of an examination system where all exam stages are performed electronically.

The problems associated with the conduct of e-examination are impersonation in the examination hall and repetition of examination sessions by students.

In the context of examination, impersonation is writing examination on behalf of someone else. From research, it has been deduced that the main reason why impersonation has become very rampant in the educational system is because of the desire of student to pass at all cost. In traditional paper-based examination system, exam takers are usually identified by examiners. It is the responsibility of examiners to make sure that each person checked into the examination hall

is really supposed to be taking the examination he has been checked in for. The aim of doing this is to reduce this impersonation problem but it has been discovered, however, that the problem still persists owing to human (examiner) error. The human error comes into play when examiners cannot distinctively identify each student (e.g. in the case of twins).

In order to solve these problems a two-tiered strategy to e-examination system was proposed. The structure of the new design is such that the bulk of questions in the database are shuffled and a specific number of questions are selected and presented before each student. Each of the students has different or variant type of questions and fingerprint biometric is used as a suitable solution for rapid authentication of users in accessing the questions in the exam via the use of biometric devices because fingerprints have permanent attribute unique to an individual. [15] Pointed out that fingerprints have been universally acceptable in the legal system worldwide. It has also been proved over the years that fingerprints of each and every person are unique [9]. So it helps to uniquely identify the students.

2.0 LITERATURE REVIEW

Electronic examination is of great interest from both educational and pedagogical points of view. It is aimed to resolve many questions and limitations in the conventional or traditional examination. It is flexible and handy to use with complete question types and excellent security strategy so as to make e-examination automatic and reduce the cost. It can be used for all kinds of different scaled e-examinations in different subjects in primary, secondary and tertiary institutions and in class examinations or exercises.

Due to the fact that examination is used as a means of determining student's ability, it is therefore, paramount to continue to improve on the previous method of conducting examination so as to have 100% secured examination. From literature, It is clear that students who were electronically examined performed better than those conventionally examined [12]. This leads to this research work which is to design a two-tiered strategy to e-examination system that will authenticate users and give legitimate user the right to access the exam questions and also handle multiple choice examination questions and administer an examination process effectively without any impersonation or examination malpractice.

There are many biometrics that can be utilized for some specific systems but the key structure of a biometric system is always the same [13]. Biometric systems are basically used for one of the two objectives: identification [10] and verification [6]. Some of the most commonly used biometric systems are (i) iris recognition (ii) facial recognition (iii) fingerprint identification (iv) voice identification (v) DNA identification (vi) hand geometry recognition (vii) gait recognition (viii) signature verification [8]. Among biometric trait, fingerprint is widely accepted for person identification because of its uniqueness and immutability [2].

[11] maintained that fingerprints biometric scans are the most commonly used biometric solution as they are less expensive compared with other biometric solutions. According to [5], a fingerprint is a unique "pattern of ridges and furrows on the surface of a fingertip, the formation of which is determined during the fetal period". Fingerprints are unique for each individual, where even identical twins have different fingerprints. Several scholars documented the increase popularity of fingerprint biometric-based systems and their decline in costs [11]. Similarly, fingerprints can be used for authenticating students' submissions of exams via the use of biometric devices. Furthermore, [15] pointed out that fingerprints have been universally acceptable in the legal system worldwide. Fingerprints are a permanent attribute unique to an individual. Fingerprints can be scanned, transmitted and matched with the aid of a simple device. [7] pointed out that biometric have been commonly employed in replacing conventional password systems. Biometric devices enable portable scanning and rapid identification. Thus, finger biometric can be a suitable solution for rapid authentication of users. Using a portable device, users can scan their fingerprints and send a print image via

the Internet to the University's network. The network will consist of an authentication server that will house a database of students' fingerprints images. The server will then process the matching of the transmitted print image with a stored copy of the fingerprint (called "template"). Following that, the server will generate a matching result. Thus, [7] predicted that fingerprints based biometric would become a household activity in the near future. [16] proposed a secured technique for matching fingerprints in a biometric system. Similarly, to [7] they argued that biometric systems enhance security far more than the current systems. Biometric systems are more accurate as well as simpler to operate compared with passwords systems. [16] described a fingerprint based biometric system in which the fingerprint template is kept in a server during initiation. Upon scanning the finger, an input device scans a biometric signal and transmits it to a server where it is processed for matching. In an effort to shield the system against security compromises, they recommended processing the matching of fingerprints images in an embedded device rather than the server and only transmitting the results to the servers. Furthermore, they suggested encrypting the fingerprint template prior to storing it on the server. Fingerprints templates can be decrypted whenever a matching process occurs. [16] provided additional solutions useful for building up multiple layers of security in fingerprint based biometric systems.

3.0 DESIGN METHODOLOY

3.1 Architecture for a New Two-Tiered Strategy to E-Examination System (2TISES)

Architecture for a new two-tiered strategy to e-examination system is shown in figure 1. The main task of the 2TISES is to authenticate user and give legitimate user/candidate the right to access the exam questions and also to generate multiple question types for a given subject and map these types to each candidate using the monotonic (1:1) mapping scheme such that candidates adjacent to each other do not have the same type even if they are taking the same subject examination. Under the procedure design, 2TISES is made up of three main components: pre-registration phase, configuration phase and examination phase.

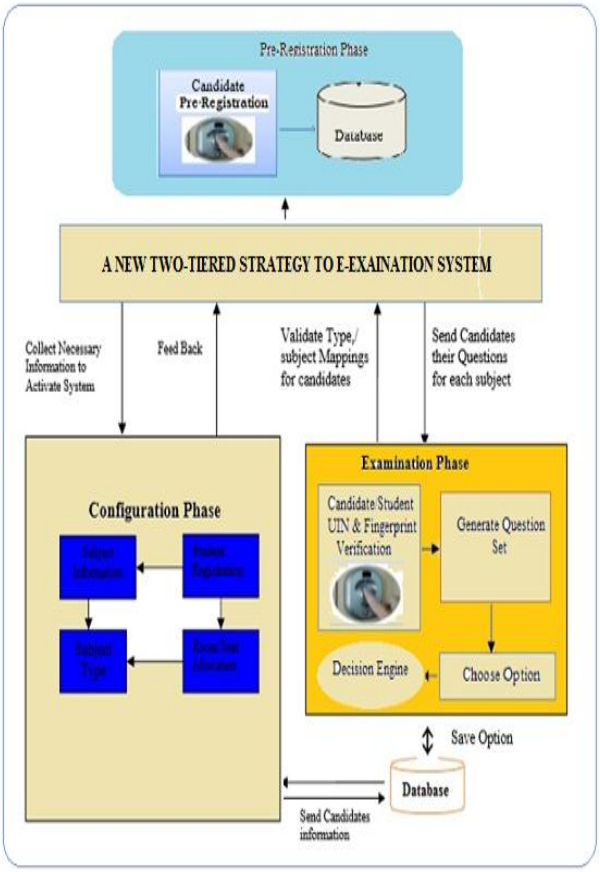


Figure 1: Architecture for a New Two-Tiered Strategy to E-Examination System (2TISES)

Pre-Registration

At this stage, candidates/students will be required to register their data such as name, unique identification number (UIN), courses/subjects offered and their fingerprint will be captured. Image of each student's fingerprint is captured using fingerprint sensors in fingerprint identification devices. The information collected from each student is afterwards stored in the database. The essence of storing the collected information in the database is to be able to recall them for comparison during examination phase.

Configuration Phase

The system will first have to be configured before it can be used for administering any examination. Firstly, the subject interface will appear where the examiner(s) will enter the number of subjects with subject identification for each subject that students will be examined on. Next is the Question/ Type Interface where the number of Questions for each subject as

well as the number of types they wish the system to generate will be entered into the system. For example, an examiner may have a subject A, with 30 questions and would like the question types to be 8.
When this is done, the next step will be an Interface that allows the examiner to specify either static/dynamic allocation of types to each candidate. If static is chosen, then the examiner will have to supply each candidate's Unique Identification Number (UIN) together with the subject(s) to be taken by each candidate. If Dynamic allocation is selected, then it is the user that will supply this information during the examination phase. When this is done, the system will now generate a type-set for all subjects to be examined and allocate subject type to each candidate and this information will be stored in a database.

The configuration phase on the client-side involves four major sub-phases namely;

(a) **Students' Pre-Registration**: this function will authenticate users whenever they login and verify if they have previously participated in the examination process. If verified, the student/candidate is allowed into the examination phase.

(b) **Room Information Entry/Seat allocation**: This function will dynamically allocate seats to each registered candidate thereby filling up the room. The function will detect if a room has been filled up so that it can automatically move to the next room.

(c **Subjects' Information Entry:** This function will allow candidates to select their pre-registered subjects and take part in the examination for that subject after a successful login. The routine just maps the type allocated to that students to the original question format in order to display the correct question. This engine also provides for a candidate who is seating for multiple subjects to switch between subjects while the time slot for the entire examination is still valid. The time remaining is displayed on top of the Title Bar for the Engine's interface so that the candidate is abreast of time used. The engine automatically shuts out the student when the time slot expires.

(d) **Type Generation:** This is the main component of the Secured N-Type Fingerprint-based E- Examination System. This function is capable of creating an N-Type Question Set for each subject registered by an examiner. In designing this engine, the following constrains will be enforced:

(i) Types will be created based on the maximum number specified by the examiner.
(ii) Each Type generated will be allocated a Unique Type Identifier (UTI).
(iii) Each subject type will be generated based on a pivot key supplied by the examiner. The Value of the pivot key will be in the neighborhood of the midpoint of the total number of questions available for that subject.
(iv) Type keys will be filtered into even, odd and prime number arrays.
(v) Type keys will now be re-arranged into the type key list.
(vi) Each type key will be saved along with the corresponding type list.

Within this engine, we define the degree of closeness between 2 types to mean the ordinal similarity between corresponding questions using their positional reference. Also, we define the Type Variance of a particular type, which is the overall effect of the degree of closeness of that type against all other generated types within the Type Set for that subject. Mathematically, this can be expressed as Type Variance = $\sum$ degrees of closeness/(n-1) Where n is the type cardinality. A Type variance Value μ for a particular type key is regarded as healthy if and only if $\mu < 0.01$. This means that if we have One Hundred (100) Question types in a type set for a subject that 100 candidates are to take, the likelihood of two students having the same type will be less than one (1%) percent.

Examination Phase

Students gain entry into this stage haven passed through the registration module. There will be a login interface where users (candidates) are able to access the examination

module. The typical security for this area will be that each student is expected to login with a Unique Identification Number (UIN) and an authenticated fingerprint access.

In this stage, students will first be requested to fill in their unique identification number, after which they will be required to place their right thumbs on a fingerprint identification device. The purpose of having them place their fingerprints on the device is to make sure that the candidate writing the examination is actually the one that is supposed to write it. A candidate will only have been able to successfully pass through this stage if he has passed through the initial stage (pre-registration stage). In this phase, there exists an interconnection structure between the fingerprint identification device and database. The fingerprint identification device is responsible for capturing input fingerprint patterns and presenting them for comparison. It compares the input pattern with that of the database to see if the input fingerprint matches one already existing in the database. If a student's fingerprint pattern matches an already existing one, the student is allowed access to the examination module. In case where no match was found, the student is denied access into the examination module. Every student that gets access into this phase is considered a legitimate student, that is, the student's information already exists in the database.

Examination phase is where students are faced with the set of questions they are expected to find answers to. If during the configuration phase the examiner selected dynamic allocation, then when the user gains access to the system, the subject interface will appear where the user will select the subject(s) to be taken. After this interface, the system now dynamically maps subject(s)-type(s) to the candidate and the examination clock begins to count down as the student answers each question. Since the Format for each Question is Multiple Choice Answer, the Candidate just enters an appropriate letter corresponding to an option in a white coloured answer box. This box automatically shades itself when an option is entered and reverses back to white when the option is erased. The user may go back and forth to view each question for each subject and may submit all answers whenever it is desired. However, if the Examination Clock Located on the Title Bar of the Window expires, then the user will be automatically disabled from doing any more work while the Test-Engine collects all answers supplied by the user.

Database

The database is responsible for storing all information such as the Questions, all Questions type-set generated by some internal functions and all users unique identification number (UIN) with their subject type(s), answers and scores. Each entry made by the each candidate for all subjects both correct and incorrect options are collated and stored in the database. The database consists of various relational tables used for data storage and management.

4.0 IMPLEMENTATION AND RESULTS

4.1 Implementation Procedure

The two (2) categories of software used are the system software and the application software. The system software consists of the operating system which was Windows XP professional Service Pack 2. The application software architecture is further subdivided into three categories; the programming language aspect in which Visual Basic 6.0 was used to build the client side architecture, the Microsoft Access database system used to develop the Server Side architecture and Structured Query Language (SQL) used as the Major Link Platform between the Client Side and Server side. The data used for the implementation was collected from the past questions of Universities Matriculation Examination (UME) conducted by Joint Admission and Matriculation Board (JAMB), Nigeria. Figure (2) shows the flowchart/design view of the new two-tiered strategy to e-Examination System

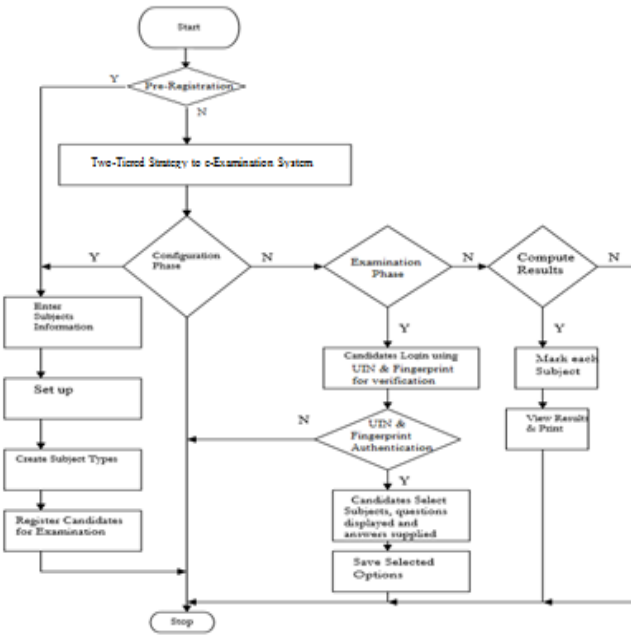


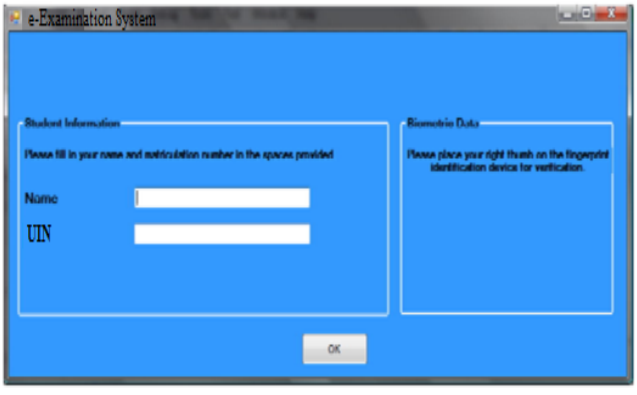
Figure 2: Flowchart for a New Two-Tiered Strategy to E-Examination System

4.2 Results and Discussion

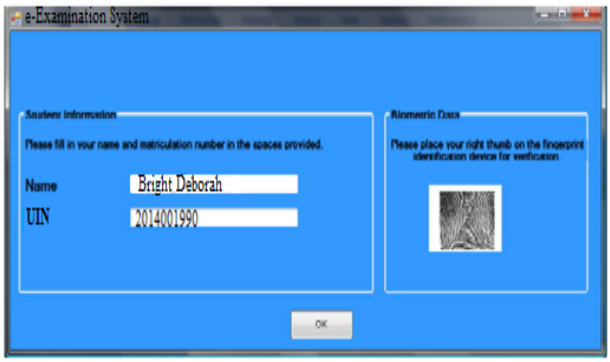
The design was run and tested and the interface designs are shown.

Interface Design

Figures 3 a & b show the candidate console. Before a Candidate can access the e-examination area; he/she must have been authenticated by the system. Authentication requires the placing of the candidate's right thumb on the fingerprint biometric device. On the Console in Figure 3a, the student/candidate is expected to enter his/her name, unique identification number (UIN) and then place his/her right thumb on the fingerprint identification device attached to the system for verification. In figure 3b, the fingerprint pattern shows up immediately the student clicks OK in the former stage.



(a)



(b)

Figure 3(a & b): Candidate Console for Verification

Figure 4 shows that the student is a registered candidate and the information about the candidate such as the name, UIN and photograph are in the database

Figure 4: Verified Candidate

Figure 5 shows the candidate console. On the Console, there are various tasks that have been labeled to guide the user. Each task is initiated by clicking on the arrow pointing to it. Here, the candidate select his/her registered subject which are displayed on the console.

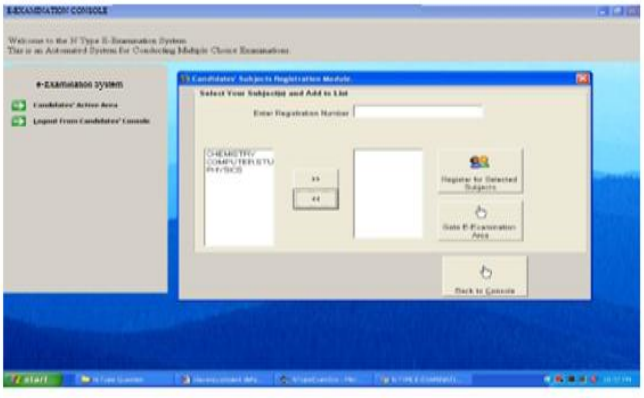


Figure 5: Candidate Console

Figure 6 is a message interface that shows up after the student select one of the registered subject and then clicks on continue in the previous stage. Clicking OK in this stage will prompt the student to the set of examination questions.

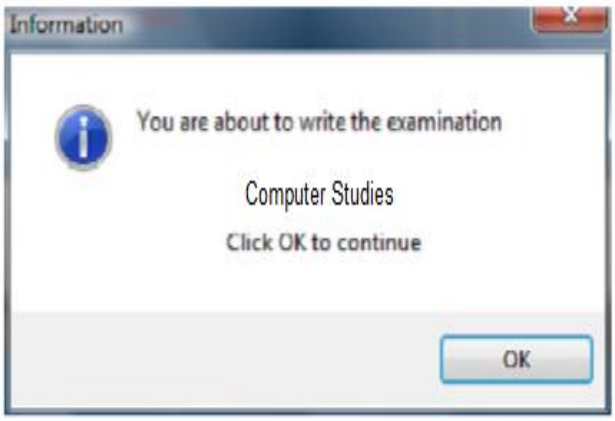


Figure 6: Prompt Message

Figure 7 shows the internally generated N-types examination questions set for the subject computer studies

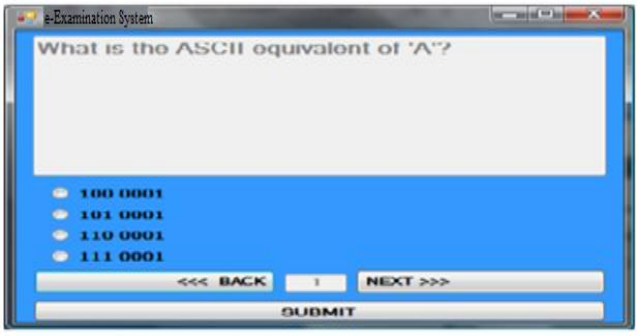


Figure 7: Question Set

5.0 CONCLUSION

This work presents a solution to the problems of traditional paper-based examination. We have been able to design and implement a new e-Examination System (2TISEES), a system that is able to:

(a) address major issue of impersonation and cheating in e-Exam.
(b) generate different variants of a question set for each subject.
(c) promote transparency of an examination process by making available a real time analysis of student's performance.
(d) develop a process that will evaluate each candidate's performance and save information.

The performance results indicate that using a two-tiered strategy to e-Examination System can address the major problems of traditional paper-based examination.

REFERENCES

[1] C. K. Ayo, I. O. Akinyemi, A. A. Adebiyi and U. O. Ekong. "The prospects of e-examination implementation in Nigeria." Turkish online Journal of Distance Education-TOJDE, 8(4), pp126, 2007.

[2] S. A. Daramola and C. N. Nwankwo. Algorithm for fingerprint verification system. Journal of emerging Trends in Engineering and Applied Sciences, 2(2), pp355-359, 2011.

[3] A. Fluck, O. S. Adebayo and S. M. Abdulhamid. Secure E-Examination Systems Compared: Case Studies from two Countries. Journal of Information Technology Education: Innovations in Practice, 16, pp107 – 125, 2017.

[4] A. J. Ikuomola and T. A. Olayanju. N-Types electronic examination system: An effective approach for combating examination malpractice. *Journal of Natural Sciences, Engineering and Technology, 9(2) 2010,* 2010.

[5] A. Jain, L. Hong and S. Pankanti. Biometric identification. *Communications of the ACM, 43*(2), pp91–98. 2000.

[6] A. Jain, A. Ross and S. Prabhakar. An introduction to biometric recognition, circuits and systems for video technology, IEEE Transactions, 14(1), pp4-20, 2004.

[7] M. McGinity. Staying connected: Let your fingers do the talking. *Communications of the ACM, 48*(1), pp21-23, 2005.

[8] T. Nawaz, S. Pervaiz, A. Korrani A. and Azhar-ud-din. Development of academic attendance monitoring system using fingerprint identification. International Journal of Computer Science and Network Security, 9(5), 2009.

[9] Ortega-Garcia, J. Bigun, D. Reynolds and Gonzalez-Rodriguez. Authentication

gets personal with biometrics, Signal Processing Magazine, IEEE, 21(2), pp50-62, 2004.

[10] S. Pankanti, S. Parabhakar and A. K. Jain. On the individuality of fingerprints, IEEE Transaction on Pattern Analysis and Machine Intelligence, 24(8), 2002.

[11] A. P., Pons. Biometric marketing: targeting the online consumer. *Communications of the ACM, 49*(8), pp60-65, 2006.

[12] C. Stergiopoilos, P. Tsiakos, D. Triantis, M. Kaitsa. "Evaluating Electronic Examination Methods Applied to Students of Electronics". IEEE International Conference on Sensor Networks, Ubiquitous and Trustworthy Computing, 2: pp143-151. ISBN:0-7695-2553-9, 2006.

[13] A. C. Weaver. Biometric Authentication, Computer, 39(2), pp96-97, 2006.

[14] Wikipedia. Examination. http://en.wikipedia.org/wiki/Test_%28assessment%29, 2014.

[15] J. M. Williams. New security paradigms. *Proceedings of the 2002 Workshop on New Security Paradigms*, Virginia Beach, Virginia, pp97-107, 2002.

[16] S. Yang and I. M. Verbauwhede. A secure fingerprint matching technique. *Proceedings of the 2003 ACM SIGMM workshop on Biometrics methods and applications*, California, USA, pp89-94, 2003.

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

EXTENDED HYBRID CONJUGATE GRADIENT METHOD FOR UNCONSTRAINED OPTIMIZATION

I. A. Osinuga[1], I. O. Olofin[2]
[1,2]Federal University of Agriculture, Abeokuta, Ogun State, Nigeria.

[1]osinuga08@gmail.com; [2] olofiniyiola@ymail.com

ABSTRACT
In this paper, a new search direction vectors are defined for BFGS-CG proposed by Ibrahim et al. by combining it with a term from the search direction vector expression proposed in modified PRP scheme of Zhang et al. in order to keep the descent property of the scheme. In addition, an update parameter of PRP is proposed to improve the performance of the algorithm. This new scheme known as Extended Hybrid BFGS – CG (EHCG) method is globally convergent with Armijo-type line search. Preliminary numerical results show that the method is efficient when subjected to comparison with classical PRP, modified PRP and conventional BFGS – CG algorithms.

Keywords: Global convergence, Hybrid Conjugate Gradient Method, Sufficient Descent Condition, Unconstrained Optimization.

1.0 INTRODUCTION

CONJUGATE gradient (CG) method is one of the many tools used to solve large scale unconstrained optimization problems. The linear CG method was proposed by Hestenes and Stiefel in 1952, as an iterative method for solving linear systems of equations with positive definite matrices. In 1964, Fletcher and Reeves introduced the first CG method to solve general unconstrained optimization problems. The attractive features of these algorithms are that they require no matrix storage and their programs are relatively simple. We consider the CG methods for solving the following unconstrained optimization problem:

$$\min f(x), \ x \in \mathbb{R}^n \qquad (1)$$

where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuously differentiable function and g(x) is the gradient of the objective function $f(x)$.

The CG method generates a sequence $\{x_k\}$, starting from an initial point $x_0 \in \mathbb{R}^n$, using the recurrence relation

$$x_{k+1} = x_k + s_k, \ k = 0, 1, 2, \ldots \qquad (2)$$

where x_k is the current iterate, $s_k = \alpha_k d_k$, and α_k is a step length which is calculated by a line search and d_k is a search direction given by

$$d_k = \begin{cases} -g_{k,} & k=0 \\ -g_k + \beta_k d_{k-1} & k \geq 1 \end{cases} \qquad (3)$$

where g_k is as defined above and β_k is a CG method update parameter such that the method reduces to the linear conjugate gradient method

in the case when f is strictly convex quadratic function and the line search is exact. In this paper, we compute α_k using the Armijo line search by:

$Given\ s > 0,\ \beta_k \in (0,1),\ \rho \in (0,1)$ (4)

$\alpha_k = \max\{s,\ s\beta, s\beta^2, s\beta^3,\ s\beta^4, \ldots\}$ (5)

such that

$f(x_k) - f(x_k + \alpha_k d_k) \geq -\rho\alpha_k {g_k}^T$ (6)

Then, the sequence $\{x_k\}_{k=0}^{\infty}$, converged to the optimal point, x^* which minimizes (1). Some well-known pioneer CG methods usually employed for solving large - scale unconstrained optimization problem of the form (1) includes;

$$\beta_k^{FR} = \frac{{g_k}^T g_k}{\|g_{k-1}\|^2}$$

$$\beta_k^{PRP} = \frac{{g_k}^T(g_k - g_{k-1})}{\| g_{k-1} \|^2}$$

$$\beta_k^{LS} = \frac{-{g_k}^T y_{k-1}}{{d_{k-1}}^T g_{k-1}}$$

$$\beta_k^{HS} = \frac{{g_k}^T(g_k - g_{k-1})}{(g_k - g_{k-1})^T d_{k-1}} \qquad (7)$$

$$\beta_k^{DY} = \frac{-{d_{k-1}}^T(g_k - g_{k-1})}{\|g_{k-1}\|^2}$$

$$\beta_k^{CD} \frac{-{d_{k-1}}^T g_{k-1})}{\| g_{k-1} \|^2}$$

$g_k - g_{k-1}$ and $\|\cdot\|$ denotes Euclidean norm of vectors.

The remaining part of this paper is organized as follows. In the next section, we discussed related works. Section 3 is devoted to the methods and global convergence of the method. Numerical results are reported in section 4, and we end the paper with conclusion in section 5.

2.0 Related work

It has been shown in the literature that the choices of β_k affect the numerical performance of the method, and hence, many researchers studied choices of β_k (see [7-8, 22-23 and references therein). The CG methods $\beta_k^{FR}, \beta_k^{CD}$ and β_k^{DY} possess strong global convergence properties, but less computational performance [23]. On the other hand, the $\beta_k^{PRP}, \beta_k^{HS}$ and β_k^{LS} methods in general, may not be convergent, but they offer better computational performances [22, 23]. The CG algorithms, according to formula β_k computation, can be classified as classical, hybrid, scaled and parametric [8]. The classical algorithms are defined by (2) and (3), where the CG coefficient is computed as shown in (7). Several modified classical methods are found in literature (see [1-3, 13, 16-17, 22, 28, 30, 36] and references therein). The next categories otherwise known as hybrids are derived to exploit the exciting features of the classical algorithms. The first class of the hybrids combines in a projective manner the classical CG algorithms while the second class considers linear and convex combinations of classical schemes. The latter are recently established in literature. For instance, among the earliest developed hybrid methods may be found in Touti - Ahmed and Storey [32]

$\beta_k^{TaS} = max\{0, \min\{\beta_k^{PRP}, \beta_k^{FR}\}$ (8)

Gilbert and Nocedal [22] extended this result to the case that

$\beta_k^{GN} = max\{-\beta_k^{FR}, \min\{\beta_k^{PRP}, \beta_k^{FR}\}$ (9)

Dai and Yuan [18] proposed a hybrid HS-DY, that is

$\beta_k^{DY} = max\{0, \min\{\beta_k^{HS}, \beta_k^{DY}\}$ (10)

A growing idea of constructing hybrid methods is the use of linear and convex combinations of classical algorithms to develop more robust and efficient schemes. Based on this idea, Xu and Kong [34] suggested a linear combination of β_k^{DY} and β_k^{HS} methods. The parameters β_k is formulated as

$$\beta_k^{(1)} = \begin{cases} a_1\beta_k^{DY} + a_2\beta_k^{HS}, & if\ \| g_k \|^2 > | {g_k}^T g_{k-1} | \\ 0, & otherwise \end{cases}$$

$$\beta_k^{(2)} = \begin{cases} a_1\beta_k^{FR} + a_2\beta_k^{PRP}, & if\ \| g_k \|^2 > | {g_k}^T g_{k-1} | \\ 0, & otherwise \end{cases}$$ (11)

where a_1 and a_2 are non – negative parameters.

The idea of convex combination of CG parameters started with Andrei [11] and further utilized in [4-7, 20, 21, 29]. In the former, the β_k is calculated as

$\beta_k = (1 - \theta_k)\beta_k^{PRP} + \theta_k\beta_k^{DY}$ (12)

while the latter are computed as

$\beta_k^{H1} = (1 - \theta_k)\beta_k^{HS} + \theta_k\beta_k^{DY}$

$\beta_k^{H2} = (1 - \theta_k)\beta_k^{LS} + \theta_k\beta_k^{CD}$ (13)

respectively.

The notion of combining the classical algorithms and quasi-Newton methods was started by Buckley [15]. Several others hybrid CG methods in this category can be found in [24 – 27, 31, 33]. Ibrahim et al. [24] gave the hybrid method known as BFGS - CG and stated the search direction d_k as

$$d_k = \begin{cases} -H_k g_{k,} & k=0 \\ -H_k g_k + \eta(-g_k + \beta_k d_{k-1}) & k \geq 1 \end{cases} \quad (14)$$

where $\eta > 0$, β_k is CG coefficient given by

$$\beta_k = \frac{{g_k}^T g_{k-1}}{{d_k}^T g_{k-1}} \quad (15)$$

and H_k is the approximate Hessian.

In this paper we focus on hybrid CG methods. The other conjugate gradient algorithms classified as scaled and parametric can be found in [9-10, 14, 35]. However, one of the earliest developed three term CG method may be found in Beale [13] as another important innovation to CG methods. Recently, Babaie - Kafaki and Ghanbari [12], gave an extension of the three - term CG method proposed by Zhang et al. [36]. AbbasH Taqi [1], developed a three - term CG algorithm for training feed - forward neural networks which was a vector based training algorithm derived from Davidon-Fletcher-Powell (DFP) quasi - Newton and has 0(n) memory. Application of the three - term CG method to regression analysis was reported by Aliyu et al. [3].

In this paper, a new hybrid nonlinear conjugate gradient method that combines the features of two CGMs proposed by Ibrahim et al. [24] and Zhang et al. [36] is presented.

3.0 Proposed Approach

In this section, we describe the EHCG method. In order to introduce our method, let us simply recall the Broyden-Fletcher-Goldfarb-Shanno conjugate gradient (BFGS – CG) method [24] in which the update parameter β_k is defined by (10). The authors proved the BFGS - CG method can always generate descent directions which satisfy the descent condition

$${g_k}^T d_k < 0 \quad (16)$$

Zhang et al. [36] proposed a three - term CG (MPRP) method as

$$d_k = \begin{cases} -g_{k,} & k = 0 \\ -g_k + {\beta_k}^{PRP} d_{k-1} - \vartheta_k y_k & k \geq 1 \end{cases} \quad (17)$$

where ${\beta_k}^{PRP} = \frac{{g_k}^T y_k}{\|g_{k-1}\|^2}, y_k = g_k - g_{k-1}$ and $\vartheta_k = \frac{{g_k}^T d_{k-1}}{\|g_{k-1}\|^2}$. The authors proposed MPRP method to overcome the drawback of PRP as one of the CG methods that may not always converge for the general objective functions. It was reported in [36] that the parameter ϑ_k as stated above guaranteed that d_k provides a descent direction of f at x_k. Hybridization of various CG methods spawned a new era in CG methods involving large scale unconstrained optimization problems, and this constitute an excellent choice for the solution of unconstrained optimization problem (1). Motivated by the exciting performance of the methods [24, 36] we propose a hybrid search direction d_k that combines the concepts of BFGS - CG and MPRP formulas to come up with

$$d_k = \begin{cases} -H_k g_k, & k = 0 \\ -H_k g_k + \eta\left(-g_k + {\beta_k}^{PRP} d_{k-1}\right) - \vartheta_k y_k, & k \geq 1 \end{cases} \quad (18)$$

where $\eta > 0$, $\vartheta_k = \frac{{g_k}^T d_{k-1}}{\|g_{k-1}\|^2}$, $y_k = g_k - g_{k-1}$, ${\beta_k}^{PRP}$ is a scalar called the CG (update) parameter. This new formula possesses the good properties of the BFGS - CG and also that of MPRP methods. The standard CG coefficient: PRP was used to establish EHCG. Consequently, we summarize this obvious result as the following theorem.

Theorem 3.1: Let d_k be defined by (18). Then, d_k satisfies the descent condition (12).

Based on theorem 3.1, a CG algorithm to implement the proposed method is given as follows:

Algorithm 3.2 (EHCG Method)

1. Given a starting point x_0 and $H_0 = I_n$, choose values for s, β, and σ and set k=1
2. Terminate if $\| g(x_{k+1}) \| < 10^{-6}$ or $k \geq 10{,}000$
3. Calculate the search direction by (18).
4. Calculate the step size α_k by (5).
5. Compute the difference between $s_k = x_k - x_{k-1}$ and $y_k = g_k - g_{k-1}$.
6. Update H_k by (8) to obtain H_{k+1}.
7. Set k = k + 1 and go to Step 1.

3.3 Convergence Analysis

To establish the global convergence of our method, we make the following basic assumptions on the objective function which have been used in literature.

Assumption 3.1: Consider the following.

A: f is bounded below on the level set $S = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ where x_0 is the starting point.

B: In some neighborhood N of S, the function f is continuously differentiable and its gradient, $g(x) = \nabla f(x)$, is Lipschitz continous, i.e. there exist a constant L > 0 such that

$$\| g(x) - g(y) \| \leq L \| x - y \| \quad (19)$$

for all $x, y \in$ N

Under assumption (A) and (B) on f, we state the following famous Zoutendijk condition in [37] as a lemma.

Lemma 3.2: Suppose that assumptions (A) and (B) hold. If the CGM satisfies $g_k{}^T d_k \leq -(\lambda + \eta + \frac{\eta-1}{4\mu}) \| g_k \|^2$ and the step length α_k satisfies the Armijo inexact line search (6), then

$$\sum_{k=0}^{\infty} \frac{(g_k{}^T d_k)^2}{\|d_k\|} < +\infty \quad (20)$$

A straight forward proof of this lemma can be found in [16]. From lemma (3.2) we have the following theorem which presents the global convergence of the proposed method.

Theorem 3.3 (Global Convergence): Suppose that assumption (A) and (B) hold. Let $\{d_k, x_k, \alpha_k\}$ be generated by algorithm 3.1 and α_k determined by the Armijo line search (6), then,

$$\lim_{k\to\infty} \inf \| g_k \| = 0 \quad (21)$$

Proof: Let

$$\lim_{k\to\infty} \inf \| g_k \| \neq 0 \quad (22)$$

then $\| g_k \| > 0$, there exists a constant n > 0 such that $\| g_k \|^2 > n; \forall k$. Since

$$d_k + g_k = \beta_k d_{k-1}$$

$$\| d_k + g_k \|^2 = \beta_k{}^2 \| d_{k-1} \|^2 \quad (23)$$

$$\| d_k \|^2 + 2 d_k{}^T g_k \| g_k \|^2 = \beta_k{}^2 \| d_{k-1} \|^2 \quad (24)$$

$$\| d_k \|^2 = \beta_k{}^2 \| d_{k-1} \|^2 - \| g_k \|^2 - 2 d_k{}^T g_k \quad (25)$$

Divide through by $(d_k{}^T g_k)^2$

$$\frac{\|d_k\|^2}{(d_k{}^T g_k)^2} = \frac{\beta_k{}^2 \|d_{k-1}\|^2}{(d_k{}^T g_k)^2} - \frac{\|g_k\|^2}{(d_k{}^T g_k)^2} - \frac{2 d_k{}^T g_k}{(d_k{}^T g_k)^2} \quad (26)$$

$$\leq \frac{\|d_{k-1}\|^2}{(g_{k-1}{}^T g_k)^2} - \frac{\|g_k\|^2}{(d_k{}^T g_k)^2} - \frac{2}{(d_k{}^T g_k)} \quad (27)$$

$$= \frac{\|d_{k-1}\|^2}{(g_{k-1}{}^T g_k)^2} - \left(\frac{1}{\|g_k\|} + \frac{\|g_k\|}{(g_{k-1}{}^T g_k)^2}\right) + \frac{1}{\|g_k\|^2} \quad (28)$$

$$\leq \frac{\|d_{k-1}\|^2}{(g_{k-1}{}^T g_k)^2} - \frac{1}{\|g_k\|^2} \quad (29)$$

Since $\frac{\|g_k\|^2}{(d_k{}^T g_k)^2} = \frac{1}{\|g_k\|^2}$, then

$$\frac{\|d_{k-1}\|^2}{(g_{k-1}{}^T g_k)^2} \leq \sum \frac{1}{\|g_k\|^2} \leq \frac{k}{n} \forall\, k \quad (30)$$

This implies

$$\sum_{k=0}^{\infty} \frac{(g_k{}^T d_k)^2}{\|d_k\|} = \infty \quad (31)$$

which contradicts lemma 3.2. Hence, the result is proved.

4.0 Numerical Results and Discussion

In this section, we consider some test problems to validate the numerical strength of our method versus some methods in existence such as PRP, BFGS - CG and MPRP. The test problems are from the unconstrained optimization problems in [7] with dimensions varying from 2 to 1000. For the Armijo inexact line search, we use $\sigma = 0.1$, the stopping criteria used are $\| g_k \| \leq 10^{-6}$ or the number of iterations exceeds a limit of 10,000. To better compare the numerical performance of our proposed method against some known CG methods, performance profiles of Dolan and More [19] was used. Performance profiles were drawn for the above methods. In general $\rho(\tau)$ is the fraction of problems with performance ratio τ thus, a solver with high values of $\rho(\tau)$ is preferable. The implementation of the numerical tests was performed on Samsung Notebook PC, Windows 8 operating system, and Matlab 2013 languages.

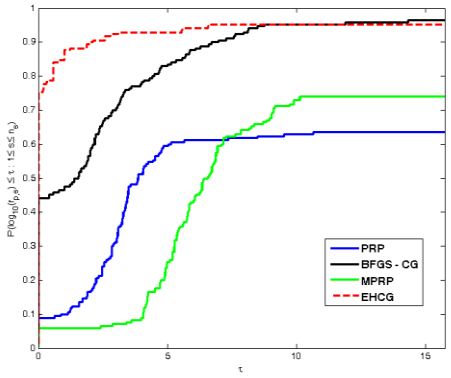


Figure 4.1: Performance Profile based on

numbers of iteration for EHCG versus PRP, BFGS – CG, and MRP.

examined. EHCG performed much better than the others did. It outperforms the PRP, BFGS - CG, MPRP for over 90% of the test problems. In addition, it is also the fastest solver. MPRP method presents the worst performance, since it does not solve all the test problems for both performance of iteration numbers and CPU times.

No	**Test Problems**	**Dim**	**Initial points**
1.	Extended Matyas	2	[1, 1], [5, 5], [10, 10], [50, 50]
2.	Extended Booth	2	[10,10], [20,20], [50,50], [100, 100]
3.	The six-hump	4	[1,1], [2,2], [5,5], [10,10], [-10,-10], [8,8], [-8,-8]
4.	Extended Wood	4	[-1,-1,-1,-1], [-3, -1, -3, -1], [-2, -1, -2, -1], [-4, -1, -4, -1]
5.	Ext. Freud. & Roth	2,4	[2,2], [-2,-2], [5,5], [-5,-5], [8,8], [-8,-8], [10,10], [-10,-10]
6.	Quadratic	2, 4, 10	[5,5] , [20,20], [23,23], [50,50]
7.	Extended Maratos	2, 4, 10	[0,0], [0.5,5], [10, 0.5], [70,70]
8.	Raydan 1	2, 4, 10, 100	[1,1,1]
9.	Quadratic QF1	2, 4, 10, 100	[5,5], [7,7], [10,10], [100,100]
10.	White & Holst	2, 4, 10, 100, 500, 1000	[-3,-3], [6,6], [10,10], [3,3]
11.	Diagonal 4	2, 4, 10, 100, 500, 1000	[2,2], [5,5], [10,10], [15,15]
12.	Ext Rosenbrock	2, 4, 10, 100, 500, 1000	[13,13], [16,16], [20,20], [30,30]

Figure 4.2: Performance Profile based on CPU time for EHCG versus PRP, BFGS – CG, and MRP.

We compared the performance of the new method EHCG with those of PRP, BFGS - CG and MPRP, using the Armijo inexact line search. Figures 4.1 and 4.2 above, show the performance profiles of each method based on the number of iterations and CPU time respectively. Table 4.3 list the problem functions, and table 4.4 show our numerical results by numbers of iteration and CPU time respectively. With respect to the number of iterations and CPU time, for the test problems

TABLE I: LIST OF PROBLEM FUNCTIONS

TABLE II: NUMERICAL RESULTS OF PRP, BFGS - CG, MPRP, EHCG

Problem	Dim	Initial point	PRP	BFGS - CG	MPRP	EHCG
Quadratic	2	[5,5]	31/0.3360	58/0.0633	325/5.2992	17/0.0318
Quadratic	2	[50,50]	35/0.0401	54/0.4104	711/11.7380	6/0.0069
Quadratic	4	[20,20]	25/0.0146	30/0.1573	228/3.7896	10/0.0124
Quadratic	10	[20,20]	72/0.0445	135/0.2195	1736/28.2087	6/0.0087
Quadratic	10	[50,50]	98/0.0601	30/0.0912	715/8.6668	7/0.0084
Quadratic	1000	[23,23]	NAN/NAN	3/0.4231	NAN/NAN	6/2.1984
Quadratic	1000	[50,50]	NAN/NAN	3/0.9195	NAN/NAN	2/0.5595
Diagonal 4	2	[2,2]	33/0.0580	23/0.0438	184/1.8226	6/0.0087
Diagonal 4	2	[10,10]	67/0.0945	35/0.0850	132/1.3688	7/0.0107
Diagonal 4	4	[2,2]	33/0.0612	23/0.1238	184/2.0710	6/0.0082
Diagonal 4	4	[10,10]	69/0.0848	36/0.0908	132/1.5361	7/0.0116
Diagonal 4	10	[2,2]	35/0.0787	24/ 0.1784	184 /2.1314	6/0.0113
Diagonal 4	10	[10,10]	73/0.1057	37/0.0918	132/1.5249	7/0.0238
Ext. Himmelblau	2	[200, 200]	22/0.0345	73/0.1606	713/15.4565	6/0.0227
White & Holst	2	[3,3]	NAN/NAN	NAN/NAN	2627/52.7431	8/0.0239
White & Holst	2	[-3,-3]	NAN/NAN	NAN/NAN	989/17.289	8/0.0651
White & Holst	2	[6,6]	NAN/NAN	NAN/NAN	NAN/NAN	7/0.0232
White & Holst	2	[10,10]	NAN/NAN	NAN/NAN	NAN/NAN	6/0.0479
Ext. Beale	2	[3,3]	57/0.1153	48/0.2374	492/18.8939	6/0.0420
Ext. Beale	2	[15,15]	58/0.1029	21/0.0706	251/9.3354	8/0.0420
Ext. Beale	2	[30,30]	98/0.1886	64/0.3261	333/12.8411	9/0.0336
Ext. Beale	4	[3,3]	57/0.1135	44/0.1078	492/21.3876	6/0.0150
Ext. Beale	4	[15,15]	60/0.0954	21/0.0895	251/9.7683	8/0.0286
Ext. Beale	4	[30,30]	100/0.2294	64/0.3241	333/12.6049	9/0.0396
Ext. Beale	100	[3,3]	65/0.3404	1494/10.6575	490/18.3003	6/0.0305
Ext. Beale	100	[15,15]	58/0.2104	29/0.4808	251/9.7950	8/0.0648
Ext. Beale	100	[30,30]	102/0.3540	41/0.3616	333/12.9407	9/0.1005
Extended Hiebert	10	[500,500]	NAN/NAN	118/0.6289	3170/24.0182	6/0.0688
Extended Hiebert	10	[1000,1000]	NAN/NAN	97/0.3222	2958/18.3399	6/0.0106
Ext. Maratos	2	[0,0]	50/0.0815	3/0.0998	1260/24.5174	3/0.0208
Ext. Maratos	2	[10,0.5]	108/0.1430	22720/32.6646	502/8.7722	6/0.0319
Ext. Cliff	2	[10,10]	100/0.1235	464/1.6146	345/8.4100	6/0.0204
Ext. Cliff	2	[500,500]	190/0.2588	33/0.1253	325/6.1384	19/0.0742
Ext. Cliff	2	[1000,1000]	161/0.1792	32/0.0652	286/6.5715	11/0.0095
Cube	2	[10,10]	88/0.1009	154/0.3169	705/8.1842	6/0.0229
Cube	2	[1000,1000]	976/1.3534	2/0.1010	657/11.5127	6/0.0263
Powell Badly Scale	2	[10,10]	33/0.0832	2142/9.7200	159/4.8948	6/0.0223
Powell Badly Scale	2	[100,100]	33/0.0622	4/0.0288	469/16.1062	6/0.0369
Powell Badly Scale	2	[1000,1000]	39/0.0950	4/0.0224	117/3.1630	7/0.0414
Ext. Quad. Penalty QP1	2	[500,500]	33/0.0321	1146/2.2348	467/7.9047	11/0.0377
Ext. Quad. Penalty QP1	2	[1000,1000]	33/0.0364	1572/3.9050	359/6.0646	8/0.0302
Arwhead	2	[100,100]	36/0.0422	98/0.1485	173/3.4828	11/0.0236
Arwhead	2	[500,500]	52/0.0582	57/0.0868	196/3.9045	11/0.0164
Arwhead	2	[1000,1000]	50/0.0627	45/0.1254	199/3.8135	11/0.0306

Dim: dimension of the test problem. The detailed numerical results are listed in the form NI/CPU, where NI denotes the number of iterations and CPU time in seconds, respectively

5.0 Conclusion

In this paper, we have proposed a new algorithm called extended hybrid BFGS - CG method (EHCG) that possesses the descent condition and also globally convergent. This new hybrid conjugate gradient method was combined with a standard CG coefficient which has been used extensively by numerous researchers. The method was tested on a number of unconstrained optimization problems, and the results show that the proposed method is quite efficient. As part of our future research, numerical computations of our proposed method will be carried out with other line search procedures.

Acknowledgement

The authors would like to thank the anonymous referees for giving us many valuable suggestions and comments, which improve this paper greatly.

REFERENCES

[1] AbbasH Taqi, "Improved Three - Term Conjugate Algorithm for Training Neural Network", *Journal of Kufa for Mathematics and Computer*, 2015, vol. 2, no. 3, p.p 93 – 100.

[2] Adeleke, O. J. and Osinuga, I. A.,"A five-term hybrid conjugate gradient method with global convergence and descent properties for unconstrained optimization problems", Asian Journal of Scientific Research, 2018, DOI: 10.3923/ajsr.2018.

[3] Aliyu U. M., J. L. Wah, S. Ibrahim, "On the Application of Three - Term Conjugate Gradient Method in Regression Analysis", *International Journal of Computer Applications*, 2014, vol. 102, no. 8.

[4] Andrei N., Another hybrid conjugate gradient algorithm for unconstrained optimization. *Numer. Algorithms*, 2008, Vol. 47, no. 2, pp. 143 – 156.

[5] Andrei N., Hybrid conjugate gradient algorithm for unconstrained optimization. *Journal of Optimization Theory Applications*, 2009. Vol. 141, no. 2, pp. 249 – 264.

[6] Andrei N., Accelerated hybrid conjugate gradient algorithm with modified secant condition for unconstrained optimization. *Numer. Algorithms*, 2010, vol. 54, no. 1, pp. 23 – 46.

[7] Andrei N., "Open Problems in Nonlinear Conjugate Gradient Algorithms for Unconstrained Optimization", *Bulletin of the Malaysian Mathematical Sciences Society*, 2011, vol. 34, no. 2, pp. 319 – 330.

[8] Andrei N., "Numerical comparison of conjugate algorithms in unconstrained optimization", Studies in Informatics and Control, 2007, 16, 333-352.

[9] Andrei N., "Scaled memoryless BFGS preconditioned conjugate gradient algorithm for unconstrained optimization", Optimization Methods and Software, 2007, 22(4), 561-571.

[10] Andrei N., "A scaled BFGS preconditioned conjugate gradient algorithm for unconstrained optimization", Applied Mathematics Letters, 2007, 20, 645-650.

[11] Andrei N., "New hybrid conjugate gradient algorithms as a convex combination of PRP and DY for unconstrained optimization". ICI Technical Report, October 1, 2007.

[12] Babaie – Kafaki S., and Ghanbari R., "An Extended Three - Term Conjugate Gradient Method with Sufficient Descent Property, *Miskole Mathematics Notes*, 2015, vol. 16, no. 1, pp. 45 – 55.

[13] Beale E. M. I., "A derivative of conjugate gradients in Numerical Methods for Nonlinear Optimization", F.A. Lootsma, Ed., 1972, pp. 39 – 43, *Academic Press, London, UK*.

[14] Birgin E. G. and Martinez J. M.," A spectral conjugate gradient method for unconstrained Optimization", Applied Mathematics and Optimization, 43, 117 – 128.

[15] Buckley A. G., "A combined conjugate-gradient quasi-Newton minimization algorithm, Mathematical Programming, 1978, 15, 200 – 210.

[16] Can Li, "A Modified Conjugate Gradient Method for Unconstrained Optimization", *TELKOMNIKA*, vol. 11, no. 11, 2013, pp. 6373 – 6380.

[17] Dai Y. H. and Yuan Y., "A nonlinear conjugate gradient with a strong convergence property", *SIAM J. Optim*. 1999, vol. 10, pp. 177 – 182.

[18] Dai Y. H. and Yuan Y., "An efficient hybrid conjugate gradient method for unconstrained optimization,". Annals of Operations Research, 2001, 103,33- 47.

[19] Dolan E., More J. J., "Benchmarking optimization software with performance profile, Mathematical Programming", 2002, vol. 91, pp. 201 – 213.

[20] Djordjevic S. S., "New hybrid conjugate gradient method as a convex combination of FR and PRP Methods", Filomat, 2016, 30(11), 3083 – 3100.

[21] Djordjevic S. S., "New hybrid conjugate gradient method as a convex combination of LS and CD Methods", Filomat, 2017, 31(6), 1813 – 1825.

[22] Gilbert E. G. and Nocedal J., "Global convergence properties of conjugate gradient methods for Optimization, SIAM Journal of Optimization, 1992, 2, 21 -42.

[23] Hager W. W. and Zhang H.,"Asurvey of nonlinear conjugate gradient methods", Pacific Journal of Optimization, 2006, 2, 35 – 58.

[24] Ibrahim M. A. H., Mustafa M., Leong W. J., "The Hybrid BFGS - CG Method in Solving Unconstrained Optimization Problems",. *Abstract and Applied Analysis*, 2014, Vol. 2014 Article ID 507102, 6pp.

[25] Ibrahim M. A. H., Mustafa M., Leong W. J., Azfi Z. S., "The Algorithms of Broyden – CG for Unconstrained Optimization Problems", *International Journal of Mathematical Analysis*, 2014, vol. 8, no. 52, pp. 2591 – 2600, http://dx.doi.org/10.12988/ijma.2014.49272.

[26] Ibrahim M. A. H. B., (2014). The CG – BFGS method for unconstrained optimization problems. AIP Conference Proceedings, 2014, 1605, 167 – 172.

[27] Ibrahim S. M., Mustafa M., Kamil U. K., Abdelrhaman A., "A New Hybrid WYL – AMRI Conjugate Gradient Method With Sufficient Descent Condition for Unconstrained Optimization", *International Journal of Technical Research and Applications*, e-ISSN: 2320 - 9163, 2014, 14 – 17.

[28] Li X., Zhao X., "A Hybrid Conjugate Method for Optimization Problems", *Natural Science*, 2011, 3, 85 – 90.

[29] Li J. K. and Li S. J., "New Hybrid Conjugate Gradient Method for Unconstrained Optimization", *Applied Mathematics Computation*, 2014, 245, 36 – 43.

[30] Li, Q. and Li, D. H. "A class of derivative - free methods for largescale nonlinear monotone equations", *IMA Journal of Numerical Analysis*, 2011, vol. 31, 1625 – 1635. doi:10.1093/imanum/drq015.

[31] Mamat M., Ismail M., Leong W. J., Yosza D., "Hybrid Broyden Method for Unconstrained Optimization", *International Journal of Numerical Methods and Applications*, 2009, 121 – 130.

[32] Touati - Ahmed D., Storey C., "Efficient hybrid conjugate gradient techniques", *J. Optim. Theory Appl*. 1990, vol. 64, no. 2, pp. 379 – 397.

[33] Wan Osman W. F. H., Ibrahim, M. A. H., Mamat M., " Hybrid DFP-CG method for solving Unconstrained optimization problems", Journal of Physics: Conf. Series, 2017, 890, DOI: 10.1088/1742-6596/890/1/012033.

[34] Xu X. and Kong F., "New Hybrid Conjugate Gradient Methods with the

generalized Wolfe Line Search". *Springerplus*, 2016, 5:881.

[35] Yu – Hong Dai, "A Family of Hybrid Conjugate Gradient Methods for Unconstrained Optimization", *Mathematics of Computation*, 2003, Vol. 72, no. 243, pp. 1317 – 1328.

[36] Zhang L., Zhou W. and Li D., "A descent modified Polak – Ribiere – Polyak conjugate method and its global convergence", *IMA J. Numer. Anal*., 2006,Vol. 26, no. 629 – 649.

[37] Zoutendijk G., "Nonlinear programming, computational methods in Integer and Non- linear Programming", *J. Abadie, ed., North - Holland, Amsterdam*, 1970, pp. 37 – 86.

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

AN ONTOLOGY BASED APPROACH FOR IMPROVING JOB SEARCH IN ONLINE JOB PORTALS

O. S. Dada[1], A. F. D. Kana[2], S. E. Abdullahi[3]

[1,2,3]Ahmadu Bello University, Zaria

[1]saintdada2000@gmail.com; [2] donfackkana@gmail.com

ABSTRACT
Internet has become the primary medium for Human Resource Management, specifically job recruitment and employment process. Classical job recruitment portals on the Internet rely on the keyword based search technique in plain text to locate jobs. However, this technique results in high recall, low precision and without considering the semantic similarity between these keywords. Many researchers have proposed semantic matching approaches by developing ontologies as a reference to determine matching accuracy qualitatively, however these approaches do not quantify how closely matched applicants and employers are, based on core skills. This paper proposes a technique that uses an ontology based approach to enhance keyword searching by leveraging on the similarity between concepts in the ontology, which represent core skills needed and required for a job in order to determine how closely matched an applicant is to a job advertisement and vice-versa. This was achieved by developing a CV Ontology based on core skills, annotating applicant profiles and job profiles using a common vocabulary and modifying the semantic concept similarity algorithm to accurately compute and rank matching score between profiles when a query is performed. The results showed improvements of 54% and 36% for Recall and F-measure respectively, over [21].

Keywords: Ontology, Semantic, Algorithm, Core Skills, OWL.

1.0 INTRODUCTION

The Internet has become the primary medium for recruitment and employment processes. Jobberman's Online Recruitment Service Report (2015) claims that applications on its job recruitment portal increased by over 50% between May and September 2015. This clearly indicates an upward trajectory in online job portals being a major player in contemporary job recruitment process. The relevance of the Internet in job recruitment process cannot be overemphasized, more than three-quarter of the age class qualified for recruitment are active internet users and there is an increasing number of companies that publish their job vacancies on the web [17]. Most classical search engines and search mechanism adopted by online job recruitment portals rely heavily on containment

of keywords in free text before search results are returned. This may produce a lot of result from a submitted keyword or phrase but many of these results may be irrelevant to user's need. Therefore, a user may have to navigate through a large number of results to find a domain specific results. Many researchers have proposed several semantic matching approaches and have developed prototype job portals to effectively match job seekers with corresponding job postings [15]. They achieved this by developing human resource or Curriculum vitae (CV) ontologies using controlled vocabularies to determine how applicants are closely related to job positions advertised. however, how to quantitatively and precisely match job seekers with available job postings based on semantic similarity between their core skills and competences relative to the core skills and competences required for the advertised jobs and also ranking the search according to the semantic closeness between applicant profile and job profile relative to their respective core skills set was not provided. Lack of such matching may lead to imprecision and lack of overall effectiveness in matching between available jobs and qualified candidates. This paper therefore seeks to address the above mentioned problems by developing a CV Ontology based on core skills, annotating applicant profiles and job profiles using a common vocabulary and modifying and implementing the semantic concept similarity matching algorithm to accurately compute and rank matching score between profiles when a query is performed.

2.0 Related work

Several works have been carried out in recent times in order to improve the quality of online job recruitment using ontologies in particular and semantic web technologies in general. A system that focuses on the semantic modeling of online recruitment documents was proposed in [8], the author developed an ontology for the database field. The proposed ontology is inspired from the common parts, which were considered significant to CVs and job offers in the field of databases. Furthermore they bring clarifications regarding the essential concepts for the semantic modeling of online recruitment systems, field ontology, semantic annotation, semantic indexing, and semantic association of documents.

[10] Proposed a framework for building intelligent interoperable applications for employment system. This was achieved by collaborating between distributed heterogeneous data models using semantic web technologies. The objective of their work is to provide a better inference system for the query against dynamic collection of information in collaborating data model. Their employment exchange system provides interface for the users to register their details thereby managing the knowledge base dynamically. Semantic server transforms the queries from the employer and jobseeker semantically for possible integration of the two heterogeneous data models to drive intelligent inference. The semantic agent reconciles the syntax and semantic conflicts which exists among the contributing ontologies in different granularity levels and performs automatic integration of two source ontologies and gives response to the user.

[11] Proposed an approach to job matching with user provided information, referred to as parameters. Common parameters for job matching includes domain of job, job title, position, knowledge, experience, location, salary etc. Predefined rules assign weighting factor to each parameter and defining how matching results could be filtered and ranked to produce job matching results. The used an auto-filling technique in places where a candidate has missed out certain important information in their resume. The auto-filing utilized the self-learning engine to collect information, analyses the data and auto generates standard template for different categories group. The standard template is categorized based on some parameters such as qualification, education background and job experience. Their proposed self-learning engine

uses the advantage of ontology to make inference from data in order to discover missing parameters as well as new relationship among the parameters. The inference techniques also improve the possible inconsistencies of various parameters. The system then performs a final job matching based on direct parameters extracted from user input and dynamically populated parameters from matched standard template.

[22] Proposed a qualitative assessment of resumes on the basis of different quality parameters using a simple text analytic based approach for a resume collection. The resume collection is assessed for two qualitative aspects, coverage and comprehensibility; and these ratings are transformed into a comprehensive quality rating. All the quality parameters are collectively measured into a combined 1 to 5 rating scale for determining the quality metric for the resumes. While for coverage, it is simpler; but in case of comprehensibility, it is a bit complex and tricky to transform computed values to 1 to 5 scale rating. Nevertheless, the algorithmic formulation was used in an annotation and recommendation system.

[21] Proposed a technique that uses ontologies to implement a query augmentation that improves defining the context through users adding suggestions of relevant keywords which represent core skills needed for a particular job or task. The intuition is that by searching based on core skills, the context narrows, making it easier to search for any consultant matching a specific assignment or a job seeker searching for a particular job. In his work he claims that the Human Resource Management Ontology was better suited for storing information rather than applying relations between the information. The author created an ontology based on core skills which gives more freedom when querying with SPARQL.

3.0 Proposed Approach

In This section we discuss the system architecture and the modified semantic concept similarity Matching Algorithm

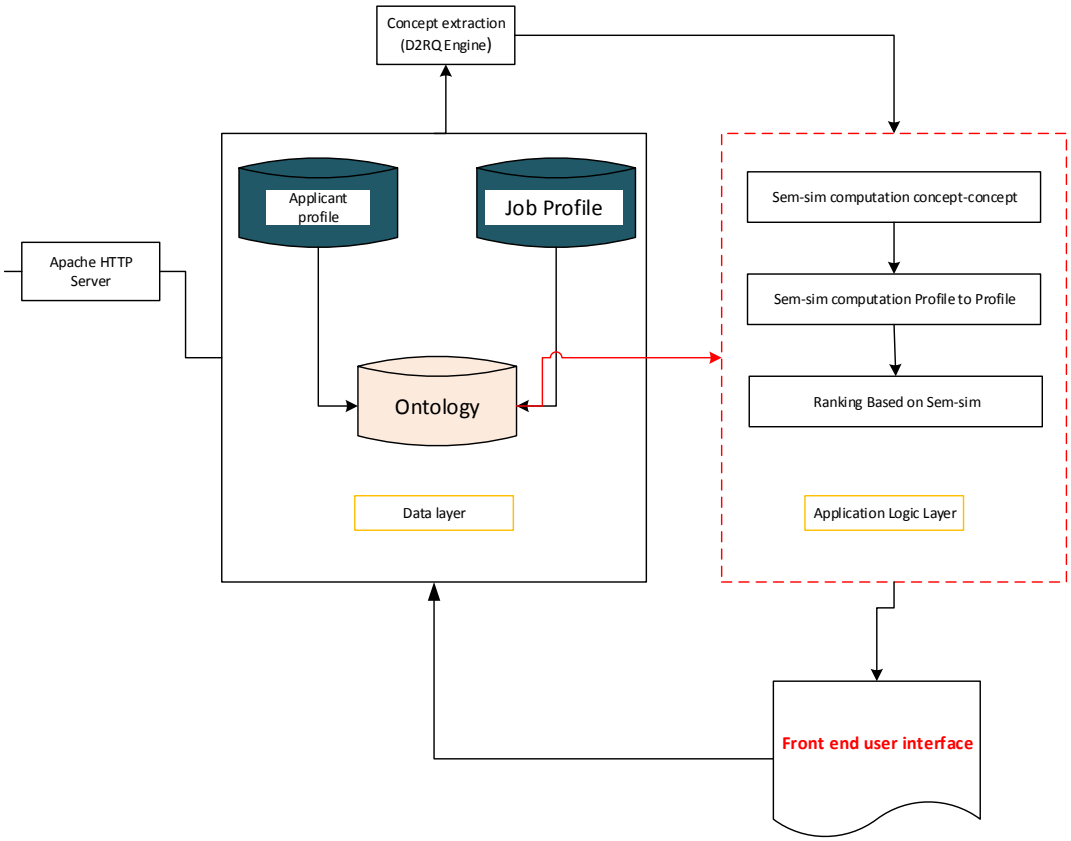


Figure 1: System Architecture

3.1 System Architecture

The system architecture illustrates the different levels of functionality of the proposed ontology based job recruitment portal. The system comprises of three major components or layers.

3.1.2 Data Layer

At this layer, the job applicant's full profile and the Employer profiles with job adverts originally stored in a MySQL database are converted to RDF format using the D2RQ platform. The D2RQ engine, is a system for accessing relational databases as RDF graphs and also allows data stored in RDF to be queried using SPARQL. When a query is performed, core skills are extracted from the applicant and job profiles. Core skills represents concepts on the ontology. These core skills are accessed by the application logic engine from D2RQ for semantic similarity computation

3.1.3 Application Logic Layer

At this layer, classes which represent core skills required in the programming domain initially implemented in an OWL file, are accessed in Java using the OWL2Java framework. The Modified Semantic similarity matching Algorithm is deployed to the ontology where the weight values are assigned to the edges between classes. Furthermore the semantic matching of job profiles and applicant profiles is performed by computing the semantic similarity of the core skills and further ranking based on semantic similarity in relation to the ontology. .

3.1.4 User interface/Front end layer

On top of the architecture is the front end layer which offers a browser-based user interface, which accepts input from users by forms: These forms are used for login and entering details on user profiles. Ranked query results are also displayed to the users.

3.2 The Modified Algorithm

The modified concept similarity matching algorithm based on semantic distance adopts the approach used in [9] where four macro steps are considered in computing semantic similarity between two concepts. However, in this work the semantic similarity is not between only two concepts, but rather a job profile and an applicant profile which both contain multiple skills. The algorithm therefore computes the semantic similarity between all the skills in a job profile with all the skills in an applicant profile in pairs and obtains a total. For a particular applicant profile, it repeats this against all the available job profiles and vice versa on the portal. It further sorts and returns in ascending order all the jobs for which a particular candidate is qualified for or all the available candidates qualified for a particular job which represents a ranking mechanism based on semantic similarity.

The algorithm allocates the weight value to the edge between concept nodes using the weight allocation function defined in [9]. Given two concepts c_1 *and* c_2, the weight allocation function is given as.

$$\mathbf{W[sub(}c_1, c_2\mathbf{)] = 1 + } \frac{1}{k^{depth(c_2)}} \qquad \mathbf{(1)}$$

Equation 1 is the weight allocation function for Parent-child node and sibling nodes. Where, depth(C) is the depth of concept C from the root concept to node., K is a predefined factor larger than 1 indicating the rate at which the weight values decrease along the ontology hierarchy. The equation has two fundamental properties: (a) The semantic differences between upper level concepts are higher than those between lower level concepts, in other words, two general concepts are less similar than two specialized ones. (b) The distance between sibling concepts is greater than the distance between parent and child concepts. Specially, the depth of root is zero and the depth of other concepts is equal to their path length to root concept node.

ALGORITHM: Modified Concept Similarity Matching Algorithm for Profile Matching and ranking

Algorithm Searching (P, $Z\{\ z_0 . . . z_n\ \}$)
// P is the profile to search
// Z is the set of profiles being searched
// f [0 . . . n] is an array of skills of P
// k [i] is an array of skills of z
// n_0 is the root node
// v is the total number of concepts on the ontology
// H [][] is an array which hold weights between node edges
1 H [][] ← Weights
2 For each $z\ in\ Z$
3 k [ ] ← core skills of z
4 Sum ← 0
5 For i = 0 to k.length - 1
6 For j = 0 to f.length - 1
7 If k[i] = f[j] then
8 Sem_Dis (k[i], f[j]) = 0
9 Else if there exists the direct path relation between k[i] and f[j]
10 Sem_Dis (k[i], f[j]) = W_d [sub (H, k[i], f[j])]
11 Else if there exists the indirect path relation between k[i] and f[j]
12 Sem_Dis (k[i], f[j]) = $\sum W_i$ [sub (H, k[i], f[j])
13 Else
14 Sem_Dis (k[i], f[j]) = min {Sem_Dis($k[i], n_0$)} + min {Sem_Dis(f[j], n_0)}
15 End for loop
16 Sum+= Sem_Dis
17 End for loop
18 Q[]. Profile = Z
19 Q[]. Sem_Dis = Sum
// sort Q based on the value
20 End for each
21 Q[].Sort (Sem_Dis)

3.3 Designing the Ontology

The first step requires that the domain and scope of the ontology to be designed must be determined. For this ontology, the domain is the computer programming languages and the scope is to model the basic information as regards core skills/concepts that are associated with Job positions in the aforementioned

domain. Two super classes were created for the ontology in this work. Firstly is the "Person class" which comprises of two subclasses which are the employer and the applicant class. The second is the "Programming language class" which comprises of two major subclasses which are, Hardware development based programming languages subclass and Software development based programming language subclass. Hardware development based programming languages subclass: This class consists of programming languages often required by employers who need programmers for the purpose of hardware development. These programs are commonly used in coding electronic circuits and digital logic circuits. Some of the sub classes used under this category include Assembly, Lisp, Matlab, Pascal and also Hardware Development Language (HDL). HDL class comprises of two sub classes which are Analog and Digital. Software development based programming languages subclass: This class was populated with major programming languages required in the software development circles relative to the needs of the employer. They include C++, C#, Delphi, Java, Ruby, Perl, Objective C, Python, MacRuby, Swift and Visual Basic. The Software Development class had one major sub-class which is Web development based programming which is further broken down into two other major sub-classes these are Classical web programming languages and Semantic web programming languages. Classical web based programming languages are programming languages required by employers for the development of classical web systems and applications, This was further subdivided into two main classes often times as required by employers, which are Front-end Path and Full-Stack Path. These two subclasses were populated with sub classes as shown in the figures 2 and 3. Semantic Web based programming languages includes Languages often required by employers for the development of semantic web based systems and applications basically OWL RDF and XML.

It is important to note that the classes and sub-classes created in this ontology which represents core skills and concepts required for employment in the programming languages domain are not exhaustive but good enough for our proposed modified concept similarity matching algorithm to operate upon. The ontology was developed in Protégé.

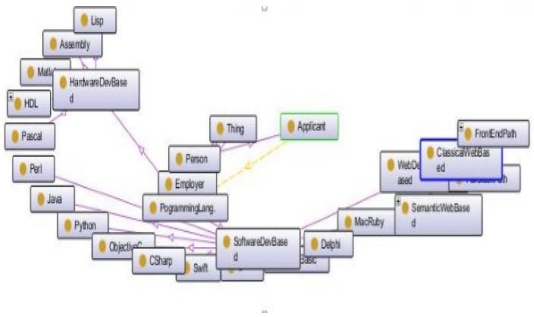


Figure 2: View of CV ontology on Ontograph

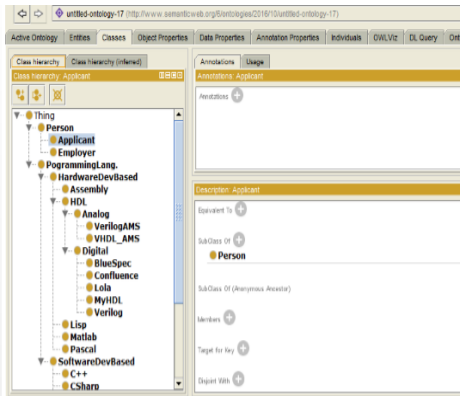


Figure 3: View of CV ontology on Protégé

4.0 Results and Discussion

The following performance metrics were used to determine the performance of our system:

i. Precision
ii. Recall.
iii. F-Measure

Comparison with work by [21] using measure (i), (ii) & (iii) based on the number of web pages to be used for evaluation.

Precision, also known as Positive Predictive Value (PPV), primarily it is used to show how useful search results are. Precision and recall are units of measurement that show an indication of how well a system performs when querying for relevant documents. Given a set of documents divided into relevant and non -relevant documents, Precision measures how many of the documents retrieved by a search are relevant [21]. It's given by Eq. (2)

$$\frac{relevant\ documents\ \cap\ retrieved\ documents}{retrieved\ documents} \quad (2)$$

Recall, sometimes also known as Sensitivity or True Positive Rate (TPR), is similar to precision but looks at how many of the relevant documents are actually retrieved[21].It's given by Eq. (3)

$$\frac{relevant\ documents\ \cap\ retrieved\ documents}{relevant\ documents} \quad (3)$$

F-measure is a combination of both precision and recall, and is a harmonic mean of both the measurements. It tries to give a better measurement of "effectiveness" than recall and precision alone are able to accomplish
F-measure, sometimes also called F-score, has multiple definitions but its main one is defined as follows [21]. It's given by Eq. (4)

$$\textbf{F-measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

4.1 Data Set

A Total of two hundred and forty (160) applicant and employer profiles were created. There were eighty (80) profiles for two core skills, three, and four respectively, all relative to the ontology. A total of sixty (60) profiles were used to test the Precision, Recall and F-measure of the search. The same set of data set were also used to test the performance of [21].

Table 1: Results Obtained for Proposed Portal

No of Skills	Precision	Recall	F-Measure
Two skills	0.91	1	0.95
Three Skills	0.90	1	0.94
Four skills	0.90	1	0.94
Average	**0.90**	**1**	**0.94**

Table 2: Results Obtained (Tran, 2016)

No of Skills	Precision	Recall	F-Measure
Two skills	1	0.22	0.36
Three Skills	1	0.56	0.71
Four skills	1	0.60	0.74
Average	**1**	**0.46**	**0.60**

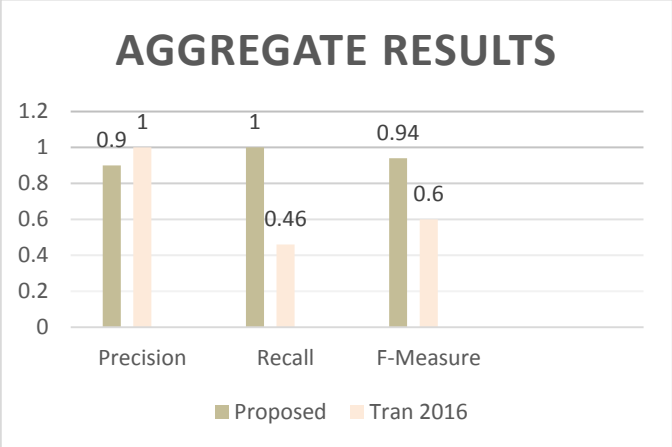


Figure 4**:** The aggregates of the results obtained in this work as compared to Tran (2016) shows that an improvement of 54% and 36% were obtained over Tran (2016) for Recall and F-Measure Respectively

Some of the observations made from the results obtained are summarized as follows:
i) Our experiments indicate that the search and matching is performed to specification since it was able to get a ranked list of all jobs available on the portal relative to the ontology according to the core skills of the applicant or employer.

ii) From the results obtained from this work there were relative improvements of 54% and 36% over Tran (2016) for Recall and F-Measure respectively these improvements were obtained as a result of the implementation of the concept similarity matching algorithm deployed on the ontology as against just leveraging SPARQL query used in [21]. However in the work of [21] a 100% Precision is recorded as against an average of 90% in this work. This is because SPARQL query retrieves only profiles that exactly match the query, thereby making all retrieved profiles relevant to the search.

iii) Query relaxation was achieved as evident in the recall rate, which is 100%. This indicates that all candidate profile on the portal are compared against all the jobs and are indexed in the search result. This is contrary to [21] where only best fit candidates are returned.

5.0 Conclusion

In this research work, a modified concept similarity matching algorithm was deployed to a CV ontology in order to match core skills of applicants to employers more accurately. From the results of experiments carried out on the implemented system, the searching and matching technique recorded an improvement in accuracy of matching. Results affirm the relevance of the concept similarity matching algorithm in determining ontolological closeness or similarities between concepts on an ontology and how it can be applied to improve the employment process.

REFERENCES

[1] Abrouk, L. (2006). Annotation de documents par lecontexte de citation basée sur une ontologie. M.Sc. Thesis, Universite Montpelier II, France. Available from http://tel.archives-ouvertes.fr/docs/00/14/25/68/PDF/these.pdf Retrieved May 17, 2016

[2] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Available from http://www.scientificamerican.com Retrieved March 14, 2016

[3] Bizer, C., Heese, R., Mochol, M., Oldawaski, R., Tolksdorf, R., & Ecstein, R. (2005). The Impact of semantic web technologies on job recruitment process. In: proceedings of International Conference Wirtschaftsinformatik (WI), Bamberg, Germany.

[4] Colucci, S., Di Noia, T., Di Sciascio, E., Donini, F. M., Mongiello, M., Mottola, M. A. (2003). Formal Approach to Ontology-Based Semantic Match of Skills Descriptions. Journal of Universal Computer Science., 9(12), 1437-1454.

[5] Dafoulas, G. A., Nikolaou, A. N., Turega, M., & (2003). E-Services in the Internet Job Market. In: proceedings of 36th Hawaii International Conference on System Sciences, Hawaii, USA.

[6] Fazel, Z., & Fox, M. (2009). Semantic Matchmaking for Job Recruitment: An Ontology Based Hybrid Approach. In: proceedings of the 3rd International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web at the 8th International Semantic Web Conference (ISWC), Washington D.C., USA.

[7] Guissé, A., Lévy, F., Nazarenko, A., & Szulman, S. R. (2009). Annotation sémantique pour l'indexation de règles métiers. Available from

http://www.irit.fr/TIA09/thekey/articles/guisse-levy-nazarenko-szulman.pdf Retrieved March 16, 2016

[8] Ionecusu, B., Ionecusu, I., Florescu, V., & Tinca, A. (2012). Semantic Annotation and Association of Web Documents: A Proposal for Semantic Modelling in the Context of E-recruitment in the IT Field. Journal of accounting and management information system, 11(1), 76-96.

[9] Jike, G., & Yuhui, Q. (2008). Concept Similarity Matching Based on Semantic Distance. In: proceedings of Fourth International Conference on Semantics, Knowledge and Grid. IEEE computer society, Beijing, China.

[10] Kavidha, A., & Saradha, A. (2013). Intelligent interoperable application for employment Exchange using ontology. Webology, 10(2).

[11] May, F., & Yew, C. (2015). Intelligent job matching with self-learning recommendation engine. In: proceedings of 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences,

[12] Mochol, M., Jentzsch, A., & Wache, H. (2007). Suitable employees wanted Find them with semantic techniques. In: proceedings of Workshop on Making Semantics Web For Business at European Semantic Technology Conference (ESTC2007), Vienna, Austria.

[13] Mochol, M., & Paslaru, B. (2006). Practical Guidelines for Building Semantic eRecruitment Applications. . In: proceedings of International Conference on Knowledge Management, Special Track: Advanced Semantic Technologies, Berlin, Germany.

[14] Mulder, W. (2010). Personal informations system e-Entwicklungsstand, Funktionalitat und Trends. Wirtschafts informatik. Journal of Universal Computer Science., 16(2), 98–10.

[15] Niaphruek, P. (2012). Job Recruitment System Using Semantic Web Technology. In: proceedings of 15th International conference of international academy of physical science Pathumthani, Thailand.

[16] Noy, N. F., & McGuinness, D. L. (2016). Ontology Development 101: A Guide to Creating Your First Ontology. Available from URL: http : / / protege . stanford .edu/publications/ontology%7B%5C_%7Ddevelopment/ontology101-noy-mcguinness.htm Retrieved May 5, 2016

[17] Report., J. s. O. R. S. (2015). Available from http://www.jobberman.com/INFOGRAPHIC-Online-Recruitment-Services-Report-2015.htm Retrieved April 21, 2016

[18] Schmidt, A. (2005). Bridging the Gap Between E-Learning and Knowledge Management with Context-Aware Corporate Learning Solutions. Professional Knowledge Management. . In: proceedings of Third Biennial Conference WM, Revised Selected Papers, Lecture Notes in Artificial Intelligence (LNAI).

[19] Schmidt, A. (2006). Context-Aware Workplace Learning Support: Concept, Experiences, and Remaining Challenges In: proceedings of European Conference on Technology-Enhanced Learning (EC-TEL 06).

[20] Sheth, A., & Thirunarayan, K. (2013). Multimedia Semantics empowered Web 3.0 Available from https://books.google.com.ng/books?id=IK4v4fluDJgC&pg=RA2PA7&lpg=RAA7 Retrieved April 20, 2016

[21] Tran, H. (2016). Human Resource Matching Through Query Augmentation for Improving Search Context. (M.Sc. thesis), Linkopings University Sweeden. Available from https://liu.diva-portal.org/smash/get/diva2:941053/FULLTEXT01.pdf

[22] Vinaya, R., Munjula, R., & Sidna, N. (2015). An unstructured text analytics approach for qualitative evaluation of resumes. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 8(2), 2349-2163.

[23] Yahiaoui, L., Boufaida, Z., & Prié, Y. (2006). Semantic Annotation of Documents Applied to e-recruitment. Available from https://www.semanticscholar.org/paper/Semantic-Annotation-of-Documents-Applied-to- E-Yahiaoui-Boufa/pdf Retrieved April 14, 2016

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

FACE DETECTION SYSTEM FOR STUDENTS ATTENDANCE USING PRINCIPAL COMPONENT ANALYSIS (PCA)

Akinrotimi A.O[1], Oladele R.O[2]
University of Ilorin,
Department of Computer Science,
Faculty of Communication and Information Sciences,
Ilorin,
Kwara State.

Abstract: Face detection technology has widely attracted attention due to its enormous application value and market potential, such as face recognition and video surveillance system. Real-time face detection not only is one part of the automatic face recognition system but also fast becoming an independent research subject. As such, there are many approaches to solve face detection problems. This paper introduces a new approach in automatic attendance management systems, extended with computer vision algorithms. We propose using real time face detection algorithms integrated on an existing Learning Management System (LMS), which automatically detects and registers students attending on a lecture. The system represents a supplemental tool for instructors, combining algorithms used in machine learning with adaptive methods used to track facial changes during a longer period of time. This new system aims to be less time consuming than traditional methods, at the same time being non-intrusive and does not interfere with the regular teaching process. The tool promises to offer accurate results and a more detailed reporting system which shows student activity and attendance in a classroom.

Keywords: Face Recognition, Attendance, Principal Component analysis, Feature Extraction, Machine learning.

1.0 INTRODUCTION

The consequences of low attendance are serious and not just affect the students who miss school but also affect the community. The attendance rate tells us the average percentage of students attending school in each day in the given year. Sometimes students and parents might question why school attendance is so important. Parents might think it is not worth fighting with their child to get out of bed and make them to go to the school.

Traditionally student's attendance is taken manually by using attendance sheet, given by the faculty member in class. The Current attendance marking methods are monotonous & time consuming. Manually recorded attendance can be easily manipulated. Moreover, it is difficult to individually identify each student, in a large classroom environment with distributed branches whether the authenticated students are actually responding or not. Hence the paper is proposed to tackle these crucial issues.

Facial recognition or face recognition as it is often referred to, analyses characteristics of a person's face image input through a camera. It measures overall facial structure, distances between eyes, nose, mouth, and jaw edges. These measurements are retained in a database and used as a comparison when a user stands before the camera. One of the strongest positive aspects of facial recognition is that it is non-intrusive. Verification or identification can be accomplished from two feet away or more, without requiring the user to wait for long periods of time or do anything more than look at the camera.

Face recognition is as old as computer vision, both because of the practical importance of the topic and theoretical interest from cognitive scientists. Despite the fact that other methods of identification (such as fingerprints, or iris scans) can be more accurate, face recognition has always remains a major focus of research because of its non-invasive nature and because it is people's primary method of person identification. Face recognition technology is gradually evolving to a universal biometric solution since it requires virtually zero effort from the user end while compared with other biometric options. Biometric face recognition is basically used in three main domains: time attendance systems and employee management; visitor management systems; and last but not the least authorization systems and access control systems.

Maintaining the attendance is very important in all the institutes for checking the performance of employees. Every institute has its own method in this regard. Some take attendance manually using the old paper or file based approach and some have adopted methods of automatic attendance using some biometric techniques. But in these methods, employees have to wait for long time in making a queue at time they enter the office. Many biometric systems are available but the key authentications are same is all the techniques. Every biometric system consists of enrolment process in which unique features of a person is stored in the database and then there are processes of identification and verification. These two processes compare the biometric feature of a person with previously stored template captured at the time of enrollment. Biometric templates can be of many types such as, Fingerprints, Eye Iris, Face, Hand Geometry, Signature, Gait and voice. Our system uses the face recognition approach for the automatic attendance of employees in the office room environment without employees' intervention [1].

2.0 LITEARURE REVIEW

The advancement of technology today has immersed itself towards education. The presence of technology has reached its maximum of providing sustainable technology towards quality education through delivery and effective learning [2]. Attendance is very important for every

student, a single absent is big difference in performance in the school. Most students of high school are prone to absences due to sometimes finding the class boring, being lazy to attend the classes and also some students prefer going to computer shops playing games rather than being in the class while some students cannot refuse the influence of a friend inviting to go with them during class period. Some of this reasons are not reported to the parents or guardians because the way of informing them is the traditional way which often involves inviting the parents to the school and informing him or her about the absenteeism of the student. This process takes a long time and sometimes parents are not able to come because of some reasons such as being busy at work or having some other important matters to take care of. As such, sometimes the individual access management and access control. It is also used to identify persons in groups that are under observation.

2.1.2 Security parameter of biometric technique

The probability of 2 people giving out the same biometric data is virtually zero. Different biometrics systems have been developed around unique characteristics of individuals. Since a biometric belonging is a unique property of an individual, it is particularly difficult to duplicate or share (you cannot give a copy of your appearance or your hand to someone!).

2.2 Survey of Different Attendance Tracking Systems

Following traditional systems are used to mark attendance in the teaching process

.

2.2.1 Computerized attendance system

Sharma et al. [3], develop a desktop application in which all the list of registered students in a particular course will be displayed when the lecturer start the application. The attendance is done by clicking a check box next to the name of the students that are present, and then clicked on register button to mark their presence. But in this parents are not informed about the absenteeism of the students.

2.1 Face Detection System

Due to its enormous application value and market potential, such as face detection and video surveillance system, face detection equipments has widely attracted attention. Face detection is a technology that determines the location and sizes of human faces in an image. It detects faces and ignores everything else.

2.1.1 A Biometric Approach

Biometrics consists of method for individually recognizing humans based upon one or more unique physical or behavioral qualities. In computer science, biometrics is used as a form of

also, human involvement for attendance tracking is needed.

2.2.2 Bluetooth based attendance system

Vishal Bhalla [4], have proposed the attendance system which can take attendance using Bluetooth. In this project, attendance is being taken using instructor's mobile phone. Application software is installed in instructor's mobile telephone. This enables it to query student's mobile telephone via Bluetooth connection and perform a transfer of student's mobile telephone Media Access Control (MAC) addresses to the instructor's mobile telephone. The problem of this proposed system is student's phone is required for attendance. If student didn't carry the mobile phone with him without mobile phone his presence will not considered in Bluetooth Based Attendance System. The second problem of this proposed system is, in case of students' absent if his mobile is given to his friend then also present is marked, so presence of student is not necessary only phone should be in coverage area.

2.2.3 NFC based attendance system

An implementation of an (AMS) Attendance Management System, based on Bluetooth and

NFC technologies in a multiuser environment is presented in [1]. It uses fingerprint & the Bluetooth address of the NFC enabled phone of the user to authenticate the identity of the user. A Java based desktop application receives the NFC tag IDs, other information associated with the mobile phone and the user and submits them to an analyzer for the interpretation of the user's behavior. But in this case, student must be having NFC enabled phone to mark presence in the class room.

2.2.4 Fingerprint based attendance system

An employee attendance system using fingerprint is presented in [5]. This system checks one fingerprint template with all templates stored in the database. The main problem in this case is it is very time consuming as it check one fingerprint with all the temple stored in the database.

Fingerprint recognition based identification system is designed in [6] for student identification. This system is being designed for taking attendance in institutes like NIT Rourkela. In this system, fingerprint template matching time is reduced by partitioning database. In this system all students of every class has to stand in a long queues to make attendance, again this system is suffering from fingerprint device , and one most important disadvantage is that it is work within short distance.

2.2.5 Iris Based Attendance System

A wireless iris recognition attendance management system is designed and implemented [7] using Daugman's algorithm [8]. This system based biometrics and wireless technique solves the problem of spurious attendance. It can take the users' attendances more easily and effectively. The system is based on RF wireless technique, however it is very expensive. In this system, all the students of every class has to stand in a long queues to take attendance, and most important disadvantage is that it is too capital intensive.

3.0 METHODOLOGY

3.1 System Algorithm

This section describes the software algorithm for our proposed system. The algorithm consists of the following steps:

1. Database Making
2. Feature Extraction
3. Face Recognition
4. Attendance Marking
5. View Attendance

3.1.1 Database Development

This is the first step in which information are computed into the database. Details of student such as student ID, matric number, department, level etc. are added in database along with their passport images. The database is developed using MySQL server.

3.1.2 Feature extraction

In the proposed system algorithm, features of the face image are extracted using Principal Component Analysis (PCA). Principal component analysis (PCA) is a statistical dimensionality reduction method, which produces the optimal linear least-square decomposition of a training set. Kirby and Sirovich applied PCA for representing faces and Turk and Pentland extended PCA for identifying faces. In applications such as image compression and face recognition a helpful statistical technique called PCA is utilized and is a widespread technique for determining patterns in data of large dimension. PCA commonly referred to as the use of Eigen faces.

The PCA approach is then applied to reduce the dimension of the data by means of data compression, and reveals the most effective low dimensional structure of facial patterns. The advantage of this reduction in dimensions is that it removes information that is not useful and specifically decomposes the structure of face into components which are uncorrelated and are

known as Eigen faces. Each image of face may be stored in a 1D array which is the representation of the weighted sum (feature vector) of the Eigen faces.In case of this approach a complete front view of face is needed; or else the output of recognition will not be accurate.The major benefit of this method is that it can trim down the data required to recognize the entity to 1/1000th of the data existing.

The estimate covariance matrix to represent the scatter degree of all feature vectors related to the average vector is also calculated. The covariance matrix *C* is defined by:

$$CV = \lambda V \qquad (\ , \neq 0) \qquad (3)$$

Where *V* is the set of eigenvectors matrix *C* associated with its eigenvalue λ. Project all the training images of i^{th} person to corresponding Eigen-subspace:

$$y_k{}^i = w^T(x_i)(i = 1, 2, 3 \ldots\ldots\ldots\ldots N) \qquad (4)$$

Where the y_k are the projections of x and called the principal components also known as Eigen faces. The dimensionality can be reduced by selecting the first Neigenvectors that have large variances and discarding the remaining ones that have small variance.

Steps involved in PCA are:

Step 1: Get image data in form of matrix as A1,A2….Am.
Step 2: Calculate the mean Ø as Σ Al /M.
Step 3: Subtract mean from original data (image data) A1- Ø.
Step 4: Calculate covariance matrix C= ATA.
Step 5: Calculate eigenvalues of the covariance matrix.
Step 6: Choose components with highest information and form a feature vector.
Step 7: Derive new data set and project the eigenfaces image.

The following steps summarize the process PCA. Let a face image $X(x, y)$ be a two dimensional *m* x *n* array of intensity values. An image may also be considering the vector of dimension mn, so that a typical image of size 112x92 becomes a vector of dimension 10304. Let the training set of images $\{X_1, X_2, X_3 \ldots X_N\}$. The average face of the set is defined by:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (1)$$

$$C = \frac{1}{N}\sum_{i=1}^{N}(\bar{X} - X_i)(\bar{X} - X_i)^T \qquad (2)$$

The Eigenvectors and corresponding Eigen-values are computed by using

4.0 SYSTEM IMPLEMENTATION

4.1.1 Database Structure

Frontal faces of students obtained from University of Ilorin students were used to construct the image database of the system. For effective result production the images loaded to the database are images that have passed through geometric normalization. The system was developed with ninety six sample images of twenty four individuals making four images per individual.

Figure 2.0: Sample Facial Database

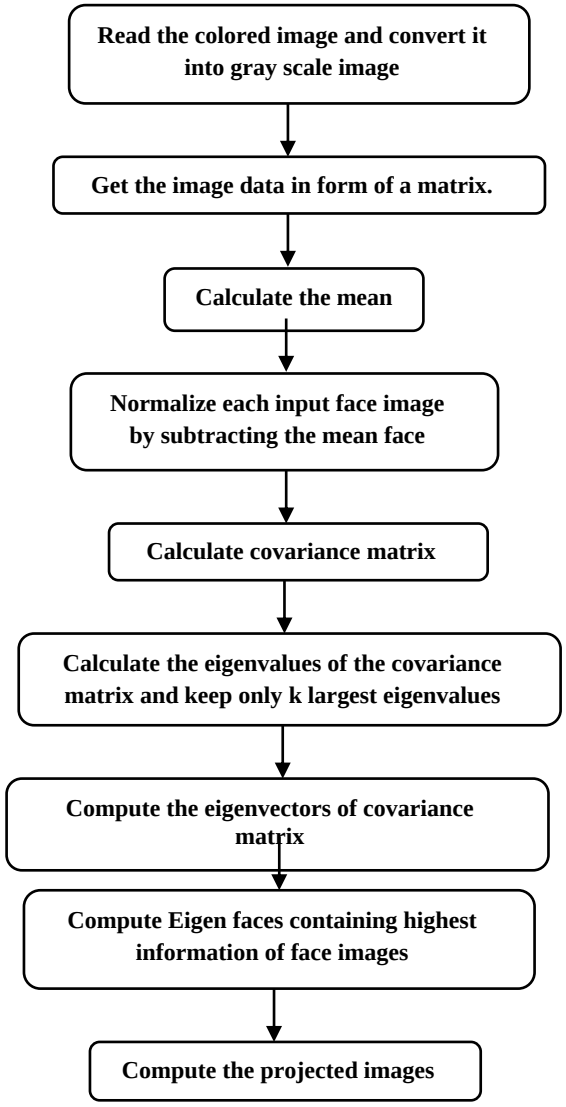
4.1.2 Face Normalization

The face database were normalized geometrically by cropping and resized to 100 x 100 dimension in Matlab Simulink environment. The resizing dimension of the images had a little effect in performance of the recognition accuracy. There are total number of 96 images for the facial database, the images were divided into 75% for

training and 25% for testing making a sum of 72 images for training and 24 images for testing.

4.1.3 Photometric Face Normalization

The facial images after geometric normalization were further subjected to the contrast limited adaptive histogram equalization technique so as to improve contrast on all parts of the face image and make for adjustment in the intensity of the face images



.

Figure 1.0: Sequence of events for performing PCA

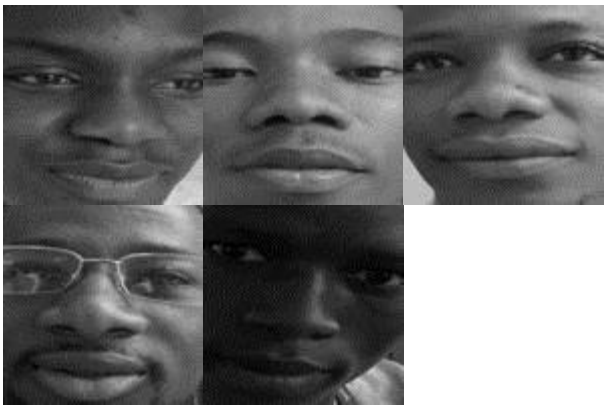


Figure 3.0: Sample Face Image after photometric normalization

4.1.4 Feature extraction

After the pre-processing stage, the normalized face image is given as input to the feature extraction module to find the key features that will be used for classification. The module composes a feature vector that is well enough to represent the face image. The feature extraction algorithm is the principal component analysis. The second method for extracting gainful information from the preprocessed and normalized images so as to generate a reduce

4.2 Experimental Results

With the help of a pattern classifier, the extracted features of face image are compared with the ones stored in the face database. The Distance for comparing/matching of the test and trained images. In the testing phase each test image is mean centered. The test image is then projected into the same Eigen space as defined during the training phase. This projected image is now compared with projected training image in Eigen space. Images are compared with similarity features for training the system the observed extracted feature vector is the normal dot principal component analysis and the observed result is shown in figure 4.0.

Figure 4.0: PCA Feature Vector

face image is then classified as either known or unknown. After feature extraction step next is the classification step which makes use of Euclidean

measures. The training image that is the closest to the test image will be matched and used to identify the individual. Also the relative Euclidean distance between the testing image and the reconstructed image of i^{th} person is calculated and the minimum distance gives the best match.

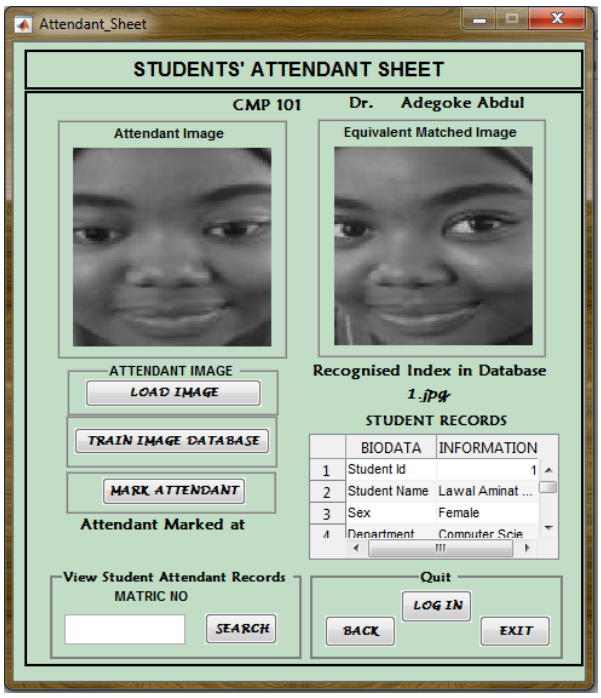


Figure 4.3: Student's Face Recognition Interface

4.2.1 Attendance marking

As soon as a face is recognized, attendance is marked in the database (corresponding to the matched face as the information is already stored in database in the first step). If an unknown face is tested the result shows no match found.

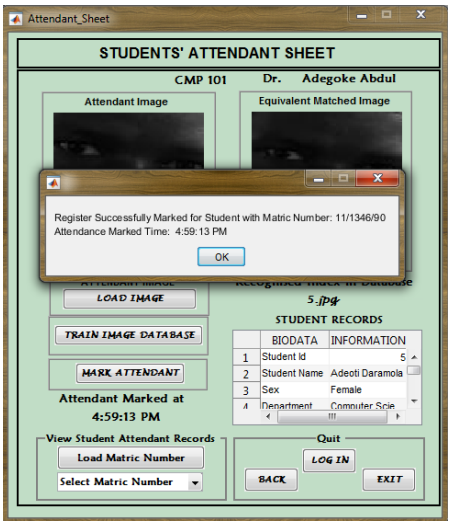


Figure 4.3: Student's Attendance Marking Interface after face recognition

4.3 Facial Attendance System Evaluation

This option lets the user to view his attendance in the database. By entering the value in the option given for dates the student can see his attendance for that period of time. On clicking on the name of a student, his attendance percentage can be seen. If the percentage is less than 75% then a message will be displayed that his attendance is short.

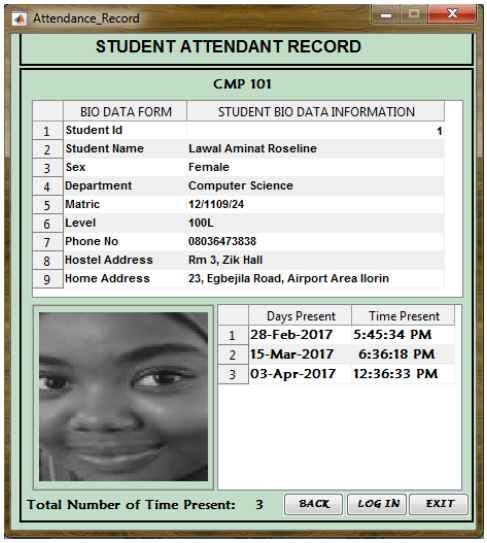


Figure 4.3: Student's Attendance Record Interface

5.0 CONCLUSION

An automatic attendance management system is a necessary tool for any institution. Most of the existing systems are time consuming and require for a semi manual work from the teacher or students. This approach has attempted to solve the issues by integrating face recognition in the process. Even though this system still lacks the ability to identify each student present on class, there is still much more room for improvement. Since we implement a modular approach, we can improve different modules until we reach an acceptable detection and identification rate. Another issue that has to be taken in consideration in the future is a method to ensure users privacy. Whenever an image is stored on our servers, it must be impossible for a person to use that image.

REFERENCES

[1]. Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. (2012) Surf: Speeded up robust features. Computer Vision and Image Understanding (CVIU), 110(3):346–359.

[2]. Raymond H. Chan, Chai J.F. Shen Z. (2008). A framelet-based image in painting algorithm. Applied and Computational Harmonic Analysis 24(2), 131-149.

[3]. S. K. Jain, U. Joshi, and B. K. Sharma (2010). Development of student attendance management system using RFID and face recognition 1428-1435.

[4]. V. Bhalla, T. Mittal, S. Saroha, V. Khullar, H. Tyagi, RA Taylor, TP Otanicar (2013). Numerical study of solar photovoltaic/thermal (PV/T) hybrid collector using Nano-fluids ASME Paper No. MNHMT2013-22090.

[5]. Seema Rao and Prof. K.J. Satoa (2013). International journal of Engineering applied Science and Technology 1(8) 287-293, ISSN No. 2455-2143

[6]. Neha Verma, Komal Sethi and Megha Raghav, (2013) An Efficient Automatic Attendance System Using Fingerprint Reconstruction Technique. International Journal of Advance Research in Science and Engineering. IJARSE, 2(3).

[7]. Seifedine Kadry and Mohamad Smaili, (2010). Wireless attendance management system based on iris recognition.

[8]. Daugman .J, (2003). The importance of being random: Statistical principles of iris recognition, Pattern Recognition 36(2): 279-29.

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

E-IASAODV: An Enhanced Framework for Preventing Ad-hoc on-demand Distance Vector (AODV) Multiple Black Hole Attack in MANET

S. M. Mubarak[1], A.A. Obiniyi[2], S. Aliyu[3]

[1,2,3,] Faculty of Physical Science, Department of Computer Science, Ahmadu Bello University, Zaria. Nigeria.

[1]musanimu@gmail.com; [2]A.A.Obiniyi@gmail.com; [3]Salis_001@yahoo.com

ABSTRACT
Mobile Ad-hoc networks (MANETS) are collection of mobile nodes that dynamically change the network topology in which nodes can join and leave the network at any point of time. Due to fundamental characteristics of MANETS, such as open medium, dynamic topology, and distributed cooperation; it creates several security vulnerabilities to its security design. Security is an essential requirement in mobile Ad-hoc networks to provide secure communication between mobile nodes. Ad-Hoc On-demand Distance Vector (AODV) routing protocol is a routing protocol in MANET that broadcast the network with a route discovery message anytime a node is seeking for a route to a destination, any node that have a route to that destination will reply to the route discovery request, which provides a vulnerability to the routing protocol by making it open to black hole attack which is one of the most common attacks in MANETs. A Black Hole is a malicious node that falsely replies to any route requests without having active route to specified destination and drops all received packets. This work, which is an enhancement of Intrusion Avoidance System for Ad-Hoc on-demand Distance Vector (IASAODV), a framework developed in 2015 to prevent black hole attack, presents a new framework that prevents the security threats of AODV multiple Black Hole attack with better Packet Delivery Ratio (PDR). This framework tackled the problem by making nodes monitor the activities of their neighbors by collecting Route Request (RREQ) messages sent by nodes and keeping in a table. To justify the solution, we made appropriate implementation and simulation using Network Simulator NS-2.35. The conducted experimental result shows an improvement in Packet Delivery Ratio (PDR) compared to that of IASAODV routing Protocol with the proposed framework having 100%, 100%, 99% and 98% for 1, 3, 5 and 7 malicious nodes respectively compared to the existing system with 88%, 79%, 61% and 57%, for 1,3,5 and 7 malicious nodes respectively

***Keywords*:** MANETs, AODV, IASAODV, RREQ, RREP, NS2, HRRT

1.0 INTRODUCTION

Wireless mobile ad-hoc network (or simply MANET) is a self-configuring network which is composed of a lot of movable user equipment. These mobile nodes communicate with each other without any infrastructure. Furthermore, all the transmission links are established through the wireless medium. MANET is widely used in places such as military, disaster area and personal area network [1]. However, there are still many open issues about MANET, such as security problem, finite transmission bandwidth, abusive broadcasting messages, reliable data delivery, dynamic link establishment and restricted hardware caused processing capabilities [6]. The security threats have been extensively discussed and investigated in the wired and wireless networks, the correspondingly perplexing situation has also happened in MANET due to the inherent design defects. There are many security issues which have been studied in recent years. For instance, snooping attacks, wormhole attacks, black hole attacks [8], routing table overflow and poisoning attacks, packet replication, denial of service (DoS) attacks, distributed DoS (DDoS) attacks [2], especially, the misbehavior routing problem which is one of the popularized security threats such as black hole attacks. Securing Ad-Hoc routing faces difficulties which do not exist in wired networks, nor in infrastructure-based wireless networks. These difficulties make trust establishment among nodes virtually impossible. Among these difficulties are the wireless medium itself and its physical vulnerability, the lack of centralized control and permanent trust infrastructure, the cooperation of nodes, restricted power and resources, highly dynamic topology and short-lived connectivity and availability, an implicit trust relationship between neighbors and other problems associated with wireless communication.

1.1 Ad-Hoc on-demand Distance Vector (AODV) Routing protocol

AODV routing protocol is based on DSDV and DSR algorithm and is a state-of -the -art routing protocol that adopts a purely reactive strategy. It sets up a route on demand at the start of communication and uses it till it breaks after which a new route setup is initiated [13]. AODV routing protocol is a reactive or on-demand routing protocol, which means that a route between two nodes will be determined only when there is data to be transmitted [3]. Each node's routing table only contains the next hop to a destination, so the information on the route to be traversed by a packet is distributed to all the nodes on the path. Neighbor connectivity is established with periodic Hello Messages. Routes are found by the flooding of route request (RREQ) messages as can be seen in Figure 1. As each node receives and retransmits the RREQ it records the previous hop in its routing table. In AODV, when a source node S wants to send a data packet to a destination node D and does not have a route to D, it initiates route discovery by broadcasting a route request (RREQ) to its neighbors. A timer called route reply waiting time (RREP_WAIT_TIME) is started when the RREQ is sent. The immediate neighbors who receive this RREQ rebroadcast the same RREQ to their neighbors. This process is repeated until the RREQ reaches the destination node. Upon receiving the first arrived RREQ, the destination node sends a route reply (RREP), as can be seen in Figure 2, to the source node through the reverse path where the RREQ arrived. The destination node will ignore the same RREQ that arrives later. In addition, AODV enables intermediate nodes that have sufficiently fresh routes (with destination sequence number equal or greater than the one in the RREQ) to generate and send an RREP to the source node. Once the source receives the

first RREP message, it starts the data transmission along the path traced by the RREP packet [9].

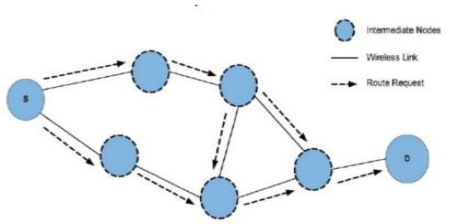


Figure1: AODV RREQ [9]

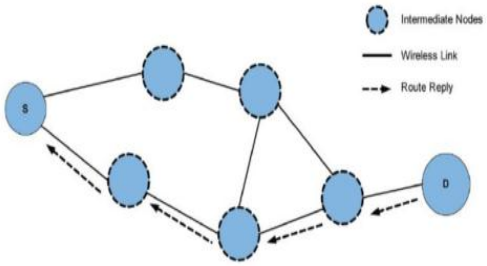


Figure2: AODV RREP [9]

AODV provides a rapid, dynamic network connection, featuring low processing loads and low memory consumption. AODV uses a node sequence number to distinguish whether the routing message is fresh or not. Node sequence numbers serve as time stamps and allow nodes to compare how fresh their information on the other node is. However, when a node sends any type of routing control messages such as RREQ and RREP it increases its own sequence number. Higher node sequence number means that the fresh route to the destination can be established over this node by other nodes. The sequence number is a 32-bit unsigned integer value (i.e., 4294967295) [9].

1.2 Black Hole Attack

Routing protocols are exposed to a variety of attacks. Black hole attack is one such attack and kind of DOS, in which a malicious (fake) node makes use of the vulnerabilities of the route discovery packets of routing protocols to advertise it as having the shortest path and higher sequence number to the node whose packets it wants to intercept (Nirali and Gupta, 2014). This attack aims at modifying the routing protocol so that traffic flows through a specific node controlled by the attacker. During route discovery phase, the source node sends the RREQ packet to the intended destination. Malicious nodes respond immediately to the source nodes as these nodes do not refer to the routing table. The source node assumes that the route discovery phase is complete and ignores other RREP messages from other nodes and selects the path through the malicious node to route the data packets. The malicious node does this by assigning a high sequence number to reply packets. Consider figure3, a malicious node M reply to the source node S as having the best route to D. In figure 4, the source node after receiving an RREP from the malicious node is now sending the packet to the node that will either drop or consume it.

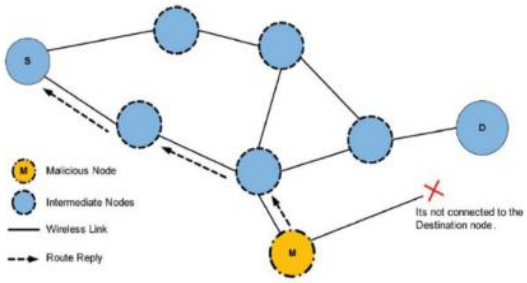


Figure3: A malicious node sending false RREP to the source node [9]

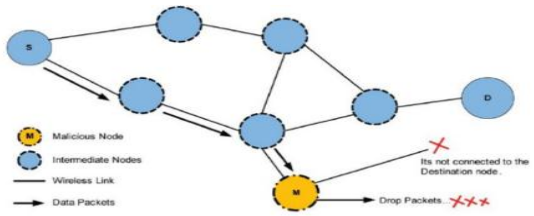


Figure 4: A malicious node dropping data packets as the source node is unaware [9]

2.0 Related work

Zapata [15] proposed a secure ad-hoc on-demand distance vector protocol (SAODV), to reduce the attack, this algorithm proposes to wait and check the replies from the entire neighboring node. The source node will store sequence number and the time at which the packet arrives in a Collect Route Reply Table (CRRT). If more than one path exists in CRRT, then it randomly chooses a path from the CRRT. This reduces the chance of black hole attack. Unfortunately, selecting path randomly gives multiple black hole nodes room to collaboratively attack and be randomly picked, secondly, each node maintaining a CRRT will cause a big overhead that will reduce the performance of the network. Chin *et al.* [4] in their study, proposed a specification-based intrusion detection system that can detect attacks on the AODV routing protocol. They used finite state machines for specifying correct AODV routing behavior and distributed network monitors for detecting run-time violation of the specifications. In addition, they proposed one additional field in the protocol message to enable the monitoring. Their work suffered from too much end-to-end delay due to distribution of network monitors all over the network and system overhead due to the provision of additional field in the protocol.

In another study, [10] proposed a mechanism called Anti-Black Hole mechanisms (ABM), which are intrusion detection systems (IDS) deployed in MANETs to detect and prevent selective black hole attacks. The nodes must be in sniff mode to perform the so-called anti-black hole mechanism function, which is mainly used to estimate a suspicious value of a node according to the abnormal difference between the routing messages transmitted from the node. When a suspicious value exceeds a threshold, an IDS nearby will broadcast a block message, informing all nodes on the networks, asking them to cooperatively isolate the malicious node. Ira and Chaki [7] in their research tried to prevent the black-hole attack in the network using the concept of clustering. According to the characteristics of black-hole node deployment, the black-hole nodes come in the path from source node to the destination node and it only receives data packets but never forward to further destination nodes. They proposed a friendship table for cluster heads. Their scheme provided a solution to only external nodes joining a cluster; all internal nodes in each cluster are measured as trusted nodes. The friendship table is used by each cluster head to identify the relationship between cluster heads and neighbors which are measured in term of "friend" or "stranger". Their work failed to consider internal malicious nodes and malicious cluster heads. Swadas and Raj [11], Proposed a new control packet called ALARM is used in DPRAODV, while other main concepts are the dynamic threshold value. Unlike normal AODV, the RREP_seq_no is extra checked whether higher than the threshold value or not. If the value of RREP_seq_no is higher than the threshold value, the sender is regarded as an attacker and updated it to the black list. Mahmoud *et al.* [9], Proposed a new Intrusion Avoidance System (IASAODV) which is a modification of the AODV protocol. They consider that the

source node must wait a time equals the double value of traditional AODV RREP_WAIT_TIME, before sending data, to receive more RREP messages. Once a source receives the RREP messages it will store its sequence number and the time at which the message arrives at a table. They refer to this table by Route Reply Table (RRT). Route Replies that have a sequence number that is greater or equal to and the hop count in that route reply is less than the hop count in source routing table are added into the (RRT). When the timer, RREP_WAIT_TIME expires, and then their proposed algorithm will check the number of RREP messages in RRT. If RRT contains one RREP message and its sequence number is less than the maximum sequence number value, then the source sends the data to the destination. If the RRT contains more than one RREP messages, then the RRT is considered to have a black hole node(s). To detect this black hole node(s), the algorithm compares the nodes sequence number of each RREP messages which exist in the RRT. The nodes with the maximum sequence number will be considered black hole node and will be inserted into the blacklist table. The limitation observed in their work is; The use of sequence number only for decision making will not be much efficient in preventing multiple black hole attack because the nodes can act in a collaborative fashion and guess the given sequence number.

3.0 Proposed Algorithm

The proposed algorithm is designed to prevent any alterations in the default operations of either the intermediate nodes or the destination nodes. After analyzing the effect of Black Hole attack in MANETS, we modify the IASAODV Protocol and propose a technique that will be able to prevent multiple black hole attack. Our proposed framework will monitor nodes participation in the network by collecting the messages sent my nodes within the network, and saving that information including the identity of those that sent the messages in a table called History Route Request Table (HRRT). We considered that the proposed framework must wait a route reply time double that of the traditional AODV routing protocol in other to receive multiple replies, including that of the black hole nodes if there is any in the network. Once a source node receives a route reply, it will go into its HRRT and check whether the nodes that replied have been active in the network, i.e it has ever received an RREQ from that same node. If yes, then it will unicast the data packet to it, if no, it will mark it as a black hole node and remove it out of the HHRT.

The proposed framework consists of two components;

- Collection of RREQ
- Initiating Route Discovery, collecting RREP and black hole node detection.

```
1. Begin
2. While (received_RREQ)
3. {
4.    If HRRT contain (received_RREQ)
5.   Discard RREQ
6.    Else
7.  Add recived_RREP_nodeID into
    HRRT
8. }
9. End
```

Figure 5: Algorithm for collecting Route Request

```
1.  Begin
2.  Replying node = NULL
3.  Reply_count = 0
4.   Initiate route Discovery (RREQ)
5.   Set Time (RREP_WAIT_TIME)
```

6. Set Maximum Sequence Number

7. While received Route Reply
8. If Destination sequence number in RREP is ≥ Node source Sequence Number AND hop_count in RREP is < node_hop_count in source routing table
9. Add this RREP Message into HRRT
10. Else

11. Discard RREP Message.
12. End if
13. Reply_count = Reply_count+1.
14. If reply count >= 1
15. Check HRRT for the History of the received replies
16. If received RREP_NODE_ID is a friend
17. Send Data packet to the route specified in the RREP
18. Else

19. Remove This RREP Message from HRRT
20. End if.
21. Else
22. Go back to line 13
23. End if
24. End

Figure 6: Algorithm for collecting Route Replies and checking for malicious nodes.

4.0 Simulation Results and Analysis

The conducted simulation experiments were performed using NS-2 Ver. 2.35 simulator run on Linux Ubuntu version 14.04. The simulation models a network of 50 mobile nodes migrating with square area size 750 x 750 m2. The mobility model uses the random waypoint (RWP) model in the considered area. In this, each node is randomly placed in the simulated area and remains stationary for a specified pause. We chose out traffic sources to be Constant Bit Rate (CBR). Each CBR packet size is 512 bytes. Table 1 summarizes the used simulation parameters.

Table 1: Simulation Parameters

PARAMETERS	**VALUES**
Simulator	NS-2 (Ver. 2.35)
Simulation Time S	500 secs
Number of mobile nodes	50
Number of Black Hole nodes	1,3,5 and 7 nodes
Simulation area	750 m X 750 m
Transmission range	250 m
Routing Protocol	IASAODV and the proposed
Traffic Type	Constant Bit Rate (CBR)
Maximum Speed	20 m/s
Packet size	512bytes
Mobility model	Random Waypoint
Data Rate	4Mbps

RREP_WAIT_TIME	2sec

To evaluate the performance of our proposed framework, we compared it with IASAODV protocols. We used the packet delivery ratio PDR and Average End-to-End Delay as performance metrics. PDR is defined as the ratio of the data packets received at the destination station compared to the total of data packets transmitted by the source node. The Average End-to-End Delay is defined as the average time employed for a data packet to be delivered from the source node to the destination node.

We consider Four simulation scenarios:

a. One black hole attack scenario.
b. Three black hole attack scenario.
c. Five black hole attack scenario.
d. Seven black hole attack scenario.

Figure 7 and 8 represent the simulation results of the considered protocols. In each Figure, the number of black hole nodes is considered versus packet delivery ratio and average end-to-end delay respectively for both the proposed and the existing frameworks.

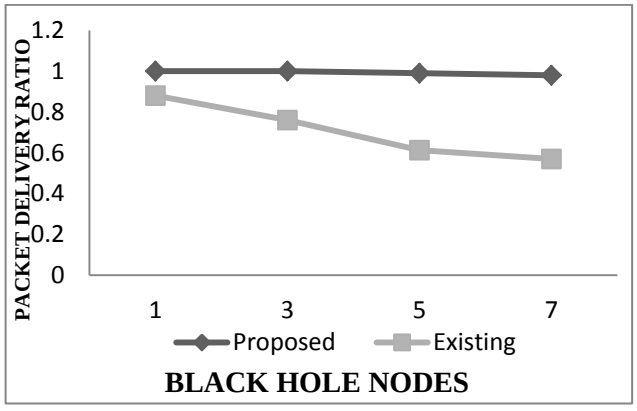


Figure7: Packet Delivery Ratio VS No. of black hole nodes

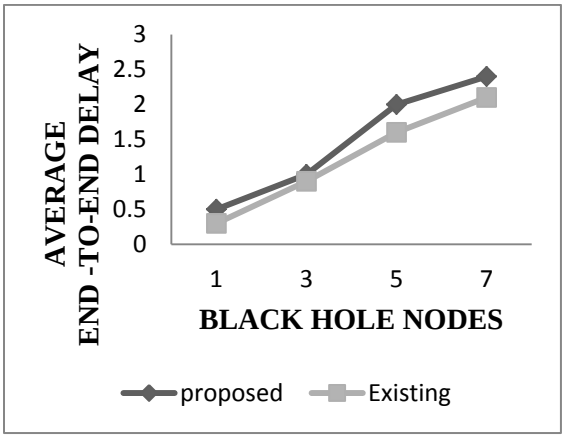


Figure 8: Average End-to-End Delay VS No. of black hole nodes.

Figure 7 shows the graphical representation of packet delivery ratio and how much packets are lost in the existing system compared to less packet been lost in the proposed system, which results in the proposed system becoming better than the existing system in term of packet delivering. Figure 8 shows the graphical representation of the average end-to-end delay experienced in both the existing and proposed system, with the proposed system, unfortunately, having a higher end-to-end delay.

5.0 Conclusion and future work

This work provided an analysis for the effect of the Black Hole in an AODV Network by simulating the AODV protocol under multiple Black Hole attack using NS-2. We worked on improving the IASAODV protocol characteristics, which is considered an improvement of the AODV protocol under multiple Black Hole attack. As mentioned earlier, we considered four different scenarios; the first scenario has one Black Hole node, the second has three Black Hole nodes, the third has five Black Hole nodes and the fourth having seven black hole nodes. Each one was implemented on IASAODV and the proposed algorithm. The simulation results showed that the packet delivery ratio of the existing framework (IASAODV) protocol is better than

the proposed. The proposed framework provided better results in term of Packet Delivery Ratio for 1,3,5 and 7 malicious nodes which are 100%, 100%, 99% and 98% respectively compared to the existing with 88%, 79%, 61% and 57%, for 1,3,5 and 7 malicious nodes respectively.

5.1 Recommendations for future work

a. The proposed framework was not deployed for different types of attacks such as wormhole and greyhole attacks.
b. Study can also be made to improve on the delay in end-to-end communication.
c. Also, for future work, research can be done so as a possible way on how to provide restrictions that will only allow nodes that want to communicate to access the network

REFERENCES

[1] Burbank J.L. Chimento P.F Haberman B.K., & Kasch W.T., (2009)" Key Challenges of Military Tactical Networking and the Elusive Promise of MANET Technology''. IEEE Communication Magazine 44(11):39–45. doi:10.1109/COM- M.2006.248156.

[2] Bing W., Jianmin C., Jie W and Michael C., (2006) "A Survey on Attacks and Countermeasures in Mobile Ad Hoc Networks," Springer, 2006.

[3] Charles E.P. Elizabeth M.B.R. SamirDas. (1999). "Ad-Hoc On-Demand Distance Vector (AODV) Routing," Proceedings of IEEE WMCSA'99, New Orleans, 1999.

[4] Chin-Yang T., Poornima B., and Calvin K., (2003)"A Specification-based Intrusion Detection System for AODV," in Proceedings of 1st ACM Workshop on Security of Ad Hoc and sensor networks, California, Davis.

[5] Djenouri, D. Badache, N. (2008) Struggling Against Selfishness and Black Hole Attacks in MANETs. Wireless Communications & Mobile Computing 8(6):689–704. doi: 10.1002/wcm. v8:6

[6] Dow C.R. Lin P.J. Chen S.C. Lin J.H. & Hwang S.F., (2005) "A Study of Recent Research Trends and Experimental Guidelines in Mobile Ad-hoc Networks". Paper presented at the IEEE 19th International Conference on Advanced Information Networking and Applications, Tamkang University, Taiwan, 28-30 March 2005.

[7] Ira N. & Chaki R. (2012) "New black hole attack prevention system in clustered MANET" international journal of advanced research in computer science and software engineering, vol. 2 (8), 113-121] Umang S. Reddy B.V.R. & Hoda M.N., (2010)."Enhanced Intrusion Detection System for Malicious Node Detection in Ad Hoc Routing Protocols using Minima EnergyConsumption,"IETCommunications4(17) :20842094.doi10.1049/ietcom.2009.0616.

MANET". International Journal of Computer Science 2:54 59. Doi: abs/0909.2371

[9] Mahmoud T.M. Aly A.A. & Makram O.(2015)"Avoiding Black Hole attack of AODV routing protocol in MANET (IASAODV)"Int.J.onNetworkSecurity,Vol.

[10] Ming Y.S., (2011) "Prevention of selective black hole attacks on mobile ad hoc network through intrusion detection systems," Elsevier Computer Communications 34, pp. 107-117.

[11] Swadas PB and Raj PN, (2009) DPRAODV: "A Dynamic Learning System against Blackhole Attack in AODV based

[12] Tamilselvan, L. Sankaranarayanan, V.(2007) Prevention of Blackhole Attack in MANET.Paper presented at the 2nd International Conference on Wireless Broadband and Ultra-Wideband Communications, Sydney, Australia

[13] Nirali, M & Gupta, V. K. (2014). "Prevention of black hole attack using AODV Routing Protocol in MANET," International

Journal of computer science and information technologies, vol.5 (3), 3254-3258.

[14] Zapata M.G., (2002) "secure ad-hoc on-demand distance vector (SAODV) routing". ACM- SIGMOBILE Mobile Computing and Communication, 6(3), 106-10

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

NAIJASPELL: SPELLCHECKING FOR NIGERIAN PIDGIN

F. Tanshi[1], T. Adegbola[2]

[1]Federal University of Petroleum Resource, Effurun.
[2]African Languages Technology Initiative, Ibadan.

[1] foghor.ai@gmail.com; [2] taintransit@hotmail.com

ABSTRACT
Current state of the art spellchecking techniques are based on an efficiently stored list of correct spellings of words in a language against which wrongly spelt words are checked. However, Nigerian Pidgin does not have a compiled list of such proofed spellings which is required by these techniques. As a result, people generally prepare writings in Nigerian Pidgin using different spelling styles, leading to inconsistency each time a word is spelt. To solve this problem which also holds for many other resource-scarce languages, this paper presents a machine learning approach to spellchecking that does not require an existing word list. In this approach, the correct spelling of a word is learnt based on the relative frequencies of various renditions of the spelling of the word in a document. That is, the technique flags spelling errors by depending only on words within the document that is being edited.

Keywords: Edit distance, Orthography standardisation, Spellchecking, Unigram Probabilities.

1.0 INTRODUCTION

Nigerian pidgin (popularly referred to as Naijá) is an English based pidgin. Like other pidgins in climes where English is not native to its users, Naijá is a contact language that emerged from the fusion of many indigenous languages (such as Yoruba, Igbo, Itsekiri and Urhobo) and some foreign languages (including English, Portuguese and Spanish). It obtains its vocabulary (meanings and contexts may vary) from all its substrate languages. Some Naijá words and their meanings in English include;

Ben-dan (bend down)

Ben-ben (ambiguous, zig zag)

Beleful (satisfied, satiated)

Kompound (compound)

The Naijá Langwej Akedemi (NLA) is a non-governmental organisation with a vision to develop corpus and document the structures of the language. NLA was established in 2009 after a group of Nigerian linguists held a conference on Nigerian Pidgin and adopted the name Naijá as the official name for the language. This is because Naijá has creolised throughout the country and its functions have surpassed that of a pidgin. In view of this, the NLA and many linguists have made efforts to describe the grammatical structure of Naijá and to develop a spelling guide [1, 2, 3].

Naijá is the most widely spoken language in

Nigeria with more than 80 million speakers – spanning several regional, social and ethno-linguistic groups [4, 5, 6]. However, it is regarded by the minority ruling elite class as an inferior form of English, to be used only by the uneducated [7]. This is despite the fact that members of this elite class also sometimes communicate in formal settings using Naijá. As a result, the Language has experienced little official recognition and very limited documentation of its vocabulary and structure. Consequently, this has inhibited Human Language Technology (HLT) research into the language. By virtue of the fact that Naijá is the most widely spoken language in Nigeria, it deserves much better treatment which is the impetus for this work.

Some of the most basic tools needed for the development of HLT in any language are the tools required for the development of corpora in the language. Amongst these tools, the spellchecker is one of the most important because it provides a facility for automatic editing of written texts which are usually employed as the basis for developing morphological, syntactic and semantic models for a language [8]. In addition, considering that most of these models are statistical in nature, the corpora needs to be voluminous and the data needs to be as accurate as possible, hence the vitality of the spellchecker.

For English and other languages which have standard orthographies, mono-lingual dictionaries and adequately sized corpora, it is relatively easy to develop human language technologies. This is not the case with languages like Naijá which is known to have several unofficial orthographies because of the general ad hoc approach to its writing and almost no official documents prepared in it, making it difficult to estimate the size of the Naijá vocabulary in digital media.

2.0 RELATED WORK

Naijá like many other African languages is a resource-scarce language because every day use of the language is not documented. Hence, it does not have a documented list of correct spellings that may be used to edit corpora, which is a basic requirement for the development of Natural Language Processing (NLP) techniques. Present state of the art spellchecking techniques such as implemented in Hunspel and Ispell are rule based [9, 10, 11, 12, 13]. That is, they require an efficiently coded database comprising a list of correct spellings in Naijá in order to detect spelling errors. Furthermore, other spellcheckers such as discussed in [14, 15, 16, 17, 18] also follow this general pattern. In other words, without voluminous corpora, from which a word list can be extracted, creating a spellchecker is difficult and without a spellchecker creating voluminous corpora will be laborious and time-consuming to clean. Thus in order to address this chicken and egg situation, it became necessary to develop a spellchecking technique that does not depend on an existing wordlist. This technique implemented in Naijaspell detects non-word errors by depending only on words within the document being edited.

3.0 THEORY AND APPROACH

Spellings, very much like the words they represent are based on conventions in which for reasons of history, structure or cognitive efficiency, if a certain word is frequently spelled in a particular way, such spelling, sooner or later becomes standard. Examples abound in the silent conflict between British and American versions of the spellings of words such as programme versus program and colour versus color. Today due to the use of these words in computer related environments and hence the relatively higher frequency of encountering the American versions of these spellings, they now tend to appear less awkward to English speakers who were weaned on British spellings.

To this end therefore, Humans are sometimes capable of identifying misspellings even in words they have never encountered. That is, when they find two or more occurrences of

words that have very similar spellings, (e.g. pilla and pila) they consider the possibility that the word with the least frequency of occurrence is a spelling error of the one with the higher frequency. If this human behaviour can be expressed quantitatively, the resulting features can be used to teach a machine to perform spellchecking without a pre-compiled wordlist.

3.1 Data Preparation

A list of word tokens was extracted from text obtained from a Naijá New Testament Bible and its alphabet and orthography (though excluding diacritic marks used to indicate tonality) are very similar to that provided by the NLA [1]. This study is therefore limited to the Naijá dialect used in the aforementioned bible.

The extracted words were sorted alphabetically and a table comprising all unique word pairs and the various features defined in section 3.2 below was created from the wordlist. The table had over 3.5 million rows. Considering the sheer volume of the data, only word pairs within a Damerau-Levenshtein Distance of 1 were isolated. It is known that 80% of human spelling errors lie in this region [19]. The pairs were manually labelled as shown in Table 1, where 1 indicates spelling error and -1 indicates an independent pair or non-spelling error.

In the extracted table which is comprised of 2395 rows, the number of non-error pairs (2185 in total) was more than twelve times that of error pairs (174 in total). A well-established fact in machine learning is that training with such a data would result in a decision boundary that is biased towards non-spelling errors [20]. As expected, in a preliminary inside test, when the extracted table was fed to a perceptron learning algorithm, it predicted all inputs as non-spelling errors. To cater for this anomally, an equal number of spelling error pairs and non-spelling error pairs (170 each) were randomly selected and used for inside testing to select the best set of features.

Since the input features are measured in different units (edit-distance and occurrence frequency), it became necessary to make them all dimensionless by use of appropriate scaling factor in order to avoid bias towards a measurement unit with larger values. This also makes learning algorithms converge quickly.

It was assumed that the longer a word is, the higher the chances of an error occurring in its spelling. Thus the edit distance features were normalised by the maximum length of a pair, making the value assigned to the edit distance of longer word pairs similar to those of shorter word pairs. This is necessary because they both indicate the same error.

In accordance with the assumption in the theoretical background proposed in Section 3.0 above, it was discovered that one word in an error pair usually has a much higher frequency relative to the other. Thus the difference in occurrence frequency feature was normalised with the higher frequency.

Table 1: Word Pairs and Their Target Labels

word1	word2	Target
Sulphur	Sulphur	1
Boat	Boastin	-1
Gallon	Gallon	1
ben-dan	ben-ben	-1
ben-ben	beleful	-1
Kompoun	kompound	1

3.2 Feature Extraction Algorithms

Word similarity is often expressed quantitatively using string metrics such as Difference-in-length, Hamming Distance, Levenshtein distance, Damerau-Levenshtein Distance and other derivatives of the aforementioned. These metrics assign a score to word pairs based on the number of characters that need to be changed in order to convert one word in the pair to the other. The closer the words are, the lower the score assigned. Difference-in-length considers the difference in

the number of characters of a word pair without regard to the constituent characters. Even though the Hamming distance considers the minimum number of substitutions that is required to convert one word to another, it is not suitable for comparing versions of spelling of words because it requires the two strings to be compared to be of the same length. The Levenshtein Distance however, assigns a score by computing the total number of insertions, deletions and substitutions that is required to convert one string to another irrespective of their lengths. Then the Damerau-Levenshtein Distance includes transposition (swapping) of two adjacent characters in addition to the allowable operations of the Levenshtein distance.

In order to determine the most suitable choice of features based on which spelling errors in Naijá can be learnt, edit distance metrics as well as the relative occurrence frequencies of versions of word spellings in a document were tested. Inside testing provided insights into the behaviour of various candidate features and the features with the best performance were used to implement a 10-fold cross validation outside testing.

3.2.1 Normalized Difference in Length (NDL).

Word pairs having the same length are more likely to be spelling errors compared to words with different lengths. The NDL is derived by dividing the Difference in length between a word pair by the length of the longer word of the pair as described by the algorithm below.

NDL Algorithm:

Input words: *w1, w2*
NDL =
(Len(w1)-Len(w2))/max(Len(w1),Len(w2))
Return NDL

3.2.2 Normalized Levenshtein Distance (NLD).

This is based on the total number of single character insertions, deletions and substitutions that is required to transform one word to another [21]. The NLD is derived by dividing the Levenshtein distance (LD) between a word pair by the length of the longer word of the pair. That is, ∀ i ∈ len(w1), ∀ j ∈ len(w2), L[i,j]: ld ∈ $R^{m\times n}$ is the Damerau-Levenshtein distance between strings (words) w1 and w2.

NLD Algorithm:

Input words: *w1, w2*
Comment: *∀ i, j, ld[i,j] is the Levenshtein distance between w1 and w2.*
Int ld Є $R^{m\times n}$
Comment: *initialize each element in ld to zero*
m = len(w1), n = len(w2)
for j =1 to n:
for i = 1 to m:
if w1[i] == w2[j]:
score := 0
else:
score := 1
d[i, j] :=
Comment: deletion
min(ld[i-1, j] + 1,
Comment: insertion
ld[i, j-1] + 1,
Comment: substitution
ld[i-1, j-1] + score)
NLD = d[m, n]/max(len(w1),len(w2))
Return *NLD*

3.2.3 Normalized Damerau- Levenshtein Distance (NDLD)

This is based on the total number of single character insertions, deletions, substitutions and transposition required to convert one word to another [19]. The NDLD is derived by dividing the Damerau-Levenshtein distance (DLD) between a word pair by the length of the longer word of the pair. That is, ∀ i ∈ len(w1), ∀ j ∈ len(w2), ld[i,j]: ld ∈ $R^{m\times n}$ is the Damerau-Levenshtein distance between strings

(words) w1 and w2.

NDLD Algorithm:

Input words: w1, w2
Int ld ∈ $R^{m \times n}$
m = len(w1), n = len(w2)
for j =1 to n:
for i = 1 to m:
if w1[i] == w2[j]:
score := 0
else:
score := 1
d[i, j] :=
min(ld[i-1, j] + 1, $ deletion
ld[i, j-1] + 1, $ insertion
ld[i-1, j-1] + score) $substitution
d[(i,j)] = min (d[(i,j)], d[i-2,j-2] + score) $transposition
NDLD = d[m, n]/max(len(w1),len(w2))
Return NDLD

3.2.4 Normalised Difference in Frequency (NDF)

In spelling error pairs, one of the words usually has a much larger frequency compared to their spelling errors. The NDF is derived by dividing the difference in occurrence frequency of a pair by the maximum frequency.

NDF Algorithm:

Input words: *w1, w2*
Compute the frequency of w1, w2
NDF = (freq(w1)-freq(w2))/max(freq(w1),freq(w2))
Return *NDF*

3.2.5 Lower Frequency Normalised by highest pair frequency or total number of words (LF).

It was reasoned that in addition to the relative frequency between error pairs, the absolute frequency of the word with the lower frequency is also indicative of a spelling error pair. This frequency may be normalised by the frequency of the more frequent word in the pair or the total number of words in the document.

LF Algorithm 1:

Input words: w1, w2
Compute the frequency of w1, w2
LF1 = min(freq(w1),freq(w2))/ max(freq(w1),freq(w2))
Return LF_1

LF Algorithm 2

Input words: *w1, w2*
Compute the frequency of w1, w2
LF_2 = min(freq(w1),freq(w2))/ cardinality(wordlist)
Return LF_2

4.0 RESULTS

4.1 Inside Testing and Analysis of Data Performance on Perceptron Algorithm

A set of features comprising an edit distance metric and word pair frequencies were fed to a bipolar activated perceptron model which then learns to distinguish error pairs from independent pairs. By so doing, the algorithm classifies words with small edit distances and high difference in frequency as spelling errors. In other words, as long as a word is written consistently in a certain way in the document, the system tends not to flag it as a spelling error since it is not likely to have a counterpart that may constitute a spelling error.

Table 2: Inside Testing Data for Features NDL, NLD, NDLD, NDF, LF.

Inputs	Accuracy (%)
NDL, NDF	52.94
NLD, NDF	65.88
NDLD, NDF	77.06
NDLD, NDF, LF_1	77.06
NDLD, NDF, LF_2	77.06

Table 3: Inside Testing Results of Combinations of Feature NDL, NLD, NDLD, NDF, LF_1, LF_2.

W1	W2	Tag	NDL	NLD	NDLD	NDF	LF
sulfur	sulphur	1	0.143	0.286	0.800	0.0002	0.0007
boat	boastin	-1	0.429	0.429	0.975	0.0013	0.0052
gallon	galon	1	0.167	0.167	0.500	0.0001	0.0003
ben-dan	ben-ben	-1	0.000	0.286	0.000	0.0000	0.0001
also	alredi	-1	0.333	0.667	0.973	0.0009	0.0049
ben-ben	beleful	-1	0.000	0.714	0.938	0.0003	0.0021

The inside testing results displayed in Table 3 clearly shows that the Damerau-Levenshtein distance feature (NDLD) yields better accuracy than the Difference in Length (NDL) and Levenshtein Distance (NLD) features. In addition, it is seen that the Lower Frequency features (LF_1 and LF_2) are redundant to the Difference in Frequency (NDF) feature. While Difference in Length and Levenshtein Distance features are redundant to the Damerau-Levenshtein distance. Therefore, in order to ensure the best possible accuracy and time efficient performance, the only features retained are Damerau-Levenshtein Distance and Difference in Frequency.

As presented in Figure 1, the data shows some consistency in pattern. The line of separation is around a Normalised Damerau-Levenshtein Distance of 0.2, there is a small area of overlap between the two classes of data and a few points straying away from this boundary which implies that the features are not sufficiently discriminative to render the two classes linearly separable.

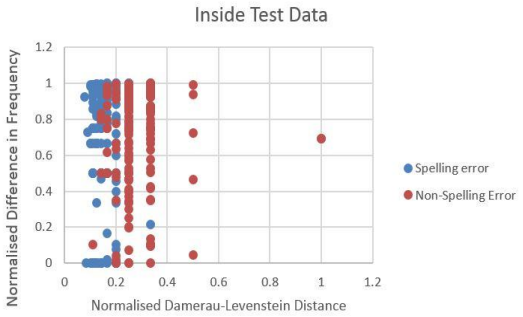


Figure. 1: Damerau-Levenshtein Distance and Difference in Frequency Features

In addition, among the false positives predicted by the model, there were also correct spellings as well as plurals and other affix forms of valid word pairs e.g. glass and grass, ship and ships, tish and tisha. This is due to their small Damerau-Levenshtein distance, thus implying an exception to the rule. Therefore, in order to address this exception, another feature that could possibly clarify the difference between words that have common morphemes was introduced.

4.2 Morphology Feature Definition and Performance on Perceptron Algorithm

This feature attempts to clarify valid pairs that have common morphemes from those that are actually spelling errors. For example, “woka” and “wokas” are valid words which were misclassified as spelling errors because of their common morpheme “woka”.

The morphology feature of a pair is assigned using an algorithm that computes the probability that their common morpheme (whether free or bound) is a true morpheme. This feature is named probability of true morpheme (PTM) as indicated in the following algorithm.

PTM Algorithm:

Input words: w1, w2
Obtain morpheme w1 and w2 have in common
Compute freq(morpheme with other free or bound morphs)

Compute cardinality(morph)

If freq(morpheme with other free or bound morphs) ≥ 2

PTM(morph| w1,w2)=(freq(morph with other free or bound morphs)) ÷ (cardinality(morph))

If more than one morpheme occurs for a pair:
PTM(morph_1, …, morph_n| w1,w1)= PTM(morph_1 | w1,w2) × … PTM(morph_n| w1,w2)

Return PTM

In other words, for a pair such as "wok" and "woka" since their common morpheme is "wok", the algorithm computes the number of times that "wok" occurs with other valid bound morphemes such as; "a" and "as" as in "woka" and "wokas" and divides it by the number of times that work occurs with both valid and invalid bound morphemes such as "a", "as", "k". In addition since this decision is based on only the document being edited, a valid morpheme was defined as one that occurs with at least two or more other free or bound morphemes. That is, for example in the document, "a" and "as" are considered valid bound morphemes because they occur with other free morphemes such as "pray" and "prish" unlike "k" that does not.

The morphology feature was computed for both prefixes (whether free or bound) and suffixes (whether free or bound) increasing the number of features to four. In addition, when pairs have more than one morpheme in common, the probabilities of the morphemes are multiplied to obtain the feature value for that pair since it is believed that they are independent events.

The morphology feature values of non-error pairs such as; "tish" and "tisha" mostly ranged between 0.5 and 1 while those of actual spelling error pairs such as "huri" "hurri" mostly ranged between 0 and 0.6 (besides exceptions such as deserv and deserve), thus further distinguishing the two classes. Then the inside test accuracy obtained after passing the four features through the perceptron algorithm increased to 79.41%.

Since spelling error data resource was low, it was not possible to create a test set that was independent of the initial training set used for inside testing. Therefore, the same data which is comprised of 340 word pairs was rotated in a 10-point cross validation process. This process also helps to check sampling error in the data.

Table. 4: Word pairs and Damerau-Levenshtein Distance, Difference in Frequency and Morphology Features.

W1	W2	Tag	NDLD	NDF	PTM-P	PTM-S
tish	tisha	-1	0.200	0.634	0.667	1.000
mi-sef	misef	1	0.167	0.923	0.200	0.615
ju	tu	-1	0.500	0.465	1.000	0.052
deserv	deserve	1	0.143	0.500	1.000	1.000
min	tin	-1	0.333	0.949	1.000	0.164
aminadab	amminadab	1	0.111	0.500	0.008	1.000
again	agian	1	0.200	0.997	0.023	1.000
clay	play	-1	0.250	0.733	1.000	1.000

The average accuracy obtained from the perceptron cross validation was 46.47%.

Furthermore, many valid pairs such as grass and glass were still misclassified as errors. The data was then passed to a Support Vector Machine algorithm and the performance was observed.

4.3 Support Vector Machines (SVM): Inside Testing, Cross Validation and Analysis

Inside testing accuracy obtained was 84.71% using a linear SVM and 99.12% using a Gaussian SVM.

As shown in Table 5, the maximum and average accuracies obtained from a 10-fold cross validation process were 66.67% and 50.91% (with a Standard deviation of 0.08305.) which is a 4.44% increase from the average performance of the perceptron. In addition, the average percentage of true positives was 50.30%, average percentage of false positives was 48.79%, the average percentage of true negatives was 0.61% and the average percentage of false negatives was 0.303%.

Table 5: Cross validation results for Support Vector Machines (SVM)

ACC(%)	PREC	REC	F1-SCO
66.67	0.667	1	0.8
54.55	0.545	1	0.706
48.48	0.469	1	0.638
45.45	0.438	1	0.609
63.64	0.636	1	0.778
45.45	0.455	1	0.625
51.52	0.531	0.944	0.68
45.45	0.455	1	0.625
42.42	0.424	1	0.596
45.45	0.455	1	0.625
50.91	0.51	0.99	0.67

As a spellchecker, although the accuracy improved, the morphology feature did not sufficiently solve the problem, since many valid pairs such as “grass” and “glass”, “ship” and “ships” were still misclassified as errors. For cases of pairs with common morphemes such as “ships” and “ships”, it is believed that this will automatically resolve itself when more data with rich morphology becomes available for training. More data which will be crowd-sourced using the selected features would hopefully improve accuracy.

However, as a standard wordlist generation tool, an average 0.303% false negative rate implies that the classifier correctly predicted most smelling errors in the test sets. Then the false positives (which on average accounts for 48.79% of all cross validation predictions) will count as new entries for a standard wordlist for Nigerian pidgin after being manually ratified by linguists. This spelling list when updated extensively in the future can then be used with state of the art spellchecking techniques and in order words, solves the chicken and egg problem. That is, the method presented in this study will half the time required to generate a standard spelling list for Nigerian pidgin by clustering errors and thus providing a much smaller error search area.

5.0 CONCLUSION

The SVM algorithm obtained almost perfect prediction during inside testing and improved accuracy by 4.44% during cross validation. This indicates that given more data, it will yield a much higher average accuracy during cross validation and outside data test.

Furthermore, this method will potentially aid the standardisation of orthographies as well as the development of corpus for many of the world’s resource-scarce languages. That is, since available writings in such languages like Naijá appear in various spelling styles, this technique can be used to aid the development of a standard orthography and spelling list from crowdsourced text.

Due to the performance obtained, the approach presented in this study may not offer a compete replacement of the exiting methods of spell checking, but it certainly offers a simpler means by which some of the otherwise laborious aspect of corpus development (which is often manual) for resource-scare languages can be semi-automated in a bootstrapping phase.

Naijá is a particularly resource scarce language and this study in itself suffered from the

resource scarcity in the sense that little data was available for it. In particular, the editing of the bible from which corpus for this study was extracted also suffered evidently from this resource scarcity. Thus, having demonstrated that this proposition is viable, the present study serves as motivation for a detailed proposal for raising necessary resources for gathering of adequately sized data for a future study to standardize the orthography of the language. And due to recently launched Naijá data generation projects by different media organizations, several potential data sources will include British Broadcasting Corporation (BBC) pidgin channel, The Jehovah's Witness Organization and bible translation projects.

Out of the more than 6000 languages spoken in the world today, a very large percentage constitutes resource-scarce languages, which are lacking in BLARK – the basic language resource kit required for the development of NLP in such languages. As already pointed out, these resources first and foremost depend on corpora which need spellcheckers to automate their development. Yet, the development of spelling checkers too requires corpora. Methods such as presented in this study can be used to breach such a cycle.

ACKNOWLEDGEMENT

The authors would like to thank the Riverine Gospel Mission, Nigeria for granting permission to use their Naijá New Testament Bible text as data for this research.

REFERENCES

[1] C. I. Ofulue and D. O. Esizimetor, Guide To Standard Naijá Orthography. An NLA Harmonized Writing System For Common Naijá Publications, 2010. [Online]. Available: http://www.ifra-nigeria.org/naija-corner/naija-languej-akedemi. [Accessed 26 February 2018].

[2] C. I. Ofulue, Towards the Standardisation of Naija: Vocabulary Development and Lexical Expansion Processes, in Conference on Nigerian Pidgin, Ibadan, 2010.

[3] D. O. Esizimetor, What Orthography for Naijá? in Conference on Nigerian Pidgin, Ibadan, 2010.

[4] K. U. Ihemere, A Basic Description and Analytic Treatment of Noun Clauses in Nigerian Pidgin, Nordic Journal of African Studies, vol. 15, no. 3, p. 197, 2006.

[5] F. P. C. Endong, The Use of Nigerian Pidgin English Print Advertising: Deviation from Standard Orthography and Effectiveness, International Journal of Art, Culture, Design and Language Works, vol. 1, no. 1, pp. 1-7, June 2015.

[6] U. O. Nneka and M. Okitikpi, Age variation in the use of Nigerian Pidgin (NP): A case, International Journal of English and Literature, vol. 8, no. 2, pp. 16-25, February 2017.

[7] T. A. Balogun, In defense of Nigerian pidgin, Journal of Languages and Culture, vol. 4, no. 5, pp. 90-98, 2013.

[8] K. Oflazer, Developing Computational Morphology for Low- and Middle-Density Languages, in International Workshop on Finite-State Methods and Natural Language Processing, Pretoria, 2009.

[9] N. Laszlo, Hunspell, 20 March 2015. [Online]. Available: http://hunspell.sourceforge.net/.

[10] G. Kuenning, International Ispell, 1996. [Online]. Available: http://www.cs.hmc.edu/~geoff/ispell.html#authors. [Accessed 3rd March 2015].

[11] N. Laszlo, T. Victor, P. Halacsy and A. Rung, Leaveraging the Open Source Ispell Code Base for Minority Lanaguage Analysis, in SALTMIL Workshop at the 4th International Conference on Language Resources and Evaluation, Centro Cultural de Belem, 2004.

[12] Nemeth Laszlo H. Peter, K. Andras, T. Viktor, Open Source Morphological Analyser, 2008.
[13] L. Al-Hussaini, "Experience: Insights into the Benchmarking Data of Hunspell," ACM Journal of Data and Information Quality, vol. 8, no. 3-4, June 2017.
[14] K. Kukich, "Techniques for Automatically Correcting Words in Text.," ACM Computing Surveys, vol. 24, no. 4, pp. 377 - 439, 1992.
[15] M. Lewellen, "Neural Network Recognition of Spelling Errors," in 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Quebec, 1998.
[16] M. Renato Cordeiro de Amorim, "Effective Spell Checking Methods Using Clustering Algorithms," in Recent Advances in Natural Language Processing, Hissar, Bulgaria, 2013.
[17] J. Victoria J. Hodge, "A Novel Binary Spell Checker," Lecture Notes in Computer Science, vol. 2130, pp. 1199 - 1204, 2001.
[18] J. p. Victoria J. Hodge, "A comparison of a novel neural spell checker and standard spell checking algorithms," Elsevier, Pattern Recognition, vol. 35, no. 11, p. 2571 – 2580, 2002.
[19] F. J. Damerau, "A technique for computer detection and correction of spelling errors," Communications of the ACM, vol. 7, no. 3, March 1964.
[20] A. Ng, Feature Scaling, Machine Learning Course, Coursera, 2014.
[21] V. L. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," Soviet Physics - Doklady, vol. 10, no. 8, pp. 707 - 710, 1966.

nigeria computer society
ncs
www.ncs.org.ng

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

EVALUATING INDUSTRIAL CONTROL SYSTEM (ICS) SECURITY VULNERABILITY THROUGH FUNCTIONAL DEPENDENCY ANALYSIS

U. D. Ani[1], N.C. Daniel[2] and S. E. Adewumi[3]

[1]Cranfield University, United Kingdom

[2]University of Greenwich, United Kingdom

[3]Federal University Lokoja, Lokoja -Nigeria

[1]u.p.ani@cranfield.ac.uk; [2] n.c.daniel@greenwich.ac.uk; [3]sunday.adewumin@fulokoja.edu.ng

ABSTRACT

Industrial Control Systems are highly interconnected infrastructures and networks that are used to control and manage industrial processes. Irrespective of the form assumed; Supervisory Control and data Acquisition System (SCADA), Distribution Control System (DCS), Process Control System (PCS), etc., a form of dependency exist within the setup, contextualised defined as the connection between two or more assets or infrastructure, such that the state of one can influence unilaterally, or correlate to the state of the other. This phenomenon introduces security threats, vulnerabilities, and risks in the emerging setup where IT are combined with OT for improving operational performance and productivity. However, since the criticality of cyber-attack impacts on ICS infrastructure can be quite huge, rapid, and damaging; there are needs for responses that can leverage new security concepts and approaches and expedite security assurance. In this paper, a functional dependency modelling perspective is considered, and a cascading impact scoping approach is presented for determining the potential impact of exploiting security vulnerabilities on targeted ICS infrastructure. The outcome can be used to influence security decision-making for improved cybersecure ICS. The proposed technique is validated using real cyber-attack and vulnerability analysis scenario on an assembly-line ICS testbed. The proposed approach offers episteme into the various destructive capacities possible from the failure of functionally dependent ICS components. A cascading impact value (CIV) metric is also proposed which can be adopted when evaluating an industrial system's security in a much quicker decision-making and response order, to avert potential damages and help improve cyber security in the ICS environment.

Keywords: Cybersecurity, Functional dependency, Security Interdependencies, Cyber-Physical Interdependencies, Industrial Cyber Security.

1.0 INTRODUCTION

Industrial Control Systems (ICS) have emerged as the fundamentals of modern industrialization and precisely manufacturing, proffering

solutions to the complexities in human systems. ICSs are key components of an organization's industrial infrastructures, which contribute greatly to process management. ICSs are usually situated and used to control, monitor, and manage large industrial service systems often in critical infrastructure areas like manufacturing, electric power, transportation, chemical, energy, oil and gas, water, and wastewater industries. Depending on the environment, targeted objectives, and the processes required, ICSs could be implemented in three varied forms; Supervisory Control and Data Acquisition System (SCADA), Process Control System (PCS), and Distribution Control System (DCS) [1]. All ICS forms at some points in their functionalities gather partially or fully automated information about industrial processes and states from a variety of endpoint devices. They are able to initiate interactions based on automated logic, setting off alerts on events needing user (operator) responses, reporting and storing system changes [1]. The significant social importance of ICSs and their criticality to the functioning of everyday activities imply that ample safety and security measures need to be identified and engaged to reduce the potentials for failure [2], [3]. The need is quite timely, now that the industrial sector seem to have become attractive targets of cyber-attacks [4], causing disturbing impacts that happen so rapidly that before responses and remediation are conceived and engaged, weighty damages have occurred on infrastructures, people, businesses, and the environment.

Essentially, ICSs consist of a collection of devices connected to form a network of systems, working harmoniously to further a predefined operational and (or) process goal. This harmonious interoperability expresses a form of (inter)dependency amongst the individual components of the ICS structure and architecture, such that the functionality of one component directly depends on the resulting and transferable output from another connecting component. Accordingly, ICSs consist of three broad component building blocks; field station, control station, and communication networks [5] – presented in Figure 1. The field station comprises of instrumentation devices (sensors and actuators) directly connected to physical equipment (pumps, valves, conveyor belts systems, robots, etc.,) and Remote Terminal Units (RTU). Physical system dynamics are measured by sensors, and resulting values used to determine the initiation of actions by actuators. This continuous bidirectional data communication is enabled by the RTU, which provides the exchange interface to the instrumentation devices. For example, devices that can function as RTUs include; Programmable Logic Controllers (PLCs) and Intelligent Electronic Devices (IEDs). The controller station comprises of Master Terminal Unit (MTU) devices like Control Servers, Historians, and Human Machine Interface (HMI). The MTU operates as the soul of the ICS (SCADA, PCS, DCS), where command and control instructions are initiated, and activity supervisions are engaged over physical processes. The MTU typically relies on RTU data for its operations, while the HMI provides visual representations of the processes to operators. In some cases, the HMI allows flexibility for operators to manipulate process settings. The communication network comprises of devices and tools that make up, and enable the exchange of data across stations and components (devices). It typically includes various communication media and gateways like dial-up (telephone line), radio, fibre optics, and satellites. Efficiency of communication only requires that data be encapsulated within the required ICS protocol (Modbus, Profibus, DNP3, etc.) designed for the system. As shown in Figure 1, ICS building blocks are typically represented in a layered linking structure to show the interactive disposition between any connected layers. It also shows the component-to-component connectivity and dependency that drives the basic functionalities of the system. This connectivity and dependency forms the basis of this study.

The analysis of cascading impacts is a significant issue in ICS security because cyber-attacks on ICS systems can result to common-cause [6] and undesirable cascading impacts, which can lead to damaging consequences on multiple system components. Therefore, the criticality of cyber-attacks impacts on ICS infrastructure can be quite huge, rapid, and hurtful, and calls for responses that can expedite security assurance. It is pertinent to have handy; security concepts and approaches that can help foster effective security. This paper focuses mainly on cyber adversarial actions that can cause common-cause effects; and by extension cascading impacts on multiple connected components and infrastructure. This problem is of significant magnitude given that most ICS components and infrastructure control, or are bound to physical processes for the formation and delivery of services (products); locally and globally.

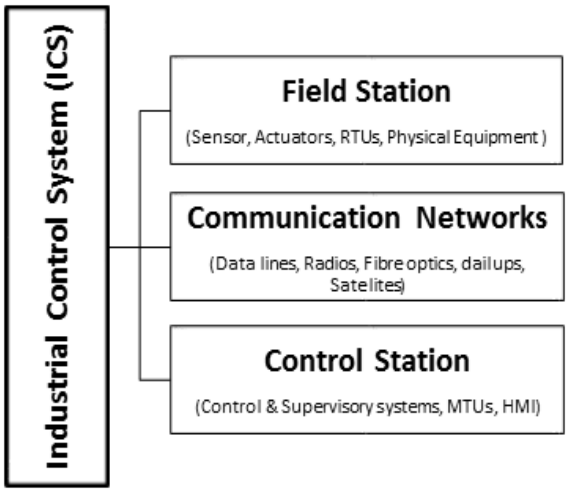


Figure 1: Industrial Control System Building Blocks

Typically, the immediate cause of cyber adversarial actions on ICS infrastructure are functionality disruptions or destructions. Connectivity and dependency in the ICS domain means that functionality disruption effects can travel long and wide through complex functional dependency chains beyond the initial target's domain. It can extend directly or indirectly to physical, institutional, political, and even economic effects [6]. An emulator approach is adopted to explore a realistic scenario of cyber dependency impact analysis in the context of security criticality through cascading effect evaluation, and protective prioritization. Security criticality is used to imply the degree of impact to services, operational, and societal disruptions or destructions due to cyber-attacks. Cascading impact scoping explores the extent to which impact on one device or component ripples down to another connecting component. Thus, the potentials for estimating the scope of negative impacts as a proactive measure towards engaging relevant cyber security capacity is explored. Understanding the span of impacts can support the prioritization decision required to identify and select what infrastructure technology device(s) bear the greatest impact scope and susceptibility, and requiring a more prompt and decisive security response.

1.1 Security and Dependency in Perspective.

Threats and attacks to the functionality of ICS normally target the violation of ICS core security objectives - availability, integrity, and confidentiality [7], [8], and ancillary security objectives – authenticity, accountability, graceful degradation, timeliness, non-repudiation, veracity, etc. [9]–[12]. Widely held successful cyber incidents on ICSs clearly demonstrate that faults and flops in ICS networked components can greatly pose substantial ripple effects on other connected devices. This can lead to other risks such as; health and safety of humans, severe impairment to the environment, and economic impacts like production losses, injury to the industry and by extension the nation's economy, and illegal disclosure of proprietary information [13]–[15]. Earlier, this security stakes were non-existent; as ICSs were operated in isolation; what is technically called *'air-gapping'* [16]. This meant that ICSs were completely secluded from other digital operating domains; particularly the IT domain. Today, the security-by-isolation concept has disappeared [17]. Industrial procedures and practices have transformed with management and control of most industrial

network infrastructures moving over to the internet domain [18]. Computerized industrial control, automations, and management has been noted to be a necessary apparatus for the sustenance and improvement of business/market relevance, industrial performance, competitiveness, profitability, and cost savings for both ICS vendors and users [4]. These have continued to drive investment motivations away from highly proprietary, custom-built, stand-alone components, and towards commercial-off-the-shelf (COTS) system components (hardware and software). The trend is that of convergence of ICS on to the same operating platform as the IT systems; exposing the security weaknesses of the ICS domain, and also transferring to the ICS domain, inherent IT security vulnerabilities and risks. For one of several threat-enabling reasons; ICS devices and infrastructure were not initially designed with security in mind. It suffices to note that advances in the digitization of industrial processes, and the ease of getting through with planned or premeditated attacks explains why cyber adversaries have focused their nefarious activities on industrial networks. Adversarial moves are exploiting organized systems whose business models depend on the availability, exchange and transmission of digital infrastructures and content. Industrial organizations store and transmit intellectual properties in digital forms, use computer-aided processes to develop products, services and manage their operations networks online [19]. It is noted that an estimated 10% of all IP-enabled devices in existence today are ICS devices, and projected to attain a compound increase to about 30% by 2020. This is interpreted to amount to about 7 billion devices expected to be on IP-based connections and platforms [20], [21], and expressing the envisioned characteristics of industrial internet of things (IIoT) [22]. This also translates to clear exposures to as much IP-based threats and vulnerabilities [23]–[25], and bringing about malicious cyber-attacks, and cyber dependency failures [26], [27]. However, as industrial needs and reliance intensify, it is not only important that industrial actors become aware of the cyber-attack threats and the full extent of their consequences [4], but also understand the extent of dependency-driven impacts expressed by varied system components that would help the determination of security criticality and control of attacks.

Dependency is described as the connection between two or more assets or infrastructure, such that the state of one can influence unilaterally, or correlate to the state of the other. Interdependency refers to a multi-directional influence, where devices are linked at multiple points, such that a bidirectional dependency exists between the states of one or more given pairs of infrastructure [25], [28]. In this work, dependency is used to imply both *'dependency'* and *'interdependency'* interpretations accordingly. Real incidents have shown that dependencies exist among interconnected infrastructure components that make up an ICS. A broader but good example of this impact potentials is seen in the Italian electrical blackout incident of 2003. The initial effect was the shutting down of power stations. This led to the failure of nodes in the internet communication network, and further triggered the breakdown of power stations [29]. Another incident is the 2003 electrical blackout in Midwest and Northeast United States and Ontario, Canada [27]. Presumably, the resultant aggregated impacts of this incident could have started with the impairment of perhaps a single system or device within the electric power ICS network. However, the impairment could have sent ripple impacts to devices directly connected to it, which in turn also exerted surging effects on devices connected to them. In the end, the impairment is replicated on all devices down the tree according to their dependencies for data on pre-affected devices. A way dependency manifests is in the disruption of one system which can spread to other systems in a cascading fashion. A second mark of dependency is an event triggering contrary effects on multiple

components simultaneously, and a third is the undesirable effects of a single device or infrastructure; building up over time, and causing complications for other systems [30]. While some of the impacts could be rapid; especially in the case of direct connections or dependence, or gradual; in the case of indirect connectivity or dependency. However, the common ground is the reality that resulting impacts can affect overall system functionality, operational goals, services, health and safety, and economy, and in worse cases; extending ripple effects over large geographical areas that could cut across national or international boundaries [31].

2.0 Related work

The unifying idea in the plethora of researches related to dependency classifications, modelling, and cascading effects analysis is that they are all about infrastructures with multipart, adaptive, and dependent systems with the potentials for transferring impacts across a chain of connection.

Dependency could be either unidirectional or bidirectional [32]. The researchers noted dependency to be unidirectional or bidirectional, such that direction specification is not as necessary a requirement as the dependency's existence and multiplicity itself. We concur that dependencies may occur between two components, within a single infrastructure itself, and may include loops (components causing impairment in another connected component, which again causes additional impairment in the first component), or dual/multiple (one component cases impairments in two or more connected component). Dependency can also be distinguished from the perspective of direct (first order) or indirect (second order) [33]. It follows that if for instance an ICS component *i* depends on component *j*, and component *j* depend on component *k*, then there is a second order (indirect) dependency between *i* and *k*. However, it is noted that such extended level dependencies tend to be difficult to observe and identify depending on the perspective upon which the mappings are drawn.

Another classification is presented in [34], unveiling the aspects of spatial and functional dependencies. While spatial interconnectedness implies closeness between components being the most important relationship between systems, functional dependency refers to a scenario in which one component's operation is necessary for the operation of another component. For example, an MTU device like HMI server relying on the functionality of certain communication gateway (router) to pass on command instructions to an RTU device (PLC). If the gateway does not function as desired, then the RTU would not functions as desired. In [30], this dependency classification is further expanded to cover a wider transactional sphere covering; functional, physical, budgetary, market, and economic interdependencies.

Dependency has been categorized into four; physical, cyber, geographical, and logical [31]. Two devices are physically interdependent if the state of each is dependent on the material output(s) of the other. Accordingly, perturbations in one device can ripple over to another device; implying that the deviation from normal operating conditions in one device can be a function of the deviation in a second device. Cyber dependency is established when resultant states depend on data or information transmitted through the information infrastructure. This type of interdependency is relatively new, and predominantly encountered in electronic and computing environments where control and automations are instantiated and maintained to ensure the normal running of certain systems. Examples are in manufacturing, transport, telecommunications, oil and gas, energy, water and waste water treatment, etc. Infrastructures that run or control processes in these industrial sectors are only able to function because of interconnectivity amongst varying devices, enabling the sending, reception, and exchange of digital data. Geographic interdependency is

established where a local environment event can create state changes amongst locally close objects; within close spatial proximity. Logical interdependency comes into play if a state in one device depends on the state of the other through a mechanism that is neither physical, cyber, or geographic. A typical example is the human decision factor. Researchers have further complemented the earlier classification [31] with one additional class referred to as; *social* interdependency [35]. This class have been described as dependency impact relationship enabled by the spreading of disorder or faults due to human activities on the normal operations of an infrastructure. A similar classification is adopted for interdependencies. These include; physical. Informational, geospatial, policy/procedural, and societal interdependencies [36]. Researchers has also outlined couple of interdependency identification modelling techniques like; aggregate supply and demand tool, dynamic simulations, agent-based models, physic-based models, population mobility models, and Leontief-based models [37]. Some of these are variants of other methods earlier adopted, like the simulation-based approach input-output, and generalized network and system-of-systems approach [30].

3.0 Proposed Model

To develop a functional dependency and cascading impact assessment method that takes into account existing security deployments in an ICS environment, and guide the development of a cyber security plan, a structured approach is presented. This approach takes into consideration the layers of a typical ICS network infrastructure, which is an already supported concept in earlier works [38], [39]. The generalized notion involves understanding how micro-level dependencies affect macro-level structures or system, and vice versa. These may easily be grasped through the analysis of multi-level hierarchies following predefined objectives and requirements of a decision-maker. However, unlike earlier layered (inter)dependency works which do not directly trail the path of risk assessment, the method presented in this paper adopts an indirect approach. It assumes a metric-like dimension such that the quantities derived could in themselves be useful for high-level quick decision-making, and can as well be adopted as potential metric inputs in a larger security risk assessment goal.

The layered approach is thus adopted in the examination of criticality and cascading impact because of cyber-attack or incident on the ICS system through strategic targeting of specific system-constituent devices. Essentially sharing the views in [6] about cascading effects potentials, only that this work mainly focuses on the scope of impact spread and consequences on the system. The key here is to understand what devices has the potential to amass a wider and larger scope of ripple effects on the entire system when attacked and caused to fail or malfunction. Accordingly, we break down the typical ICS network infrastructure into five functional layers; Field Execution, Data and Process Execution and Monitoring, Data and Process Control, Gateway Communications, Network-Based Defence Mechanism/Setup. This functional layered concept is abstracted from the popular Purdue Enterprise Reference Architecture (PERA) [1], which recognizes that network structures must manage assets (PLCs, Controllers, Historians, Servers, etc.) at varied levels of dependency, rapidity, and assurance so as to achieve tolerable response, resolve, trustworthiness, and reparability. Following this, we define functional layers to mean the various actionable roles, responses and behaviours expressed by devices on the ICS network, and essentially follow a bottom-up approach to examine the functional dependencies between proximity layers, the impact and extent of cascading effects that are feasible from attacking any layered device. Accordingly, this work proposes a three-staged Cascading Impact Scoping (CIS) Scheme driven by a functional dependency relationship modelling approach, which is represented in Figure 2.

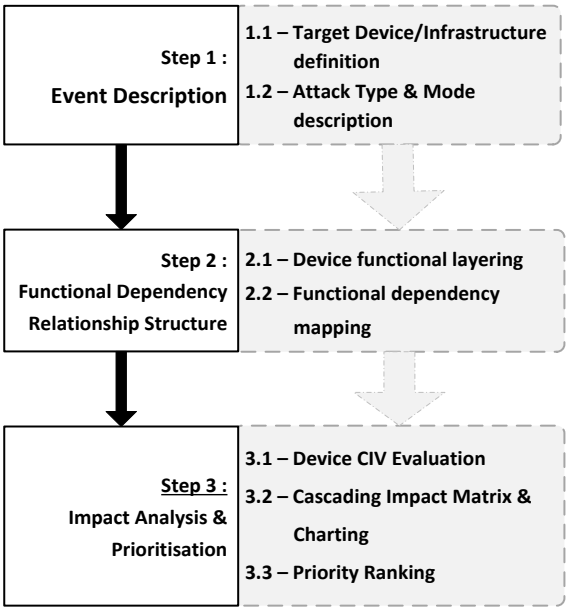


Figure 2: Cascading Impact Scoping Scheme

3.1 Event Description

This initial stage of the CIS scheme involves the clear identification and description of the system being analysed, and the type/mode of attack from which the functional dependency is analysed. Thus, two sub-steps are required in this stage.

3.1.1 Target device/infrastructure definition

This requires clearly identifying the scope of system being analysed and the specific device targeted in the scoped system. Scope implies an indication of how wide in network coverage the analysis intends, e.g., a local network of ICS, or a segmented network of specific process out of a larger area-wide ICS network. The specific device to be targeted would also be indicated, e.g., an industrial switch, PLC, HMI, etc. This will usually be from the list of enumerated functional and connected devices on the ICS network.

3.1.2 Attack type / mode description

This basically involves relating the specific type of attack intended against any stated device(s), e.g., Denial of service, System Hijacking, etc. This is usually influenced by the type of vulnerabilities discovered in the system and networked devices. This step is considered very important because it is believed that different attack types would potentially exert different impairments impacts if executed on the same devices. It is therefore pertinent to clearly distinguish the impact levels of different attacks forms on a certain device, and not assume that all attacks will yield the same results.

3.2 Functionality Dependency Relationship (FDR) Structure.

This stage involves developing a layered structure of the ICS system in line with the way the physical networked has been designed and configured to be operational, and doing a functional dependency mapping of the devices on the ICS network.

3.2.1 Device Functional Layering.

The layer representations are done according to typical operational functions within basic ICS platforms [1], [25]. Table 1 shows a summarised description of the ICS functionality classification layers used in this work. Accordingly, there exist some form of dependency within close (1st order) spatial proximity layers, and extended (2nd order) spatial proximity layer. 1st order imply next-to-next layer closeness, while 2nd order implies more than a single layer closeness relationships. Enumerated device on the network are positioned in the layered class description to which they belong to capture the nature of dependencies each has with its corresponding proximity device and layer.

3.2.2 Functional Dependency mapping:

This involves using directional arrow symbols (functional dependency connector) to connect one device to another in their order of functional dependency; following a bottom-up approach. Accordingly, an arrow from $i \rightarrow j$ denotes a directional functional dependency, i.e., the functionality of device *i* aptly depends on the functionality of device *j*. hence, if there abound

any form of impairment due to cyber-attack on device *j*, the impact of such is assumed to ripple down to device *i*; altering its functionality or operations as well. A bi-directional arrow $i \leftrightarrow j$ depicts a bi-directional functional dependency, and by extension; cascading impact This is not isolated from the potentials for an impairment in a single device to exert cascading impacts on multiple layers and devices, which forms the basis for the prescribed cascading impact value (CIV) for every individual device on the ICS network. This could be easily described by determining; the layer upon which the said device resides, and the number of underlying devices whose functionality depends on the said device's normal operation

Table 1: Industrial Control System Functionality Classification Layers

Layer (Functional Dependency)	Description	Example Devices
Layer 1: Physical execution functions	Outlines physical devices that execute physical processes activated by connected sensors and actuators.	Sensor & actuator driven valves, pumps, motors, transmitters, assembly system, etc.
Layer 2: Logical execution & monitoring functions	Outlines devices responsible for sensing and manipulation of physical processes. Encompasses continuous closed-loop, sequences, batch, and discrete controls, and process monitoring.	PLCs, IEDs, RTUs, HMIs, DCS systems, etc.
Layer 3: Control and Workflow Management functions	Outlines devices with multitasking process capacities. Includes devices that handle workflow management for product realisation	Engineering workstation, Data Historian, process & production management, scheduling, control optimisation, etc.
Layer 4: Communication Gateway functions	Outlines devices responsible for enabling exchange and passage of instruction/command set, data, and information across stations, sub-station, and all devices on the ICS network, locally or area-wide.	Dial-up (telephone line), radios (routers & switches), fibre optics, etc.
Layer 5: Boundary-based security functions	Delineates network-linking device(s) usually position at boundary points between two or more network segments, e.g. between ICS network and corporate or internet network. Security typically considered due because most devices at this position usually assume security by isolation approach either as a key role, or part of a list of roles undertaken by the device(s).	Hardware Firewalls, Intrusion Detection Systems, etc.

3.2 Impact Analysis and Prioritization

This stage involves the process of determining the CIV and range for all the devices on the ICS infrastructure, the development of a cascading impact charting, and matrix-like structure to indicate the position of each device with respect to functional dependency, and the computation of CIV Probabilities and ranking same in order of priority responses.

3.3.1 Device CIV Evaluation

We present a cascading impact array of values, which describes the summative impact scope of a vulnerability or attack mode relative to a specified asset or device on the ICS network. This outlines the physical extent to which a vulnerability if exploited successfully could exert cascading impacts on the whole system. From the FDR structure, it has been earlier noted that two quantities are derivable that could be used to describe a potential cascading impact value of a specific device; (i) Functional layer

placement (FLP_k), and (ii) Number of devices whose impairment are functionally dependent on specified device (ND_k). Both quantities could take values from a set; $A = \{x_1 ... x_p\}$ according to the number of devices registered on the ICS network, and set $B = \{y_1, ... y_p\}$; the number of functionality dependent devices. However, the rang of these quantities define the minimum and maximum values possible for x and y. The minimum value for x in A is 1; i.e., the least number layer on the FDR structure, while the maximum x_{max} is the highest possible number layer that could be achieved during the FDR structure development process. It might not be a member of set A, but the definitive value of x would increment by 1 if the attacked device is also impaired. The minimum value of y in B is 1; i.e., implying the existence of a single device on the network, while the maximum value of y is y_{max}, not necessarily a member of B, which also implies the maximum number of independent operationally integrated device on an ICS network. Thus, in set A, no potential x value could abound greater than x_{max}, and same in a set B (none greater than y_{max}). Functional dependency attributes could thus be represented as a set of Natural numbers (N; given that the members of the set are countable numbers). Such that the Range of set $A = \{x_1, ... x_p\}$, where $x_p \leq x_{max}$, and set $B = \{y_1, ... y_p\}$, where $y_p \leq y_{max}$. It follows that $\{x \in A: \forall\ x \leq x_{max}\}$, and $\{y \in B: \forall\ y \leq y_{max}\}$.

3.3.2 Cascading Impact Matrix and Priority Charting

The values x_{max} and y_{max} can be used to develop an x_{max} by y_{max} matrix structure that prescribes a resulting set C containing an outline of potential cascading impact attributes resulting from impairments due to attacks on specific devices. Each element in C is referred to as a cascading impact value (CIV) of a potential device in the ICS network. This is a Cartesian; product yielding an ordered pair of elements (values) such that in each ordered pair, the first component is an element of A, and the second component is an element of B. For instance, a layer 3 device ($x = 3$) with certain vulnerability if successfully exploited; could amass underlying effects that extend to layer 2 and rippling onto two other functionally dependent devices ($y = 2+1 = 3$). The cascading impact matrix presents a pool of cascading impact values (CIVs) within which every resultant device CIV would fall. The cascading impact matrix is thus described as a function F, which is a Cartesian Product: $A \times B$, yielding a set C.

$$C = A \times B = \{(X, Y) : (x \in A)\ and\ (y \in B) : \forall\ y \leq y_{max}\} \quad (1)$$

So that

$$CIV = (x \times y) \quad (1.1)$$

Since FLP_k implies any value of x in A, and ND_k implies any value of y in B, then, function;

$$CIV_k = f(FLP, ND)_k \quad (2)$$

$$CIV_k = FLP_k \times ND_k \quad (2.1)$$

Where: $k = k^{th}$ *particular device.*

3.3.2 Priority Ranking

With known individual CIV values, it is possible to order the impact levels of devices by their CIV values. Accordingly, it is possible to determine which devices possess the greatest cascading impact potentials depicted by the device with the greatest CIV value among the enumerated devices. This could also aid both inter and intra-layer security prioritization analysis, and the possible deduction of cascading impact trends due to cyber incidents on an ICS infrastructure.

4.0 SCENARIO TESTING

To aptly represent the concept presented, we adopt a testbed approach that involved developing an emulator ICS to mimic the basic functionalities of an industrial environment. Cyber-attack is then simulated on the built ICS

testbed with a view to underscoring the impacts on targeted devices, and the observable ripple effects on functionally dependent (directly connected) devices to the target, and how far the impacts extend to the whole entire network infrastructure.

4.1 Step 1 - Event Description (Attack type and Target)

The network consists of a PLC with extended Input/Output Modules, HMI, Industrial Switch, Programming Workstation, Attacker's Computer, a sensor-driven & motor-driven conveyor belt system, and sensor-driven & motor-driven 3D Robot system. Figure 3 shows a network architecture of the ICS emulator setup. The 3-D robot and conveyor/punching machine represent field equipment are controlled with photoelectric sensors and actuator limit switches receiving instruction sets from the PLC with its extended I-O module unit. The external I/O module, the HMI, and process control machine are connected via the Industrial Switch, which serves as the central hub for the PROFINET network. The PLC is linked to the same hub, sharing a direct communication connection on the PROFINET network. The attacker's system is used to gain access to the network via the industrial switch; being a prominent and easy entry and access point to the demonstrator network.

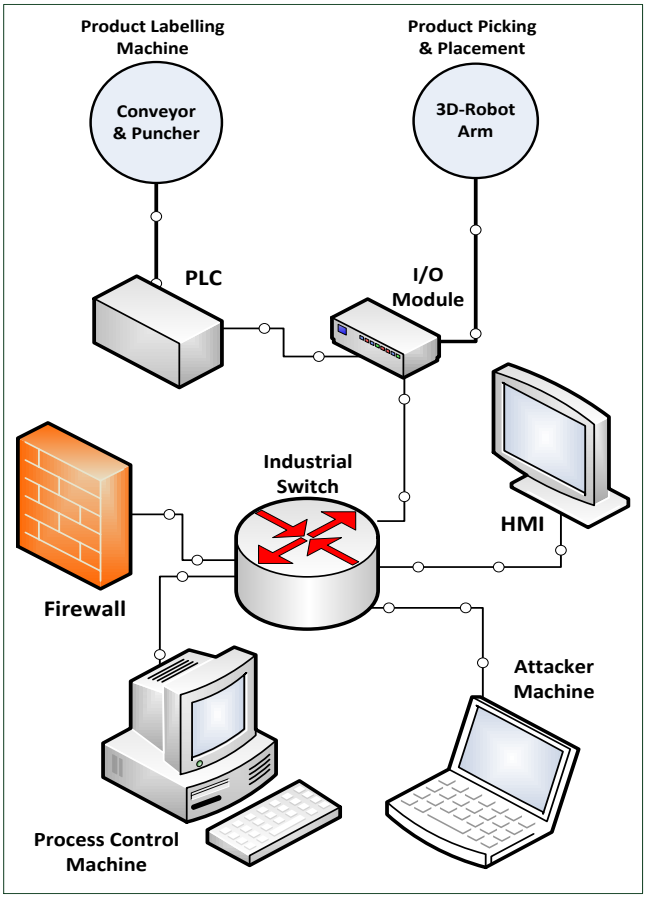


Figure 3: ICS Emulator Network Architecture for Test Scenario

Network reconnaissance was carried out (profiling and vulnerability analysis), target selection and exploitation undertaken. These were furthered using a Denial of Service (DoS) attack (using free exploitation tools available on Kali Linux Distribution) on the industrial switch.

The choice of DoS as attack mechanism is because it is the easiest form of attack on ICS that could bring about very severe impacts due to the design and computing power; emphasizing their inability to cope or withstand flooding and intermittent interruptions in communication. We thus considered the most critical security requirement of a typical ICS infrastructure; availability; which emphasizes the inability to access processing data at the needed time to accomplish pre-set task.

Exploitation of the discovered vulnerability on the industrial Ethernet switch (gateway) via a port 80, listening and accepting connection queries from a web client brought about very noticeable results, and impact on the ICS network infrastructure. To articulate further, it is essentially a buffer overflow attack executed through initiating a SYN flood command targeting the said port to achieve DoS; overwhelming the switch with numerous random packet than it can handle per time, and causing it to lag in its preconfigured status of enabling gateway data exchange amongst other devices on the network; control workstation, PLC, HMI, etc. As aimed, the attack was effective when sustained over an extended period; causing the switch to restart shutting off the flow of data on the network. Again, as a result, couple of cascading inefficiencies that rippled out to other devices on the network due to the disruption of the switch's functionality. On the immediate, there was a loss of functionality in term of monitoring capability on the HMI. The control of the PLC via the control station computer was breached and disrupted. Control of the extended I/O was temporarily lost which in turn affected the functionality of the 3D-Robot.

The outcome of the DoS attack on the prototype process ICS SCADA; specifically targeting the industrial switch, and the observation of corresponding impacts on other devices showed that clear functional (inter)dependencies exist amongst the components that made up the process control system. apparently, it affirms that ICS components are greatly dependent on the other such that an interruption due to cyberattack in one is capable of causing cascading impact on others, especially devices and assets directly connected to, and directly rely on data or instruction sets from the failed device for their normal functionality. As simply articulated in [171] in (Knowles, 2015), "what happens to one infrastructure can directly and indirectly affect other infrastructures, impact large geographical regions, and send ripple throughout the national and global economy". It is apparent that any marginal service or operational disruptions of ICS or its component(s) can induce devastating crunches on the environment and society.

4.2 Step 2 - Functionality Dependency Relationship (FDR) Structure

In line with the emulator (cyber-attack demonstrator) network architecture, the functional dependency relationship (FDR) structure was developed considering the proposed layering of devices in order of functionalities, and the mode of connection and dependencies expressed by the design.

4.2.1 Layering and Dependency Mapping:

Connector arrows were used to indicate the direction of dependency for each devices and multiple arrows directed on single devices depicted multiple dependences from underlying devices. Equation 1 above was used to compute the CIV values of individual devices as represented on the FDR structure in Figure

4.2.2 Impact Prioritization

Given that there are 5 functional layers represented in the functional dependency architecture, and a total of 8 connected devices in the architecture, equation 1 is used to compute individual device CIV values to generate an array of Cartesian product set *C* as shown in figure 5 the values shaded therein indicate the computed values of testbed ICS components on the scenario architecture. Precise CIV values for precise components are is clearly presented in

table 2., and a line chart showing the layer-by-layer CIV trend is also presented in Figure 6 4

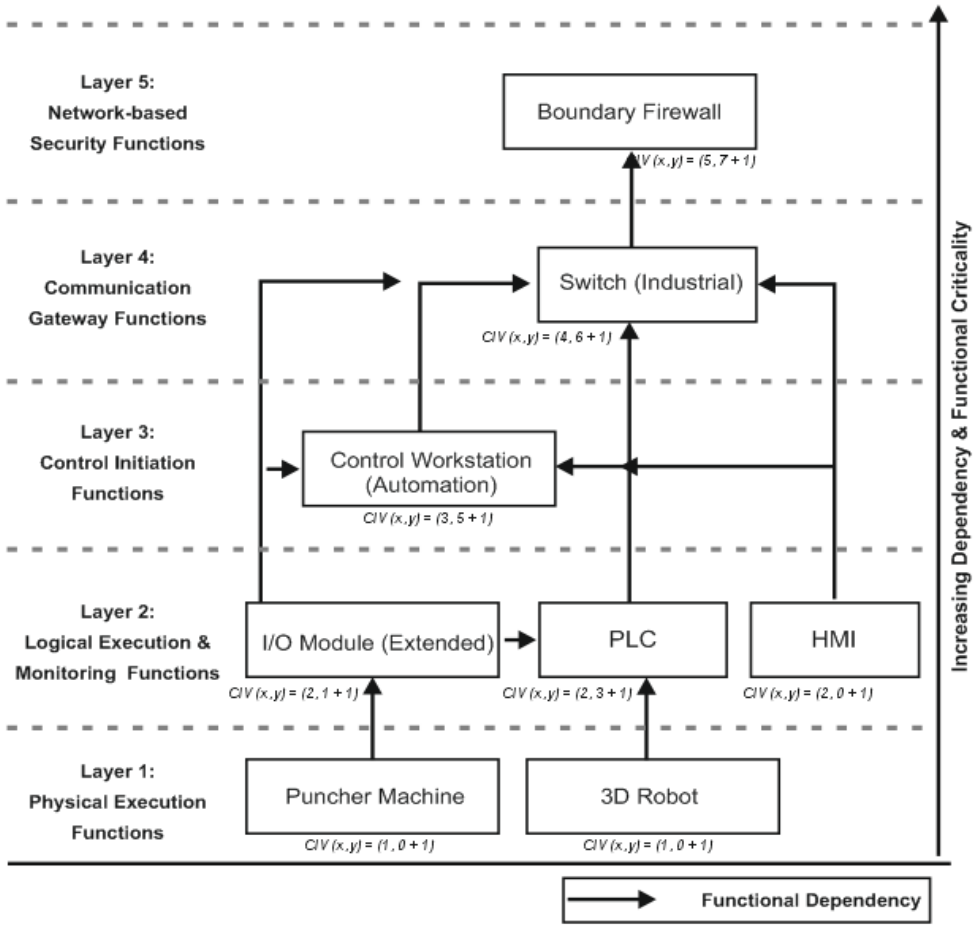


Figure 4: Functionality Dependency Relationship (FDR) Structure for Test Scenario

Functionality Layers									
	5	5	10	15	20	25	30	35	40
	4	4	8	12	16	20	24	28	32
	3	3	6	9	12	15	18	21	24
	2	2	4	6	8	10	12	14	16
	1	1	2	3	4	5	6	7	8
		1	2	3	4	5	6	7	8
		Number of Devices on Network Infrastructure							

Figure 5: Cascading Impact Value matrix (Cartesian) Representation for Test Scenario

Table 2: Precise Component Cascading Impact Values for test Architecture

Device	Layers	CIV
3D-Robot	1	1
Conveyor/Punching Machine	1	1
HMI	2	2
PLC	2	8
I/O Extension Module	2	4
Control Workstation	3	18
Industrial Switch	4	28
Industrial Firewall	5	40

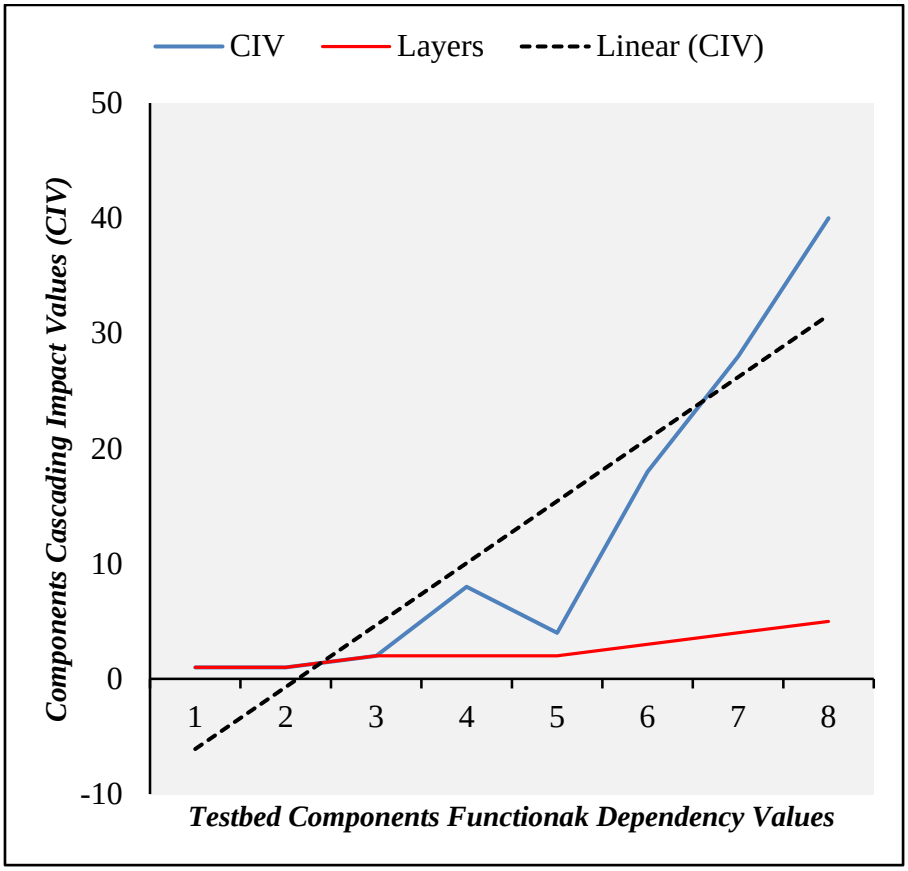


Figure 6: Cascading Impact Value Layered Trend

4.3 Analysis and Discussion

From the outputs derived, the device that takes on a highest priority for security response is interpreted to refer to the device that retains the greatest CIV, which further implies the impairment scale on device that exerts the widest rippling effect on the ICS network. The output from table 2 also makes easy the process of prioritizing the focus of security effort; the process of identifying what device(s) on the

network that assumes highest priority both from an inter and(or) and intra-layer analysis perspective. From the inter-layer analysis point for instance, the PLC is on layer 2 of the FDR architecture, and has a CIV = 8. The industrial switch is on layer 4 of the FDR structure with a CIV = 28. It implies that security attacks that impacts negatively on the standard functionality of the industrial switch would potentially amass wider, and potentially more devastating effects on the underlying directly or indirectly connected devices and potentially the whole network of ICS infrastructure. The cascading impact would be potentially greater than any similar cascading impairment resulting from the direct targeting of the PLC. If an inter-layer choice is to be made on which device of the two should be accorded first response in terms of security consolidation or hardening, certainly the industrial switch takes such priority position over the PLC. However, from an intra-layer standpoint, three devices; HMI, PLC, and I/O Extension Module, are all functionally positioned on layer 2 of the FDR structure. This presumes them to all bear equal impact exertion degree due to security attack on the infrastructure. However, their respective CIV values seem to vary due to the nature of functions played by each, and the number of lower layer devices that are functionally dependent on each of them. The HMI yields a CIV = 2, the PLC CIV = 8, and the I/O Extension module CIV = 4. Following the rule of greatest cascading impact value, the PLC takes on the highest priority position; needing most security attention and effort such that would control a wider impacts on the ICS infrastructure in the face of a cyber-attack targeting the layer 2 functional devices. Generally, following the outcomes of the Cartesian product matrix C represented in figure 5, it suggests that the industrial firewall with the greatest CIV = 40 assumes the greatest priority for security. Essentially, if such security capacity on an ICS infrastructure gets sabotaged and broken, then a potentially widest attack potential is enabled on the ICS infrastructure, and a potentially widest cascading impact is probable.

Typically, the CIV, and by extension the impact scope tends upwards up the layers of functional dependency or functionality attributes of an ICS infrastructure as seen in figure 6. It suggests that the strength and capacity of security efforts on the network infrastructure would be dependent on the position (layer) where security deployments are enabled on the infrastructure. And accordingly, as one moves upwards the FDR layers, the cascading impact of functionality impairments; depicted by CIV potentially increases.

The CIV increment by layer offers insights and justification into exploiting security optimization concepts from a network design perspective. It reaffirms already prescribed measures for improving ICS security via segmentation for networks, zones, and conduits, which describes the process of enabling physical and (or) logical demarcations, security access control, and by extension; cascading impact control. Larger networks are demarcated and broken down into smaller, more manageable, physical or logical networks with additional security controls complemented by the *Principles of Least Route.* This emphasizes that devices that do not physically or functionally belong to a zone should be denied direct connection to the zone. Since higher CIV means potentially wider cascading impacts due to security incidents, it follows that implementing least route principle would involve breaking down a network of ICS into smaller bits of functionally dependent and related devices and network, and according to the roles or functions in the actualization of predetermined processes. It would ensure that a device that is not directly connected to zone would not be allowed connection, functional dependence. Accordingly, impacts in one small zone would be controlled and not allowed to ripple out into other zones on the network. Smaller zones would bring out lower CIVs;

meaning reduced functional dependencies and cascading impacts on the network infrastructure. This work theorizes the concept '*Attack-space attenuation'*; which essentially seeks to shrink the impact space and scope of cyber-attacks on ICS networks; tightening reachability to devices, and constraining rippling effects within a small as possible area of the network.

Cascading impact scoping offers an insight into the varied damage potentials that could be sustained as a result of the failure of functionally inter(dependent) ICS components. Accordingly, this CIV offers a potential security metric quantity which could be used to assess a system or organization's security posture in a much quicker way to invoke quicker decision making and response; to avert potential damages. Aside from aiding speedy and proactive security response, this quantity as a metric could also be a subset; proffering valuable input to a larger, desirable, quantitative security metrics taxonomy. The approach presented can also be grafted into a much larger ICS critical control point security risk assessment scheme.

5.0 CONCLUSION

In the face of the numerous complexities in network architectures, infrastructure integrations, and the dependencies that abound, simple approaches are required that can help ICS designers and developers achieve more robust and secure architectures. In a modern industrial network where advancements are enabled through IT-OT infrastructure integrations, failures due to cyber-attacks on higher-level IT components could as well exert very serious devastating effects on lower-level OT, SCADA, and ICS component infrastructure. Because of the nature of connectivity and dependency among these integrated infrastructures, each component whose functionality is impaired due to cyber adversarial action, would to some degree lead to multiple cascading chains of negative impacts on its functionally dependent components. Very needful are approaches that can help both technical and non-technical users attain easy and quick strategic prioritization, recommendation, and placement of security features and capabilities for enhanced or hardened security posture. Functional dependency modelling explored proffers a high-level approach to achieving these gains via cascading impact scoping; an approach for assessing the span to which impairments can spread over a targeted network. If meticulously engaged, this approach would offer insight to the effectiveness and efficiency of any intending network segmentation and improvement activity geared towards enhancing security of a pre-evaluated ICS network. It will allow for multiple tests and optimization potentials without necessarily engaging the physical and financial rigours of real environment testing. It also helps to understand and determine which out of several discovered vulnerabilities; if exploited, bears the potential to causes the greatest impact or harm to a network infrastructure, and aid prioritization for response. Cascading impact value can also be considered a subclass of any security metrics taxonomy, and potentially an attribute in a larger security risk assessment framework. Consequently, infrastructure owners (especially the non-technical section) can achieve speedy and timely decision-resolve that can bring about improved cyber security on ICS infrastructure; in response to known or discovered vulnerabilities and their potential impacts.

REFERENCES

[1] T. Macaulay and B. L. Singer, "ICS Vulnerabilities," in *Cybersecurity for Industrial Control Systems: SCADA, DCS, PLC, HMI, and SIS*, Boca Raton, FL: CRC PRESS : Taylor &Francis Group, 2012, pp. 81–124.

[2] A. a. Ghorbani and E. Bagheri, "The state of the art in critical infrastructure protection: a framework for convergence," *Int. J. Crit.*

Infrastructures, vol. 4, no. 3, pp. 215–244, 2008.

[3] S. Delamare, A. A. Diallo, and C. Chaudet, "High-level modelling of critical infrastructures' interdependencies," *Int. J. Crit. Infrastructures*, vol. 5, no. 1/2, pp. 100–119, 2009.

[4] L. J. Wells, J. A. Camelio, C. B. Williams, and J. White, "Cyber-physical security challenges in manufacturing systems," *Manuf. Lett.*, vol. 2, no. 2, pp. 74–77, Apr. 2014.

[5] A. Tesfahun and D. L. Bhaskari, "A SCADA testbed for investigating cyber security vulnerabilities in critical infrastructures," *Autom. Control Comput. Sci.*, vol. 50, no. 1, pp. 54–62, 2016.

[6] P. Kotzanikolaou, M. Theoharidou, and D. Gritzalis, "Cascading Effects of Common-Cause Failures in Critical Infrastructures," in *Critical Infrastructure Protection VII*, Series Vol., vol. 417, no. 2003, J. Butts and S. Shenoi, Eds. Berlin Heidelberg: Springer Berlin Heidelberg, 2013, pp. 171–182.

[7] R. D. Larkin, J. Lopez Jr., J. W. Butts, and M. R. Grimaila, "Evaluation of Security Solutions in the SCADA Environment," *Data Base Adv. Inf. Syst.*, vol. 45, no. 1, pp. 38–53, 2014.

[8] R. K. Shyamasundar, "Security and Protection of SCADA : A Bigdata Algorithmic Approach," in *Proceedings of the 6th International Conference on Security of Information and Networks*, 2013, pp. 20–27.

[9] E. K. Wang, Y. Ye, X. Xu, S. M. Yiu, L. C. K. Hui, and K. P. Chow, "Security Issues and Challenges for Cyber Physical System," in *2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*, 2010, pp. 733–738.

[10] W. Stallings, *Cryptography and network security: principles and practice*, 5th Editio. Upper Saddle River, NY: Pearson Education, Inc., publishing as Prentice Hall. l, 2010.

[11] D. Gollmann, "Veracity, Plausibility, and Reputation," in *Information Security Theory and Practice. Security, Privacy and Trust in Computing Systems and Ambient Intelligent Ecosystems, Volume 7322 of the series Lecture Notes in Computer Science*, I. Askoxylakis, H. C. Pöhls, and J. Posegga, Eds. Egham, UK: Springer Berlin Heidelberg, 2012, pp. 20–28.

[12] B. Zhu, A. Joseph, and S. Sastry, "A taxonomy of cyber attacks on SCADA systems," in *Proceedings - 2011 IEEE International Conferences on Internet of Things and Cyber, Physical and Social Computing, iThings/CPSCom 2011*, 2011, pp. 380–388.

[13] N. Chandrika, "Cyber Security in the UK," *POSTnote*, no. 389, pp. 1–4, 2011.

[14] CPNI, "Cyber-attacks : Effects on UK Companies," *Oxford Econ. CPNI Publ.*, no. July, 2014.

[15] S. Radack, "Protecting Industrial Control Systems - Key Components of Our Nations Critical Infrastructures," *ITL Bulletin*, no. August, Gaithersburg, Maryland, pp. 1–7, Aug-2011.

[16] E. Leverett and D. F. S. & P. J. Crowcroft, "Quantitatively Assessing and Visualising Industrial System Attack Surface," *Comput. Lab.*, vol. MPhil, no. June, p. 54, 2011.

[17] J. F. Brenner, "Eyes wide shut: The growing threat of cyber attacks on industrial control systems," *Bull. At. Sci.*, vol. 69, p. 15, 2013.

[18] C. Alcaraz and S. Zeadally, "Critical infrastructure protection: Requirements and challenges for the 21st century," *Int. J. Crit. Infrastruct. Prot.*, vol. 8, pp. 53–66, 2015.

[19] Roland-Berger, "Think Act: Cyber-Security, Managing threat Scenarios in manufacturing companies," Munich, 2015.

[20] K. Yoo, "A Glimpse into Delphi: The World's First Industrial Cyber Security Vulnerability Database," *Website Article*. Wurldtech Security Technologies, Vancouver, p. 17, 2009.

[21] L.-O. Wallin and N. Jones, "The M2M Market Evolution: Growth Attracts Everyone," 2011.

[22] L. Thames, "The IIoT_ Fueling a new Industrial Revolution," *The State of Security*. TRIPWIRE, INC., 2015.

[23] T. Lu, X. Guo, Y. Li, Y. Peng, X. Zhang, F. Xie, and Y. Gao, "Cyber-Physical Security for Industrial Control Systems Based on Wireless Sensor Networks,"

Downloads.Hindawi.Com, vol. 2014, 2014.

[24] H. Chae, A. Shahzad, M. Irfan, H. Lee, and M. Lee, "Industrial Control Systems Vulnerabilities and Security Issues and Future Enhancements," *Adv. Sci. Technol. Lett.*, vol. 95, no. Cia 2015, pp. 144–148, 2015.

[25] K. Stouffer, V. Pillitteri, S. Lightman, M. Abrams, and A. Hahn, "Guide to Industrial Control Systems (ICS) Security - NIST.SP.800-82r2," 2015.

[26] H. Li, G. W. Rosenwald, J. Jung, and C. C. Liu, "Strategic power infrastructure defense," *Proc. IEEE*, vol. 93, no. 5, pp. 918–933, 2005.

[27] R. Ebrahimy, "Investigating SCADA Failures in Interdependent Critical Infrastructure Systems," 2014.

[28] W. Knowles, D. Prince, D. Hutchison, J. F. P. Disso, and K. Jones, "A survey of cyber security management in industrial control systems," *Int. J. Crit. Infrastruct. Prot.*, vol. 9, pp. 52–80, 2015.

[29] S. V Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, no. 7291, pp. 1025–1028, 2010.

[30] P. Zhang and S. Peeta, "A generalized modeling framework to analyze interdependencies among infrastructure systems," *Transp. Res. Part B Methodol.*, vol. 45, no. 3, pp. 553–579, 2011.

[31] S. M. Rinaldi, J. P. Peerenboom, and T. K. Kelly, "Identifying, understanding, and analyzing critical infrastructure interdependencies," *IEEE Control Syst. Mag.*, vol. 21, no. 6, pp. 11–25, 2001.

[32] R. Setola, S. De Porcellinis, and M. Sforna, "Critical infrastructure dependency assessment using the input-output inoperability model," *Int. J. Crit. Infrastruct. Prot.*, vol. 2, no. 4, pp. 170–178, 2009.

[33] J. Johansson and H. Hassel, "An approach for modelling interdependent infrastructures in the context of vulnerability analysis," *Reliab. Eng. Syst. Saf.*, vol. 95, no. 12, pp. 1335–1344, 2010.

[34] R. Zimmerman, "Decision-making and the vulnerability of interdependent critical infrastructure," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2004, vol. 5, pp. 4059–4063.

[35] S. De Porcellinis, G. Oliva, S. Panzieri, and R. Setola, "A holistic-reductionistic approach for modeling interdependencies," in *IFIP Advances in Information and Communication Technology*, vol. 311, S. Shenoi and C. Palmer, Eds. Springer Berlin Heidelberg, 2009, pp. 215–227.

[36] D. D. Dudenhoeffer and M. R. Permann, "Proceedings of the 2006 Winter Simulation Conference," 2006, no. Clinton 1996, pp. 478–485.

[37] S. M. Rinaldi, "Modeling and simulating critical infrastructures and their interdependencies," in *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*, 2004, vol. 0, no. C, pp. 1–8.

[38] Y. Haimes, J. Santos, K. Crowther, M. Henry, C. Lian, and Z. Yan, "Risk analysis in interdependent infrastructures," in *IFIP International Federation for Information Processing*, vol. 253, C. Palmer and S. Shenoi, Eds. Springer Berlin Heidelberg, 2007, pp. 297–310.

[39] M. Theoharidou, P. Kotzanikolaou, and D. Gritzalis, "A multi-layer Criticality Assessment methodology based on interdependencies," *Comput. Secur.*, vol. 29, no. 6, pp. 643–658, 2010.

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June 2018

UGMAP: A WEB AND MOBILE APPLICATION FOR NAVIGATION AND NOTIFICATION THE UNIVERSITY OF GHANA, LEGON CAMPUS

W. Owusu-Banahene[1], A. A. Amihere[2]

[1,2] University of Ghana, Computer Engineering Department, Legon, Accra, Ghana

[1]wowusubanahene@gmail.com; [2] rajoa30*@gmail.com*

ABSTRACT

Getting directions from one place to another on the University of Ghana, Legon campus can be very exhausting. In this project, a more suitable way to locate venues using a web and mobile platform was developed. This research aimed at developing a 'social-academic media' platform on top of a geo-information service using the power of mobile and web technologies in disseminating information to a wider audience. A web and mobile based 'social-academic media' platform referred to as UGMAP was developed. UGMAP allows a system administrator to create events and send notifications (alerts) to users through the web and also to their handheld devices. The mobile module of UGMAP is implemented in Android to provide the different routes that the users can take to navigate to their given destinations. This 'social-academic media' platform enables a system administrator to make modifications of events taking place within the University of Ghana, Legon campus to users via web and or mobile technology. This research shows the role of location based services in a social media context. Future additions to the work will focus on enabling the users to get dynamic route mapping, where the routes on the maps move as the user moves, and to enable any information concerning an event to be automatically fed into the application.

Keywords: Android application, location, mobile, web, geoinformation service.

1.0 INTRODUCTION

The task of finding places is not an uncommon phenomenon in human societies. People usually get to an event late, get stressed out by the time they arrive, or miss the event altogether because they do not know the location of an event beforehand. Also, some students on the University of Ghana Campus hardly ever get to know of events that happen due to busy schedules. Students and staff who are new to the campus such as distance students, international students, first year students and students on exchange programs as well as newly recruited staff need a better way of getting directions. There is a problem of people missing their way to venues and sometimes even arriving late at the events when proper directions are not given. These challenges inspired the development of the new way to get to know about events, where they are located and how to get there.

It is therefore necessary to have a dedicated application that can provide navigational information faster and more efficient way.

This research aimed at developing a platform referred to as UGMAP that would enable a system administrator to input data from the back-end and allowing for notifications to be sent to users in a convenient way. Also, the time of users which would otherwise be used to ask for directions would be saved. The next objective is to display the route from where the user is to where the user is going on a mobile device. Finally, the application is meant to display information about events at the respective venues.

The research described in this paper was implemented for notification and alert with geoinformation as the backbone for users within the University of Ghana Campus community. And it targets the students, visitors and staff, especially those who are new to the campus where rapid dissemination of information is always essential. UGMAP has an integrated web and mobile modules. The web module was developed using Ruby on Rails [[1],[2],[3]] for the web application and Android java for the mobile application. The research can be adapted in any given geographical space. The data from the back-end include addition of categories, venues and events that would be happening on campus. This would save the time of the users by showing routes on the hand-held mobile device of the user and displays information concerning events happening on the campus. The Rails and Android programs works with Ajax and RESTful services to implement the mobile module which run on a GPS-enabled device (such as a phone or laptop) to present [[4],[10]]. UGMAP does not only give the user directions to their destinations, but also sends notifications to their mobile device or phone, prompting them of upcoming events such as lecture times, exams, public events, career fairs and many others. The mobile application also routes the different paths the user can take to reach their given destination.

The remaining paper is structured as follows. Section 2 presents a background to UGMAP and a review of related work. In section 3, the system design and implementation are presented. Section 4 presents results and discussions. The paper ends with conclusions and future work in section 5.

2.0 BACKGROUND AND RELATED WORK

This section contains a brief description of the background to UGMAP, the web and mobile application for navigation and notification and related work.

One common class of service available through mobile phones centres on finding where one's location is and seeking out the nearest facilities such as cash machines, restaurants and taxis [10]. Mobile communication entails transmission of data to and from handheld devices such that least out of the two or more communicating devices, one is handheld or mobile [11]. Location-based services (LBS) are the delivery of data and information services where the content is tailored to a mobile user's location and context and thrives on mobile phones, Internet, the World Wide Web and

Global Positioning System (GPS) and geographical information (GI) systems/services. [10]. Web services enables disparate applications running on different machines to exchange data and integrate with one another without requiring additional third party software or hardware [12]. The Google Maps API has spawned a whole class of web-based applications that would have been impossible to create without it. Rails and Google Maps enable developers to build impressive web-based applications [13].

[14] introduced a way in which a user can access resources on the internet in a language that they understand. From the paper, it was noted that translation engines which aid in translation of web pages over the internet work very well with static pages but poorly with dynamic pages. The project was aimed at helping in the communication with the user through dynamic web pages in a language they understand, with the returned resultant web pages in the same language.

The paper, [15], sought to solve this issue using two main modules; The RTR (Retrieve, Translation and Render) which allows the system to get a web page for the user, translate it into the user's choice of language and render an appropriate page to the user's machine and the IHDD (Input Handler and Data Dispatcher) which is needed for the conversion of the user's input from their native language to English language which is forwarded to the web server.

A study was conducted in this paper by Lok Fang et al. [16] to evaluate the maintainability of web applications developed on J2ee, .NET and Ruby on Rails (RoR). The basis for the test is its ability to be modified, tested, understood and portability. It was discovered that RoR was more modifiable, testable and understandable. J2EE was however more portable. This test was done for Small Scale Web Applications. This project aims at getting a decision matrix for programmers to evaluate which platform would be necessary for a given web application. In conclusion, RoR was adjudged more maintainable as compared with J2EE and .NET when used on relatively small Web applications with basic CRUD operations.

[17] discussed how semantic ITSs (Intelligent Transportation Systems) can be implemented in practice using Ruby on Rails (for user interactions), Protegè (for data management) and Sesame (for user-data processes). ITSs are meant to monitor live current user location and traffic situations and to access different data from various sources. A set of databases including geospatial, administrative and business information which are linked by a known data model to allow mobile users to receive information through a semantic interface [18]

The need of Tourists to find locations using applications inspired Aleksander et al. [19] to develop a Google maps based technology with Ajax to meet this need. In the wake of increasing demand for modern websites with customised data services, a fuzzy clustering technique of building of ideology-based user profiles is proposed. The similar features in clusters are used to determine the distance between them. Rapid Web Development implies fast and effective web application development which is aimed at meeting customer deadlines, and allows for fast prototyping. It uses already available technologies (open source software, frameworks, APIs libraries, etc.) and brings them together (as a platform for running a Web 2.0 service) to help in the application construction [20].

Google maps API which has functionality for very responsive visual interface and inbuilt Ajax technologies. It allows programmers to include mapping features into their websites with their own coordinates [21]. This project was developed for a popular tourist destination Palic. The database used was able to store user and object profiles. The training of the Cluster was with a static dataset.

The applications were developed that would help to solve the problem of finding venues on the University of Ghana campus. A web and

mobile application was developed that would help to have prior notice of events that would be taking place on campus and locate the venues where they would be held.

UGMAP is an attempt to create a 'social-academic media' platform on top of geoinformation service using the power of mobile and web technologies in disseminating information to a wider audience.

3.0 SYTEM DESIGN AND IMPLEMENTATION

This section describes how the project was carried out. The process of development of the project is the Agile Method. This method was used because it allows for the process to be adaptable. It also allows the programmer incrementally develop the program. This enables faster delivery of the product and allows the customer to be satisfied because he or she can make recommendations as to how the system should function. [22]

The Agile method is an iterative and incremental method. It can allow the system to be updated often therefore, the cost of making new modifications is less. This allows changes to be easily effected in the system. Unlike the waterfall model, less planning is needed since the system changes quite frequently. This allows the projects to start and finish at a rapid rate.

3.1 System Architecture

The system architecture of UGMAP is shown in Figure 1. By design UGMAP has two components; web and mobile.

1. The system administrator at the backend inputs the Categories of venues, (such as Schools and Colleges), Venues (Example School of Engineering) and Events that would be taking place.
2. The server the administrator runs on (Localhost) sends the modifications and updated information to a platform (Heroku) on the cloud for access by other users on the application.
3. For the user on the mobile device to have updates on the information posted, Google Cloud Messaging Service is used.
4. The user android mobile device sends a sender ID (application ID) to the GCM server
5. When successful, the GCM sends a registration ID to the android mobile device as confirmation.
6. After receiving the registration ID, the android device user sends the registration ID to the web server on which the program is running.
7. The registration ID is sent to the web server which is stored in a database.
   a. Whenever a push notification is needed, the web server sends a message to the GCM server along with the device registration ID (which is stored in the database)
   b. GCM server will deliver that message to the respective mobile device using the registration ID. [23]

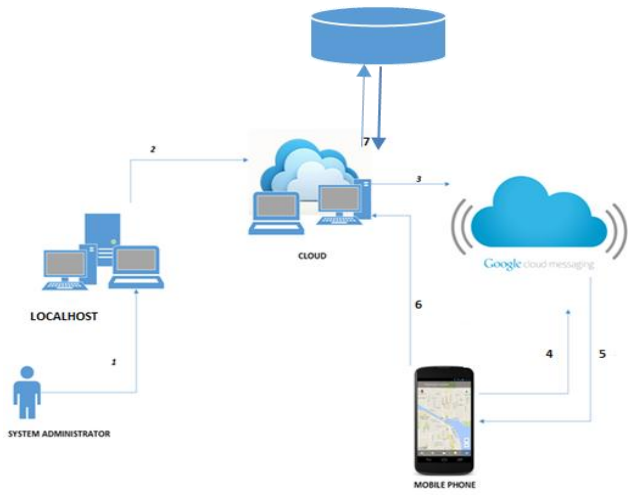


Figure 1: System architecture

3.2 System Requirements

The requirements of the system were derived as follows.

Requirement 1. *The system should be able to log coordinates of a moving bus into a database.*

Requirement 2. *The system should be able to log coordinates of a user into a database.*

Requirement 3. *The system should be able to show the location of a moving bus on a map to an administrator at any point in time.*

Requirement 4. *The system should allow an administrator to register drivers.*

Requirement 5. *The system should allow an administrator to assign a bus to a registered driver.*

Requirement 6. *The system should be able to show the location of a user on a map relative to that of a bus.*

Requirement 7. *The system should be able to display the distance between the user and bus.*

3.3 Software and Hardware Reqirements

In this section the software and hardware used to develop the system are briefly presented.

3.3.1 Software Specifications

The Ruby language was released in 1995, and invented by Yukihiro Matsumoto (commonly known as 'Matz'). It is a cross-platform, interpreted language that has similar characteristics as other 'scripting' languages such as Perl and Python ([1]). Ruby is a purely object-oriented programming language because every class is represented as an object as well as every result that is obtained from the manipulation of the classes or methods in the code ([2]).

Rails is a software package library created by David Heinemeier Hansson to extend the functionalities of Ruby. It is a ruby framework for building websites, but is integrated with HTML, CSS and JavaScript to build web applications that run on a web server. Based on its ability to run on a web server, it is known as a back-end or server-side web application development platform with the web browser as the front-end ([3]).

Ruby on Rails (known among programmers as Rails) is a web development software that can be used for different varieties of projects with much ease and flexibility. It is a made up of a Rails framework which runs on ruby code ([4]). There are two main principles of Rails which are; DRY (Don't Repeat Yourself) and Convention over configuration. DRY ensures that code is not duplicated in different places. This is facilitated by the MVC (Model View Controller) architecture. The Convention over configuration means that Rails has pre-implemented 'sensible defaults' which makes code writing very easy for the programmer, and also simplifies code modifications the programmer may want to make to the code generated. It can be easily used with Ajax and RESTful services ([4]).

The main reasons why this framework was used to build this application are; Rails is agile ([3]). This means that it is simple and subtle. It implies that rails hails individuals and interactions over processes and tools. Also, a working code, customer collaboration over

contract negotiation and most importantly, the ability of the program to respond to change, rather than follow a laid down plan.

Another reason is that Ruby has terse and uncluttered syntax that does not need a lot of punctuation. Also, Ruby is a modern high level language which supports abstraction such as metaprogramming which makes it easier to develop DSL (Domain Specific Language) which customises Ruby for a particular set of uses (needed in Rails and other gems) ([3]).

Also, Ruby has Ruby Gems, a software package manager which makes it easier to manage software libraries that extend Ruby. It provides a simple system to install gems.

Last but not the least, Rails conventions are pervasive and astute. This is because all the necessary code that a programmer may need to use in a web application is pre-written, it makes collaboration easier, development is quicker and there is a larger pool of open source libraries to enhance Rails functionality ([3]).

Another reason for the choice of Ruby on Rails is its flexibility with web services and applications in the Web 2.0 ([5]) ([6]). Web 2.0 in emergence is the ready availability of various applications as APIs and Web Services. The ease of web development in Rails owes to its rapid application development, database manipulations and Ajax makes it ideal for creating front-end and back-end applications for the rapidly evolving Web applications and services.

The mobile application was developed in Android. Android is a software toolkit for mobile phones, created by Google and the Open Handset Alliance. It's inside millions of cell phones and other mobile devices, making Android a major platform for application developers ([6]).

Google maps API was also used for both the web application and mobile application. Google Maps API is a technology provided by Google based on AJAX, which powers many map-based services. The realized software uses free, public API service from Google Maps. The system utilizes a knowledge base formed by tracking user actions ([7]).

Integrating these technologies to create a web and mobile app that would aid students and general users to get their destinations with ease, and receive notifications on events and where they would be held, as and when the administrator logs them in at the back-end.

The web application was hosted on Heroku, a deployment and web application hosting platform ([8]). This allows for the app to be viewed from any machine with a URL

3.3.2 Hardware Specifications

The following hardware were required: Android mobile phone, Universal Serial Bus (USB) Cable and a Personal Computer (PC) such as laptop.

3.4 System Modelling

This section shows the UML diagrams of how the various components of the web applications and the mobile application connect, and how the components react with each other. The use case diagram, sequence diagram and *flow* diagrams are presented.

3.4.1 Use Case Diagram

Figure 2 shows the use case for UGMAP. The user has a number of actions it must perform.

The system administrator also performs some actions. There are two systems; The Map Application and the Google API.

A user can:

- Browse / surf the application
- Search for a location or information about a location
- Can get information requested for
- Can receive a displayed output.

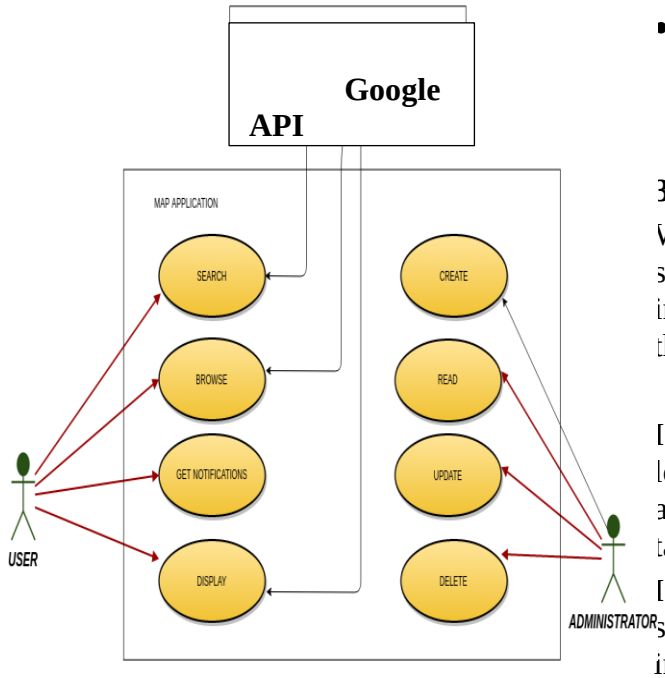


Figure 2: Use case diagram

A system administrator can perform CRUD operations that is;

- Create – Allows the system administrator to create Categories, Venues and add Events.
- Read – To see what has been added before any upload and update can be done.
- Update – To modify the existing content in the system.
- Delete – The System Administrator is the only one allowed to delete anything from the system.

3.4.2 Flow Chart and Sequence Diagram

When the program starts, it connects to the web service. If there is an error (example, poor internet connection), the program goes back to the previous step.

If there is no error, the app can display the locations on the app to the user. The user is allowed to get information about the events taking place.

If there is an error in the current step, previous step is repeated. If there is no error, the information is displayed. The program then ends.

3.5 Implementation of UGMAP

The implementation of UGMAP is presented in this section. The two main components of UGMAP, namely *web* and *mobile* modules were developed and integrated.

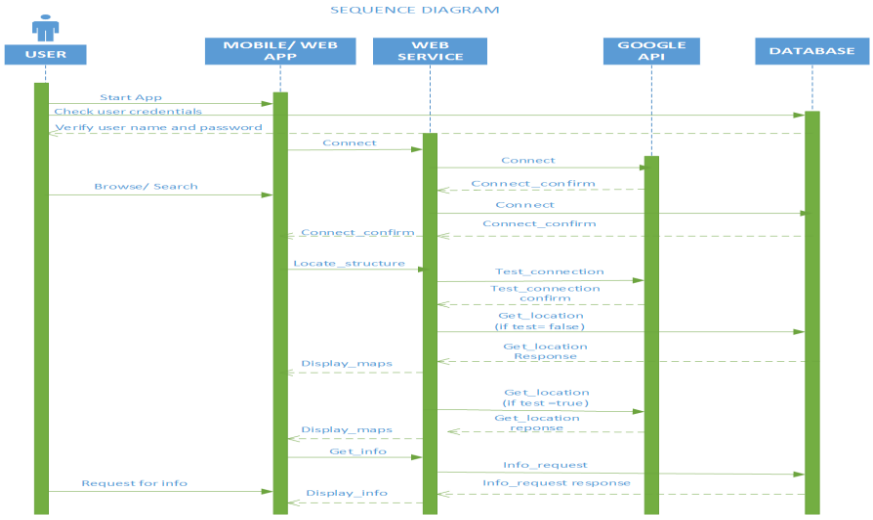


Figure 3: Sequence diagram

3.5.1 Web Application Module

The web application is divided into two parts namely back-end and front-end.

Back-end: This is the portion of the system where the system administrator is allowed access into the data and will be allowed to input new information into the system that would be displayed on the user's application.

*Front-end:*This is for the user to access an application from anywhere they are located. The application is viewed through a browser using a URL. The following components were used for the web application

- Ruby (version 4.2.5.1)
- Rails (version 2.2.3p173)
- Ruby gems (List of ruby gems and a brief description of what they do)
- gem 'sass-rails' – SCSS for stylesheets
- gem 'uglifier', '>= 1.3.0' - compressor for JavaScript assets
- gem 'jQuery-rails' – jQuery as the JavaScript library
- gem 'turbolinks' - Turbolinks makes following links in your web application faster.
- gem 'jbuilder', '~> 2.0' - Build JSON APIs with ease.
- gem 'devise' – for security features of rails objects
- gem 'gmaps4rails' – for Google maps functionality in rails
- gem 'ckeditor' - for an enhanced editor in the system administrator's window
- gem 'gcm' – for Google Cloud Messaging functionality in Rails

Ruby Gems is a package manager for the Ruby programming language that provides a standard format for distributing Ruby programs and libraries (in a self-contained format called a "gem"), a tool designed to easily manage the installation of gems, and a server for distributing them. [24]

- Google Cloud Messaging

This is a free cloud service that was used to push lightweight messages from the backend server to notify the user of new content to the apps installed on Android Devices. [25]

- Heroku Cloud Services

Heroku is a cloud platform with integrated data services for managing servers, deployment, ongoing operations or scaling. It runs the applications in containers called dynos on a reliable, fully managed runtime environment. It supports languages such as Node, ruby, Java, PHP, Python, Go, Scala, or Clojure. [26].

3.5.2 Mobile Application Module

The mobile application was developed using the following:

- Android Studio

It is a flexible Gradle-based build system with a fast and feature-rich emulator which allows for the development of all Android devices. [27]

- An Android phone with GPS

Android phones usually support GPS, a system which uses Wi-Fi or mobile data to speed up getting the information needed for an initial position fix. To benefit, turn GPS on a few minutes before it's needed -- ideally when still connected to Wi-Fi.

Leaving GPS reception on, however

drains the phone's battery more quickly. [28]

- Google play, version above 4.0

3.5.3 System Operation

The operation of the system is as follows:
The Categories of the venues (such as Schools and Colleges) and the Venues (Such as School of Engineering Sciences) are entered by the system administrator. Corresponding events (like Exams) are entered in the back-end. The user of the web or mobile application will receive an update of the event. The mobile user would receive a notification on their phone concerning the specific events and the details (such as the time, the venue, etc.)
When the app is opened, it would yield the lists of events that are happening at the various venues. A selected event would open a map activity which would have the different routes to the selected venue of the event chosen. The GPS must be turned on to show the time it would take the user to get there and the route distance in meters.

4.0 RESULTS AND DISCUSSIONS

This section presents the results of a prototype UGMAP: navigation and notification system**.** The main modules namely, *Web Application and Mobile Application API* were implemented (see sections 3.5.1 to 3.5.3) and integrated. The results of the web module and mobile modules are presented in section 4.1 and section 4.2 respectively.

4.1 Results of the Web Application Module

The results of the web module are presented from the administrator's side (sub section 4.1.1) and client side (section 4.1.2).

4.1.1 Administrator's Side

The operation of the system is as follows:

The Categories of the venues (such as Schools and Colleges) and the Venues (Such as School of Engineering Sciences) are entered by the system administrator. Figure 4 shows the administrator *Login Page* .

The Administrator enters the URL of theapp (ugcampusmap.herokuapp.com/users/sign_in) into the web portal and enters credentials to allow access into the backend of the application. After the login, the home page of the application opens to show the existing events and shows the categories, venues and events in the side bar as shown in Figure 5. The Categories page showing all existing categories entered by the system administrator is shown in Figure 6.

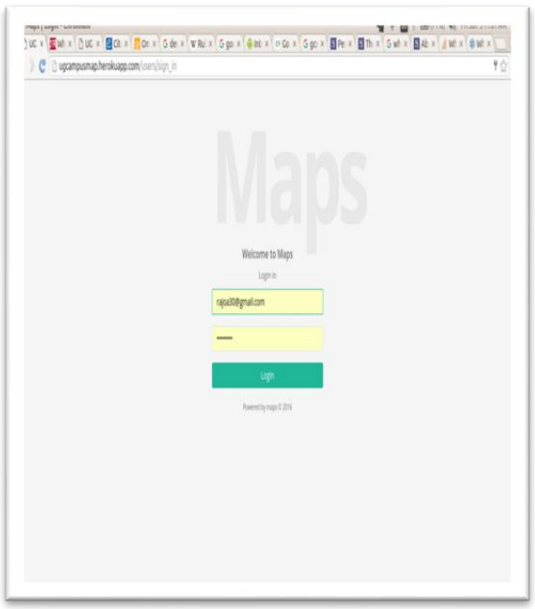


Figure 4: Administrator's Login page

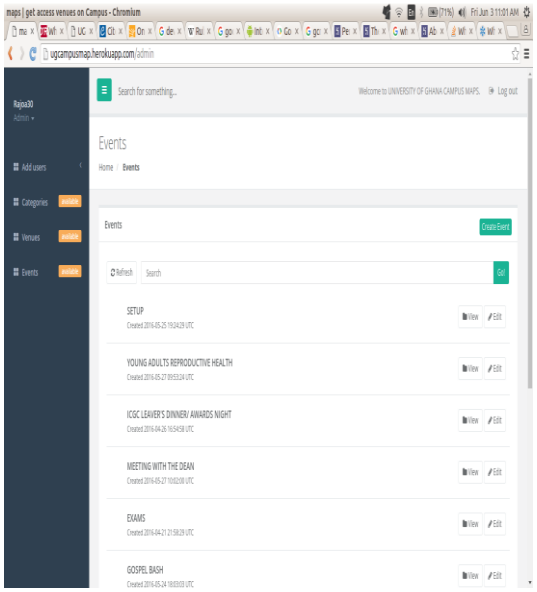


Figure 5: Events categories and venues page

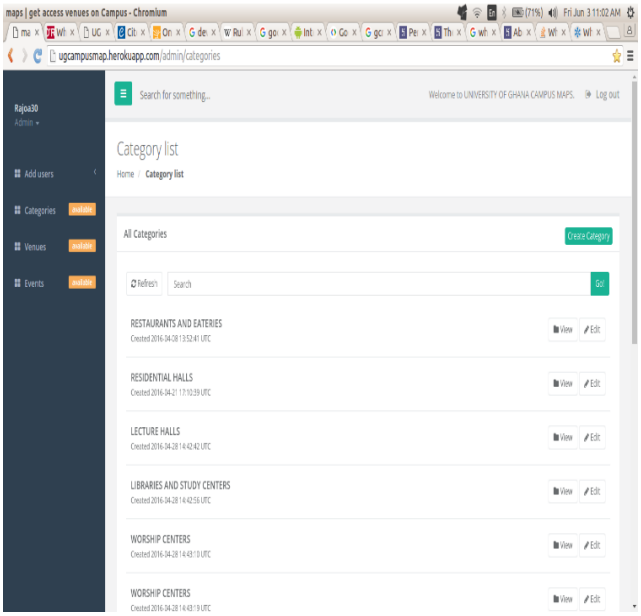


Figure 6 Categories page showing all existing categories entered by the system administrator.

The venue page displays all existing venues created by the system administrator as shown in Figure 7.

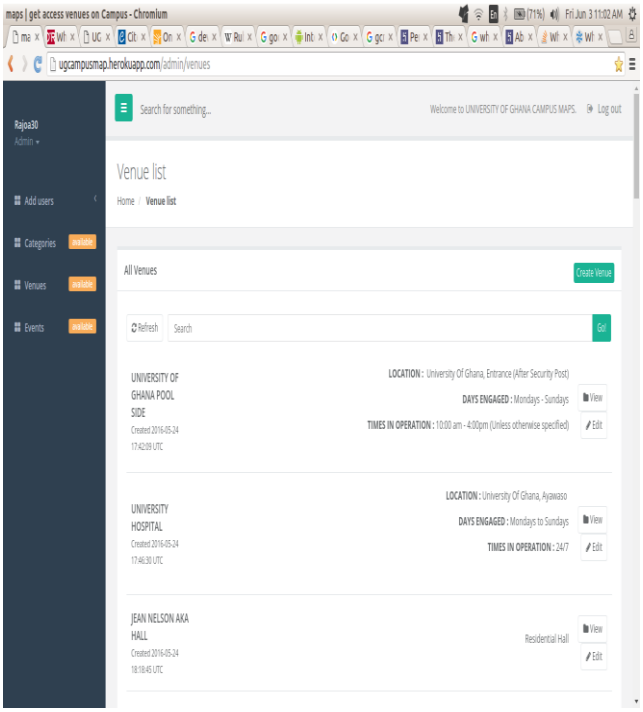


Figure 7: Venue Home Page.

When the option for creating a new venue is selected and the Name, Category and Location of the venue on the Map is selected the Longitude and Latitude points are automatically extracted on the map into the fields provided. Figure 8 shows the page for creating a new venue.

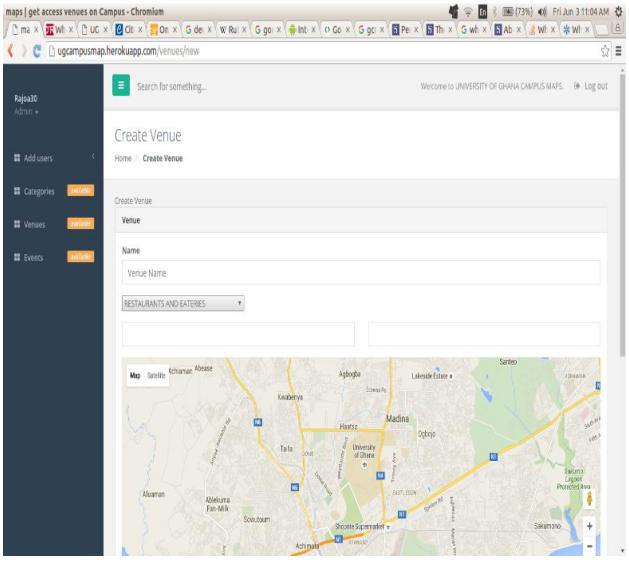


Figure 8: A Page for creating new venues

4.1.2 Client Side

A web client accesses the home page via URL http://ugcampusmap.heroku.com (this is still active at the time of writing this paper). When the said URL is accessed a page as shown in Figure 9 appears. The page also shows all the events of in the application with their dates beside them.

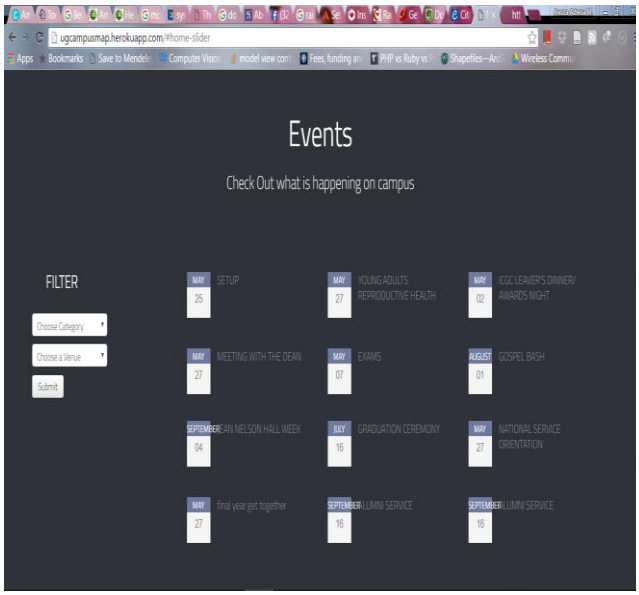


Figure 9: The Home screen of the web application

From the home screen, a user can select category. This is the first step in the process of displaying an event to the user. Figure 10 shows the selection of a category from the category drop down menu.

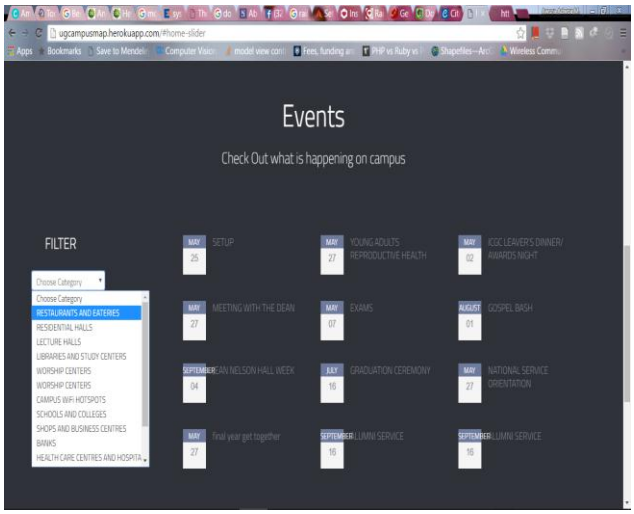


Figure 10: Selecting a category

Figure 11 shows the selection of the venue (CENTRAL CUISINE) in the selected Category (RESTAURANTS AND EATERIES). This is the second stage in the process of viewing an event. The submit button takes the user to the map and shows the location of the venue, the details as shown in Figure 12.

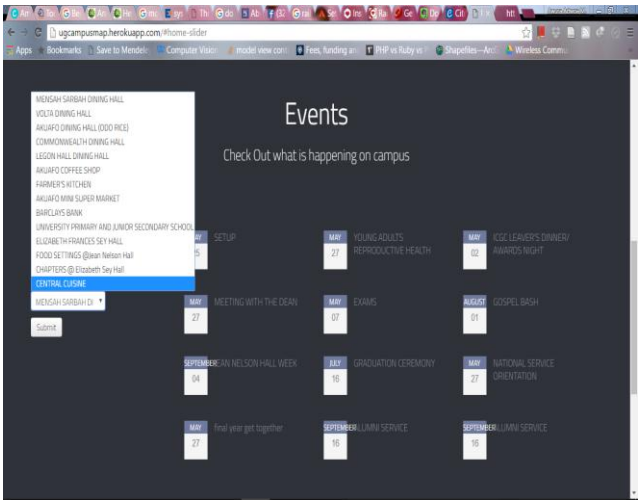


Figure 11: Selecting of the venue (CENTRAL CUISINE) in the selected Category (RESTAURANTS AND EATERIES)

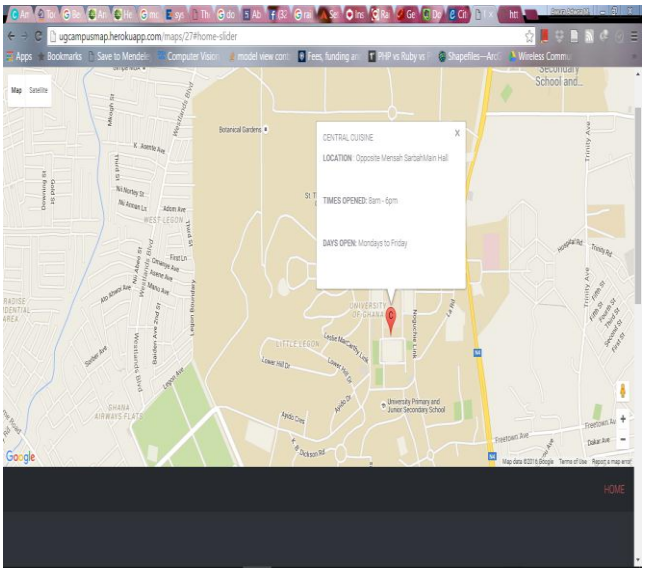


Figure 12: Display of the venue on the map

4.2 Results of the Mobile Module: Android Application

In this subsection, the results of the mobile module implemented in Android are presented as snapshots of a user's activities on an android device. Figure 13 shows the logo screen of the UGMAP mobile application whilst Figure 14 shows the notification received on the mobile device showing the recently created event (ALUMNI SERVICE).

Figure 13: The logo screen of the mobile application

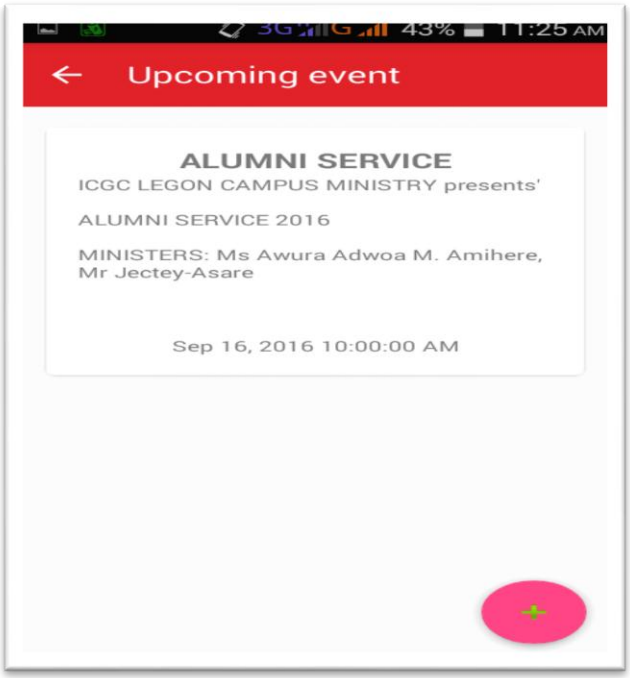


Figure 14: Notification received on the mobile device showing the recently created event (ALUMNI SERVICE).

When the notification is clicked, the details of the events are listed as shown in Figure 15.

Figure 15: The details of the events are listed when the notification is clicked.

When the icon on the bottom right of Figure 16 is clicked, the ALUMNI SERVICE is displayed in the events section as shown in Figure 17. Figure 17 displays the loading of the GPS information and calculation of distance.

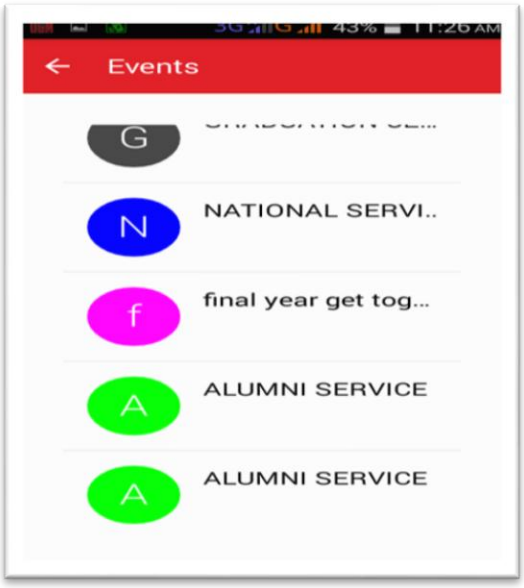


Figure 16: The ALUMNI SERVICE shown here in the events section

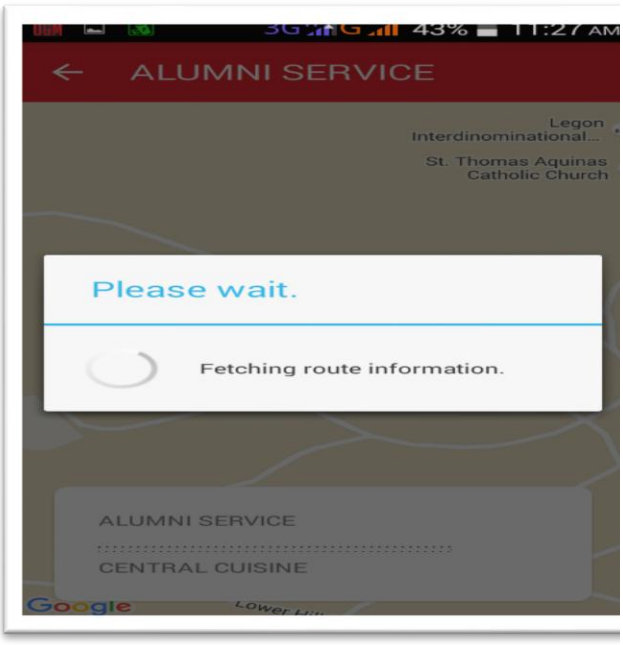


Figure 17: Loading the GPS information and calculation of distance.

The route information and distance calculations were not displayed as can be seen from Figure18. This is because the mobile device was out of the range of the University of Ghana, Legon Campus.

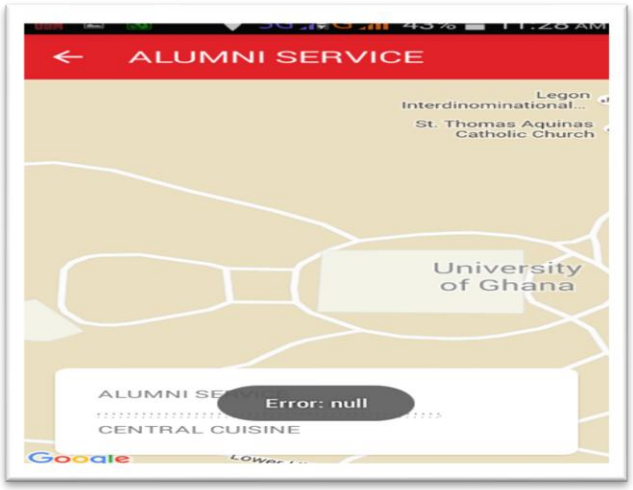


Figure 18: The route information and distance calculations were not displayed because the mobile device was out of the range of the University of Ghana Campus.

The mobile application was tested within the range of the University of Ghana campus with another event called GOSPEL BASH. As can be seen from Figure 19, the route information and distance calculations were displayed showing that the GOSPEL BASH event was scheduled to take place at the POOL SIDE within the University of Ghana,Legon Campus.

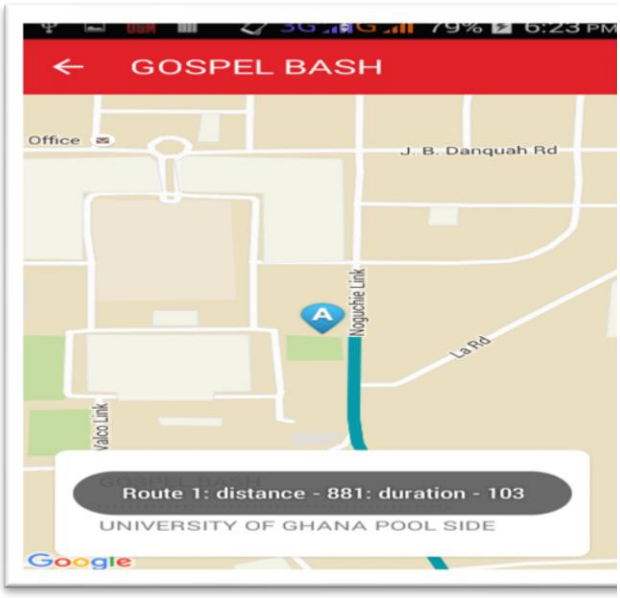


Figure 19: The display of the route in metres, and the time in seconds to get to the given location

4.3 Discussions

This system has many similarities with Google maps which include aiding the user to get directions from one place to another, showing the various locations on the map using pointers, displaying the routes on the map using polylines and showing the time it would take for the user to get to the given destination from their current location. The data from the back-end include addition of categories, venues and events that would be happening on campus. This would save the time of the users by showing routes on the hand-held mobile device of the user and displays information concerning events happening on the campus. The Rails and Android programs that implement the mobile module run on a GPS-enabled device (such as a phone or laptop). UGMAP does not only give the user directions to their destinations, but also sends notifications to their mobile device or phone, prompting them of upcoming events such as lecture times, exams, public events and career fairs and many others. This would save time hitherto wasted browsing a notice board for an event and searching for location of the event.

The mobile application also routes the different paths the user can take to reach their given destination. The web application was developed using Ruby on Rails and the mobile application was developed in Android (using mainly Java).

The differences between UGMAPS and Google Maps are that, Google maps enable the user to customize how they want the application to operate for them. However, with UGMAPS, there is a central administrator that adds events to the system. This allows all users of UGMAP to receive notifications on their devices concerning events that would be happening on the campus.

5.0 CONCLUSION AND FUTURE WORK

The research described in this paper was implemented to provide notification and alert with geoinformation as the backbone for users within the University of Ghana Campus community. The research targeted the students, visitors and staff of the University, especially those who are new on the campus where rapid dissemination of information is always taking place. The system consists of two modules namely, *Web Application*, and *Mobile Application*. All modules were successfully implemented. In conclusion, a web and mobile application to give directions to venues and notifications about events that would be happening on the University of Ghana campus was developed. A platform was designed for posting information using Ruby on Rails, with an administrator back-end for adding categories, venues and events to the app manually which is updated on the web application when refreshed. This would help to save the time of the users, reduce the stress involved in finding places on campus and increase productivity.

A mobile app was also developed in Android with Google cloud messaging to enable the user to receive updates on incoming events and providing navigational route information to aid

users get to venues of academic and social events. UGMAP further provided the distances and time required to reach those venues.

Future work will consider the using UGMAPS in a semantic web context. UGMAP can be redesigned to send news feeds to the users. The research can be scaled up to support the emergency services offered by the police, fire service and disaster management. Further due to property addressing challenges in developing countries, this research can be modified to provide alternative solution to navigation in such countries. In addition to forms and login, other features that would allow the user to enter particular details that the system administrator can use to send specific messages to the users based on the information they need can also be integrated.

REFERENCES

[1] H. Collingbourne, The Little Book of Ruby, 2nd ed. Dark Neon Ltd., 2008, p. 5.

[2] D. Thomas, C. Fowler and A. Hunt, Programming Ruby 1.9 The Pragmatic Programmers' Guide. Raleigh, North Carolina Dallas, Texas: The Pragmatic Programmers, LLC, 2009, p. 35.

[3] D. Kehoe, "What is Ruby on Rails? • RailsApps", Railsapps.github.io, 2016. [Online]. Available: http://railsapps.github.io/what-is-ruby-rails.html. [Accessed: 26- May- 2016].

[4] S. Ruby, D. Thomas and D. Hansson, Agile Web Development with Rails, 4th ed. Dallas, Texas •Raleigh, North Carolina: Pragmatic Programmers, LLC, 2011, p. xviii, xix.

[5] M. E. Maximilien, "Web Services on Rails: Using Ruby and Rails for Web Services Development and Mashups", 2006.

[6]"Web 2.0", Wikipedia, 2016. [Online]. Available: https://en.wikipedia.org/wiki/Web_2.0. [Accessed: 01- Jun- 2016].

[7] S. Lin, Y. Zhou, R. Wang and J. Zhang, "Google Map Application Development in Android Platform", AMM, vol. 513-517, pp. 466-469, 2014.

[8]"About | Heroku", Heroku.com, 2016. [Online]. Available: https://www.heroku.com/about. [Accessed: 30- May- 2016].

[9] A. MacDonald, D. Russell and B. Atchison, "Model-driven Development within a Legacy System: An industry experience report", in Proceedings of the 2005 Australian Software Engineering Conference (ASWEC'05, School of Information Technology and Electrical Engineering, The University of Queensland, 2005.

[10] Birimicombe A. and Li C. Location Based Services and Geo-information Engineering, Wiley & Sons, UK. (2009).

[11] Karujal R. Mobile Computing, Oxford Union Press India 2nd Ed. (2012)

[12] Papazoglou M.P., Web services Principles and Technology, Pearson, England.(2009)

[13] Andre L. et al. Beginning Google Maps Applications with Rails and Ajax: From Novice to Professional, Apress, USA .(2009).

[14] M. Sharma, P. Saha, S. Sarcar, S. Ghosh and D. Samanta, "Accessing Dynamic Web Page in Users Language", in IEEE Students' Technology Symposium, IIT Kharagpur, 2011.

[15] L. Stella, S. Jarzabek and B. Wadhwa, "A Comparative Study of Maintainability of Web Applications on J2EE, .NET and Ruby on Rails", no. 978-1-4244-2790-108, pp. 93-99, 2008.

[15] A. Faro and C. Spampinato, "Implementing ITS 3.0 applications by integrating Ruby on Rails, Sesame and Protegè technologies", in 2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems, 2011.

[16] A. Pejiü, B. Pejiü and S. Pletl, "An Expert System for Tourists Using Google Maps API", Subotica Tech, Department of Informatics, Subotica, no. 978-1-4244-5349-809, p. 317-322, 2009.

[17] Francis Rousseaux, Kevin Lhoste, "Rapid Software Prototyping Using Ajax and Google Map API," achi, pp.317-323, 2009 Second International Conferences on Advances in Computer-Human Interactions, 2009

[18] Martin C. Brown, "Hacking Google Maps and Google Earth", ISBN: 978-0-471-79009-9, 2006

[19]"SDLC - Agile Model", www.tutorialspoint.com, 2016. [Online]. Available: http://www.tutorialspoint.com/sdlc/sdlc_agile_model.htm. [Accessed: 30- May- 2016].

[20] R. Tamada, "Android Push Notifications using Google Cloud Messaging (GCM), PHP and MySQL", androidhive, 2012. [Online]. Available: http://www.androidhive.info/2012/10/android-push-notifications-using-google-cloud-messaging-gcm-php-and-mysql/. [Accessed: 03- Jun- 2016].

[21] S. Ruby, D. Thomas and D. Heinemeier Hansson, Agile Web Development with Rails, 4th ed. Dallas, Texas: The Pragmatic Programmers, 2011, pp. 32-34.

[22]"puma/puma", GitHub, 2016. [Online]. Available: https://github.com/puma/puma. [Accessed: 08- Jun- 2016].

[23]"Intermediate Rails: Understanding Models, Views and Controllers", BetterExplained, 2016. [Online]. Available: https://betterexplained.com/articles/intermediate-rails-understanding-models-views-and-controllers/. [Accessed: 08- Jun- 2016].

[24] "Ruby Gems", Wikipedia, 2016. [Online]. Available: https://en.wikipedia.org/wiki/RubyGems. [Accessed: 03- Jun- 2016].

[25] I. Messaging, "Integrate Google Cloud Messaging | Android Developers",Developer.android.com, 2016. [Online]. Available: https://developer.android.com/distribute/engage/gcm.html. [Accessed: 03- Jun- 2016].

[26]"The Heroku Platform as a Service & Data Services | Heroku", Heroku.com, 2016. [Online]. Available: https://www.heroku.com/platform. [Accessed: 03- Jun- 2016].

[27]"Meet Android Studio | Android Studio", Developer.android.com, 2016. [Online]. Available: https://developer.android.com/studio/intro/index.html#project_structure. [Accessed: 03- Jun- 2016].

[28] S. Handby, "Android phone GPS tips", CNET, 2012. [Online]. Available: http://www.cnet.com/how-to/android-phone-gps-tips/. [Accessed: 03- Jun- 2016].

nigeria computer society
ncs
www.ncs.org.ng

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

A PROMETHEE BASED EVALUATION OF SOFTWARE DEFECT PREDICTORS

R. G. Jimoh[1], A. O. Balogun[2], A. O. Bajeh[3] and S. Ajayi[4]

[1,2,3,4]Department of Computer Science, University of Ilorin, Ilorin
[1]jimoh_rasheed@unilorin.edu.ng; [2] balogun.ao1@unilorin.edu.ng;
[3]bajehamos@unilorin.edu.ng

ABSTRACT
A software defect is an error, flaw, mistake, or fault in a computer program or system that produces incorrect or unexpected results and the process of locating defective modules in software is software defect prediction. Defect prediction in software improves quality and testing efficiency by constructing predictive stand-alone classifier models or by the use of ensembles methods to identify fault-prone modules. Selection of the appropriate set of single classifier models or ensemble methods for the software defect prediction over the years has shown inconsistent results. In previous analysis, inconsistencies exist and the performance of learning algorithms varies using different performance measures. Therefore, there is need for more research in this field to evaluate the performance of single classifiers and ensemble algorithms in software defect prediction. This study assesses the quality of the ensemble methods alongside single classifier models in the software defect prediction using Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE), a multi criteria decision making (MCDM) approach. Using PROMETHEE, the performance of some popular ensemble methods based on 11 performance metrics over 10 public-domain software defect datasets from the NASA Metric Data Program (MDP) repository was evaluated. Noise is removed from the dataset by performing attribute selection. The classifiers and ensemble methods are applied on each dataset; Adaboostgave the best results. Boosted PART comes first followed by Naïve Bayes and then Bagged PART as the best models for mining of datasets.

Keywords: Ensemble; Classification; Software Defect Prediction; PROMETHEE; MCDM.

1.0 INTRODUCTION

A software defect is an error, flaw, mistake, or fault in a computer program or system that produces incorrect or unexpected results (Puneet&Pallavi, 2013).Human developer are responsible for software development activities in the software life cycle, however software developer and analysts are prone to error. It is therefore impossible to produce the software without errors or defects even though it's imperative to predict and fix the defects as many as possible before the product is delivered for use (Naheed&Shazia, 2011). Software defect prediction is the process of locating defective modules in software. It facilitates the improvement of the software quality and testing efficiency by constructing predictive models from code attributes to enable a timely identification of fault-prone modules (Puneet&Pallavi, 2013).Software defect prediction brings two fields of computer science together, namely; Software Engineering and Data Mining.

Software engineering is an engineering discipline concerned with all aspects of software production and development from the beginning stages of system specification through to maintaining the system after it has gone into use(Williams, 2004). Data Mining is the process of exploring meaningful information from data with different perspectives; it is a powerful tool that came to be in the middle of 1990's with the aim of evaluating and extracting valuable information from huge datasets (Sonali&Divya, 2014). Studies have shown that the results of Data Mining approaches are helping decision makers make more informed decisions. There are many data mining algorithms of various categories such as classification, regression, association and clustering are used in software quality analysis. However, in this research we used classification and ensemble (meta) approach for the prediction of defective software.Researchers in a variety of fields have created a large number of classification algorithms, such as decision tree, neural networks, Bayesian network, linear logistic regression, Naïve Bayes, and K-nearest-neighbor. How to select the most appropriatealgorithms for a given task is an important issue (Peng, Guoxun&Honggang, 2010).There is need for researches in the field to enable software developers make better informed decisions in the selection of stand-alone classifiers algorithms or ensemble method in defect prediction.This study applies attribute reduction to 10 National Aeronautics and Space Administration (NASA) datasets from PROMISE repository.4 classification algorithms and 4 ensemble learning techniques are implemented to predict defects in the datasets. Also, a multicriteria decision making technique to rank classification and ensemble algorithm is used.

The rest of this paper is organized as follows: Sections 2 defines the algorithms and multi-criteria decision-making technique used in this study;Section 3presents a review of related literature. Section 4 describesmethod usedin carrying out the research, including the design, and research method and process, also thealgorithms and MCDMmethod used respectively. Section 4 presents details of the experimental study and analyzes the results; Section 5 concludes the findings of this paper and recommendations based on the results of the study

2.1 CLASSIFIERS, ENSEMBLES AND MCDM

Software defection prediction can be modeled as a binary classification problem, where software modules are classified as either defect-prone or non-defect prone modules (Soumya& Simi, 2016; Ming, Liu,& Zhang,, 2011; Sonali&Divya, 2014). Thus, it is appropriate to use classification models to predict whether a module contains defect or not. A classification model can be trained on the data for the purpose of predicting the defect-proneness of software modules (Ming et al., 2011).

A lot of classifiers have been used for the task of predicting defect in software defects which includes Support Vector Machine (SVM), Sequential Minimal Optimization (SMO), C4.5, J48, Multi-layer Perceptron (MLP), K-nearest neighbor (KNN), PART, Naïve Bayes, Decision Stump, Decision Tree, etc.

This study is only considering Naïve Bayes, PART, SVM and Decision Stump for the base classifiers which were selected based on their characteristics for heterogeneity. As for ensemble methods, Bagging, Boosting, Voting and Stacking will be used and PROMETHEE will be used as the MCDM.

I. Naïve Bayes: The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given dataset. It is also a statistical method for classification. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable (Tina &Sherekar, 2013).
II. Partial Decision List (PART): This uses class for generating a PART decision list by adopting the divide-and-conquer technique. It builds a partial decision tree in each iteration and makes the "best" leaf into a rule for the classification (Eibe& Ian, 1998).
III. Support Vector Machines (SVM): It was developed to solve the two-classification problem but later it was formulated and extended to solve multiclass problem (Sonali&Divya, 2014). SVM divides the data samples of two classes by determining a hyper-plane in original input space that maximizes the separation between them.
IV. Decision Stump: A decision stump is a machine learning model consisting of a one-level decision tree (Iba& Langley, 1992). Class for building and using a decision stump. Usually used in conjunction with a boosting algorithm. Does regression (based on mean-squared error) or classification (based on entropy).
V. Boosting: It is an ensemble learning technique which refers to a family of algorithms that are able to convert weak learners to strong learners (Zhi-Hua,2012). Instinctively, a frail learner is quite recently marginally superior to anything irregular figure, while a solid learner is near flawless execution (Zhi-Hua,2012). In boosting, however, weights of training instances change in each iteration to force learning algorithms to emphasize on instances that were predicted incorrectly (Dietterich, 2000). There are many variants of boosting technique such as Adaboost, LogitBoost, etc. but this study focuses on Adaboost. Adaboost is the abbreviation for Adaptive boosting algorithm because it adapts to the errors returned by classifiers from previous iterations (Freund &Schapire, 1996).
VI. Bagging:The name Bagging came from the abbreviation of Bootstrap Aggregating (Zhi-Hua, 2012). As the name implies, the two ingredients of bagging are bootstrap and aggregation. Bagging adopts the bootstrap distribution for generating different base learners that (that is, it applies bootstrap sampling to obtain the data subsets for training the base classifiers) (Zhi-Hua, 2012). Bagging combines multiple outputs of a learning algorithm by taking a plurality vote to get an aggregated single prediction (Breiman, 1996). The multiple outputs of a learning algorithm are generated by randomly sampling with replacement of the original training dataset and applying the predictor to the sample (Penget al, 2011).
VII. Stacking: This is a meta-learning technique. Meta-learning means learning from the classifiers produced by the creators and from the classifications of these classifiers on training data (Lior, 2010). Stacking is a general procedure where a learner is trained to combine the individual learners. Here, the individual learners are called the first-level learners, while the combiner is called the second-level learner, or meta-learner (Zhi-Hua, 2012). Stacking is a technique for achieving the highest generalization accuracy (Wolpert, 1992). Unlike bagging and boosting, stacking can be applied to combine

different types of learning algorithms (Penget al, 2011). In stacking each base learner, also called "level 0" model, produces a class value for each instance then the predictions of level-0 models are then fed into the next level model which combines them to form a final prediction. The classifiers stacked in this study are Naïve Bayes, PART, SVM and Decision Stump.

VIII. Voting: This is an ensemble algorithm that works on nominal output (Zhi-Hua, 2012). Like stacking, voting also combines a series of classifiers together to perform the classification task. In voting, the combined classifiers vote for a class label. There are several types of voting techniques as pointed out by Zhi-Hua(2012), they include:

A. Majority voting: In this type of voting, every classifier votes for one class label, and the final output class label is the one that receives more than half of the votes, if none of the class labels receives more than half of the votes, a rejection option will be generated and then the combined classifiers will make no prediction.

B. Plurality Voting: Unlike majority voting that requires the winner class label to take at least half of the votes, plurality voting takes the class label that receives the highest number of votes from the classifiers.

C. Weighted Voting: In this of voting, all classifiers are compared, and the classifier with the strongest weight is used for the prediction of the class label.

Soft Voting: This type of voting is used to combine classifiers that produce probability outputs.

2.2 PROMETHEE

Decision-making problems usually imply the selection of the best compromise solution. Besides the real criteria values by which a decision is made, the selection of the best solution also depends on the decision maker, that is, on his individual preferences. In order to simplify the decision-making process, many mathematical methods have been suggested. Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE) represents one of the most frequently used methods of multi-criteria decision-making process (Vojislav, Zoran&Dragoslav, 2011).

In PROMETHEE, firstly the outranking method compares pairs of alternatives on each criterion, and thenit induces the preferential function to describe the preference difference between pairs of alternative on each criterion. Thus, preference functions about the numerical difference between pairs of alternatives are built to describethe preference deference from the point of the decision maker's view. These functions' value ranges from 0 to 1. The bigger the function's value, the larger the difference of the preference. When the value is zero, there is no preferential difference between pair of alternative but when the value is one, one of the alternatives is strictly outranking the other (Zhaoxu& Han, 2010).

3.0 RELATED WORKS

Software defect prediction can be tackled from a data mining point of view. The use of data mining techniques together with machine learning algorithms will help identify modules in the software that contains defects (Naheed&Shazia, 2011). Many studies have proposed the use of classification algorithms for predicting defects in software. Classification is one of the techniques of data mining which determines amongst a predefined category, which category a particular object belongs.

Issam, Mohammed and Lahouari(2014), demonstrated the positive effects of combining feature selection and ensemble learning on the performance of defect classification. Along with efficient feature selection,a new two-variant (with and without feature selection) ensemble learning algorithm was proposed toprovide robustness to both data imbalance and feature redundancy. The study carefully combines

selected ensemble learning models with efficient feature selection toaddress these issues and mitigate their effects on the defect classification performance. Theresults of forward selection showed that only few features contribute to high area under the receiver operating curve (AUC). On the tested datasets, greedy forward selection (GFS) method outperformedother feature selection techniques such as Pearson's correlation. However, ensemble learners like random forests and average probabilityensemble (APE), are not as affected by poor features as in the case of weighted support vectormachines (W-SVMs). Moreover, the APE model combined with greedy forward selection (enhancedAPE) achieved AUC values of approximately 1.0 for the NASA datasets: PC2, PC4, and MC1.This shows that features of a software dataset must be carefully selected for accurateclassification of defective components.

Penget al. (2011),assessed the quality of ensemble methods in software defect prediction using Analytic Hierarchy Process (AHP) which is a multicriteriadecision-making approach that prioritizes decision alternatives based on pairwise comparisons.Through the application of AHP, there study experimentallycompared theperformance of several popular ensemble methods using 13 different performance metricsover 10 public-domain software defect datasets from the NASA Metrics Data Program(MDP) repository. The results indicate that ensemble methods can improve the classificationresults of software defect prediction in general and AdaBoost gave the bestresults. In addition, tree and rule based classifiers perform better in software defectprediction than other types of classifiers included in the experiment. In terms of singleclassifier, K-nearest-neighbor, C4.5, and Naïve Bayes tree ranked higher than otherclassifiers.

In another study, Penget al. (2010) used a set of MCDM methods to rank classification algorithms, with empirical results based on the software defect detection datasets. Since the preferences of the decision maker (DM) play an important role in algorithm evaluation and selection, the study involved the DM during the ranking procedure by assigning user weights to the performance measures. Four MCDM methods are examined using 38 classification algorithms and 13 evaluation criteria over 10 public-domain software defect datasets. The results indicate that the boosting of CART and the boosting of C4.5 decision tree are ranked as the most appropriate algorithms for software defect datasets. Though the MCDM methods provide some conflicting results for the selected software defect datasets, they agree on most top-ranked classification algorithms.

4.0 METHODOLOGY

This research is aimed at evaluating the performance of ensembles and standalone classifiers for prediction of defects in software via the mining of datasets from a software system. Feature selection in form of attribute reduction is carried out on the datasets in order to eliminate irrelevant or noisy attributes. The feature selection technique used on the datasets is the best first search method and the classifier subset evaluation algorithm for attribute reduction. The classifiers and ensemble methods used are Naïve Bayes, PART, SVM, Decision stump, Boosting, Bagging, Stacking and Voting. The multi-criteria decision-making technique, PROMETHEE is applied to the results to generate priorities and ranking for models used.

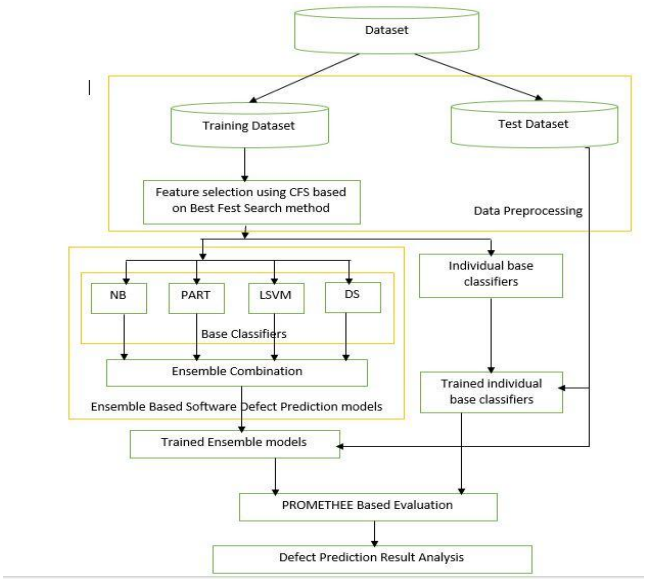
4.1 Source and Nature of Data Used

The datasets used in this study are 10 public-domain software defect datasets provided by the National Aeronautics Space Administration(NASA) Facility Metrics Data Program (MDP) repository. The datasets used in this research work are namely; CM1, JM1, KC1, MC1, MC2, MW1, PC1, PC2, PC3 and PC4 respectively. The brief descriptions of these MDP datasets are provided below.

Data Set	System	Language	Total Loc
CM1	Spacecraft Instrument	C	17K
JM1	Storage management for ground data	JAVA	8K and 25K
KC1	Storage management for ground data	C++	*
MW1	Zero -gravity experiment related to combustion	C	8K
PC1,2	Flight Software for Earth orbiting Software	C	26K
PC3,4	Flight Software for Earth orbiting Software	C	30-36K

Table 1: NASA MDP Data Sets

Figure 1: Experimental Architecture



4.2 Feature Selection

Feature selection is applied as a pre-processing method to de-noise the dataset since literatures have posited that removal of noise from dataset has positive effect of the classification (Ameen, Balogun, Usman&Fashoto, 2016). For this study, CfsSubsetEval is used for feature selection and BestFirst algorithm is used as the search technique.

I. BestFirst: It Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility (Witten & Frank, 2005).
II. CfsSubsetEval:CfsSubsetEval evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them (Witten & Frank, 2005).

4.3 Performance Criteria/Parameter

There are a number of ways to evaluate the performance of a classifier model or ensemble. Commonly used performance measures in software defect classification are accuracy, precision, recall, F-measure, AUC, and mean absolute error (Peng*et al*, 2010; Peng*et al*, 2011). The performance metrics for this study will also be based on the aforementioned metric

The objective of this research venture was to evaluate the performance of the base classifiers and ensembles methods with reduced attribute datasets, implemented with the use of PROMETHEE. This analysis was performed by executing Weka API library utilizing the Eclipse IDE. This sectionpresents the analysis result of the study. In this study, applying feature selection to the datasets increases the performance of learning models except in few cases where the situation is otherwise.The average of the 10 NASA attribute reduced datasets is represented in Table 2.

Table 2: The Average results of the classifiers and ensemble methods on the datasets

PERFORMANCE MEASUREMENT	NB	PART	LSVM	DS	Boosted NB	Boosted PART	Boosted LSVM	Boosted DS	Bagged NB	Bagged PART	Bagged LSVM	Bagged DS	Stacking	Voting
Correctly classified	86.27%	87.08%	87.35%	87.40%	87.21%	87.30%	87.34%	87.23%	87.08%	87.13%	87.18%	87.07%	87.11%	87.18%
Incorrectly classified	13.72%	12.92%	12.65%	12.60%	12.79%	12.70%	12.66%	12.77%	12.92%	12.87%	12.82%	12.93%	12.89%	12.82%
Kappa statistics	25.74%	21.19%	19.28%	15.93%	18.03%	20.34%	20.01%	20.78%	21.20%	21.32%	20.62%	21.01%	19.91%	19.68%
Mean absolute error	14.36%	16.08%	14.72%	15.68%	16.25%	15.83%	15.39%	15.79%	15.70%	15.76%	15.44%	15.41%	15.80%	15.79%
Root mean square error	33.69%	31.25%	31.42%	30.71%	30.84%	30.68%	30.82%	30.89%	31.19%	30.84%	30.86%	31.09%	30.98%	30.73%
FP rate	58.29%	67.72%	71.22%	74.98%	72.08%	70.25%	70.79%	69.50%	68.21%	68.45%	69.44%	68.48%	69.94%	70.41%
TP rate	86.23%	87.05%	87.29%	87.35%	87.16%	87.25%	87.29%	87.18%	87.02%	87.08%	87.15%	87.03%	87.05%	87.13%
Precision	85.99%	85.60%	85.41%	85.11%	84.87%	85.11%	85.04%	85.05%	85.12%	85.05%	84.96%	84.99%	84.88%	84.89%
Recall	86.23%	87.05%	87.29%	87.35%	87.16%	87.25%	87.29%	87.18%	87.02%	87.08%	87.15%	87.03%	87.05%	87.13%
F- measure	85.65%	85.29%	84.74%	84.21%	84.60%	85.02%	84.94%	85.05%	85.11%	85.21%	85.08%	85.13%	84.97%	84.94%
ROC Area	76.25%	70.57%	61.91%	63.89%	65.16%	66.14%	64.45%	65.03%	65.88%	66.82%	64.97%	65.62%	65.43%	65.99%
PRC Area	87.01%	85.45%	82.74%	81.13%	83.63%	84.05%	83.62%	83.82%	84.06%	84.31%	83.78%	83.96%	83.87%	84.00%

Table 3: Best and worst classifier performances.

Criteria/Performance	Best performer	Worst performer
Accuracy	Decision Stump	Naïve Bayes
Kappa	Naïve Bayes	Decision Stump
Mean Absolute Error	Naïve Bayes	Boosted Naïve Bayes
Relative Mean Square Error	Boosted PART	Naïve Bayes
FP Rate	Decision Stump	Naïve Bayes
TP rate	Decision Stump	Naïve Bayes
Precision	Naïve Bayes	Boosted Naïve Bayes
Recall	Decision Stump	Naïve Bayes
F-measure	Naïve Bayes	Decision Stump
AUC	Naïve Bayes	LIBSVM
PRC Area	Naïve Bayes	Decision Stump

The information presented in Table 3 is derived from Table 2

From the analysis, it is evident that based on the performance metrics used for this study, no particular classifier or ensemble method gave the best result across the performance metrics used in this study. It is not only the accuracy that should be focused on when performing classification problem. The application of MCDM to the result will put into consideration all metrics considered, prioritize and rank the respective methods based on their metric values.

With this result, it can be stated that stand alone and 8.

classifiers have more extreme values positively and negatively with Naïve Bayes leading the charge of having both the best and worst values in most of the performance values. However, we can also notice that ensemble methods mostly lack extreme values and appear to tend towards the middle in terms of their criteria values. Having applied PROMETHEE, the following results are produced. The ranking of classifiers based on their performances are presented in Table 4,5,6,7

Table 4. Ranking of Boosted Classifiers (Group 1)

S/N	Algorithms	Priorities
1	Boosted PART	0.2556
2	Boosted LIBSVM	0.0781
3	Boosted Decision Stump	0.0059
4	Boosted Naïve Bayes	-0.2911

The priorities of the classifiers in the table above in the boosted group, Boosted PART is the top ranked.

Table 5. Ranking of Bagged Classifiers (Group 2)

	Algorithms	Priorities
1	Bagged PART	0.2142
2	Bagged Naïve Bayes	-0.0402
3	Bagged LIBSVM	-0.0876
4	Bagged Decision Stump	-0.1101

In the table above Bagged PART is ranked best classifier.

Table 6. Ranking of Stacking. Voting, and individual Classifiers (Group 3)

	Algorithms	**Priorities**
1	Naïve Bayes	0.2308
2	PART	0.0805
3	Decision Stump	0.0166
4	Library Support Vector Machine	0.0107
5	Voting	-0.0497
6	Stacking	-0.3136

In the priority table above Naïve Bayes is the top ranked classifier.

Table 7. Ranking of best models from each group (Group 1 to Group 3)

	Algorithms	**Priorities**
1	Boosted PART	0.2556
2	Naïve Bayes	0.2308
3	Bagged PART	0.2142

This table shows the overall best ranked performers in each of the categories.

Table 8. Ranking of all classifiers

	Algorithms	**Priority**
1	Boosted PART	0.2556
2	Naive Bayes	0.2308
3	Bagged PART	0.2142
4	PART	0.0805
5	Boosted LIBSVM	0.0781
6	Decision Stump	0.0166
7	LIBSVM	0.0107
8	Boosted DS	0.0059
9	Bagged NB	-0.0402
10	Voting	-0.0497
11	Bagged LIBSVM	-0.0876
12	Bagged DS	-0.1101
13	Boosted NB	-0.2911
14	Stacking	-0.3136

Table 8 gives the results of all the classifiers/ensembles according to their priorities which is based on the overall performance. It can be observed that among the models Boosted PART tops them and it is an ensemble method. This shows that some ensemble methods can perform better than their stand-alone counterpart classifiers (PART) for software defect prediction. However, it is also important to note in this study that some ensembles also performed worse than single classifiers with staking at the bottom of the rank in Table 9. This beckons that the software engineers/developer and other concerned parties are to be well informed before choosing ensemble methods for defect prediction.

6.0. CONCLUSION

In this study, the efficiency of several classification and ensemble learning algorithms for software defect prediction were evaluated using the PROMETHEE (Preference ranking organization method for enrichment evaluation) multi criteria decision making technique for ranking of all algorithms in respect to performance. The classification algorithms used in this research work includes Decision Stump, Partial decision list (PART), Library Support Vector Machine (LIBSVM), and Naïve Bayes. This study employed the use of feature selection (attribute selection) to remove noisy data because it improves the classification results. The ensemble learning algorithms used includes Boosting, Bagging, Stacking, Voting. The results show that the use of ensembles can most times perform better than single classifiers in the task of software defect prediction.

It has been observed in this study that the use of ensembles can perform better than single classifiers in the task of software defect prediction, however this project limits the number of classifiers to 4 and also the number of ensembles to 4, hence it is recommended that future works should look into using more sets of ensembles/classifiers for the task of software defect prediction as this will go a long way in proving the use and selection of ensembles to perform classifications in the task of software defect prediction.

REFERENCES

Ameen, A. O., Balogun, A. O., Usman, G. &Fashoto, S.G. (2016): Heterogenous Ensemble Methods Based On Filter Feature Selection. *Computing, Information System Development Informatics & Allied Research Journals.*Vol 7 No 4.Pp 63-78.

Bauer, E.,&Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants, *Machine Learning* 36(1/2) 105–139.Boehm B., (1987). *A Spiral Model of Software Development and Enhancement*, Computer, 20(9), 61-72.

Breiman, L. (1994). Bagging Predictors, *Technical Report, Department of Statistics, University of California, Berkeley. USA.*

Dietterich, T. G.(2009). Ensemble methods in machine learning.*First International Workshop on Multiple Classifier Systems* 1–15.

Eibe, F. & Ian, H. W. (1998).Generating accurate rule sets without global optimization.*In Proc 15th International Conference on Machine Learning*, Madison, Wisconsin, pages 144-151. Department of Computer Science, University of Waikato; New Zealand.

Freund, Y. &Schapire, R. (1996).Experiments with a new Boosting Algorithm.*In Proceedings of the Thirteenth International Conference on Machine Learning*, 148-156.

Iba, W. & Langley, P. (1992).Induction of One-Level Decision Trees, *in ML92. Proceedings of the Ninth International Conference on Machine Learning*, Aberdeen, Scotland, 1–3 July 1992, San Francisco, CA: Morgan Kaufmann, pp.

233–240.

Issam, H. L., Mohammad, A.&Lahouari, G. (2014).Software defect prediction using ensemble learning on selected features.*Information and Software Technology.Information and Computer Science Department, King Fahd University of Petroleum and Minerals*, Dhahran 31261, Saudi Arabia.

Lior, R. (2010). Ensemble based classifiers.33, 1-39.doi 10.1007/s10462-009-9124-7.

Naheed, A., &Shazia, U. (2011). Analysis of Data Mining Based Software Defect Prediction Techniques.*Global Journal of Computer Science and Technology*. 11(16), 1-2.

Puneet J. K. &Pallavi, (2013), Data Mining Techniques for Software Defect Prediction, *International Journal of Software and Web Sciences (IJSWS)* 3(1), pp. 54-57

SeminPaksoy& Mehmet F T. (2017).*Investigating Banks' Performance for Turkey*: An Application of Promethee Method.C.Ü. İktisadiveİdariBilimlerDergisi, Cilt 18, Sayı 1,

SonaliA.,&Divya T. (2014). *A Feature Selection Based Model for Software Defect Prediction.*International Journal of Advanced Science and Technology.*65*(4)pp.39-58.http://dx.doi.org/10.14257/ijast.201 4.65.04

Soumya Joseph, & Simi Margaret, G.P.(2016). *Software Defect Prediction Using EnhancedMachine Learning Technique*.International Journal of Innovative research in Computer and Communication Engineering, *4*(6), 1-2.

Tina R. &Sherekar S. (2013).*Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*.International Journal of Computer Science and Applications Vol. 6, p 2.

Miao,L.,Liu,M.,Zhang, D., (2012). Cost Sensitive Feature Selectionwithapplication in software defect prediction. *International Conference on Pattern Recognition,*Tsukuba,Japan,p.967-970.

Vojislav T., Zoran M., Dragoslav J. (2011) PROMETHEE Method Implementation with Mutli-Criteria Decision. *Mechanical Engineering Vol. 9, No 2, p193 – 202.*

Williams, L. (2004). introduction to software Engineering. *Software Engineering*. Retrieved http://agile.csc.ncsu.edu

Witten I. H., & Frank E.(2005) *Data Mining: Practical Machine Learning Tools and Techniques*.

Wolpert, D. H. (1992) Stacked generalization.Neural Networks, vol 5.Pergamon Press, Oxford 241–259.

Peng, Y., Gang, K., Yong, S., Guoxun, W., &Wenshual, W. (2011). Ensemble of Software Defect Predictors: An AHP-Based Evaluation Method, *International Journal of Information Technology & Decision-Making* Vol. 10, No. 1 (2011) 187–206.

Peng, Y.,Guoxun W. &Honggang W. (2010).User preferences based software defect detection algorithms selection using MCDM,*Information Sciences. School of Management and Economics, University of Electronic Science and Technology of China, Chengdu 610054, China.*

Zhaoxu, S.& Han, M. (2010).*Multi-Criteria decision making based on PROMETHEE method*. International Conference on Computing, Control and Industrial Engineering.Pg.1. Beijing, China.

Zhi-Hua, Z. (2012). *Ensemble Algorithm: Foundations and Algorithm.* Microsoft research Limited. Cambridge,

nigeria computer society
www.ncs.org.ng

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June, 2018

Time Series Prediction using Artificial Neural network

Anifat Olawoyin

University of Winnipeg, Winnipeg R3B2E9, CANADA
anifatola@yahoo.com

ABSTRACT
This study investigated the use of Multilayer Perceptron (MLP) artificial neural network and Autoregressive Integrated Moving Average (ARIMA) models for time series prediction. The models are evaluated using two statistical performance evaluation measures, Root Mean Squared Error (RMSE) and coefficient of determination (R^2). Four different multilayer perceptron were developed and compared. The best MLP was then compared with ARIMA model. The experimental result shows that a 3-layer MLP architecture using tanh activation function in each of the hidden layer and linear function in the output layer has the lowest prediction error and the highest coefficient of determination among the configured multilayer perceptron neural network. Comparative analysis of performance result reveals that multilayer perceptron neural network MLP has a lower prediction error than ARIMA model.

Keywords:: Artificial Neural Network, ARIMA, Multilayer Perceptron, Time Series, Data Preprocessing.

1.0 INTRODUCTION

Transaction data are time-stamped data generated through business activities at no specific frequency. Common transaction data includes call center data, stock trading data, point-of-sales data and online retail sales data. Extracting meaningful knowledge from the transaction data requires some form of automation due to the large volume of transactions. To observe trends and variation over time of interest to the stakeholders, transaction data needs to be transformed (aggregated) to a time series data using statistics measures such count, mean, minimum, maximum, or summation. Time series data have specific frequency which may be hourly, daily, weekly, monthly, quarterly or yearly. This is a preprocessing stage for a Prediction problem. As more and more data are accumulated, a fast and efficient model is required for future prediction of time series data to help business in efficient allocation of resources, inventory planning among other business decisions.

Machine learning focus is on models that can iteratively learn from data to find hidden insights and patterns without being explicitly programmed. Learning methods can be supervised, semi-supervised or unsupervised. Artificial neural network is a form of supervised machine teaching model that mimic the biology nervous system. ANN can detect patterns and trends that are too complex for human or other statistical models such as non-linearity in time series data. Real world application of ANN includes pattern classification such as handwritten recognition, time series prediction [1], credit scoring for loan approval and machine control among others.

This paper designs a Multilayer Perceptron neural network for time series prediction and compares the result with Autoregressive Integrated Moving Average statistical time series prediction technique. The study varies the number of hidden layers, and investigates the best activation function for the dataset. In addition, this study explores the significance of preprocessing in time series prediction problem by transforming the dataset having 5 attributes and 1,098,044 instances to a dataset having 2 attributes and 366 instances using aggregation, equal frequency binning and feature selection techniques.

The rest of this paper is organized as follows: section 2 gives the background information and related works, section 3 presents the theoretical framework, section 4 describes the implementation details, experimental result and discussion are presented in section 5 and section 6 concludes the study.

2.0 Related work

Artificial Neural network (ANN) has been applied to time series forecasting problems by many researchers. The study in [1] employed Elman recurrent neural network (ERNN) with stochastic time effective function for predicting price indices of stock market. ERNN can keep memory of recent event in predicting the future. In addition, ERNN can learn, recognise and generate temporal and spatial pattern [1]. The study revealed that ERNN has the advantage of improving the forecasting precision when compared with the performance of Backpropagation Neural Network (BPNN) and a stochastic time effective neural network (STNN) architectures.

The study in [2] used Multilayer Feed Forward Neural Network (MLFFNN) and Nonlinear Autoregressive models with Exogenous Input (NARX) Neural Network to forecast exchange rate in a multivariate framework. Experimental findings indicated that MLFFNN and NARX were more efficient when compared with Generalized Autoregressive Conditional Heteroskedastic (GARCH) and Exponential Generalized Autoregressive Conditional Heteroskedastic (EGARCH).

An advance statistical technique for predicting future time series is the Autoregressive Integrated Moving Average (ARIMA) model. ARIMA model assumes time series data is stationary, that is, the data is not time dependent. To use ARIMA for time series prediction requires checking for stationarity, a common approach is to use augmented Dickey-Fuller test (ADF). ADF tests the presence of a unit root in a sample; if the p-value is greater than 0.05, null hypothesis is accepted. The alternative hypothesis is that the time series is stationary if the p-value is less than 0.05.

ARIMA and ANN are integrated in [3] to improve accuracy of time series prediction. The findings from the study indicated that integration of different models can be an effective way of improving accuracy of time series prediction.

3.0 THEORETICAL FRAMEWORK

3.1 Artificial Neural Network (ANN)

Artificial Neural network (ANN) are made up of series of interconnected nodes that simulate individual neurons like biological neural system. ANN can be used for classification, pattern recognition and forecasting problem in situation when the underlying processes are complex and characterized by chaotic features such as trends and seasonality observed in parking ticket data, nonlinear and non-stationary in stock market data, chaotic features in ozone concentration measurement and weather related problems with non-linear relationship between the input and the output.

The earlier ANN has a single layer and follows a local learning rule known as Widrow-Hoff or Perceptron Learning Rule (PLP) to update the weight associated with the network. A single layer neural network has no hidden layer; each input neuron has an associated weight and the output neuron uses a simple Linear Threshold Unit (LTU) activation function. The activation function commonly used in most artificial network configuration is the sigmoid function because of its ability to combine linear, curvilinear and constant behaviors and is smoothly differentiable.[1]
The single perceptron output is defined by:

$$t = w_0 + w_1x_1 + w_2x_2 \ldots + w_mx_m (1)$$

Where:
t = threshold, $w_1, w_2 \ldots w_m$ are the associated weight of the input attributes $x_1, x_2, \ldots \ldots x_m)$

[1] Sheela Ramanna, department of Applied Computer Science University of Winnipeg. Lecture note on neural network 1, 2 & 3.

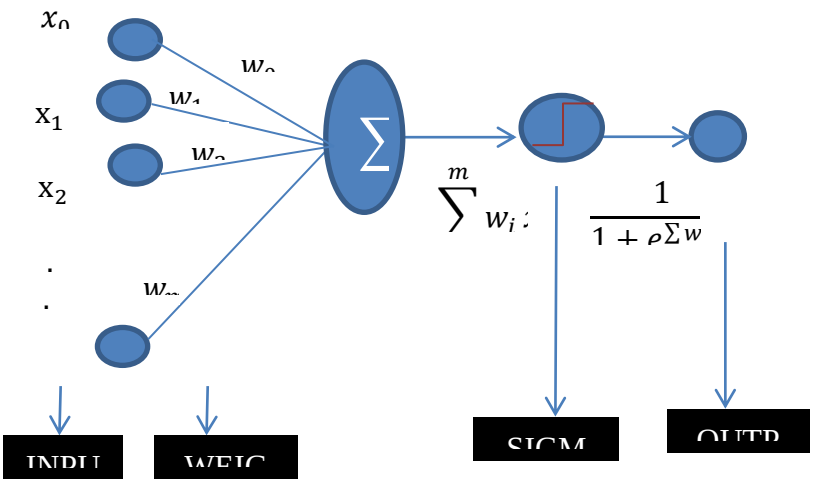


Figure 1: Perceptron Neural Network

Major drawbacks of simple neural network include:

i. Single neuron cannot solve complex tasks;
ii. It is restricted to linear calculations.
iii. Nonlinear features need to be generated by hand, an expensive operation.

The focus of this paper is the Multilayer Perceptron (MLP), a multi-layer perceptron (MLP) is a feedforward neural network consisting of a set of input, one or more hidden layers and an output layer. The layers in MLP are fully connected such that neurons between adjacent layers are fully pairwise connected while neurons within a layer share no connection.

The input represents the raw data $(x_1, x_2, x_3 \ldots \ldots x_n)$ fed into the network. The raw data and the weight are fed into the hidden layer. The input to the hidden layer is thus given as:

$$I = f(x) = \sum_{i=1}^{n}(x_i w_i) \qquad (2)$$

The hidden layer is the processing unit where the learning occurs. The hidden layer transforms the values received from the input layer using an activation function. A commonly used activation function is the sigmoid function given as

$$\sigma = {}^{1}/_{(1+e^{-x})} \qquad (3)$$

Other activation functions are:

i. tanh(x)- non-linearity activation function is a scaled sigmoid function given as:

$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1 \qquad (4)$$

tanh(x) can be expressed in form of sigmoid as:

$2\sigma(2x)-1$

ii. Rectifier Linear unit (RELU) is an activation function with a threshold of zero given as:

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

The output of the hidden layer is given as:

$$H = f(A(I)) = f(A(f(x)) = f(A(\sum_{i=1}^{n}(x_i w_i))$$

Where:
A is the activation function.
Assuming sigmoid gives

$$H = {}^{1}/_{(1+e^{-(\sum_{i=1}^{n}(x_i w_i))})} \qquad (6)$$

The output layer receives the output and the associated weight of the hidden layer neurons as input. The output Y of the output layer assuming sigmoid function is given as:

$$\boldsymbol{Y} = f\left(A(\sum_{j=1}^{m} h_j w_j)\right) \quad (7)$$

Where:

$h_j and\, w_j$ are the output and weight of individual neurons of the hidden layer.

The activation function of the output layer is commonly a linear function and depending on the task, a tanh or sigmoid function may be applicable.

A multilayer perceptron (MLP) architecture having 2 hidden layers denoted as 2-layer multilayer perceptron neural network is shown in Figure 2

The main issue with a Multilayer Perceptron neural network is weight adjustment in the hidden layer which is necessary to reduce the error at the output layer. The weight adjustment in the hidden layer is achieved using backpropagation algorithm. The back propagation takes the sequence of training samples $(\mathrm{x}_1, \mathrm{y}_1)$, $(\mathrm{x}_2, \mathrm{y}_2) \ldots\ldots, (\mathrm{x}_n y_n)$ as input and produces a sequence of weights $(\mathrm{w}_0, \mathrm{w}_1, \mathrm{w}_2 \ldots\ldots \mathrm{w}_n)$ starting from some initial weight w_0 , usually chosen at random [4]. Generally, the backpropagation rule is given as:

$$\Delta \mathrm{w} = \mathrm{w} - w_{old}$$

$$= -\eta \frac{\partial E(w)}{\partial w}$$

$$= \eta\, \partial \mathrm{x} \quad (8)$$

Where: $w - weight$

E(w) is the cost function that measures how far the current network's output is from the desired one.

∂E(w)/∂w is the partial derivative of the cost function E, that specifies the direction of the weight adjustment to reduce the error.

$\eta -$ learning rate is the step size for each iteration of the weight update equation.

The weight change for the hidden layer is given as:

$$\Delta \mathrm{w} = \eta\, \partial_j x_i \quad (9)$$

Where $\partial_j = o_j\,(1 - o_j\,) \sum w_{jz\,\partial_j}$

The weight change for the output layer is given as:

$$\Delta \mathrm{w} = \eta\, \partial_z o_j \quad (10)$$

Where $\partial_z = o_j\,(1 - o_j\,)(T - o_j\,)$

T is the target output and o_j $is\ the\ output$

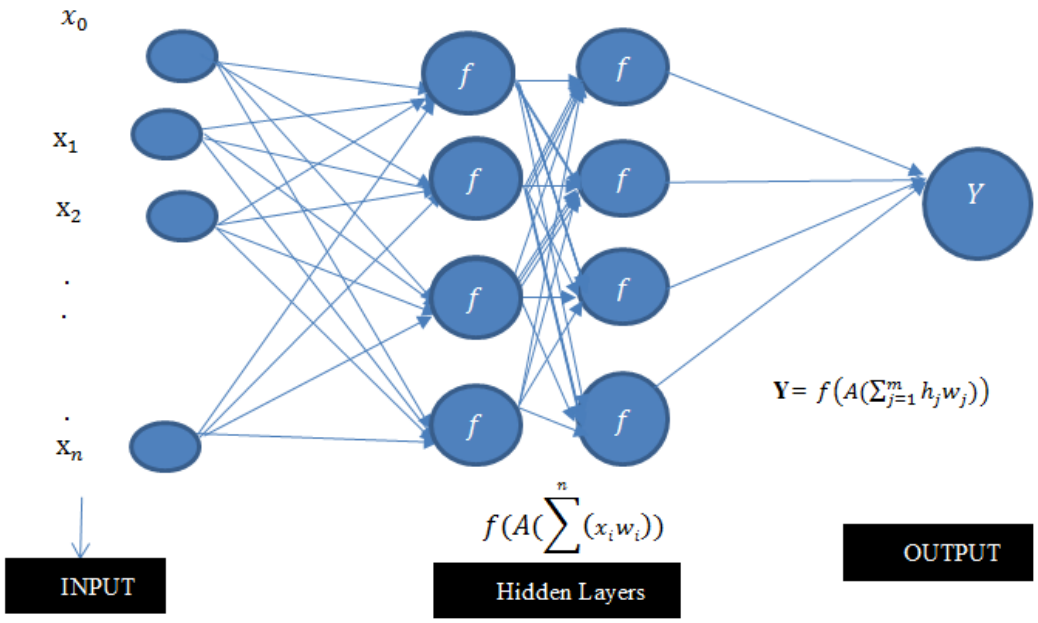


Figure 2: 2-layer Multilayer Perceptron Neural network

The network is trained by adjusting the network weights as defined in equation 8-10 above to minimize the output errors on a set of training data.
The training of a multilayer perceptron can be summarized as:

1. Given a dataset D with $(x_1, x_2, x_3 \ldots \ldots x_n)$ input and P patterns for the network to learn
2. The network with N input units is fully connected to ***h*** nonlinear hidden layers via connection weight w_{ij} associated with each input unit.
3. The hidden layer is fully connected to ***T*** output units via connection weight w_{ij} associated with each neuron in the hidden layer.
4. The training is initiated with random initial weight for each neuron in the network.
5. An appropriate error function $E(w_{jz})$, for instance Mean Square Error (MSE) to minimize by the network is predetermined.
6. The learning rate η is also predetermined.
7. The weight associated with each neuron in the hidden layer and the output layer is updated using the equation:[2] $\Delta w = -\eta \frac{\partial E(w)}{\partial w}$. Until the error function is minimized.

A momentum α is an inertia term used to diminish fluctuations of weight changes over consecutive iterations. Thus, the weight update equation becomes:

$$\Delta w = -\eta \frac{\partial E(w)}{\partial w} + \alpha \Delta w_{ij} \qquad (11)$$

3.2 The Auto Regressive Integrated Moving Average (ARIMA)

ARIMA is proposed by Box and Jenkins [2]. The model assumes that time series is stationary and follows normal distribution. To achieve the notion of stationary in time series, the model subtracts an observation at time ***t*** from an

[2] Kevin Swingler Department of Computing Science and Math, university of Sterling, Scotland. Lecture 4 Multilayer Perceptron. http://www.cs.stir.ac.uk/courses/ITNP4B/lectures/kms/4-MLP.pdf

observation at time *t-1*.
ARIMA stands for

i. Autoregressive, ***AR*** is the lag of the stationary time series data. AR is represented as ***p*** in the model
ii. Integrated, ***I*** a differencing transformation applied to time series to make it stationary. A stationary series is independent of observation time, represented as ***d*** in the model.
iii. Moving average MA is the lag of the forecast errors and is represented as, ***q*** in the model.

Thus, a non -seasonal ARIMA model can be summarized as $ARIMA(p,d,q)$ where:
p is the number of autoregressive terms
d is the number of non-seasonal differences
q is the number of moving average terms
ARIMA(p, d, q) Forecasting equation is defined with respect to the number of differencing necessary to make the time series data stationary as follows:
Let Y= original series
Let y = stationary series
No difference, d=0; then $y_t = Y_t$
First difference, d=1 then $y_t = Y_t - Y_{t-1}$
Second Difference d=2 then

$$y_t = (Y_t - Y_{t-1}) - Y_{t-1} - Y_{t-2}$$

$$= Y_t - 2Y_{t-1} + Y_{t-2}$$ [3]

This paper uses the *statsmodels* package in python to implement ARIMA model for the dataset.

4.0 IMPLEMENTATION

4.1 Development Environment and Tools

i. System: 2.4GHz Intel(R) core ™i5 laptop, 8GB installed memory, Microsoft Window 64 bits operating system
ii. Implementation programming language is Python
iii. Integrated development environment (IDE): Enthought Canopy.
iv. Machine learning tool: Scikit-learn, Keras libraries.
v. Other analysis libraries: Pandas, numpy, statsmodels and Matplotlib.

4.2 Coding

The code is divided into three different non-integrated modules due to time consuming nature of the preprocessing stage. The first module is the preprocessing module with 43 lines of code. The outputs from the preprocessing code are comma separated (CSV) files for the daily count of tickets, dataset summary and the weekly mean data which is the input to other modules.
The second module is the implementation of ARIMA(p, d, q) with 90 line of codes, the output is the evaluation result for ARIMA model save as are comma separated (CSV) file.
The third module is the implementation of the Multilayer Perceptron neural network; this module has 114 lines of code and the output is the evaluation result for the different architecture configured for the experiment, also save as comma separated (CSV) file for further analysis. All results are discussed in section 5.

4.3 Dataset

The dataset for this study is parking contravention transactions updated monthly by the city of Winnipeg on open data government license. The dataset has five attributes and over a million instances comprising of parking tickets issued between January 1st, 2010 and March 31st, 2017. For this paper, seven years' data (2010-2016) is considered**.** The description and preview of the dataset is presented in Table 1 and table 2 respectively.

4.4 Evaluation

The models are evaluated using root mean square error (RMSE) and coefficient of

[3] Robert Nau Lecture notes on forecasting: Fuqua School of Business. Duke Universityhttp://people.duke.edu/~rnau/Slides_on_ARIMA_models--Robert_Nau.pdf

determination (R^2).
RMSE is the square root of mean square error, a risk metric corresponding to the expected value of the squared error loss function. RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}} \sum_{i=1}^{n} (\hat{\sigma}_t - \sigma_t)^2$$

Coefficient of determination (R^2) is a measure of goodness of the model. It explains how well future samples are likely to be predicted by the model [4]. The value of R^2 can be negative or positive. A negative R^2 defines an arbitrary worse model. Is defined as

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1} (y_{i-\hat{y}})^2}{\sum_{i=0}^{n-1} (y_{i-\bar{y}})^2}$$

Where $\bar{y} = \frac{1}{n} \sum_{i=0}^{n-1} (y_i)$ [4]

4.5 Implementation Chart

This study preprocesses the dataset before designing the models. The output of the preprocessing will serve as input to the two major models under consideration. ARIMA model is then implemented follows by the Multilayer Perceptron neural network. Comparative analysis of the results is done after the experiment. The implementation flow for this study is presented in figure 3.

5 Experiment and Results

5.1 Preprocessing

The preprocessing stage involves aggregation of the dataset into daily count and weekly average is then calculated. Using feature selection, the end-date of each week is taken as the period and the weekly average is the time series data. Thus, after the preprocessing stage the dataset has 2 attributes and 366 instances. Sample output of the preprocessing stage is presented in table 3.
The summary statistics for the dataset presented in table 4 shows that the minimum weekly mean between year 2010 and 2016 is 178 tickets while the maximum is 1341 tickets. The graph for the dataset presented in figure 4 shows that there is a spike in ticket numbers around February each year when the snow route tickets are issued.

[4] Model evaluation: http://scikit-learn.org/stable/modules/model_evaluation.html#r2-score

Table 1: Dataset Description

Dataset Name	Parking_Contravention_Citaitons.csv
Number of attributes	5
Number of Instances	1.09M
Attributes Descriptions	
Issue Date	Timestamp
Ticket Number	Transaction unique identifier
Violation	Description of offence, (Text)
Street	Address location of offence (Text)
Location	Longitude & Latitude location of offence

Table 2: Dataset Preview

Issue Date	Ticket Number	Violation	Street	Location
12/13/2016 12:59:55 PM	70219201	01Meter Expired	Hargrave ST	(49.8884066, -97.142226)
12/13/2016 12:58:05 PM	74920668	05Overtime	Kenneth ST	(49.839005, -97.149891)
12/13/2016 12:53:09 PM	75508386	05Overtime	Girton BLVD	
12/13/2016 12:51:36 PM	73418686	01Meter Expired	Portage AVE	(49.89496, -97.136288)
12/13/2016 12:51:11 AM	73533700	13Fire Hydrant	Enfield CRES	(49.8826533, -97.11248)
12/13/2016 12:43:30 PM	73533726	05Overtime	RUE VALADE ST	(49.8852133, -97.122383)

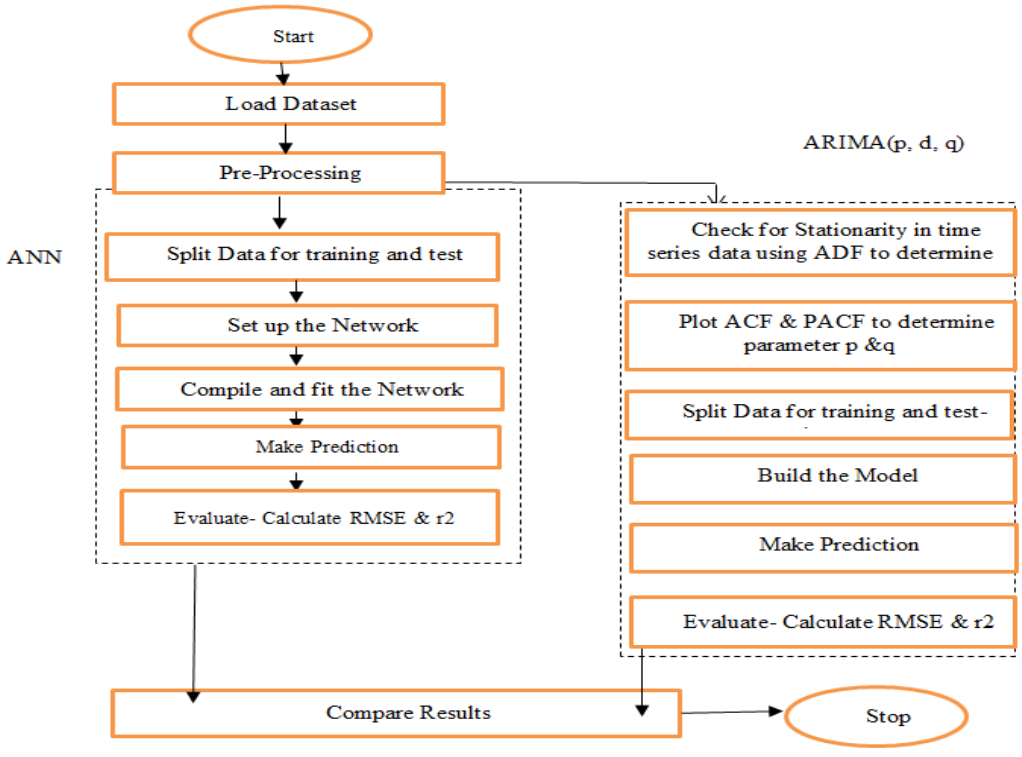


Figure 3: Implementation Chart

Table 3: Preprocessing Sample Output

Period	WeeklyMean
2010-01-03	178.33
2010-01-10	442.57
2010-01-17	483.57
2010-01-24	527.86
2010-01-31	513.43
2010-02-07	475.86
2010-02-14	502.71

Table 4: Preprocessing- Dataset Summary

Count	366
Mean	429.43
Standard Deviation	442.57
Min	178.33
Max	1340.71

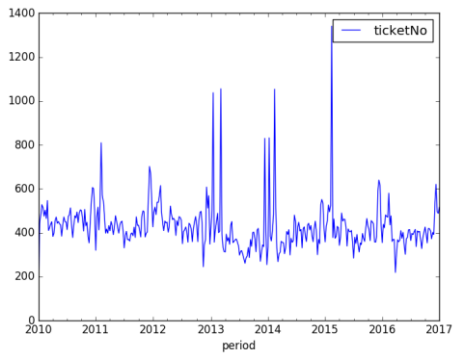


Figure 4: Dataset Trends Graph

5.2 ARIMA (p, d, q) Model

The assumption of the ARIMA model is that the time series is independent of time. Thus, Augmented Dickey-Fuller (ADF) test is performed to test for stationarity in time series data. ADF null hypothesis states that a sample data has unit root, the data is not stationary and the alternative hypothesis states that the data is stationary. If the p-value >0.05 the null hypothesis is accepted and if p-value<0.05 the null hypothesis is rejected.[5]

[5] How to Check if Time Series Data is Stationary with Python

The result of the Augmented Dickey-Fuller (ADF) presented in table 5 shows that the p-value<0.05, thus, the data is stationary and the null hypotheses is rejected.

Table 5: Augmented Dickey-Fuller (ADF) Test

ADF Test Result	
ADF Statistic:	-4.111741
p-value:	0.000926
Critical Values:	
1%	-3.449
5%	-2.870
10%	-2.571

Since the time series is stationary, the value of parameter ***d*** is assumed to be zero. $d = 0$
The significant of the preprocessing stage is observed in the result of the augmented Dickey-Fuller Test. The mean weekly ticket calculated at the preprocessing is a useful tool in transforming time series data to stationary.
Next, the log transformation is applied to the dataset for scaling and the ACF and PACF are plotted to determine the value of p and q parameters of the $ARIMA(p, d, q)$.
The Autocorrelation Function (ACF) is defined as a measure of correlation between the time series with a lagged version of itself.
Partial Autocorrelation Function (PACF) is defined as the correlation between a time series with a lagged version of itself after removing the effect already explained by previous lag. For instance, at lag 5 say X_5, the PACF is the correlation after removing the effect of x_1, x_2, x_3 and x_4.
Parameter p is defined as the point where the PACF crosses the upper confidence interval for the first time. From the result in figure 5, p =1.
Parameter q, is defined as the point where the PACF tail off. From the result in ***Figure 4, q=2.***

http://machinelearningmastery.com/time-series-data-stationary-python/

67% of the dataset is considered for training and the remaining 33% is used for testing.
The $ARIMA(p,d,q)$ is implemented using: $ARIMA(1,0,2)$, $ARIMA(2,0,2) and \ ARIMA(4,0,2)$. The result presented in table 6 shows that $ARIMA(1,0,2)$ has the lowest error (RMSE=0.181) and the highest correlation if determination $r^{2.}$ (0.141). on the average, there is a prediction error of 104 tickets per week.

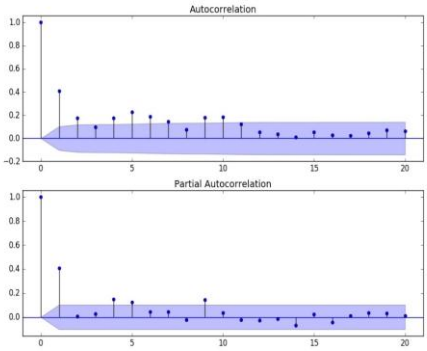


Figure 5: Autocorrelation and Partial Autocorrelation Plot

5.3 Multilayer Perceptron (MLP) Neural Network

Similar to the ARIMA model implementation, 67% of the dataset is considered for training and 33% for testing. Precisely, the training set has 245 instances while the testing set has 121 instances. Four different MLP architecture were designed in this paper; a 2-layer with one neuron in the hidden layer denoted as 2H1, 2-layer with four neurons in the hidden layer denoted as 2H4, 3-layer having four neurons each in the two hidden layers denoted as 3H44 and 4-layer with four neurons in each of the three hidden layers denoted as 4H444. All the models are separately trained for up to 900 epochs using the sigmoid activation function and a comparison is made using the tanh activation function. The relationship between sigmoid and tanh activation functions is stated in equation (4a). The optimizer selected for the training is the Stochastic Gradient Descent (SGD) optimizer with a default learning rate of 0.01 and momentum of 0.0 and the dataset is standardized using MixMaxScaler function in the range (-1,1). An attempt to use sigmoid activation function in the output layer resulted into negative r^2 (-9.67); thus, a linear activation function is used for the output layer of all the architectures. The setup is presented in table 7.
The loss function specified for all the models is the Mean Square Error (MSE), RMSE and R^2 are subsequently calculated for evaluation. The result presented in Table 8 for the sigmoid activation function shows that a 2-layer with one neuron in the hidden layer has the best goodness of fit having correlation of determination R^2 of 0.126 and an error of 0.176. Adding more hidden layer does not improve the prediction capability of the network.
The result from table 9 for the tanh function shows performance improvement of adding layer up to 3H44 where the best result is recorded. Further additions of layer beyond 3H44 add no value to the prediction capability and goodness of fit of the network. The Comparative analysis of the result presented in table 10 and figure 6 shows that 3H44 has the highest correlation of determination R^2 and the lowest error, RMSE having an average prediction error of 95 tickets per week. A 2-layer MLP with one neuron in the hidden layer also has a better performance than ARIMA(1,0,2) having an average prediction error of 102 tickets per week.

Table 6: ARIMA(p, d, q) Results-(RMSE and R^2)

Model	RMSE	R^2	RMSE (No Scaling)	R^2 (no Scaling)
$\boldsymbol{ARIMA(1,0,2)}$	**0.181**	**0.141**	**104.19**	**0.08**
$ARIMA(2,0,2)$	0.182	0.135	104.42	0.08
$ARIMA(4,0,2)$	0.182	0.132	104.38	0.08
$ARIMA(5,0,2)$	0.183	0.125	105.02	0.06

Table 7: Multilayer Perceptron architecture

Epoch=900, Optimizer=SGD, learning rate=0.01, loss function=MSE
Standardization = MinMaxScaler

Models	**No of Hidden Layer**	**No of Neuron in Hidden Layer**	**Activation Function**	**Input Dimension**	**Output Dimension**
2H1	1	1	Hidden- Sigmoid Output-Linear	3	1
2H4	1	4	Hidden- Sigmoid Output-Linear	3	1
3H44	2	4,4	Hidden- Sigmoid Output-Linear	3	1
4H444	3	4,4,4	Hidden- Sigmoid Output-Linear	3	1

Table 8: Sigmoid Activation Function Evaluation Results (RMSE and R^2)

Model	RMSE	R^2	RMSE (actual dataset)	R^2 (actual dataset)
2H1	**0.176**	**0.126**	**102.07**	**0.126**
2H4	0.181	0.067	105.45	0.067
3H44	0.177	0.115	102.67	0.115
4H444	0.183	0.052	106.30	0.052

Table 9. Tanh Activation Function Evaluation Results (RMSE and R^2)

Model	RMSE	R^2	RMSE (actual dataset)	R^2 (actual dataset)
2H1	0.172	0.165	99.77	0.165
2H4	0.166	0.217	96.61	0.217
3H44	**0.164**	**0.235**	**95.49**	**0.235**
4H444	0.167	0.211	96.94	0.211

Table 10 Comparative Analysis of Results (RMSE and R^2)

Model	RMSE	R^2	RMSE (actual dataset)	R^2 (actual dataset)
MLP tan h (3H44)	**0.164**	**0.235**	**95.49**	**0.235**
MLP sigmoid (2H1)	0.176	0.126	102.07	0.126
ARIMA(1,0,2)	0.181	0.141	104.19	0.08

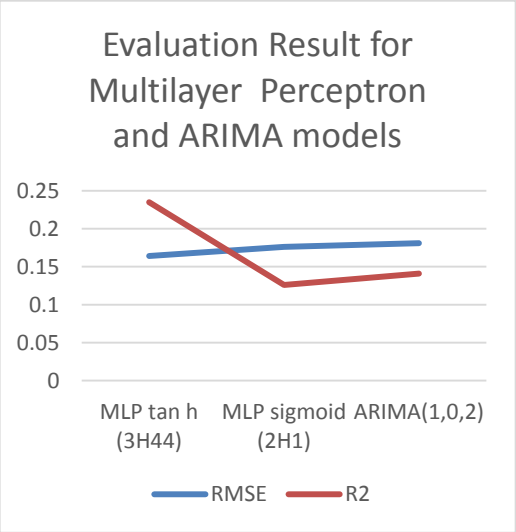


Figure 6: Evaluation Result (RMSE and R^2)

6.0 CONCLUSION

The performance of Multilayer Perceptron neural network and ARIMA models has been investigated in this study. Observations from the performance evaluation of the models as presented in table 8,9 and 10 revealed that the four MLP architectures designed using tanh activation function outperform ARIMA model with 3H44 model producing the best goodness of fit (R^2 = 0.24) and lowest prediction error (RMSE =0.16).

The effect of adding more layers on the performance of a multilayer perceptron neural network is similarly investigated. Using sigmoid activation function, a 2-layer MLP having one neuron in the hidden layer has the best performance in term of prediction error (RMSE=0.176) and coefficient of determination (R^2 = 0.126) measures. Adding more layers to a network configured using sigmoid function resulted to performance degeneration. One problem with sigmoid activation function is saturation which may lead to gradient vanishing (and /or explosion) making network training more difficult [6]. Rescaling sigmoid activation function is suggested in [6] to achieve a comparative result with tanh activation function. Like sigmoid activation function, tanh activation function also has a saturation effect, however, unlike sigmoid, the output of tanh activation function is zero-centered. Thus, adding layers to a network configured using tanh activation function can improve the performance of the network as demonstrated in this study. From the result in Table 9, it can be observed that adding more layers reduces the prediction error and improves the goodness of fit of the network up to the 3-layer network (3H44).

In addition, preprocessing of dataset is a necessity to some models like ARIMA and MLP investigated in this study. ARIMA model requires a stationary time series data. This is achieved in this study by first aggregating the ticket transaction to daily count and using equal weekly frequency grouping the mean of each group is calculated and the logarithm function is then applied to the mean. Standardization is a requirement for multilayer perceptron network to remove bias that might result from wide variation in range of values of raw data during training. From the summary of preprocessing stage in table 4, it can be observed that standardization is required since the minimum average ticket per week is 178 while the maximum is 1340. This study used the MinMaxScaler function of the Scikit-learn library to transform the dataset to a range [-1, 1]. The result from this study suggests that choosing a good activation function can significantly improve the performance of a multilayer perceptron neural network.

REFERENCES

[1] J. Wang, J. Wang, W. Fang, and H. Niu, Financial time series prediction using Elman recurrent random neural networks, Computational Intelligence and Neuroscience, vol. 2016, Article ID 4742515, 14 pages, 2016.

[2] Chaudhuri T. D. et al. Artificial Neural Network and Time Series Modeling Based Approach to Forecasting the Exchange Rate in a Multivariate Framework" Journal of Insurance and

Financial Management, Vol. 1, Issue 5 (2016), pp 92-123.

[3] Khashei, Mehdi, and Mehdi Bijari. "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting." *Applied Soft Computing* 11.2 (2011): 2664-2675.

[4] Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". *Nature*. **323** (6088): 533–536.

[5] B. Xu, R. Huang, and M. Li. Revise Saturated Activation Functions. ArXiv e-prints, February 2016.

[6] City of Winnipeg Parking contravention dataset: https://data.winnipeg.ca/Parking/Parking-Contravention-Citations-/bhrt-29rb/data

[7] Scikit-learn Machine Learning in Python http://scikit-learn.org/stable/index.html

[8] Keras Deep Learning Documentation https://keras.io/

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25, No 1, June 2018

OPTIMIZED MODEL FOR COMPARING TWO INTRUSION LOGS

[1]Dr O. J. Nehinbe and [2]U. S. Onyeabor
[1,2]Department of Computer Science, Federal University Oye-Ekiti, Nigeria
[1]nehinbe@yahoo.com; [2]uchechukwu.onyeabor@fuoye.edu.ng

Abstract- The detectors for watching, keeping and reporting records of digital activities that have the tendency to endanger the security of computer and mobile systems are greatly needed in digital security and forensics across the globe. Nevertheless, most detectors are fraught with series of challenges whenever they are concomitantly operated to detect potential intrusions within mobile and computer networks. Conventionally, analysts must correlate and aggregate alerts of such devices before well-informed decisions can be made from them. Unfortunately, correlations and aggregations will fail to produce desirable results whenever multiple pairs of alerts do not possess visible, mutual, complementary, or reciprocal relationships. Consequently, most of the existing models can suffer low efficacies whenever they are adapted to compare two intrusion logs within different time intervals. This paper presents a pragmatic and optimized approach that uses computational methods to compare a pair of intrusion logs together. Category utility and entropy are applied to respectively measure the quality of each pair of logs generated from Snort. Series of evaluations carried out using intrusion logs that are derived from synthetic and real traces demonstrate how analysts can forecast the extent of similarities and dissimilarities of two intrusion logs. The results further suggest some intrusive themes, the nature, quality, degree and significance of pairs of intrusion logs across different time intervals.

Keywords: Intrusion, intrusion detection system, detector, networks forensics.

1.0 Introduction

The techniques for monitoring various communications that migrate across Computer and mobile systems that started with key findings from researchers [1, 5], is playing significant roles in safeguarding digital resources and their respective users from intruders across the globe [10, 20]. As shown in Figure 1, Intrusion Detection System (IDS) or detector is a strong-growing Network forensic toolkit that is specially designed to technically

monitor, gather and report digital activities that may endanger the security of computer and mobile systems across the last three decades [21, 23, 19 and 8].

The central issue here is that the aforementioned mechanism has several advantages and disadvantages for using numerous attributes to clearly describe suspicious packets and for recording them as alerts in the form of intrusion logs [22, 18, 13, 15 and 17]. For instance, intrusion logs that are informative can furnish analysts with comprehensive sequence of online real-time communications and communications that have already taken place across Computer and mobile systems without raising much protesting, doubting or further objection about the validity of the recorded activities.

Additionally, detectors can log alerts that are clear to the mind of professionals. Similarly, experience has shown that the toolkits can operate in the mode to detect potential intrusions and yet they will capture and log alerts that will show certain characteristics that are distinct to mental discernment [18, 16].

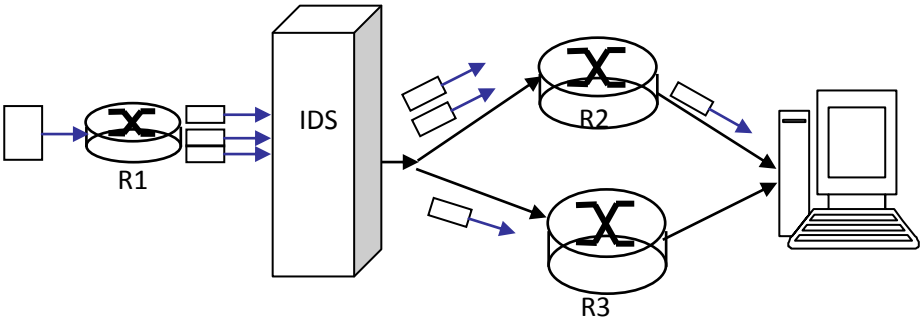


Figure 1: Illustrates some packets that can lost in transits

Conversely, the recent changes in semantics, data engineering and data analytics globally have invigorated sudden novelty in data paradigm and the potential sizes of intrusion logs within corporate networks. Thus, the requests for collaborative analyses of intrusive logs are intensified across continents [10]. As a result, the perceptibility of the quality of alerts within an intrusion log rapidly becomes difficult as the size of the log increases.

Besides, evolution of intrusion logs has proposed several methods of log analysis over the years. The research domain that has advanced with notable research findings such as by researchers [4, 6 and 7] to cite a few, has equally reached a stage whereby intrusive logs should not only be ultimately used to thwart intrusions and to litigate intruders. Experience has shown that there is need to further compare intrusive logs for a number of reasons.

In the same token, the detectors of today are expected to possess the capabilities to sniff several terabytes of data accurately and without any successful attempt to evade detection [18]. The implementation of this issue is in progress in a recent time. For example, according to [10] international bodies are planning enhanced international collaborations with the view to curtail cases of cyber threats across the globe. Conventionally, alerts of

multiple detectors are often correlated and aggregated before analysts can make well-informed decisions from them. With the recent issues surrounding big data analytics, correlations and aggregation can fail to produce the expected outcomes for a number of reasons.

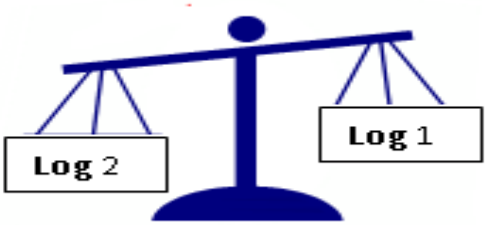


Figure 2: Comparison of digital logs

The methods can be ineffective if there are multiple alerts within two or more intrusion logs that do not exhibit substantial mutual, complementary, or reciprocal relationship. For these reasons, there are potential challenges for most analysts to draw out sound comparison between a pair of intrusion logs. As shown in Figure 2, it is possible to compare a pair of intrusion logs together. The basis for comparing a pair of intrusion logs is that analysts and forensic professionals can collaboratively deduce useful conclusions that can be useful to the cyber groups from such comparison.

There are other benefits inexplicit in comparing a pair of intrusive logs together. Such comparison will enable forensics professionals to examine and understand the degree of resemblances of specific pairs of digital logs that are obtained from computer and mobile networks. The method will explicate the underlying facts and the degree of equivalences, similarities and differences between pairs of digital logs that are under review. Qualities of intrusive logs that are comparable will be known and the scenario whereby there is no comparison between two logs can be unveiled. Additionally, the intention to compare a pair of intrusive logs can help operators to establish the essential equality, interchangeability, overestimation or underestimation of the given pairs of logs as the case may be during incident reporting of critical issues especially to the Board of Directors (BoD) in a corporate setting.

Unfortunately, analysts often face further challenges whenever detectors are homogeneously or heterogeneously deployed within the same networks and there are necessities to compare their logs across several time intervals. Secondly, Return on Investment (ROI), forecasting of countermeasures, prudent deployment of limited resources to achieve swift forensic analysis of intrusion logs and the goals of detecting potential intrusions may be defeated if such situations are not under control on time. For these reasons, this paper presents Log-evaluator, as a method for comparing two intrusion logs together.

Log-evaluator uses a computationally quick technique to systematically compare two intrusion logs together by forming clusters of the values held in particular attributes of alerts in each log in a top-down manner. The model is implemented on the platform of C++ programs running on Window Vista Operating System. Furthermore, category utility and entropy are applied to respectively measure the quality of intrusion logs that are obtained from Snort. Also, in-depth evaluations of the model are carried out using intrusion logs that are derived from synthetic and real datasets.

Additionally, one of the substantial contributions of this paper is its ability to propose a novel method for concurrently forecasting the extent of similarities, equivalences and dissimilarities of two intrusion logs, first at sight. Unlike most models, the method proposed in this paper, has the capability to reveal two logs that are equivalent to each other in some respects. The paper has used entropy and category utility to provide deeper understandings and explanations of nature of intrusions with pairs of intrusion logs whenever they are closely examined together.

The remainder of this paper is organized as follows. Section 2 gives some related works while section 3 exhaustively discusses redundancy. Section 4 gives an overview of the datasets that we used for experimental investigations. Section 5 gives a description of our approach to concurrently compare a pair of intrusion logs together. Section 6 provides elaborate account of experimental results performed with the proposed model. Section 7 gives conclusion and future research direction that can be pursued to improve the quality of the research described in this paper.

2.0 Related work

Some of the studies about log analysis have been comprehensively reported [15, 17, 19 and 8]. The reviews suggest two generic issues about IDS logs. Firstly, the analysis of IDS logs is conventionally used to aggregate and visualize intrusions, mitigate dangers; thwart online intrusions, debug IDSs, and to regulate network management procedures of corporate organizations so that firms will conform to accepted standards. Secondly, the comparison of intrusion logs is a novel research area in network forensics.

Apart from the fact that previous methods for analyzing the quality of IDS logs are models that mostly quantify one intrusion log at a time, few of them published in contemporary literatures commonly adopt clustering technique and its variants to partition intrusion logs into two or more groups [20]. Unfortunately, there are major drawbacks associated with such models. For example, critical information may be lost in the course of manipulating the subdivisions of the IDS logs. Secondly, high level of expertise and time constrain are issues that can militate against prompt harmonization of the reports obtained from such models.

In the works of a researcher [15, 17] the quality of intrusion logs and patterns of intrusions are determined by separating failed attacks from attacks that can potentially succeed within computer and mobile networks. Thereafter,

information theory and category utility are used to analyze each subgroup of the datasets. The model suggests informative attributes and attributes that best discriminate network intrusions [16]. Similarly Shui [19] evolved novel taxonomies and ideas for effective investigation of the Distributed Denial of Service (DDOS) attacks. However, despite their immense benefits with them, there are serious weaknesses identified with them as well. For instance, most DDOS attacks are not good examples of failed attacks because successful DDOS attacks start from the point at which an intruder compromises digital networks before installing the Trojan that will launch the attacks. Hence, it will be erroneous to compare such intrusive logs with logs that have been split into two groups.

3.0 Comparison of intrusion logs

In traditional network forensics and as shown in philosophical doctrine of intrusion detections, studies by different researchers [12, 11] infer that the central goals of analysts and forensic tools are to achieve abstract separation of an intrusion log into its constituent attributes. Thus, the possibility of investigating IDS logs using attributes like source addresses, destination addresses and timestamp are explored.

Some intruders may take longer durations to achieve their motives by launching cycles of suspicious packets that will complete their life cycles within a year. Hence, such attacks will elude online real-time analysis of IDS logs.

Another drawback of the traditional log analysis is that the results obtained from them are restricted to the explication of the investigations and the components of such IDS log. As a result, the degree at which such IDS log reflects analogy of intrusive events become subjective in most cases. In other words, an intrusion log is basically an explanation of the events that have happened with the networks where the digital log has been extracted. Nonetheless, if there is any request to be equivocal about the degree of similarities and dissimilarities within two intrusion logs irrespective of whether they are extracted from the same computer or mobile networks or they are extracted from different digital peripherals, the efficacy of the results obtained from traditional network forensics are not frequently opened to two or more interpretations.

Essentially, it is plausible to compare a pair of intrusion logs across different parameters. We opine that two IDS logs can be compared on a daily basis, weekly, monthly and annually. The results of such comparisons usually underpin different motives of the analysts.

3.1. Benefits of Log comparison

Logs must be compared for a number of reasons. Two intrusion logs that are compared together can help analysts and business organizations to understand the degree at which similar or closely related events are migrating across their digital networks.

In addition, comparison of two intrusion logs can suggest detectors that can be overloaded within an organization in the future time. Organizations can equally deduce if they are placing too much or too little workloads on their analysts at a given period.

Furthermore, concurrent comparison of a pair of intrusive logs is necessary to achieve adequate planning. Similarly, adequate planning about outbreak of intrusion is the bedrock upon corporate firms can achieve the best quality of services from their analysts, detectors and service providers.

The act of monitoring resource utilization, distribution by allotting tasks, partitioning and apportioning logs during collaborative intrusion analysis can be enhanced if the qualities of two intrusion logs are compared together according to feasible contingency plans.

Further still, comparison of intrusion logs can serve as substantive evidence to illuminate certain intrusive events that depart from expectations. Essentially, comparison of two intrusion logs can substantiate the levels of differences between conflicting facts, claims, personal beliefs and judgments about intrusions that are not founded on experimental proof or certainty.

Attributes of alerts are often used to categorize them into smaller conceptual schemes. All alerts within the same scheme thus indicate a collection of events sharing a common characteristic. According to Nehinbe [15] and [9], category utility is a statistical concept that can be used to estimate the essential and distinguishing attributes of clustering schemes of alerts. The measure can suggest the quality at which alerts within a clustering scheme can be practically use to replace another group of alerts within another clustering scheme.

Mathematically, suppose C_1, $C_2 \ldots C_k \ldots C_N$ are clusters with attributes A_1, A_2, … A_j . Whereby the i^{th} attribute has the values that range from V_{i1}, V_{i2},…to $V_{ij..}$ The expected value of number of correct attribute value predictions [9], a predictor can make if a predictor is told that an attribute belongs to a cluster C_k is

3.2. Category Utility

$$\sum_i \sum_j (P(A_i = V_{ij} \mid C_k))^2 \qquad (1)$$

Similarly, the expected value of number of correct attribute value predictions a predictor can make without the knowledge of the cluster to which an attribute belongs is

$$\sum_i \sum_j (P(A_i = V_{ij}))^2 \qquad (2)$$

The gain for all the clusters is

$$\sum_k P(C_k) \left\{ \sum_i \sum_j (P(A_i = V_{ij} \mid C_k))^2 - \sum_i \sum_j (P(A_i = V_{ij}))^2 \right\} \qquad (3)$$

Then, category utility score for the clustering scheme formed by an attributes A_i is

$$\frac{\sum_k P(C_k)\left\{\sum_i\sum_j (P(A_i = V_{ij} \mid C_k))^2 - \sum_i\sum_j (P(A_i = V_{ij}))^2\right\}}{N} \quad (4)$$

In equation (4) above, N is the total number of the clusters formed by an attribute A_i and it is introduced to favour an attribute that generates smaller numbers of clusters

.

Algorithm 1: Pseudo-codes to compute expected correct prediction for clustering schemes

F2: Assign attributes Ai to form cluster;

Partition D into cluster Ci , i= 1, 2, ..n based Ai;

While (there exists a cluster Ci)

Find probability of all attributes in cluster Ci , i≥1;

Sum squares of probability to get expected prediction for each attribute in the clustering scheme;

Sum expected correct prediction for all attributes to get total expected prediction for the clustering scheme T1;

End while

Algorithm 2: Pseudo-codes to compute Category utility

K_i=T- T^1; K_i= gain for cluster i, T^1 is correct prediction for entire dataset
while T'is correct prediction for cluster i
$P(G_i) * K_i$;
$V_i = ((\sum P(G_i) * K_i)/n)$;
if (attribute to assign) **then**
goto F2;
else
exit;

3.3. Entropy

Intrusion detectors generate alerts that can lack any predictable order. Hence, the quality of such alerts cannot be estimated by manual approach. Entropy or Information in this

context signifies the degrees of uniformity possessed by a collection of alerts in an intrusion log, from which conclusions can be drawn, are related.

Mathematically, and in the works of some researchers [3, 9] Entropy or Information content of an intrusion log

$$= -\sum_{i=1}^{n} p_i \log_2(p_i) \qquad (1)$$

Where: The sum is for all the clusters in the clustering scheme and $p(i)$ is the probability that an alert belongs to the i^{th} cluster

Algorithm 3: Pseudo-codes to compute inform (entropy) of intrusion log

Input D (data set);
For ($A_i \in D | i= 1, 2, \ldots, j$)
Partition D into $G_i | i= 1, 2, \ldots, n$;
end for
while ($G_i \in A_i$) do
$P(G_i) | i=1,2 \ldots n$;
$E_i = (P(G_i)*(\log_{10} P(G_i)/\log_{10}(2))$;
$\sum E_i$;
End while
exit

4.0 Evaluative datasets

Structurally, we evaluate six categories of intrusive traces that are stored in the Packet Captured (PCAP) format. The datasets are divided into three pairs. The first pair of datasets is the trace files extracted from LLDOS.1.0 and LLDOS.2.0.2. There are two categories of DARPA datasets or traces commonly known as DARPA 2000 intrusion detection scenario-specific datasets [14]. Furthermore, both datasets, attackers installed Trojans in compromised computers to maliciously launch the DDoS attacks. The first group of the datasets is labeled as LLDOS 1.0 to signify first scenario of DDoS attacks that was launched by novice attacker. The second category of DARPA 2000 datasets is labeled as LLDOS 2.0.2 [14]. LLDOS 2.0.2 dataset is the second scenario of DDoS attacks that was launched by experienced attacker [14] within the same networks.

The second pair of the evaluative dataset comprises of the DEFCON-10 and DEFCON-11 datasets. In [2], DEFCON datasets were

Internet trace files that were collected from RootFu networks during the capture the flag contest organized in different years. Each of the attacks within both datasets lasted between 24 hours to 48 hours [2]. DEFCON-10 dataset contains bad packet, attacks on administrative privilege, FTP attacks via telnet protocol, ports scan and port sweeps attacks, fragmented attacks that intended to cause buffer-overflow, directory traversal attacks and IP spoofing attacks.

The third pair of the evaluative dataset is made up of the EXPERIMENTAL and LIVE datasets. They were respectively extracted from Local Area Network (LAN) and University's networks within two different months [15]. They consist of attacks on administrative privileges; ports scan attacks, Trace route attacks, Ping attacks, UDP and TCP scanning attacks.

Essentially, interesting readers may obtain further information about the topology of the networks and other details about the above datasets from the repositories of the [14] and [2].

5.0 The Log-evaluator

In Figure 3, **Log-evaluator** has four generic components. They subsume LogFeelers, Log-Analyzers, Evaluators and Rundown engine. The model is implemented with the C++ programming language, compiled and executed on Windows Vista Operating system.

The following seven attributes viz: SI, DI, IPP, IPL, TTL, IPF and TOS are selected for forensic analysis of all the intrusion logs evaluated in this paper. One of the uniqueness of this model is it has the ability to analyze a log at a time, and its ability to concurrently analyze two logs at the same time.

Fundamentally, LogFeeler-1 and LogFeeler-2 must contain a pair of intrusion logs at the same time. For the first time, the inputs to the LogFeelers are alerts from LLDOS.1.0 and LLDOS.2.0.2 datasets that are intended to be compared together. Each ***LogFeeler*** has the same inbuilt rules that make the component sensitive to alerts. The sensitivity of the ***LogFeeler*** is similar to the sensitivity of a sensor of a detector and both logs are processed in a top-down manner.

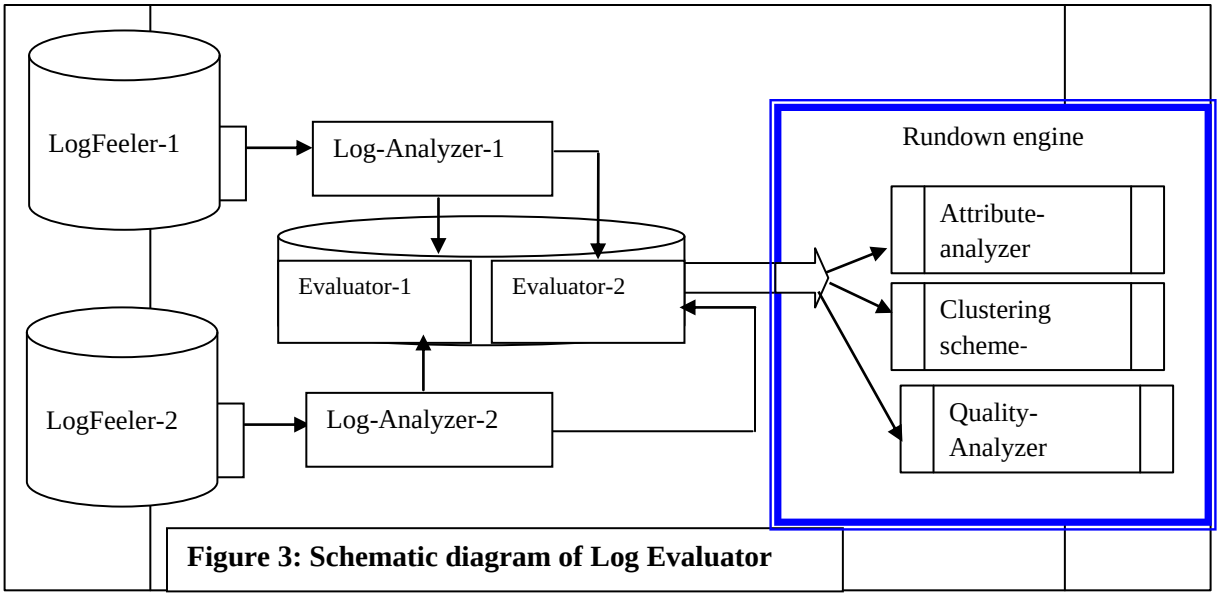


Figure 3: Schematic diagram of Log Evaluator

Technically, the ***LogFeelers*** have clustering rules that are specific to each of the above attributes. Both ***LogFeelers*** acquire alerts from the given pair of intrusion logs. They independently process them and sent them as inputs to the ***LogAnalyzers***. The ***LogAnalyzers*** have rules that are used to perform three basic functions. The rules group alerts from the intrusion logs into clusters. Subsequently, the rules forward clustered alerts to ***Evaluator-1*** and ***Evaluator-2*** for further analyses.

In a simple technical term, the evaluators use the algorithms already stated above to determine the quality of each IDS log. In other words, ***Evaluator-1*** uses seven attributes of alerts mentioned above to compute the category utility of each attribute of the log. Similarly, ***Evaluator-2*** uses the same seven attributes of the alerts from the same intrusion log to compute the entropy of each attribute and the aggregate entropy for the entire log.

The results obtained are outputs to the ***Rundown engine***, which in turn organize them in human readable formats. The above procedures are repeated for other pairs of intrusion logs described above and the results obtained are critically discussed below.

6.0 Results and discussions

This section presents the results, the discussions of the results and further intrusive themes that are extracted from the results.

6.1. Results and analysis

The table below gives the statistics of the duration of each dataset. Figure 4 up to Figure 10 are the key results obtained from the above analyses. Figure 4 illustrate the sum total of many heterogeneous attacks within each evaluative data whenever they are considered together.

Table 1: Dataset and statistical duration of events

Data Set	**Statistical duration of events**	
	Start time	**End time**
LIVE-DATA	2013-01-24 21:40:02	2013-01-25 18:21:45
EXPERIMENTAL	2013-01-24 13:27:08	2013-01-25 10:36:06
LLDDOS-1.0	2000-03-07 16:27:51	2000-03-07 16:27:56
LLDDOS-2.0.2	2000-04-16 21:06:15	2000-04-16 21:06:23
DEFCON-10	2002-08-03 00:58:03	2002-08-04 20:22:44
DEFCON-11	2003-08-02 14:59:03	2003-08-04 13:17:13

The results suggest that LIVE dataset respectively generate the largest quantity of alerts while LLDDOS-2.0.2 dataset generate the lowest quantity of alerts.

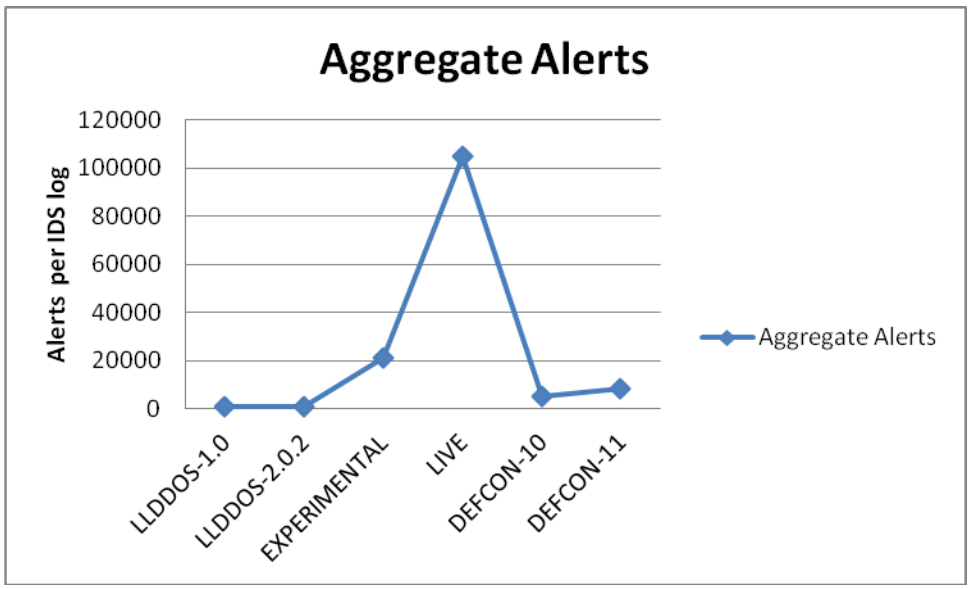


Figure 4: Aggregate alerts per IDS log

The total packets across evaluative datasets are exemplified by Figure 5. The results further illustrate that DEFCON-11 dataset generates over ten millions of packets while EXPERIMENTAL dataset generates the least packets whenever they are considered together.

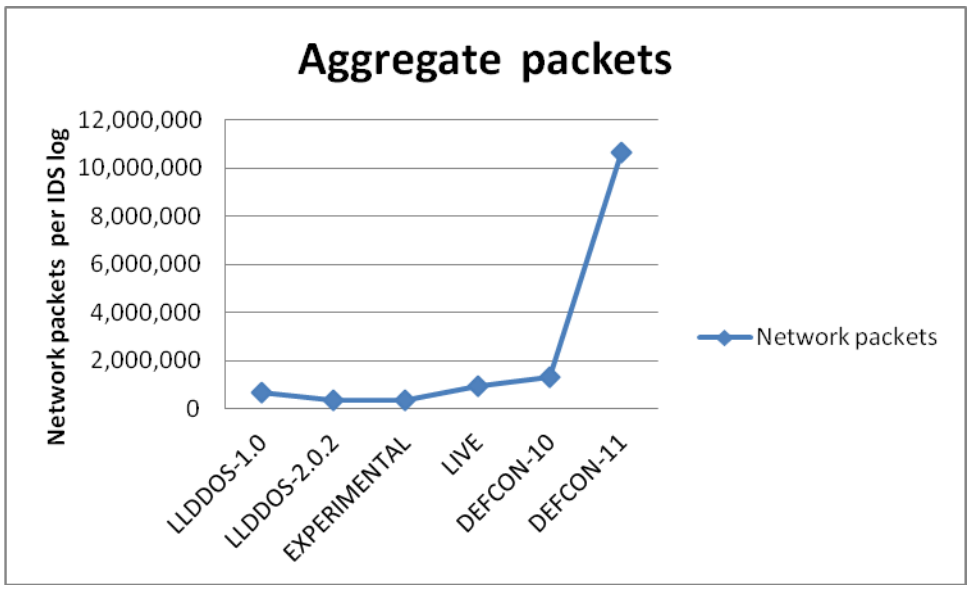


Figure 5: Aggregate packets per IDS log

The possibility of predicting the attacks within two intrusion logs have been demonstrated in Figure 6. The results imply that the intrusions within DEFCON-11 dataset will be difficult to predict when compared with other intrusion logs.

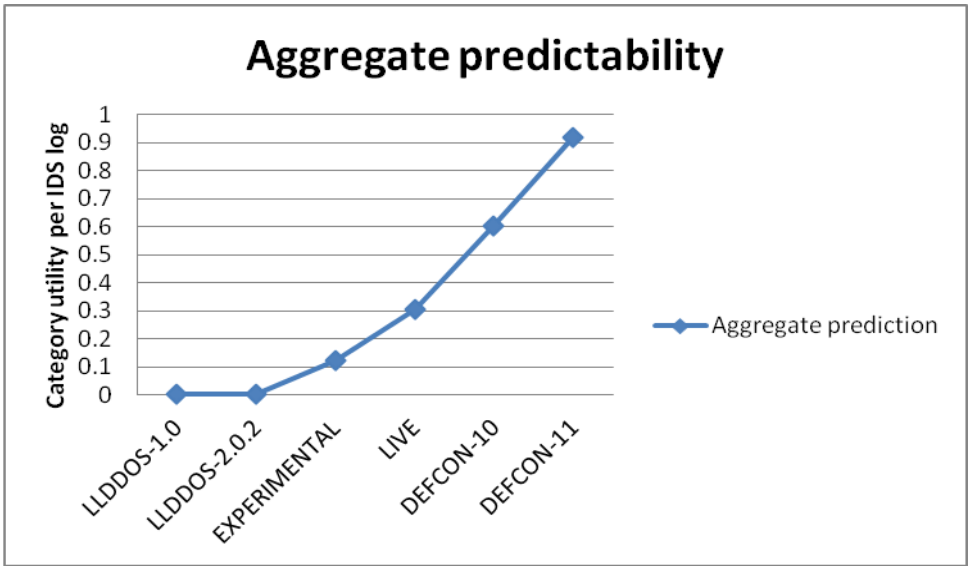


Figure 6: Aggregate predictability per IDS log

Also, the DDoS attacks within LLDDOS-1.0 and LLDDOS-2.0.2 datasets possess the quality of being easily predictable. In other words, both categories of DDoS attacks within the aforementioned datasets are certainly predictable than attacks within other datasets.

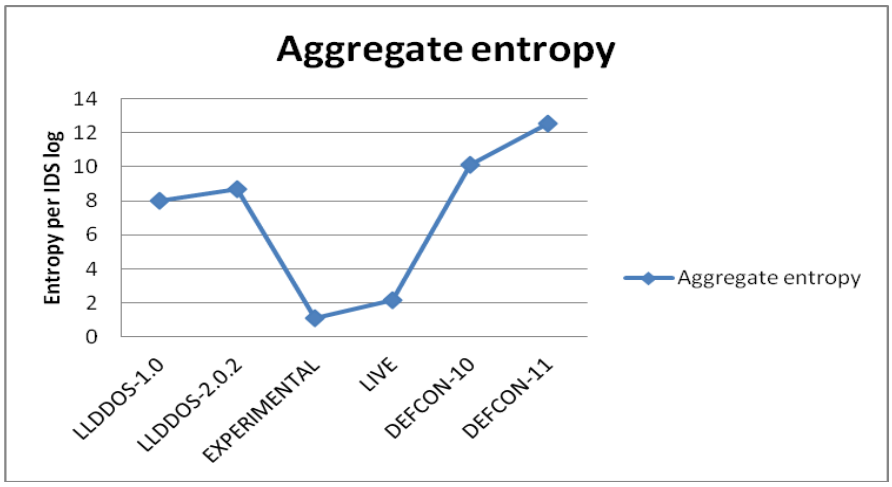


Figure 7: Aggregate entropy per IDS log

Figure 7 adds details, as to an account of evaluation carried out to clarify the degree of uniformity of the sum total of heterogeneous attacks within each evaluative data whenever all the datasets are considered together. The central findings for this evaluation show most of the alerts within DEFCON-11, DEFCON-10, LLDDOS-2.0.2 and LLDDOS-1.0 datasets are most likely to be regular and unvarying at sight.

The degree of badness of clusters formed by attributes of LLDDOS-2.0.2 and LLDDOS-1.0 convey less information. Contrarily, most of the intrusions or alerts within EXPERIMENTAL and LIVE datasets are most likely to be diverse and interesting.

The results shown in Figure 8 and Figure 9 compare the degree of goodness, badness and the degree at which it is certain to predict intrusions within each intrusion log given the above-named seven attributes. The results suggest that notion of the goodness or badness within a pair of intrusive logs depend on the preference of the analysts. That is, the intrusive log that may be good to one analysts may be regarded as bad to another analyst.

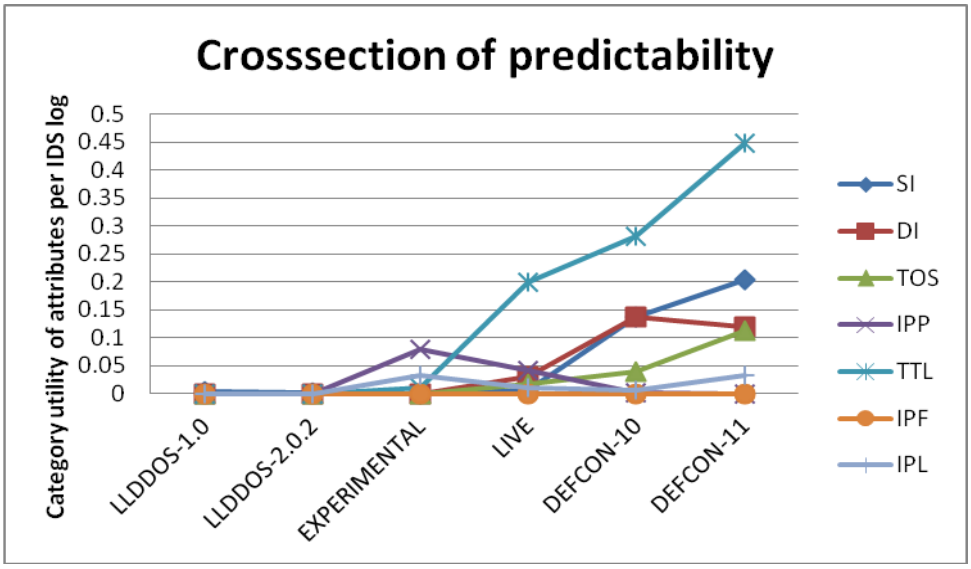


Figure 8: Crosssection of predictability per IDS log

Figure 8 says the intrusions within the LLDDOS-1.0 and LLDDOS-2.0.2 datasets are relatively easy to be predicted by their source addresses while Figure affirms that there are repeated attacks originating from the same source addresses in LLDDOS-1.0 and LLDDOS-2.0.2 datasets respectively.

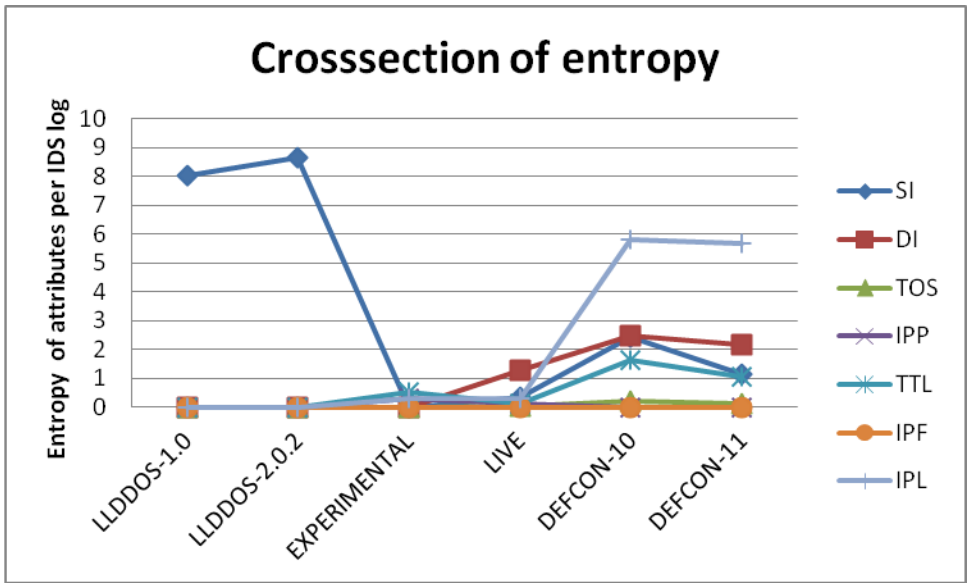


Figure 9: Crosssection of entropy per IDS log

The individual clustering schemes that are formed by each attribute across the six evaluative datasets are shown in Figure 10. The results illustrate groups of independent but interrelated components that can be used to explicate the discrepancies observed across different intrusion logs and across unified whole log.

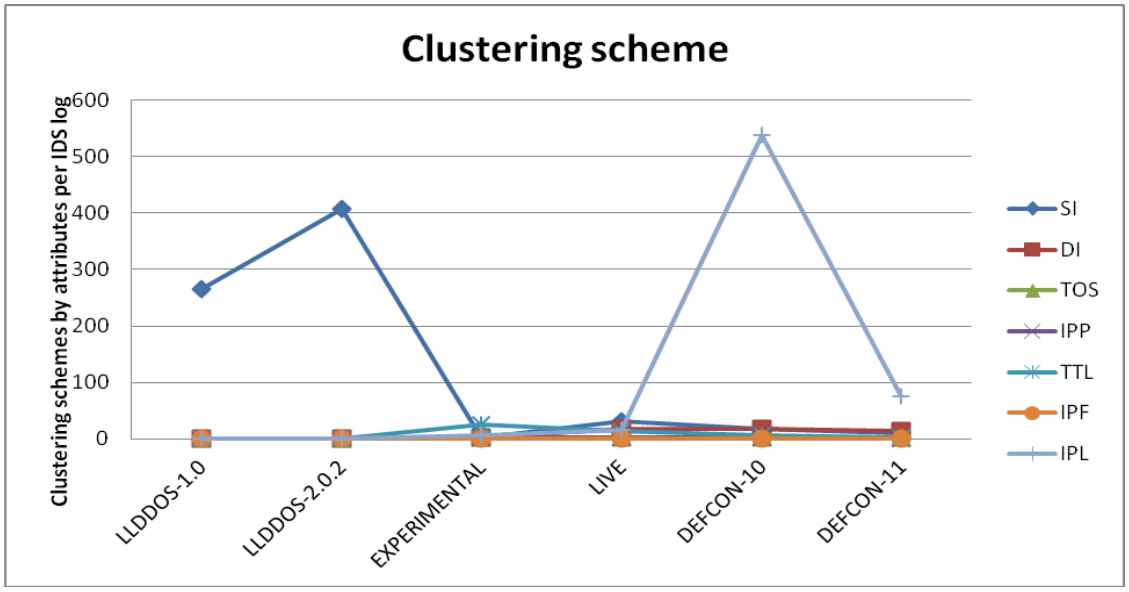


Figure 10: Clustering scheme by attributes per IDS log

In essence, the degree of the usefulness of the clustering schemes formed by the sources of the attacks within the LLDDOS-1 and LLDDOS-2.0.2 datasets suggest that both datasets generate the highest numbers of schemes whenever they are compared with other attributes of the other intrusion logs.

6.2. Discussions of the central intrusive themes

The above analyses have suggested the following central intrusive themes.

a) The uniformity of attacks or alerts within intrusion logs can be determined at sight: The degree of randomness of the contents of two or more intrusion logs can be determined to reduce skepticism and to enhance countermeasures given suitable statistical metrics like entropy.

b) ***Transferability and consolidation of countermeasures are plausible***: The plausibility for similarities to exist between two intrusion logs has been inferred above. Two intrusion logs that may be dissimilar in certain degree can still share closely related attributes

together in another perspective. Hence, analysts can consolidate their operations by reusing relevant and feasible countermeasures to improve the efficacies of measures to be taken to thwart intrusions in progress.

c) ***Predictability of quality, the degree of goodness or badness of intrusion logs is achievable***. Another central theme from the above analyses is that the quality at which an intrusion log is being predictable is plausible across certain attributes that detectors have used to describe the events within the intrusion log that is under review. Intrusion analysts can concurrently investigate some apparent discrepancies and characteristic properties that define the apparent individual nature of two intrusion logs that detectors log within the same or different computer and mobile networks.

d) ***Homogeneous attributes can be discovered within attributes of pairs of intrusion logs***. Network forensics of intrusion logs can reveal an attribute that is commonly shared across numerous attacks. Hence, the degree of regularity of the entire log measured by such attribute when it aggregates all the alerts in the log into a clustering scheme will be zero. Consequently, the above empirical studies shows that DDoS attacks as implemented in the LLDDOS-1 and LLDDOS-2.0.2 datasets form a clustering scheme by the values held in their Time to live (TTL), IPL and DI to cite a few. Conventionally, DDoS some analysts temporarily shut down the networks to thwart DDoS attacks. Better still, the above findings suggest that single countermeasure of highest efficacy can be designed using either TTL or IPL of the attacks to swiftly thwart. DDoS attacks rather than dealing with the attacks on the basis of their sources or targets.

7.0 Conclusion

Intrusion Detection System (IDS) is a strong-growing technical aspect of monitoring, gathering and reporting digital activities that have the possibility to endanger the security of computer and mobile systems. The mechanism of detection employed by IDS has quite a lot of benefits. In a standard setup, the toolkit uses several attributes to describe suspicious packets and record them as alerts in the form of intrusion logs. Nevertheless, advances over the years affirm that IDS logs and IDS alerts pose extra challenges to the analysts whenever the toolkit is operated to detect possible intrusions. Analysts face further challenges whenever two or more of such toolkits are also deployed within the same or different computer or mobile networks to maximize intrusion detections.

Fundamentally, intrusion logs are mostly used for litigations and for thwarting intrusions in progress over the years. Conventionally, alerts of intrusion detectors are often correlated and aggregated before analysts can make well-informed decisions about them. Nonetheless, this paper suggest that correlations and aggregation can fail to produce the desired results whenever multiple alerts do not possess

visible mutual, complementary, or reciprocal relationships. Hence, it is often difficult for most analysts to compare intrusion logs together.

Furthermore, most of the existing instances of applicability of intrusion logs are flawed given the sudden changes in the paradigms, semantics, bigness and analytical aspects of digital logs in recent years. For these reasons, two intrusion logs must be compared together. It is crystal-clear that there is further need for network forensics and investigators to determine the quality of intrusion logs for numerous purposes. Such determinations will help them to estimate the degree of goodness or badness of the intrusion logs under review. For examples, the measure of badness of intrusion logs help to reveal degree at which the logs are undesirable and the potential pains they can cause an organization. Also, this measure will unveil the level at which the logs are below ethical standards or expectations as of ethics or decency of network forensics.

Logs comparison with the motive of determining the degree of goodness of intrusion logs can help intrusion analysts to compare two intrusion logs to ascertain which of them is valuable, most informative or most useful. For the aforementioned reason, this paper presents a framework hat can be used to compare the quality of all alerts within two intrusion logs. The model does not partition log into two. Instead, the model uses a computationally fast method to compare two intrusion logs together by forming clusters on the basis of the values held in particular attributes of alerts within each log. Starting from the degree of predictability of both IDS logs, category utility and entropy are applied to respectively measure the quality of each log as a whole. Thereafter, series of evaluations are carried out and properly explicated using intrusion logs that are derived from synthetic and real datasets.

The results discussed above demonstrate degree of goodness and badness of six intrusion logs under reviewed. The method evolves degree of similarities and differences among the logs and some intrusive themes to guide network forensic professionals. Nevertheless, this paper has not investigated the nature, quality, extent and significance of various categories of virus attacks within several intrusion logs and across different time intervals. Hence, it is expected to pursue such research domain to enhance the quality of this paper in the nearest future research.

References

1. Bejtlich, R. (2013), *The Practice of Network Security Monitoring: Understanding Incident Detection and Response* 1st Edition, Addison-Wesley Professional

2. Capture The Flag Contest (CFC), 2016, Defaco datasets, http://cctf.shmoo.com/data/, Accessed 22/06/2016

3. Cover, T.M & Thomas, J.A., 1991, *Elements of information theory*, 2nd Edition, John Wiley and Sons, Inc., New York, USA.

4. Cuppens, F. & Miege, A, 2002, Alert correlation in cooperative intrusion detection framework, proceedings of IEEE symposium on security and privacy, May 2002.

5. Denning, D. & Neumann, P., 1987, An intrusión detección model, IEEE Transactions on Software engineering, Vol.13, No. 2, February 1987, pp. 222-232.

6. Debar, H. & Wespi, A, 2001, Aggregation and correlation of intrusion detection alerts, proceedings of international symposium on recent advances in intrusion detection, page 85-103, Davis, CA..

7. Fatima, L.S. & Mezrioui, A., 2007, Improving the quality of alerts with correlation in intrusion detection, International Journal of Computer Science and Network Security, Volume 7 No.12, December.

8. Ghorbani, A.A., Wei, L .& Mahbod, W., 2010, *Network Intrusion Detection and Prevention*, Springer,. ISBN: 978-0-387-88770-8

9. Han, J. & Kamber, M., 2006, *Data mining: concepts and techniques*, 2nd Edition, Morgan Kaufmann publisher, USA.

10. HM Government (HMG), 2016, National Cyber Security Strategy, United Kingdom

11. Jan, N.Y., Lin S.C. , Tseng, S.S. & Lin, N.P., 2009, A decision support system for constructing an alert classification model, Journals of Expert Systems with Applications February.

12. Kabiri1, P. & Ghorbani, A.A., 2007, A Rule-Based Temporal Alert Correlation System, International Journal of Network Security, Vol.5, No.1, PP.66–72, July.

13. Morin,B., Me, L., Debar, H. & Ducass, M.. 2002, M2D2,: A formal data model for IDS alerts correlation. In: Recent Advances in Intrusion Detection (RAID2002). Volume 2516 of Lecture Notes in Computer Science, Springer-Verlag, 115–137.

14. MIT Lincoln Laboratory. DARPA, 2016, Intrusion Detection Scenario Specific Data Sets, http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/2000data.html Accessed 14/06/2016

15. Nehinbe, J.O., 2011, Methods for using intrusion logs to reduce workload during investigations of intrusion logs, PhD thesis, University of Essex, UK..

16. Nehinbe, J.O., 2010, Guessing strategy for improving Intrusion Detections Computer Science and Electronic Conference (CEE2010), proceedings of IEEE, London, UK.

17. Nehinbe, J.O., 2011,. Time series analysis for forecasting network intrusion, proceedings of 10th IEEE International Conference on Cybernetics Intelligent System (CIS2011), London, UK

18. Scarfone, K. & Mell, P., 2007, Guide to Intrusion Detection and Prevention Systems (IDPS), Recommendations of the National Institute of Standards and Technology, Special Publication 800-94, Technology Administration, Department of Commerce, USA.

19. Shui, Y. (2014), *Distributed Denial of Service Attack and Defense*, Springer, ISBN: 978-1-4614-9490-4

20. Steiner, M., Barlet-Ros, P. & Bonaventure, O., 2017, *Traffic Monitoring and Analysis*, 7th International Workshop, TMA 2015, Barcelona, Spain, April 21-24, Proceedings, Springer International Publishing

21. Urko, Z. & Roberto, U., 2004, Intrusion Detection Alarm Correlation: A Survey, Computer Science Department, Mondragon University, Gipuzkoa Spain..

22. Valeur, F., Vigna,G., Kruegel, C. & Kemmerer, R.A., 2004, A Comprehensive approach to Intrusion Detection Alert Correlation, IEEE Transactions on Dependable and Secure Computing, Vol. 1, No. 3, July-September.

23. Yusof, R., Sulamat, S.R. & Sahib, S, 2008, Intrusion Alert Correlation Technique Analysis for Heterogeneous Log, International Journal of Computer Science and Network Security, Volume 8, No.9, September.

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25 , No 1, June, 2018

MEPLN E-LEARNING: MESH-BASED PEER-TO-PEER COLLABORATIVE E-LEARNING SYSTEM FOR LIMITED CAPACITY NETWORKS

O. E. Ojo[1], T. D. Ajayi[2], O. O. Orenuga[3] and A. O.Oluwatope[4]

[1,2,3]Federal University of Agriculture, Abeokuta
[4]Obafemi Awolowo University, Ile-Ife.

[1]ojoeo@funaab.edu.ng; [2]ajayitaiwo89@gmail.com;
[3]orenugaoluwafumilayo@yahoo.com ; [4]aoluwato@oauife.edu.ng.

ABSTRACT
The concept of collaborative e-learning scheme facilitates users' creativity and critical thinking abilities by creating an interactive learning environment. Although, the conventional e- learning architectures support server-based principle but the decentralized approach of peer to peer network provide better opportunity for collaboration in e-learning. In peer-to-peer (P2P) collaborative e-learning paradigm, both the trainer and the trainee can act as provider and consumer of knowledge simultaneously. Furthermore, utilization of e-learning system is very poor in most Institutions in developing countries due to poor facilities. The scalable and affordable features of P2P network make it more suitable for e-learning technology in limited capacity networks environments typical in Nigeria while maintaining quality services. To this end, we propose an effective mesh-based P2P collaborative e-learning system named MEPLN e-learning suitable for proper peer management and resource allocation in limited capacity networks using Federal University of Agriculture, Abeokuta, Nigeria as a case study. MEPLN overlay topology was formalized using temporal logic; it was further verified and validated using Simulink design verifier in MATLAB simulation tool. The MEPLN e-learning scheme was then implemented using PHP, Java and MySQL programming languages.

Keywords: Collaborative e-learning, peer-to-peer networks, limited capacity Networks.

1.0 INTRODUCTION

The rapid growth of Internet in the last two decades had directly impacted learning systems across the globe [8]. Computer-supported collaborative learning or collaborative e-learning is an integral part of the learning sciences related to studying how people can learn together with the help of computers [18]. Collaborative e-learning facilitate dialogue among students and their

lecturers with the aid of information and communication technology [20]. It supports a faster learning curve, since students can interactively customize their learning and have more control of their learning process [5]. The concept of students' engagement in peer review using collaborative e-learning has gained increased attention in several universities in developed countries. This can be attributed to the growing focus placed on collaborative learning as an aspect that supports student learning [23]. Students are provided with an opportunity to reflect on their peer's work as well as their own, thus reinforcing key learning [7, 19].

1.1 E-Learning in Peer-To-Peer Network

Research has shown that most learning systems were implemented either with client-server architecture or centralized server based [14]. Although, the centralized server approaches are metaphors of student-teacher and repository centric which reflect real world learning scenarios in which teachers act as the content producers while students act as the content consumers [6]. However, these approaches differ from the main principle of collaborative e-learning; thereby posing some challenges in adapting it to collaborative e-learning.

The advent of peer-to-peer technology offers decentralized approach which paved way for easy sharing of large volumes of data among peers without requiring high resources as compared to centralized server approaches. This brings flexibility and efficiency in network-based learning and distance education, thus encourages collaboration. Many peer-to-peer environments allow users to browse the shared files directory on another peer and this enables efficient, more effective and synchronous learning environment for collaborative e-learning [4].

Peer-to-peer networks unlike client-server and centralized approaches make each peer play as both client and server [1]. The main advantage of peer-to-peer network over client-server network is their democratic nature which allows freely sharing of opinions and any available contents [9].

The rest of this paper is organized as follows. Section 2 presents related works, the proposed collaborative e-learning P2P system is described in section 3. Section 4 discussed the implementation results and the conclusion is presented in section 5.

2.0 RELATED WORK

Over a decade now, researchers had employed collaborative e-learning scheme as an active tool for easy interaction and communication. An agent-based collaborative system to support extra-class interactions among students and teachers was presented in [21]. The system achieved rich collaborative environment.

Also, [11] proposed a two-dimensional taxonomy in which types of P2P application and knowledge networks are combined and identification of e-learning supporting models based on this taxonomy. Thus, the proposed taxonomy can be used as a decision making tool.

Another researcher studied the use of peer-to-peer networks in combination with collaborative learning. It was observed that peer-to-peer technology presents a better solution in learning environments because the architectures of peer-to-peer networks and collaborative learning are similar [3].

The potential of P2P technology for collaborative learning capabilities were also explored and a scheme for collaboration within a social network was designed [2].

An investigation was carried out in [17] to test whether P2P technology could provide an improved platform for virtual classrooms and laboratories (VCL) and e-learning settings. The results from the investigation revealed that P2P paradigm offers an enhanced platform where institutions can fulfil instructor and student needs within VCL and

deliver advanced features compared to other platforms.

Furthermore, a distributed event-based awareness approach for P2P groupware system was modelled.

. The model focuses on supporting group activities in an academic setting and provides different forms of awareness through a set of interoperating, low-level awareness services. The system achieved good performance and scalability [15, 22]. Despite the rapid growth of e-learning which allows personalization of contents and building of user profiles based on the learning behaviour of each individual user [10], developing countries like Nigeria are still lacking behind due to limited networks experienced by end-users [12]. Hence, this research presents a P2P collaborative e-learning system in an attempt to provide good quality of service and quality of experience to end-users using limited resources.

3.0 MEPLN E-LEARNING SYSTEM

A new collaborative e-learning system called mesh-based P2P collaborative e-learning systems for limited capacity networks (MEPLN e-learning) was formulated. MEPLN e-learning system as shown in Figure 1 consists of the DB (database), lecturer module, admin module and student module. The database is the repository for information on all the courses offered over a particular period. The lecturer module allows the course lecturers to upload contents that can be easily accessible by their students.

It also accommodates feedback from various students. The admin module serves as an intermediary between the lecturer and the student modules. It ensures information is disseminated appropriately to the right channel, the administrator have access to create, update and delete any record of user. This scheme was designed to accommodate both the lecturers and student representative as administrator which actually depict peer-to-peer system.

Once, the content is uploaded in to the database from the lecturer or admin module, it is transferred directly to the student module. In the student module, each peer (student) can exchange content with its nearest neighbour without any obstruction. This scheme is expected to provide adequate knowledge to learners in academic environment in low capacity networks like Nigeria using decentralized shared resources.

3.1 MEPLN Overlay Topology

The overlay topology plays significant role in P2P networks; it ensures even distribution of resources to all peers within the network. Tree-based (TB) and mesh-based (MB) are the most popular of P2P overlay topology. TB is a structured topology where the children peers depends on equivalent parent peers for resources, however, failure at the parent peer can lead to total collapse of the system. MB approach is an improved version of TB topology with unstructured peers' arrangement, formed by peers connecting to neighbours. Although, MB has proven to address the problem of rigid and tightly controlled placement policy in TB but effective and fair distribution of resources without unnecessarily delay is still challenging [13]. The MEPLN topology is a modified MB P2P topology as displayed in Figure 1 (student module). In MEPLN topology, peers are classified based on the resources (like upload/download bandwidth) they are capable of donating into the network, thereby, minimising end-to-end delay.

For each network session, the available peer within the network with the highest resources is made the super-peer while others are classified as sub-peers. The super-peer is capable of downloading content at a faster rate and exchanges the content with sub-peers; it gets information directly from the lecturer or admin module and then transfers it to the sub peers. Since, it is a mesh topology; the sub peers can also exchange information with its nearest neighbour.

3.2 Model Checking for MEPLN Topology

Model checking is the formal process through which a desired behavioural property (the specification) is verified to hold for a given model via an exhaustive enumeration either explicit or symbolic [16]. Here, the model checking of MEPLN topology is presented with an attempt to check for the consistency and completeness. MEPLN model specifications were formulated using computation tree logic (CTL). The definitions for the CTL operators are given in Table 1.

Table 1: Computation tree logic Operators

Symbol	Meaning
AG	(everywhere - along all paths)
EF	(everywhere - along some path)
AF	(somewhere - along all paths)
AX	(all successors)
A [ψ1 U ψ2]	(until – along all paths)
→	Implication
¬	Negation
∧	Conjunction
∨	Disjunction

The specifications for MEPLN topology are given as follows:

1. If peer in MEPLN is active, then peer can send or receive frames: *AG (Active → Send ∨ Receive).*
2. If peer in MEPLN is active, then peer can send and receive frames: *AG (Active → Send ∧ Receive).*
3. Whenever a peer in MEPLN is active, eventually the peer will send or receive frames: *AG (Active → AF Send ∧ Receive).*
4. If peer-classify is true in MEPLN, then if high-bandwidth(Hband) is true in any subsequent state, then super-peer will eventually become true until Hband is false : *AG (Peer-Classify → AG (Hband → A (¬ Hband U SuperPeer))).*

3.3 Verification of MEPLN Topology

All the specifications presented in section 3.2 were tested using Simulink design verifier within MATLAB R2015a simulation tool. These

specifications were verified for accuracy and validity with Boolean variables 0(false), 1(true), and values between 0 and 1 (degree of accuracy). The output results displayed using command windows. The results of verification for specifications (1-3) showed peers behaviour in the MEPLN topology for sending or/and receiving contents as displayed in Figure 3; it ascertains that the probability of peers sending content while receiving is 50%. On the other hand, the chance of peers receiving contents while sending is approximately 60%. However, few consistent interval transmission failures were also observed. This suggests that it is easier for peers to complete the receiving process before sending to neighboring peers. Furthermore, the probability of classifying a particular peer as super-peer in the MEPLN at a given time as stated in specification 4 gives 100% validation as shown in Figure 4. This shows that the MEPLN is effective for peers' placement.

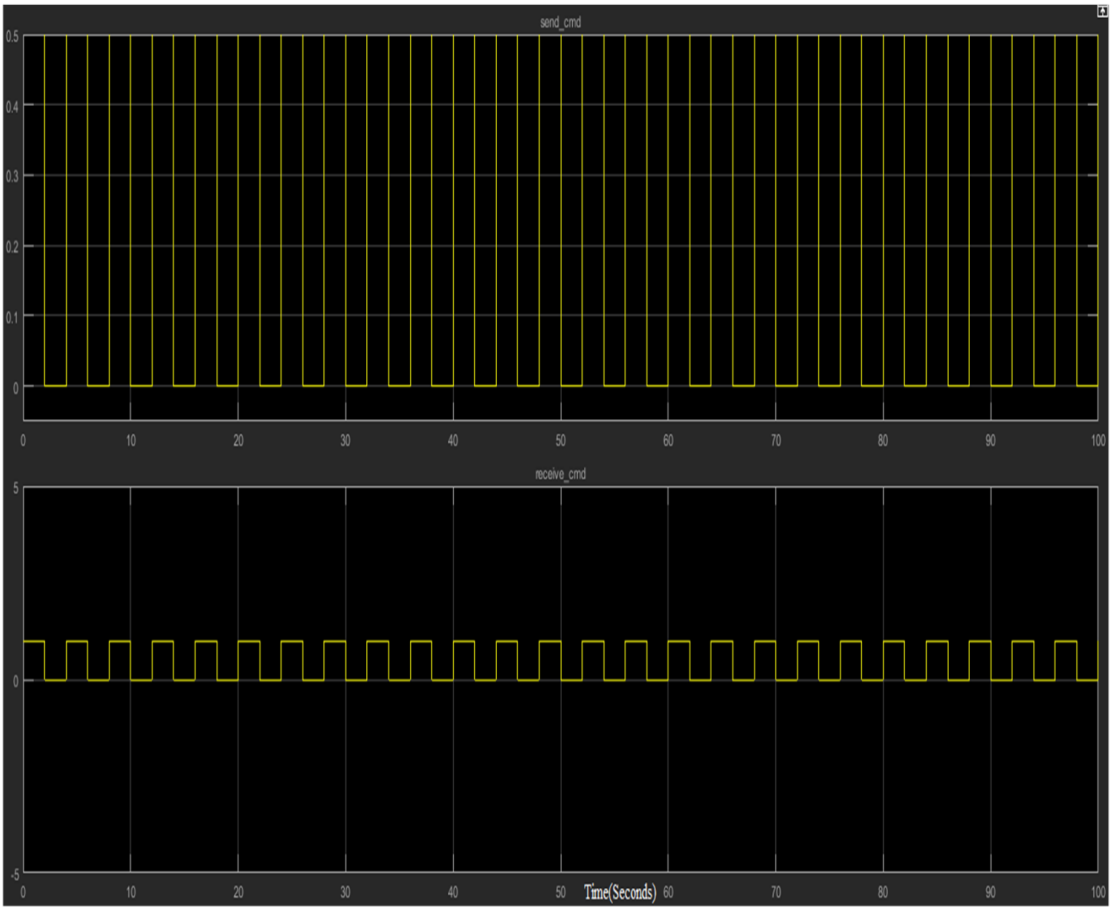


Figure 3: Peer Behaviour (Send and Receive)

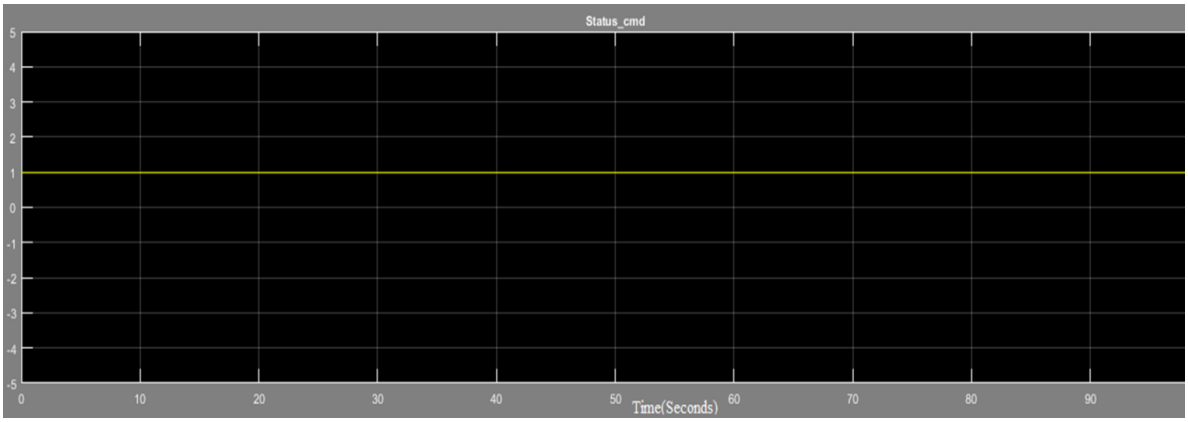


Figure 4: Peer Classification

4.0 IMPLEMENTATION

The prototype of MEPLN e-learning system was implemented on window operating system using PHP, Java and MySQL programming languages.
NetBeans IDE provides the user with the interface while MySQL WAMP SERVER database was used in developing the software. The program made use of web forms, which makes it easy to use, and protect it against accepting invalid data. The administration of modules was not restricted, both the students and lecturers can upload and download materials and thus depicts a peer-to-peer system. The front page of the MEPLN e-learning is displayed in Figure 5.

Figure 5: Login Interface

Users or peers can access the e-learning system through registered user name and password. Also, new users can as well register by clicking on sign up. A sample of lecture registration interface is shown in Figure 6 and the student registration interface is displayed in Figure 7. The sample of uploaded Interface which is easily accessible to the students that registered for the course is shown in Figure 8. Finally the sample of the database is given in Figure 9.

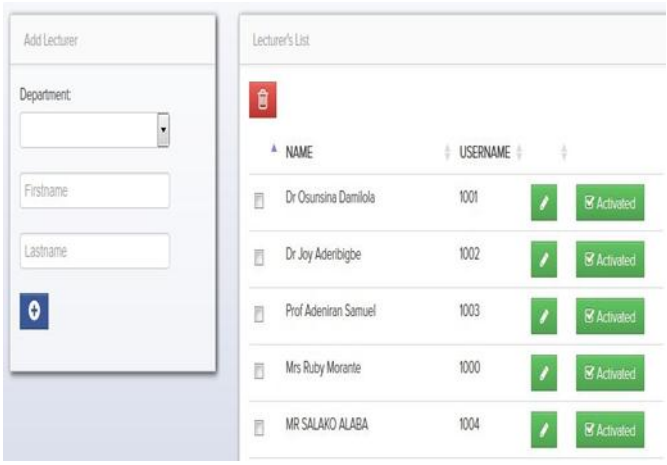


Figure 6: Lecturer Registration

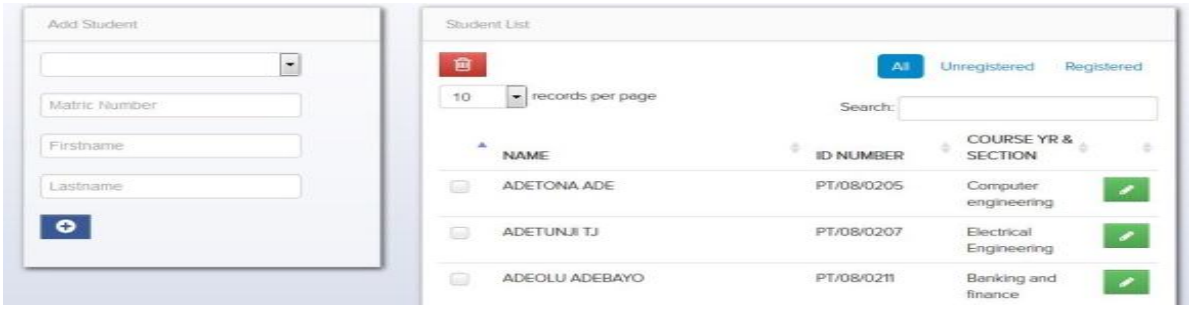


Figure 7: Student Registration Interface

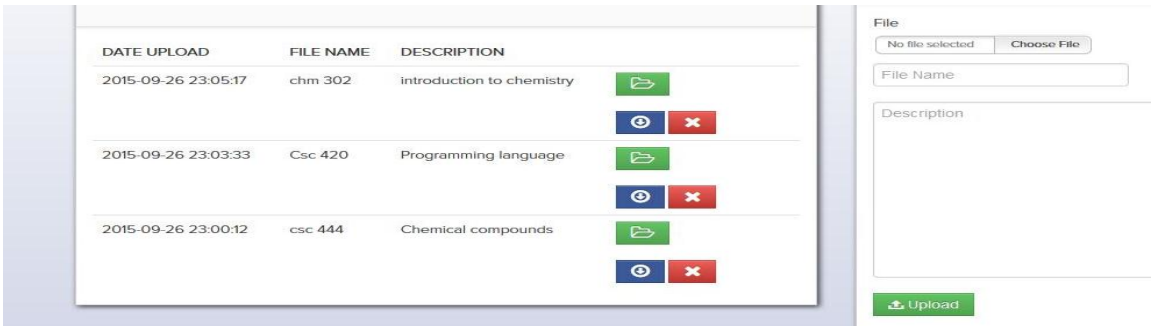


Figure 8: Upload Interface

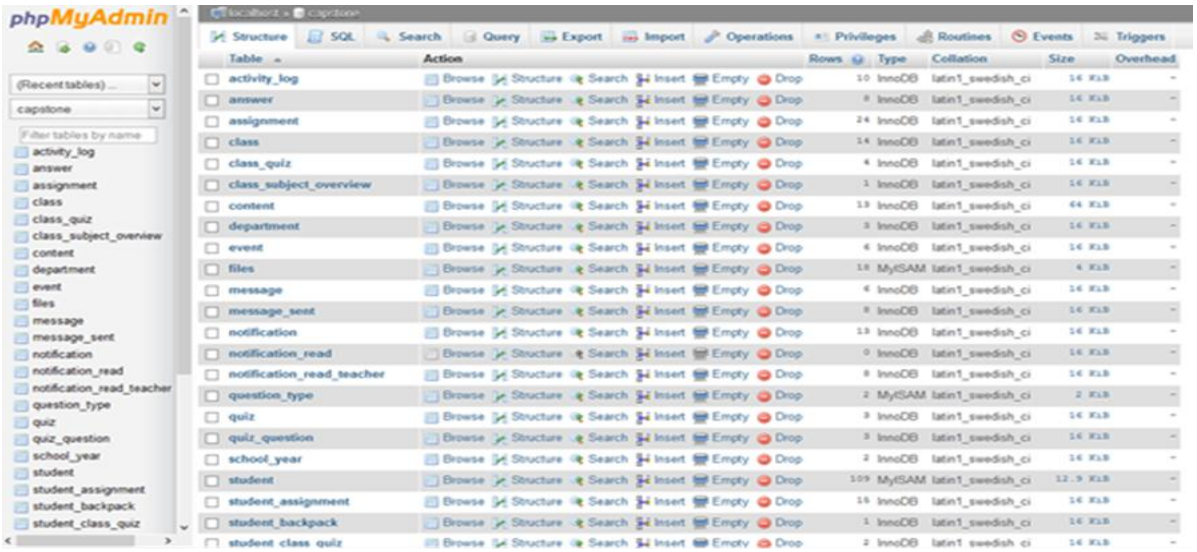


Figure 9: Sample of Database

5.0 CONCLUSION

This work presents a mesh-based peer-to-peer collaborative e-learning system for limited capacity networks (MEPLN e-learning). MEPLN e-learning was designed to ensure effective deployment of e-learning system in Nigeria universities. This

scheme was designed to work effective in both offline and online platforms. The overlay topology of MEPLN e-learning was design to accommodate sharing of resources among users within the network at a particular time by including peer classification mesh-based P2P technology thereby reducing the workload at the server end.

Also, MEPLN topology was formalized using temporal logic function to check for consistency. The topology was further verified and validated using Simulink design verifier in MATLAB simulation.

The system was implemented using specific programming languages. The testing of prototype revealed that the MEPLN is suitable for limited capacity networks. MEPLN e-learning system can be extended to cloud computing environment. The scheme can also be tested with video contents (video on demand or live video streaming).

REFERENCES

[1] M. Aberer, K. Punceva, M. Hauswirth and R. Schmidt. Improving Data Access in P2P Systems. IEEE Internet Computing, 6 (1), 58-67. 2002.

[2] R. Bostrom, S. Gupta and J. Hill. P2P technology in collaborative learning networks; applications and research issues, Interactive Journal knowledge and learning 4(1):36-57. 2008

[3] J. Biström. Peer-to-peer networks as collaborative learning environments. In HUT T-110.551 seminar on internetworking.2005

[4] C. Chang and J. Sheu. Design and Implementation of Ad Hoc Classroom and e-Schoolbag Systems for Ubiquitous Learning. IEEE International Workshop on Wireless and Mobile Technologies in Education, Växjö, Sweden (pp. 8-14). 2002.

[5] V. Cantoni, M. Cellario and M. Porta. Perspectives and challenges in e-learning: towards natural interaction paradigms. Journal of Visual Languages and Computing, 15(5), 333-345. 2004.

[6] Z. Cheng, S. Shengguo, M. Kansen, T. Huang and H. Aiguo. A Personalized Ubiquitous Education Support Environment by Comparing Learning Instructional. 19th International Conference on Advanced Information Networking and Applications, Tamkang University, Taiwan. (Vol. 2, pp. 567-573). 2005

[7] S. Cassidy. Developing employability skills: Peer assessment in higher education. Education and Training, *48*(7), 508-517. 2006.

[8] C. Greenhow and B. Robelia. Informal learning and identity formation in online social networks. Learning, Media and Technology, 34(2), 119-140. 2009.

[9] Q. Jin. Learning community meets peer-to-peer networking: Towards an innovative way of person-to-person E-learning support. Proc. of Knowledge Economy and Development of Science and Technology. (pp. 1-8). 2003.

[10] A. Klašnja-Milićević, B. Vesin, B. M. Ivanović, Z. Budimac and L. Jain. Introduction to E-Learning Systems. In E-Learning Systems, Springer International Publishing. (pp. 3-17). 2017

[11] M. Lytras, A. Tsilira and M. Themistocleous. P2P Knowledge Network and E-learning Clusters

of application, A Conceptual and Technological Approach of Potential Business Value. Ninth Americas Conference on Information System (AMCIS) proceeding. (pp. 1782-1791). 2003.

[12] O. Olayiwola and A. Oluwatope. Packet-level Simulation of Real-Time Video in Slow-Speed environment using NS-3. Proceeding of the African Conference, International Association of Science and Technology for Development, Bostwana. (pp.148-153). 2014.

[13] O.Ojo, A. Oluwatope and O. Ogunsola. UStream: Ultra-Metric Spanning Overlay Topology for Peer-to-Peer Streaming Systems. IEEE International Symposium on Multimedia, USA. (pp. 601-604). 2015

[14] J. Peng, D. Jiang and X. Zhang. Design and implement a knowledge management system to support web-based learning in higher education. Procedia Computer Science, 22(2013): 95-103. 2013

[15] A. Poulovassilis, F. Xhafa and T. O'Hagan, Event-based Awareness Services for P2P Groupware Systems. Informatica, 26(1):135-137. 2015.

[16] K. Rozier. Linear Temporal Logic Symbolic Model Checking. Computer Science Review, 5(2), 163-203. 2011.

[17] H. Rajaei and N. Hakami. P2P Grid Technology for Virtual Classrooms and Laboratories. In Proceedings of 16th Communications & Networking Symposium, (pp. 11). 2013.

[18] G. Stahl, T. Koschmann and D. Suthers. Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), Cambridge handbook of the learning sciences. Cambridge, UK: Cambridge University Press. Available at http://GerryStahl.net/cscl/CSCL_English.pdf, (pp. 409-426). 2006.

[19] S. Sahin. An application of peer assessment in higher education. The Turkish Online Journal of Educational Technology, *7*(2), 5-10. 2008.

[20] H. So and C. Bonk. Examining the roles of blended learning approaches in computer supported collaborative learning (CSCL) environments: A Delphi study. Educational Technology and Society, *13*(3), 189-200. 2010.

[21] G. Wagner, L. Aroyo and R. Guizzardi . Agent oriented modeling for collaborative environment, A P2P Helpdesk case study. http://gnutella.wego.com. 2002

[22] F. Xhafa and A. Poulovassilis A. Requirements for distributed event based awareness in P2P groupware system. Advanced Information Networking and Applications Workshops (WAINA), IEEE 24th International Conference (pp. 220-225). 2010.

[23] C. Zhu, M. Valcke, T. Schellens and Y. Li. Chinese students' perceptions of a collaborative e-learning environment and factors affecting their performance: Implementing a Flemish e-learning course in a Chinese educational context. Asia Pacific Education Review, 10(2), 225-235. 2009.

THE JOURNAL OF COMPUTER SCIENCE AND ITS APPLICATIONS
Vol. 25 , No. 1, JUNE, 2018

AN INFORMATION TECHNOLOGY (IT) EQUIPMENT FAULT PREDICTION SYSTEM FOR AN IT SUPPORT UNIT

P. S. Adamade[1], E. O. Nwachukwu[2]
[1]petersiad@yahoo.com, peter_adamade@uniport.edu.ng -Masters Degree student ;
[2]enoch.nwachukwu@uniport.edu.ng -Prof. of Computer Science.
[1,2,] Department of Computer Science,
University of Port Harcourt. Rivers State
Nigeria.

ABSTRACT

A common expectation for high-end customers in most system operated environment is that systems must never fail, hence activities should be seamless. Information Technology components (hardware and software) are inherently prone to failure.
Though these systems rarely "crack", hardware components can still fail causing software running on them to fail as well. These fault situations have high cost to management and to customers who may experience poor service. If a fault can be predicted, preventive action can be taken to mitigate the pending failure. This research work aims at developing a novel framework that applies machine learning and probability theory based on data mining techniques using significant amount of captured IT equipment fault log data from a central server to predict faults. Statistical test based on Naive Bayes and K-Nearest Neighbour classifiers were used and implemented using Rapid miner running on java programming language. The results obtained were models showing good prediction results with accuracy of 83% and 27% respectively indicating substantially, that applying data mining in equipment fault prediction is possible with datasets features that are best fit.

Keywords: Fault, machine learning, data mining, Naive bayes, k nearest neighbour(knn).

1.0 INTRODUCTION

[14] In the Proceedings of the World Congress on Engineering and Computer Science, in a paper titled "A Prediction System Based on Fuzzy Logic" opined that Prediction of an event requires vague, imperfect and uncertain knowledge. Complexity in a prediction system is its intrinsic characteristics with various techniques used to achieve the prediction.

Manually formed rules were used to build prediction systems in the early times, but with increase in the number of inputs, predicting an event became complicated and difficult. Engineering assisted in building prediction system that could adapt to the increasing number of inputs and framed rules with higher and better accuracy and speed superior to former.

Complex systems suffer stability problems due to unforeseen interactions of overlapping fault events and mismatched defence mechanisms. Hackers and criminally minded individuals invade systems, causing disruptions, misuse, and damage. Accidents result in severed communication breakdowns sometimes affecting entire regions. Natural occurrences like light sun outage, tsunami, earthquakes and landslides greatly affect communication equipment across the entire earth leading to loss of man hours and lots more. To guarantee that high confidence systems will not betray the intentions of their builders and the trust of their users are systems that are fault tolerant and predictable.

In this work, we are focusing on post-deployment failures, rather than fault densities since these failures are experienced by end users and affects reliability and work flow more directly. Fault prediction in IT equipment will be narrowed down to a specific fault user's encounter with computer system and networks during normal operations with the system.

2.0 Related work

It is not surprising that a lot of research work have been put into developing techniques to collect, study and mitigate malicious code. Some researchers have used static based approach, machine learning, controlled called graph to analyse malware. A static based approach was used to detect malware by applying ordered of weighted averaging [7]. The method uses a parameterized family of aggregate operators to select prominent features from malware

Network-based techniques have also been used to monitor the traffic produced by some categories malware [20]. The problem with network-based techniques is that since it relies on the traffic produced by the programs, it will be difficult to observe the activities of the virus directly. Also, network based techniques cannot identify malicious program that does not sends or receives data [5].

Signature-based detection have enjoyed commercial success today, nearly all anti-malware uses signature-based [13, 19]. A signature is a sequence of bytes that is present within a malicious executable and within the files already infected by that malware [21]. In order to devise a signature, expert usually collects information from an infected computer or network to determine a file signature for a new malicious executable. Using this approach, suspected files can be analysed by comparing its signature bytes with the list of already known signatures. If a match is found, the file under test will be identified as a malicious executable. This approach has proved to be effective when the threats are known beforehand. Several issues have rendered signature-based less reliable. It cannot cope with code obfuscations and cannot detect previously unseen malware [21, 8, 4].

Different techniques have been proposed to deal with unknown malware which signature-based cannot identified. One of them is the data-mining based approach, data-mining based approach uses dataset that combine both the characteristic features of malicious samples and benign samples. It uses this, to build classification tools that can detect un-seen malware.

The idea of applying data-mining models to the

detection of different malicious programs based on their respective binary codes was proposed by [22]. In [2], the authors presented an approach to detect malware using evolving clustering method for classification. The shortcoming with most of the data mining technique is that the domain used by most of the authors does not cover all the different types of malware. The use of Opcode Sequence for unknown malware detection was proposed by [21]. The authors uses data-mining algorithm to train the selected Opcode in order to detect unknown malware.

To capture the intrinsic properties of malware researchers now consider disassembling infected file for identifying malware features that are not detected ordinarily by normal scan [3, 26]. Window application programing Interface (API) calling sequence have also been used by [16, 27], to indicate behaviour of malwares attacks on PE files. A graphical-based mining approach was used by [7], to detect an unknown malware. In that approach the system disassembles the PE-file to a standard assembly code and then extracts the Control flow graph from the assembly code. The limitation with this approach is that it consumes a lot of processing time and also the issue of graph isomorphism is a known NNP complete

A new method for detecting an unknown malware that generate fuzzy rule from an evolving clustering technique was used to detect malware [2]. The approach effectively combined the information gain as a feature selector with an evolving learning classifier. The main challenge with this technique is that it cannot detect any malware that falls outside the cluster.

3.0 Proposed Approach

[11],[12],[13]Some principles or techniques in fault prediction are;

i. Multivariate state estimation technique-otherwise known as MSET utilizes multiple parameters to monitor results in predicting.

ii. Trend analysis (data mining)-This method of analysis allows traders to predict what happens to stocks in the future. It is based on historical data relating to performance given the overall trend of the market and indicators within the market.

iii. Dispersion frame technique (DFT)-This technique measures interval time between successive error events.

iv. Reliability based technique-It uses the mean values of the random system parameters as design variable and optimizes the objective function subject to predefined probabilistic constraints.

v. Artificial intelligence-AI looks at intelligence displayed by machines in contrast to the natural intelligence showed by humans.

vi. Expert system-It's considered to be a piece of software that uses databases of expert knowledge to offer advice or make decisions.

[12] The system analysis will be based on probability of naturally-occurring faults in the past which is among the most accurate ways to achieve insights into the fault behaviour of a system in the future.

Event logs have been widely adopted over the past decades where analysis of failure data in the log provides valuable information on classes and allows pinpointing dependability bottlenecks. The architecture of the fault prediction model is based on three main modules;

(i) Collection,

(ii) Filtering, and

(iii) Analysis of entries in the log

3.1 Collection: User logs of IT related issues have been collected. The logging mechanism allows users to state what their IT challenges were, state their location, specific location, classification, description of the issue and the responding part has the line supervisors and those on the responsible party escalating path and the logger will be alerted via emails. This alert will among other things indicate the maximum time the issue is expected to be

resolved.

3.2 Filtering: Given a large volume of data collected in the system, a crucial step of each log-based measurement study is determining who the responsible party is.

3.3 Analysis of the entries: Given the sieved volume of data collected from the process of filtering, the resulting models are applied to the datasets for the purpose of prediction and knowledge discovery.

The concept overview highlights the sequential relationship among them. Once an event log is collected from the target system, filtering procedures make it possible to infer failure data from the event log.

Finally, failure data are analyzed to characterize properties of interest of the system. Major tools and techniques adopted to manipulate the data at each step of the methodology.

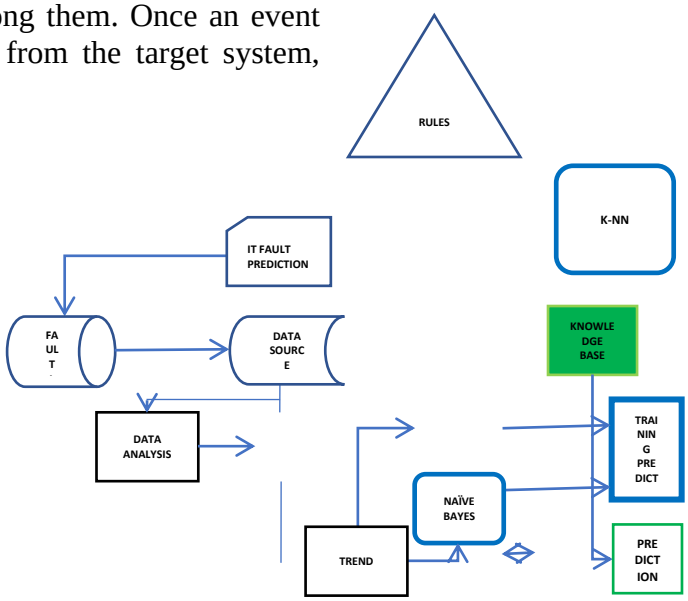


FIGURE 1:

ARCHITECTURE OF IT EQUIPMENT FAULT MODELS

3.4 Machine learning - The proposed system involves the use of machine learning and probability theory applied on historical datasets in the prediction of fault in an IT environment.

[7] "Model frameworks are a combination of "Naïve Bayes Network Technique and K-Nearest Neighbour Technique" which employs an objective supervised learning approach".

In this system, the relative probabilities and the distances of the features is used to obtain the classified data which in turn will be used for the prediction.

The aggregated dataset retrieved from the central portal over a period of time and the "weight of each statistic is determined prior to making a prediction" which involves determining it's probabilities and nearest neighbour via modelling.

3.4.1 Breakdown of proposed system

Steps employed in the proposed system;

Step 1: Problem definition

Step 2: Data collection and pre-processing

Step 3: Modelling

Step 4: Training and applying of Models to unseen data

Step 5: Performance evaluation

Step 1: Problem definition -At this level, the problem will be define considering the kind of data available from the industry.

Step 2: Data collection and pre-processing - For the purpose of this system, industrial domain historical data will be retrieved from a central sever located at the researchers' office international headquarters in Lagos Nigeria. The dataset located on a server named ntops2 will be downloaded using ftp protocol as a tool for the commencement of pre-processing activities with Microsoft excel to acquire the right features.

The purpose of pre-processing is to determine the related features to be used in the prediction. Pre-processing of data will involve critically analysing the datasets by replacing missing values, selecting roles and attributes for the purpose of transformation and use by the model. Select attribute operator is introduced to the "main process" which is connected to "read excel" and to the example operator. when executed a result of the example sets of attributes is displayed.
Step 3: Modelling - The modelling phase is sub-divided into build and execute models.
Build: In the build phase, the "Naïve Bayes and K-Nearest Neighbour techniques" are brought or introduced to the main process area.
Execute: These enables the running of the process containing the models and other parameters that provides the expected outcomes

3.4.2 Naive Bayes

It helps in reducing complexities "by making conditional independence that dramatically reduces the number of parameters to be estimated when modelling". it looks at the probabilistic relationship among set of variables.

Algorithm: Naïve Bayes

Input:

Attributes X1, X2…. Xn

Naïve Bayes process:

P (X | ϑ , S) = Y n i=1

P (Xi | pai , ϑ i , S)

Output:
Y= Dependent variable representing resulting weights of the attributes.

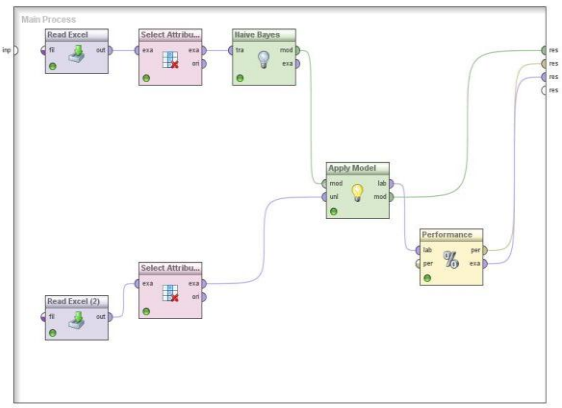


Figure 2. Naive Bayes Main Process

3.4.3 K-Nearest Neighbour (K-NN)

[8] K-NN is gotten by considering each of the characteristics in our training set as a different dimension in some space, and take the value an observation has for this characteristic to be its coordinate in that dimension, so getting a set of points in space. We can then consider the
similarity of two points to be the distance between them in this space under some appropriate metric".

Input (Training set):
This provides for the features used in the model Ticket, Sol Id, Name, State, Region, Country, HotSpot, Class, ClassType, LogType, Notification, Responsible, Start Time, Time Closed and Closed By
Output:
Predicted results for unseen fault datasets

Step 4: Training and applying of KNN Model to unseen data
Step 5: Validation
Evaluating the model for performance means to estimate the model whether it meets the expectations or not. If the

model does not fit the original expectations, they go back to the modelling phase and rebuild the model by changing its parameters until optimal values are achieved. To calculate the performance of the model, the accuracy must be obtained. To calculate accuracy, Accuracy: Acc = (Number of correctly classified examples / No of examples) X 100

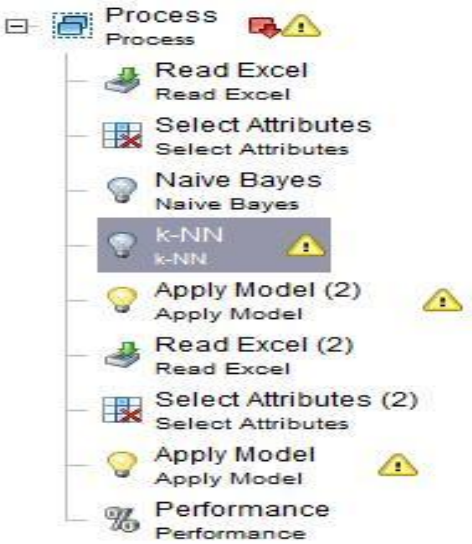


Fig.3 Tree view of Main Process(Naive Bayes and K-nn)

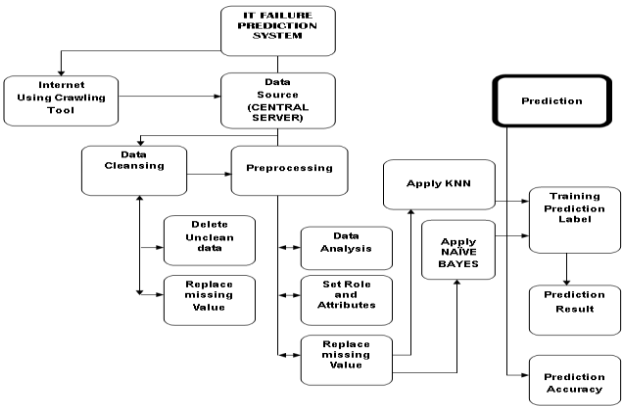


Fig. 4 Implementation algorithm

3.6 RESULT

In the analysis of the fault log data pooled over 90,000 datasets, the results obtained significantly shows that using relevant features provides prediction of fault in IT environment. The wrongly predicted represent poorly classified data by the KNN and Bayesian models. The Bayesian network also prevents noise from the datasets which justifies the high prediction accuracy obtained.

The result also shows how effective Naïve Bayesian network manages text datasets and can be used in text mining.

Model Prediction Accuracy:
When the testing set was applied to the Naïve Bayesian model, a prediction accuracy of 83%, which is substantially higher than the output when the prediction sets were applied to the K-NN.
This indicates that improved prediction data, using the best fit attributes would lead to more accurate predictions.

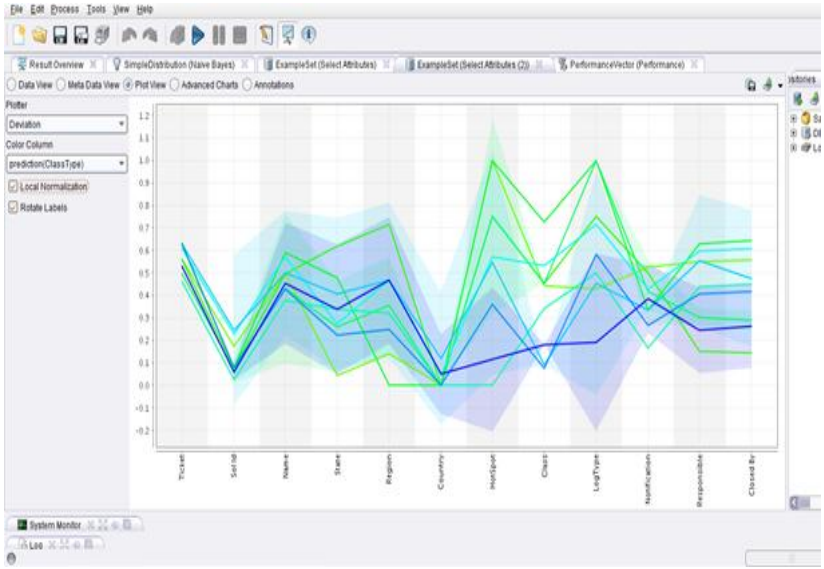


Figure 5. Result Chart (Naive Bayes and K-nn)

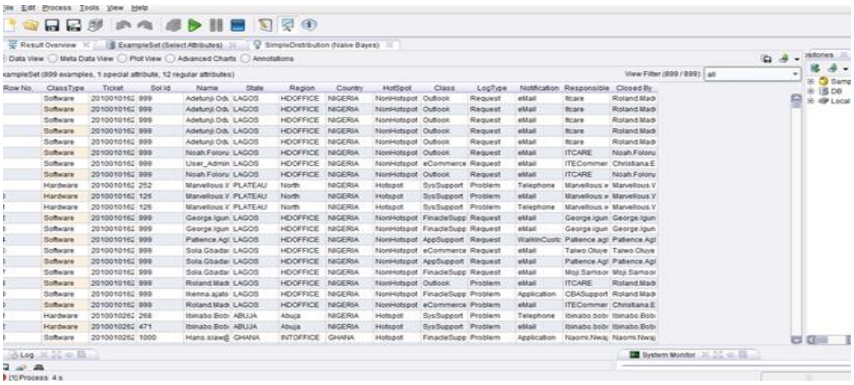


Figure 6. Naive Bayes dataset

3.5 RESULT COMPARISON

Model – NAÏVE BAYES	**Model – K-NN**
YIELDED Prediction Accuracy: 83%	YIELDED Prediction Accuracy: 27%
Intelligent learner	Lazy learner
Classifier returns probabilities	Classifier returns K- distance of the nearest neighours
Naive Bayes classifiers are computationally fast when making decisions as training of datasets was completed in 1sec	K-NN classifiers Are computation-ally SLOWER than the NAÏVE BAYES when making decisions. Training of datasets was completed in 3sec
Learning done by statistical inference	Learning done by analogy
Uses Laplace correction	Uses K = 1 – N
Manages large datasets efficiently	Poor with large datasets
Resulting model smoothened	Resulting model noisy
Efficient in handling text data sets	Poor in handling text data set

4.0 CONCLUSION

System methodology used is "Rapid Prototyping" which entails going through the system developmental cycle of analysis, design, implementation, and post-delivery maintenance with the ability to change requirements at the various stages except the retirement stage where the excellence of necessity and specifications was achieved with better and amplified user participation.

In this final chapter, a total recap of the research work was done with reference to the earlier chapters with thorough synopsis on the complete investigation of the set goals and objectives. It further elucidates how this research work contributes to knowledge and ascertains projections for further research.

The Prediction System explores the use of machine learning techniques in the field of predictive analysis (IT equipment fault log). Driven by the irresistible increase in the repository of existing data in the domain, datasets were collected, data mining and machine learning techniques were productively used in diverse aspects of the research work.

Predictive methods that empowers us to discover out more treasured datasets to improve the overall performance and to make exact choices. In several incidences, predicting the outcomes of procedures has always been a challenge and in some cases a rewarding experience. Consequently predicting problem reveals an increasing need to conduct experiment in this area.

Historical comprehensive and statistical data have been reserved to assist the proceedings in the work environment. The members of staff, the equipment and the fault log all presents various forms of these statistical evidences over a period of time This pool of information from the refined datasets will keep inspiring different assemblages, ranging from public domain, statisticians and information technology enthusiasts to discover embedded discrete knowledge in it.

The work can be continuously reused by future research work and finally, can be well-thought-out as a fruitful examination and survey that can provide a respectable spine for further studies.

ACKNOWLEDGMENT

We hereby acknowledge all those that provided the platform for this study all of which are appropriately referenced.

REFERENCES

[1] Mahmoud O. E., Mojeeb A. R., Al-Khiaty and Mohammad A., (2013). An Exploratory Case Study of aspect-oriented metrics for fault proneness, content and fixing effort prediction. *International Journal of Quality and Reliability Management Vol. 30, No. 1, pp 80 – 98.*

[2] Ahmad A., Iman P. and Min T., (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. *International Journal of Advanced Computer Science and Applications,* Vol. 4, No. 11, pp 33-38

[3] Amornchai S., Suphakant P. and Chidchanok L., (2009). Tennis Winner Prediction based on Time-Series History with Neural Modeling. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 127-132

[4] Sayanta M., and Gaétan H., (2011). Resource Prediction Model for Virtualization Servers *Proceedings of the First International Conference on Information and Communication Technologies for Sustainability ETH Zurich,* ISBN I978-3-906031-24-8, DOI 10.3929/ethz-a-007337628

[5] Damir M., Koraljka B. and Davor S., (2004). Fault Analysis and Prediction in Telecommunication Access Network. HT . *Croatian Telecom Non-voice Services Division HR-51000 Rijeka*.

[6] Nagappan N., Murphy B., and Basili V., (2006). The Influence of Organizational Structure on Software Quality: An Empirical Case Study. *In Proceedings of the 30th International Conference on Software Engineering,* ACM, pp 521--530.

[7] Hardik M, Mosin I. and Komal P., (2011). Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis. International Conference on Web Services Computing (ICWSC). *Proceedings published by International Journal of Computer Applications® (IJCA*). pp 22-27

[8] Sutton O.,(2012). Introduction to K Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction. Retrieved 28/07/2014

[9] Retrieved - 21/12/2015: http://rapidminer.com

[10] David H. (1995). Learning Bayesian networks: The combination of knowledge and statistical data, pp 197 – 243, Retrieved 23/09/2014: https://bib.irb.hr/datoteka/499595.Melecon_paper.pdf

[11] Mitchell T. M., (2015). Textbook on Machine Learning vol 2, pp 3 - 17

[12] Nguyen G. et al.; (2008), *"A supervised learning approach for imbalanced data sets,"* in International Conference on Pattern Recognition, 2008, pp. 1-4. Retrieved 23/09/2014: http://ro.uow.edu.au/infopapers/3155

[13] Widman L., Loparo K. and Nielsen N., (1989). *Artificial Inteligence, Simulation and Modeling, Wilsey Inc.* Retrieved: 23/6/2014: http://www.tmrfindia.org/eseries/ebookv1-c4.pdf

[14] Vaidehi V., Monica S., Mohamed S., Deepika M. and Sangeetha S., (2008). A Prediction System Based on Fuzzy Logic. *Proceedings of the World Congress on Engineering and Computer Science*, ISBN: 978-988-98671-0-2, pp 2

[15] Darong H., (2009). Fisher Discriminance of Fault Predict for Decision-making Systems. *Second International Symposium on Knowledge Acquisition and Modeling*, vol 3, pp 37-40, ISBN: 978-0-7695-3888-4, DOI: 10.1109/KAM.2009.138.

16] Xiaozhen Z., Shanping L., and Zhen Y., (2013). A Novel System Anomaly Prediction System Based on Belief Markov Model and Ensemble Classification. Mathematical Problems in Engineering. vol. 2013, Hindawi Publishing Corporation. Retrieved - 23/6/2014: http://dx.doi.org/10.1155/2013/179390