An Open Access Journal Available Online

Covenant Journal of Informatics & Communication Technology (CJICT)

Vol. 5 No. 2, Dec., 2017

A Bi-annual Publication of the Departments of Computer Information Science, and Electrical & Information Engineering. Covenant University, Canaan Land, Km 10, Idiroko Road, Ota, Ogun State, Nigeria.

Editor-in-Chief: Prof. Sanjay Misra
sanjay.misra@covenantuniversity.edu.ng

Managing Editor: Edwin O. Agbaike
me@covenantuniversity.edu.ng

Website: http//Journal.covenantuniversity.edu.ng/cjict/

ISSN - Print: 2354-3566
- Electronics: 2354 – 3507

Published by Covenant University Journals,
Covenant University, Canaanland, Km 10, Idiroko Road,
P.M.B. 1023, Ota, Ogun State, Nigeria

Printed by Covenant University Press

Articles

Covenant Journal of Informatics & Communication Technology. Vol. 5 No. 2, Dec, 2017

An Open Access Journal Available Online

Soft Computing Techniques for Stock Market Prediction: A Literature Survey

Isaac Ibidapo, Ayodele Adebiyi & Olatunji Okesola

Department of Computer and Information Sciences,
Covenant University, Ota, Nigeria

isaacferanmi@gmail.com,
ayo.adebiyi@covenantuniversity.edu.ng,
olatunji.okesola@covenantuniversity.edu.ng

Abstract: Stock market trading is an unending investment exercise globally. It has potentials to generate high returns on investors' investment. However, it is characterized by high risk of investment hence, having knowledge and ability to predict stock price or market movement is invaluable to investors in the stock market. Over the years, several soft computing techniques have been used to analyze various stock markets to retrieve knowledge to guide investors on when to buy or sell. This paper surveys over 100 published articles that focus on the application of soft computing techniques to forecast stock markets. The aim of this paper is to present a coherent of information on various soft computing techniques employed for stock market prediction. This research work will enable researchers in this field to know the current trend as well as help to inform their future research efforts. From the surveyed articles, it is evident that researchers have firmly focused on the development of hybrid prediction models and substantial work has also been done on the use of social media data for stock market prediction. It is also revealing that most studies have focused on the prediction of stock prices in emerging market.

Keywords: Stock market prediction, Soft computing, ANN, SVM, Hybrid prediction models.

1. Introduction

The stock market is a market in which company stocks and derivatives are traded at an agreed price; they are refer to securities listed on stock exchanges and those traded privately (Dase et al., 2010). Globally, the stock market has attracted a large number of investors and economists (Agrawal et al., 2013). This is because it has the opportunity of highest return over other schemes and is a key source of fund raising for companies through initial public offer (IPO) (Sureshkumar and Elango, 2012). As stated in (Chakravaty and Dash, 2012), several techniques have been employed for stock market prediction of which statistical methods have been extensively used. Some of the statistical techniques that have been used for stock market prediction include autoregressive conditional heteroscedasticity (ARCH), autoregressive integrated moving average (ARIMA), autoregressive moving average (ARMA) amongst others. However, these models can predict linear patterns only while the stock market returns change in a non-linear pattern (Vaisla and Bhatt, 2010). The stock market is dynamic, evolutionary, complex and non-linear in nature thus several non-linear approaches to stock market prediction such as generalized autoregressive conditional heteroskedasticity (GARCH) have been proposed (Liu et al., 2012). Forecasting of stock returns is difficult because of the need to capture market volatility to implement prediction models (Atsalakis &Valvanis, 2009).

Recently, a lot of research has been carried out to using various soft computing approaches for stock market price forecasting (Liu et al., 2012). Soft computing techniques offer useful tools in forecasting noisy environments and can capture their nonlinear behaviour (Atsalakis & Valvanis, 2009). This paper carries out a literature survey of 120 published articles that focus on the application of soft computing techniques for stock market prediction. The result of the review is presented in three summary tables. The first table presents a list of stock markets that each author modeled for prediction, summary of the objective of each article as well as the experimental data used in each study. The second table summarizes information about the modeling techniques employed in each reviewed article. The third table presents the models that were compared with each proposed prediction model, the performance measures employed for comparison and the result of the comparison.

This research cohesively presents information on the various soft computing techniques that have been employed to model and predict different stock markets. This paper would help researchers to know the current state of the art in stock market prediction, facilitate comparative studies as well as spot current research opportunities.

The remainder of this paper is organized as follows: A background of the research work and related works is presented in Section 2, in Section 3 a detailed description of the methodology employed for the research is presented as well as the various summary tables, discussion points are presented in Section 4 and Section 5 contains concluding remarks.

2. Related Work

Investment in the stock market is regarded as high risks and high gains as such investors and researchers alike have sought for tools and methods that would increase their gains as well as minimize their risks (Agrawal et al.,

2013). Soft and evolutionary computing methods for stock price prediction have become hugely popular for accurate prediction of stock market behavior because of their ability to handle the uncertain, chaotic and non-linear nature of the stock market (Chakravaty et al., 2011).

In (Atsalakis et al., 2009) a survey of articles whose focus is on neural and neuro-fuzzy techniques for stock market predictions was carried out. Summary tables were presented in terms of input variable choices, comparative studies, modeling techniques, performance measures and surveyed stock markets. (Agrawal et al., 2013) studies stock prediction techniques applied to the Indian stock market and presents the advantages and disadvantages of the methods. Additionally, it presents a comprehensive review of the significant developments in the field of stock prediction of Indian stock market. (Dase&Pawar, 2010) presents a review of literature on the application Artificial Neural Network for stock price prediction. (Santosh et al., 2013) presents a literature review outlining the various application areas of soft computing techniques.(Sheng & Subhash 2012) discusses and presents various realms that can be predicted with social media. It then presents a discussion of the various prediction methods used with social media. (Wu et al., 2010) carried out a comparison study of 5 bankruptcy prediction models drawn from literature and then developed a model that aggregates the key variables from the models as well as adding a new variable. It is reported that the new model outperformed the existing models.

(Li & Ma, 2010) carried out a literature survey of the application of ANNs in a number of aspects of financial economics namely: stock price forecasting, option pricing, exchange rate forecasting and the prediction of banking and financial crisis. (Bahrammirzaee, 2010) carried out a comparative literature review of artificial intelligence applications in finance focusing on the application of ANNs, expert systems and hybrid intelligent systems in the areas of credit evaluation, portfolio management, financial prediction and planning. (Preethi & Santhi, 2012) carried out a literature survey of the use of ANN, Data mining, Hidden Markov Model and Neuro-Fuzzy systems for stock market fluctuation prediction. (Hajizadeh et al., 2010) carried out a literature survey of data mining techniques applied to data from various stock markets. (Nikfarjam et al., 2010) carried out a literature survey of research works that focused on the mining of text from financial news for stock price prediction. A presentation of the main components of text mining systems was made as well as how each component was implemented in the reviewed papers.

3. Methodology

To carry out this study, published articles that focused on the application of neural networks, neuro-fuzzy and other soft computing methods were reviewed. Articles that focused on the use of social media data with soft computing methods for stock market prediction were also reviewed. Results of the study are presented in three summary tables. The stock market surveyed, experimental data used, methodology employed and summary of the objective of each reviewed article is presented in table 1. Table 2 presents information of the modeling techniques

employed in each reviewed article in terms of input variables, prediction model, training method, network layers, and data preprocessing. Table 3 presents the prediction models against which each article's prediction model was compared, the performance measures that formed the basis of evaluation, as well as the result of the comparison.

Table 1: Summary of the reviewed articles

Article	Market Surveyed	Experimental Data (Training Data)	Summary
Chakravarty et al., (2012)	S&P 500, DJIA Index, BSE	S&P, BSE, DJIA dataset	Proposes a hybrid model of integrated functional link interval type 2 fuzzy neural system (FLIT2NS) for stock price prediction
Fenghua et al., (2014)	Shanghai Stock Exchange	SSE data	proposes a hybrid model of Singular spectrum analysis and support vector machine for stock price prediction
Kara et al., (2011)	Instanbul Stock Exchange	ISE National 100 index dataset	Predicted direction of movement in the daily ISE National 100 index using ANN and SVM and compares the performances of the models
Hsieh et al., (2011)	TAIEX, DJIA	TAIEX,DJIA, Nikkei, FTSE dataset	Proposes a model that uses wavelength transforms and RNN based on ABC algorithm for stock price prediction
Sureshkumar et al., (2012)	National Stock Exchange	NSE Dataset	
Bollen et al., (2010)	DJIA	Twitter feeds	Investigates whether public sentiment in daily twitter posts can be used for stock price prediction
Cheng et al., (2010)	Taiwan Stock Exchange	TSMC stock data	Proposes a hybrid model based on Rough set theory and genetic algorithm forstock price prediction
Dai et al., (2012)	Japanese Stock Market, China Stock Market	Nikkei 225 stock data (80%)	Proposes a hybrid model of NLICA and Neural Network
Wei et al., (2011)	Taiwan Stock Exchange		Proposes an ANFIS model for stock market prediction
Vaisla& Bhatt, (2010)	Indian stock market	NIFTY data	Employed ANN for stock market prediction and compared the result with that of statistical forecasting
Liu et al., (2012)	TAIEX, NASDAQ		Type 2 neuro-fuzzy modelling was applied to stock price prediction
Merh et al., (2010)	Indian stock market		1Developed two hybrid models ANN-ARIMA and ARIMA-ANN and comparison made between the

			two
Zarandi et al., (2012)		IBM, Dell Corporation, British Airways, Ryanair stock data	Proposes a hybrid AI model based on the coordination of intelligent agents for next-day stock price prediction
Kazem et al., (2013)	NASDAQ	80%	Proposes a prediction model based on chaotic mapping, firefly algorithmand SVR for stock market prediction
Asadi et al., (2012)	TSE, TEPIX, DJIA, NASDAQ		Proposes a stock prediction model that uses GA and LMBP to predict stock market indices
Wang et al., (2012)	SZIL, DJIA		Proposes hybrid model of ESM-ARIMA and BPNN for stock market prediction
Ballini et al., (2010)	Brazilian Stock Market	Ibovespa stock data	Proposes a class of neuro-fuzzy network and a constructive learning method for stock market prediction
Wei (2013)	TAIEX	TAIEX dataset	Proposed hybrid of adaptive expectation genetic algorithm and ANFIS for stock price prediction
Majhi et al., (2013)	S&P 500, DJIA Index	S&P 500, DJIA dataset	developed a hybrid prediction model that uses RBF NN and non-dominated sorting multi-objective GA-2
Enke et al., (2011)	S&P 500		Proposes a hybrid 3-stage stock market prediction system
Cai et al., (2013)	Taiwan Stock Exchange	TAIEX dataset	Proposes hybrid of Fuzzy time series Genetic Algorithm for stock price prediction
Anbalagan & Maheswar (2015)	Bombay Stock Exchange	TCS, RIL dataset (80%)	Proposes a Fuzzy Metagraph based model for stock market prediction
Babu& Reddy (2015)	Indian stock market	SBI, Tata Steel dataset	Proposes a hybrid ARIMA-GARCH prediction model
Hadavandi et al., (2010)	IT Sector, Airline Sector	IBM, Dell Corporation, British Airways, Ryanair stock data	Proposes a hybrid model based on genetic fuzzy systemsand ANN for stock price prediction
Hafezi et al., (2015)	German stock market	DAX price dataset	Proposed a hybrid bat-neural network multi-agent system for stock price prediction
Ticknor (2013)		Microsoft Corp, Goldman Sachs Group Incstock	Proposes a Bayesian regularized ANN for financial market behavior prediction

		dataset	
Chang & Fan(2011)	TAIEX	TAIEX stock dataset	Proposes a hybrid ANFIS stock prediction model based on AR and volatility
Hsu et al.,(2009)	TAIEX	TAIEX-FISI dataset	Proposes a hybrid model of SOM and Genetic programming for stock price prediction
Yeh et al., (2010)	TAIEX	TAIEX stock dataset	Proposed a stock prediction model based on SVR with multiple-kernel learning algorithm
Boyacioglu & Avci (2010)	Instanbul Stock Exchange	DJI, DAX, BOVESPA indices, Macroeconomic Indicators, ISE dataset	Proposes the use of ANFIS for stock price prediction
Zahedi et al., (2015)	Tehran Stock Exchange	Tehran Stock Exchange dataset	Applies ANN and Principal Component method using 20 accounting variables for stock price prediction
Tsai et al., (2010)	Taiwan Stock Exchange	Taiwan Economic Journal dataset	Compares the selection methods PCA, GA, and CART then combines them based on union, intersection, and multi-interaction approaches to analyse prediction accuracy and errors.
Atsalakis et al., (2011)	Greece stock market	National Bank of Greece stock dataset	Proposes a stock prediction system that is based on a neuro-fuzzy architecture which uses Elliot Wave Theory
Ballings et al., (2015)		Amadeus Database dataset	Benchmark study of ensembles and single classifier models
Feng & Chou (2011)	Taiwan Stock Exchange	TAIEX stock dataset	Developed an ANN prediction system with the combinations of SRA, Dynamic learning, and Recursive based PSO Learning algorithms
Liao et al., (2010)	Chinese stock market, HIS, DJI, IXIC, SP 500	Chinese stock market dataset	Proposed an improved NN model by introducing a stochastic time effective function
Mostafa (2010)	Kuwait stock Exchange (KSE)	KSE dataset	Forecasts KSE movements using 2 NN architectures: MLPNN and Generalized regression NNs
Shen et al., (2010)	Shanghai Stock Exchange	SSE dataset	Proposes RBF-NN optimizedby Artificial fish swarm Algorithm for stock price prediction
Wang et al.,	Shanghai	Shanghai	Proposed a hybrid stock prediction

(2012)	Stock Exchange	Composite Index closing prices	model based on Wavelet De-noising-based Back propagation NN
Chen et al., (2010)	Taiwan Stock Exchange	TAIEX dataset	Proposes a hybrid model which improves NGBM by Nash equilibrum concept
Anish & Majhi (2015)	DJIA, S&P500	DJIA, S&P 500 Dataset	Proposes a hybrid model of a feedback type of functional link ANN for stock prediction
Lu (2010)	Taiwan Stock Exchange	TAIEX closing cash index, Nikkei 225 opening cash index	Proposes an integrated ICA-based denoising scheme with NN for stock price prediction
Luo & Chen (2012)	Shanghai Stock Exchange	SSE dataset	Proposes a stock prediction model that integrates PLR and WSVM for stock price prediction
Tsai et al., (2010)	Taiwan Stock Exchange	Taiwan Economic Journal dataset	Examines the applicability of classifier ensembles by constructing the homogenous and heterogeneous classifier ensembles for stockprice prediction
Desai et al., (2013)	Indian stock market	S&P CNX Nifty 50 dataset	Proposes the use of ANN for predicting S&P CNS Nifty 50 Index
Yixin & Zhang (2010)	Chinese stock market	Chinese stock market dataset	Uses BP NN for stock market prediction
Wang et al., (2011)	Shanghai Stock Exchange	Shanghai Composite Index dataset	Proposes an ANN stock prediction model based on HLP
Rounaghi et al., (2015)	Tehran Stock Exchange	Tehran Stock Exchange dataset	Uses multivariate adaptive regression splines (MARS) model and semi-parametic splines technique for stock price prediction
Park & Shin (2013)		KOSPI listed companies' stock prices	Proposes a stock prediction model that uses graph based SSL for stock prediction
Ni et al., (2011)	Shanghai Stock Exchange	SSECI dataset	Used of fractal feature selection based on fractal dimension and ant colony algorithm and SVM for stock price prediction
Cocianu& Grigoryan(2015)	Bucharest stock exchange	SNP stock dataset	Proposes a feed-forward NN architecture with gradient descent with adaptive learning rate variant of BP Algorithm for stock price prediction
ZheGao & Yang(2014)	Shanghai and Shenzhen	Shanghai-Shenzhen 300	Proposes a hybrid SVR with hierarchical clustering for stock

	Stock Exchanges	index dataset	price prediction
Abraham AuYeng(2011)	NASDAQ	Nasdaq-100 index, S&P CNX Nifty index dataset	Investigates the representationof stock markets using ensembles by employing an ensemble of ANN-LSM, SVM, Neuro-fuzzy model and Difference boosting NN
Olatunji et al., (2013)	Saudi stock market	Saudi stock market dataset	Proposes the use of ANN model for the prediction of Saudi stock market
Suwandi& Santica(2014)	Indonesian stock exchange	Jakarta composite index dataset, gold fixing price, WTI crude oil price	Developed a least square SVM model to predict daily close price of Jakarta composite index
Nguyen & Le(2014)		IBM, AppleInc, S&P 500, DJI stock dataset	Proposes a stock price prediction model based on the combination of SOM and fuzzy SVM
Adebiyi et al., (2014)	New York Stock Exchange	Dell Inc stock data	Compares the performance of ARIMA and ANN for stock price prediction
Hegazy et al., (2014)	Bombay Stock Exchange	BSE Sensex index dataset	Optimizes FIS parameters by an adaptive network. The optimized model is optimized using Quantum GA for forecasting accuracy refinement
Adebiyi et al., (2012)	Nigerian Stock Exchange	NSE Companies' stock price dataset	Uses a fuzzy-neural network fed with hybrid market indicators for stock price prediction
Chet et al., (2014)	Colombo Stock Exchange	CSE Dataset (80%)	Employed the use of BP ANN for stock price prediction
Neenwi et al., (2013)	Nigerian Stock Exchange	Access bank, First bank, UBA stock dataset	Uses ANN for stock price prediction
Isenah& Olubusoye(2014)	Nigerian Stock Exchange	NSE dataset	Develops two ANN based stock prediction models and compared their performances with an ARIMA model
Magaji& Adeboye (2014)	Nigerian Stock Exchange	Cowry, CashCraft and BGL dataset	Implemented the logistic function on BP algorithm for ANN for stock price prediction
Akintola et al., (2011)	Nigerian Stock Exchange	Intercontinental Bank stock prices	Proposes a neural network based model for stock price prediction
Bola et al., (2013)	Nigerian Stock Exchange	Technical Indicators	Carries out a study of ANN and Bayesian network for stock price

			prediction
Subhabrata et al., (2014)	National Stock Exchange (India)	102 stocks dataset	Proposes a SOM based hybrid clustering technique with SVR for stock price and volatility predictions and for portfolio selection.
Dash & Dash (2016)	Bombay Stock Exchange, US Stock Market	BSE SENSEX stock index, S&P 500 stock index	Proposes a self-evolving recurrent neuro-fuzzy inference system with Modified Differential Harmony Search (MDHS) for stock price prediction.
Lahmiri (2014)	US stock market	S&P 500, Hewlett-Packard, IBM, Microsoft, and Oracle datasets	Proposes the use of low and high frequency components with BP ANN for stock price prediction.
Babu & Reddy (2014)	National Stock Exchange (India)	Simulated dataset, Sunspot data, Electricity price data, L&T company stock	Proposed a hybrid ARIMA-ANN stock prediction model that employs the use of a moving-average filter for one-step ahead and multi-step ahead predictions.
Chang & Liu (2008)	Taiwan Stock Exchange	Taiwan Electronic shares	Developed a TSK type fuzzy rule based system for stock price prediction
Guresen et al., (2011)	NASDAQ Stock Exchange	NASDAQ index dataset	Evaluates the effectiveness of various neural network models in predicting stock market index

Table 2: Summary of Prediction Methodology Employed in each Article

Article	Prediction Model	Input Variables	Training Method	Layers (Neurons)	Data Preprocessing
Chakravarty et al., (2012)	FLIT2NS	Daily closing prices, minimum and maximum price of dataset	BP Algorithm, PSO Algorithm	5	
Fenghua et al., (2014)	Hybrid SSA-SVM	Predictive value closing price	SVM		Yes (SSA)
Kara et al., (2011)	ANN and SVM	Daily closing prices, minimum and maximum price of dataset	BP Algorithm	3(10,-,1)	Yes
Hsieh et al., (2010)	ABC-RNN	10 Technical Indicators, open, close, highest, lowest price	ABC Algorithm		Yes (Wavelet transform)

Sureshkumar et al (2012)		previous close price, open price, high price, close price			
Bollen et al., (2010)	Self-organizing Fuzzy Neural Network	Past 3 days DJIA values and mood values of past 3 days	SOFNN	5	Yes
Cheng et al.,., (2010)	RST-GA	Technical Indicators	RS algorithm, GA algorithm		Yes (CPDA,MEPA)
Dai et al.,., (2012)	NLICA-ANN			3 (4-9-1)	Yes (NLICA)
Wei et al.,., (2011)	ANFIS	Opening price, highest price, lowest price, closing price, trading volume	Least squares method and BP gradient descent method	5	Yes (Correlation matrix, subtractive clustering)
Vaisla & Bhatt (2010)	ANN	Closing price, exchange rate, FII purchase, FII sales			
Liu et al., (2011)	T2NFS	closing stock prices	PSO, Least square estimation	4	Yes (Self constructing clustering method)
Merh et al.,., (2010)	ANN-ARIMA	Daily opening price, high,low and closing prices	BP Algorithm	3	Yes
Zarandi et al.,., (2012)	Fuzzy Multiagent System	index opening, closing price, daily highest and lowest values	Genetic Fuzzy System	4	Yes (Stepwise regression analysis and SOM neural network clustering)
Kazem et al., (2013)	SVR-CFA	Daily closing prices	SVR		Yes
Asadi et al., (2012)	PELMNN	7 Technical Indices	GA and LM	4 (2-4-4-1)	Yes
Wang et al., (2011)	PHM	Opening and Closing index	GA	3 (12-9-12)	
Ballini et al., (2010)	NFN		On-line Learning	5 ,2,1	Yes (First differencing, Logarithmic Transformation)
Wei (2013)	ANFIS-	7 Technical	Least		Yes

	AEGA	Indicators	squares method and BP gradient descent method		
Majhi et al., (2013)		10 Technical Indicators			Yes
Enke et al., (2011)	Fuzzy Type 2 ANN	CDR3 rate, PPI, M1 level index price level, IP reading	Differentia l Evolution Algorithm	5	Yes
Cai et al., (2013)	FTSGA	Closing index prices	GA		Yes
Anbalagan et al., (2015)		Technical Indicators	FM Learning Algorithm	4	Yes
Babu&Redd y (2015)	ARIMA-GARCH	closing stock prices			Yes
Hadavandi et al., (2010)	CGFS	Open,close,high,lo w prices	SOM		Yes (Stepwise regression analysis)
Hafezi et al., (2015)	Hybrid BNNMA S	Daily stock data, news	Bat Algorithm	3 (-, 6, 1)	Yes
Ticknor (2013)	Bayesian ANN	Daily stock prices,6 Financial Indicators	Minimizati on of the mean squared error	3 (9, 5, 1)	-
Chang et al., (2011)	ANFIS based on AR and Volatility	Different-order AR model, different-order momentum	BP Algorithm, Least square method		Yes
Hsu et al., (2011)	SOM-GP	Technical Indicators, daily closing prices			
Yeh et al., (2010)	MKSVR	Daily closing prices, Technical Indicators	SMO, Gradient Projection method		
Boyacioglu & Avci (2010)	ANFIS	6 macroeconomic variables, 3 indices	FIS	5	-
Zahedi et al., (2015)	ANN-PCA	20 accounting variables	LVM	3 (-,10,-)	Yes
Tsai et al., (2010)	ANN	Fundamental Indices,	BP Algorithm	3	Yes (PCA,GA,CART

		macroeconomic indices			)
Atsalakis et al., (2011)	WASP	EWO, Oscillator lags, moving averages of 5 and 35 days	Least squares method and BP method		
Ballings et al., (2015)		Technical Indicators			
Feng & Chou (2011)	ANN	2 Technical Indexes, 5 MA, 6 RSI	PSO-RLS Learning Algorithms		Yes
Liao et al., (2010)	The stochastic time effective neural model	Daily opening price, high,low, closing prices, and trade volume	BP Algorithm	3 (5-20-1)	Yes
Mostafa (2010)	ANN	Daily closing prices	Quasi-Newton training algorithm, EBP Algorithm, Conjugate gradient descent Algorithm	3 MLP, 4(GRNN)	
Shen et al., (2010)	RBFNN-AFSA	4 Technical Indicators	AFSA	3	Yes
Wang et al., (2011)	WDBPNN	closing prices	BP Algorithm	3 (3-10-3)	Yes (Wavelet transform)
Chen et al., (2010)	NNGBM	daily stock prices	Least square method		Yes
Anish & Majhi (2015)	FA-FFLANN-RLS	3 Technical Indicators	Recursive Least square training, LMB Algorithm	3	Yes (FA)
Lu (2010)	ICA-BPN	Technical Indicators, Futures prices	BP Algorithm	3 (6-13-1)	Yes (ICA)
Luo & Chen (2012)	PLR-WSVM	11 Technical Indicators	WSVM		Yes (PLR)
Desai et al., (2013)	ANN	Closing prices	BP Algorithm	3 (-,10,-)	Yes (Logarithmic

					First Differencing)
Yixin & Zhang (2010)	ANN	Technical Indicators	BP Algorithm	3 (21-3-1)	Yes
Wang et al., (2011)	ANN	daily closing price	BP Algorithm	4 (-,1,1,1)	Yes (HLP)
Rounaghi et al., (2015)	MARS	30 Accounting variables,10 Economic variables			Yes
Park & Shin (2013)	SSL	7 Technical Indicators, 16 Financial Economical indexes	Graph-based SSL		Yes
Ni et al., (2011)	SVM	Technical Indicators	SVM		Yes
Cocianu et al., (2015)	ANN	opening, closing, highest, lowest price, 7 Technical and Fundamental Indicators	BP Algorithm (Adaptive learning rate variant)	2	-
ZheGao & Yang (2014)	HC-SVR	22 Technical Indicators, 4 CSI300 Index futures	SVM (Grid search parameter algorithm)		Yes
Abraham et al., (2011)		opening,closing, highest, lowest price			Yes
Olatunji et al., (2013)	ANN	closing price	BP Algorithm	3	-
Suwandi et al., (2014)	LSSVM	Open, closing, high and low price, Gold fixing price	Grid search technique		Yes
Nguyen et al., (2014)	SOM-F-SVM	EMA 100, RDP-5, RDP-10, RDP-15, RDP-20,RDP+5	Fuzzy Inference System		Yes
Adebiyi et al., (2014)	ANN, ARIMA	open, low, high, close prices and volume traded	BP Algorithm		Yes
Adebiyi et al., (2012)	ANN	10 Technical variables, 8 fundamental analysis variables	BP Algorithm	3 (18-24-1)	-
Sakarya et al., (2015)	ANN	gold price, oil price, interest rate,	BP Algorithm	4 (7-9-7-2)	Yes

		CPI, exchange rate, money supply, BIST volume			
Hegazy et al., (2014)	NFIS-QGA	daily open, close, highest, lowest prices	DCQGA		Yes
Adebiyi et al., (2012)	Neuro-fuzzy model	Technical, fundamental indicators, 5 expert opinions variables	BP Algorithm	3 (21-26-1)	Yes
Chet et al., (2014)	ANN	Daily ASPI, ASTR, PER, PBV data	BP Algorithm	3 (4-8-1)	Yes
Neenwi et al., (2013)	ANN	4-day price movements	BP Algorithm	3	Yes
Isenah et al., (2014)	ANN	Technical Indicators		3	Yes
Magaji (2014)	ANN	Previous day index value, previous day's NGN/USD exchange rate	BP Algorithm	3 (4-2-1)	Yes
Akinola et al., (2011)	ANN	Daily closing price	BP Algorithm	4 (4-4-4-1)	Yes
Guresen et al., (2011)	ANN	Previous 4 days index values	BP Algorithm	-	Yes
Chang & Liu (2008)	TSK-type Neuro-fuzzy network	8 Technical Indices	Simulated Annealing	-	Yes
Babu& Reddy (2014)	ARIMA-ANN	Closing prices	BP Algorithm	-	Yes
Lahmiri (2014)	ANN	Wavelet low and high frequency components	BP Algorithm	3 (2-4-1)	Yes
Dash & Dash (2016)	SERNFIS	Open, high, low and closing stock prices, Technical Indicators	MDHS	7	Yes
Subhabrata et al., (2014)	SOM-K-means-SVR	Closing prices, intra-day volatility	Grid and Pattern Search Algorithms	-	Yes

Table 3: Summary of Comparative Studies

Article	Modeling Benchmark	Performance Measure	Result
Chakravarty et al., (2012)	FLIT2FNS, LLWNN, FLANN, Type 1 FLS	MAPE, RMSE	FLIT2FNS model was superior in terms of prediction accuracy and error convergence speed over other models
Fenghua et al., (2014)	SVM, EEMD-SVM	MSE, MAPE	SSA-SVM model resulted in better prediction accuracy than other models
Kara et al., (2011)	ANN, SVM	RMS	The average prediction performance of ANN model was significantly better than that of the SVM model
Hsieh et al., (2010)	BP-ANN, Fuzzy Time Series, ANFIS	RSME, MAE, MAPE	It's performance was superior to other models
Cheng et al., (2010)	RST,GA, Buy-and-Hold approach	Accuracy, Stock return	The hybrid model outperformed other models
Dai et al., (2012)	BPN, ICA-BPN,PCA-BPN	RMSE, MAD, MAPE	NLICA-ANN model provided better forecasting results evidenced by lower prediction error and higher prediction accuracy
Wei et al., (2011)	Fuzzy time series models	RMSE	Proposed system performs better, has the smallest average and variation of RMSE
Vaisla& Bhatt, 2010	Multiple regression technique	MAE, MSE, RMSE	ANN performs better than statistical forecasting
Liu et al., (2011)	Conventional Regression, ANN, Fuzzy Time Series, SVR	RMSE, MAE, MAPE	T2NFS performs best and has the least RMSE and average RMSE
Merh et al., (2010)	ARIMA-ANN	AAE, RMSE,MAPE, MPSE	Prediction of hybrid ANN-ARIMA model were better than hybrid ARIMA-ANN
Zarandi et al., (2012)	HMM, HMM-ANNGA, HMM-FL, ARIMA, ANN	MAPE	By MAPE evaluation, FMAS outperformed other models
Kazem et al., (2013)	SVR-GA, SVR-CGA, SVR-FA,	MSE, MAPE	SVR-CFA outperformed other models having the

	ANN, ANFIS		least average errors for MSE and MAPE.
Asadi et al., (2012)	BPNN, PENN, PEBPNN, ARIMA	MAPE, POCID, U of Theil, ARV	PELMNN outperformed other methods and improved prediction accuracy
Wang et al., (2011)	ESM, ARIMA, BPNN, RWN, EWH	MAE, RMSE, MAPE, ME, DA	Proposed model provided better forecasting results than other models in terms of prediction errors and accuracy
Ballini et al., (2010)	ANN, ARIMA	RMSE, MAPE, Associated residuals pattern, POCID	The neuro-fuzzy network provided more accurate forecasting
Wei (2013)	Fuzzy time series models	RMSE	The proposed model provided superior prediction accuracy
Majhi et al., (2013)	RBF based forecasting model	MAPE, DA, Theils U, ARV	Proposed model provided superior performance in all cases than the RBF forecasting model
Enke et al., (2011)	Fuzzy type 1 approach	RMSE	The proposed model produced better prediction accuracy
Cai et al., (2013)	Fuzzy time series models	RMSE	The proposed model bears the smallest RMSE, and has the best directional accuracy of forecast results
Anbalagan et al., (2015)	RW model, ANN, SVM	Hit Ratio (%), RMSE,MMRE	Outperformed other models
Babu& Reddy (2015)	ARIMA, GARCH, Wavelet-ARIMA, ANN	MAPE, MaxAPE, MAE, RMSE	It rendered better prediction accuracy
Hadavandi et al., (2010)	HMM, HMM-ANN-GA, HMM-FL, ARIMA, ANN	MAPE	Proposed model outperformed other models
Hafezi et al., (2015)	GA-ANN,GRNN, ERBN	MAPE	Proposed model outperformed other models in terms of prediction accuracy
Ticknor (2013)	Fusion model with weighted average, ARIMA	MAPE	Proposed model performs as well as the more advanced models
Chang et al., (2011)	Conventional Fuzzy time series model,	RMSE	The proposed model outperformed other

	Weighted Fuzzy Time series model		models
Yeh et al., (2010)	SK-SVR, ARIMA, TSK-FNN	RMSE	Proposed model outperformed other models
Atsalakis et al., (2011)	A buy and Hold Strategy	Hit-rate	The proposed system outperformed the Buy and Hold strategy
Ballings et al., (2015)	SVM,AB,RF,KF,LR, NN,KN	Area Under Curve (AUC)	Random forest proved to be the top predictor amongst others followed by SVM, KF, AB, NN, KN & LR
Feng & Chou (2011)	Standard PSO, Recursive-based PSO	RMSE, MAD, MAPE, CP, CD	Proposed system produced the most efficientprediction process
Shen et al., (2010)	RBF-GA, RBF-PSO, ARIMA, SVM, BP	Average Error	The proposed model proved to be useful for parallel computation
Wang et al., (2011)	BP Neural Network	MAE, RMSE, MAPE	Proposed model outperforms conventional BP model
Chen et al., (2010)	GM, NGBM	RPE, ARPE	The proposed model gives more precise results
Anish & Majhi (2015)	PCA-FFLANN-LMS, PCA-FFLANN-RLS, DWT-FFLANN-LMS, DWT-FFLANN-RLS, FA-FFLANN-LMS	MAPE, DA,U of Theil, ARV	The proposed model drastically reduced computation and produced better prediction results
Lu (2010)	RWMmodel, BPN model, Wavelet-BPN	RMSE,MAPE,DA	The proposed model produced the smallest value of RMSE and MAPE and provides better forecasting results
Luo & Chen (2012)	PLR-BPN, BHS	ACC	The proposed model achieved the best prediction accuracy
Tsai et al., (2010)	Ensemble of NNs, decision trees, Logistic regression	Average prediction accuracy, Type 1 and Type 2 errors, Return on Investment	The heterogeneous ensembles performed better than the homogenousones
Desai et al., (2013)	BHS	Accuracy	The proposed model produced high prediction accuracy

Park & Shin (2013)	ANN, SVM	AUC, ROI	The proposed model outperformed other models
Ni et al., (2011)	Information Gain, Symmetrical Uncertainty, Relief F, Correlation-based feature selection, OneR feature selection methods		The proposed feature selection method gives higher prediction accuracy thanthe others
Cocianu et al., (2015)	ARIMA models	MSE	The proposed model produced better results
ZheGao& Yang (2014)	PCA-SVR, GA-SVR	RMSE, NMSE, MAE, DS	The proposed model outperformed other models
Abraham et al., (2011)	SVM, NF, ANN,DBNN, E-1, E-2	RMSE,CC, MAP, MAPE	The ensemble approach based on direct error measure outperformed others
Suwandi et al., (2014)	SVM,ARIMA	RMSE, MAE, MAPE	The proposed model produced better prediction accuracy
Nguyen et al., (2014)	SOM-SVM, RBN, ANFIS	NMSE, MAE, DS	Proposed model produced more accurate results
Adebiyi et al., (2014)	ARIMA, ANN	MSE	Performance of ANN was better in terms of forecasting accuracy
Adebiyi et al., (2012)	ANN with only technical Analysis variables	Hit rate	The hybridized approach produced better prediction accuracy
Hegazy et al., (2014)	ANFIS	MSE	Proposed model produced higher prediction accuracy
Adebiyi et al., (2011)	ANN with only technical Analysis variables, FL-ANN wit single analysis variables, FL-ANN with hybrid market indicators	Hit Rate	The proposed model produced better results
Isenah et al., (2014)	ARIMA	RMSE, MAE,NMSE	ANN based models outperformed the ARIMA model
Bola et al., (2013)	ANN, Bayesian network		Bayesian network outperformed the ANN in terms of prediction power
Dash & Dash	RCEFLANN,	RMSE, MAPE,	The proposed model

(2016)	CEFLANN, ANFIS, RSEFNN	MAE	produced superior prediction performance
Lahmiri (2014)	ARIMA, RW, ANN	MAE, RMSE, MAD	The proposed model outperformed other models in terms of prediction accuracy
Babu& Reddy (2014)	ARIMA, ANN, other variants of ARIMA-ANN	MAE, MSE	The proposed model outperformed other models in terms of prediction accuracy for both one-step and multi-step predictions.
Chang & Liu (2008)	ANN, Multiple Regression Analysis	MAPE	The proposed model outperformed the other models
Guresen et al., (2011)	MLP-ANN, DA-ANN, GARCH-ANN, EGARCH-ANN		MSE, MAD

4. Discusion

Soft computing techniques had been widely used for stock price prediction as reported in the extant literature. This is because of its ability to provide better prediction accuracy above other predictive approaches. ANN continue to be the most popular in the stock price prediction efforts with several types of artificial intelligence algorithms proposed, some with feature selection and parameter setting techniques in a bid to improve its prediction accuracy. Furthermore, several studies have focused on the development of hybrid models for stock price prediction with the view of leveraging the advantages of each constituent technique for better prediction performance. Also, there were considerable efforts to make use of hybrid input variables particularly the use of both technical and fundamental analysis variables resulting in good prediction performance of prediction models. From the different articles reviewed the place of input parameters to the models developed was very significant to the outcome of the models.

The findings from table 1 indicate the various stock market exchange from which experimental data for prediction were obtained for each article reviewed. It also shows the predictive models proposed and developed by each researcher. In table 2, most of the predictive models used by each author require data preprocessing. The prevalent training algorithm used in all the reviewed articles was Backpropagation algorithm. The findings from table 3 presents the outcome of proposed model in each article reviewed in comparison with modelling benchmark techniques in extant literatures for comparative analysis. The performance measure frequently used are RMSE and MAPE respectively.

The findings of this literature survey clearly show that integrating of two or more soft computing techniques with

improved selection of input parameters would continue to be a way in which researchers can continue to explore in order to improve predictive models of the stock price prediction.

Conclusion

This study has surveyed published papers that focused on the application of soft computing techniques for stock market predictions. The survey has been presented in three summary tables. Table 1 presents the stock market surveyed, experimental data used, and summary of the objective of each reviewed article. Table 2 describes the prediction methodology employed for each article. Table 3 presents articles that carried out comparative studies of various prediction models in terms of the prediction models that were compared, the performance measure(s) employed for comparison and the result of the comparison study. This study would be useful and guide future researchers appropriately the application of soft computing model to stock price prediction.

References

Abraham A., & Auyeng A. (2011). *Integrating Ensemble Of Intelligent Systems For Modelling Stock Indices.* Usa: Department Of Computer Science, Oklahoma State University.

Adebiyi A.A., Adewumi A.O., & Ayo C.K. (2014). Stock Price Prediction using the ARIMA Model. *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.*

Adebiyi A.A., Adewunmi A.O., & Ayo C.K. (2014). Comparison of ARIMA and Artificial Neural Network Models for Stock Price Prediction. *Journal of Applied Mathematics Article ID 614342*.

Adebiyi A.A., Ayo C.K., & Otokiti S.O. (2011). Fuzzy-neural model with hybrid market indicators for stock forecasting. *International Journalof Electronic Finance Vol 5*, 286-297.

Adebiyi A.A., Ayo C.K., Adebiyi M.O., & Otokiti S.O. (2012). Stock Price Prediction using Neural Network with Hybridized Market Indicators. *Journal of Emerging Trends in Computing and Information Sciences Vol 3 No 1, 1-9.*

Agrawal J.G., Chourasia V.S., & Mittra A.K. (2013). State of the Art in Stock prediction techniques. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol 2, Issue 4*.

Akintola K.G., Alese B.K.,& Thompson A.F. (2011). Time series forecasting with Neural Network: A Case study of Stock Prices of Intercontinental Bank Nigeria. *IJRRAS 9 (3) www.arpapress.com/volumes/Vol9Issue3/IJRRAS_9_3_16.pdf*, 467-472.

Anbalagan T., & Maheswar S.U. (2015). Classification and prediction of stock maket index based on fuzzy metagraph. *Procedia Computer Science 47*, 214-221.

Anish C.M., & Majhi B. (2015). Hybrid nonlinear adaptive scheme for stock market prediction using feedback FLANN and factor analysis. *Journal of the Korean*

Statistical Society, 1226-3192`. http://dx.doi.org/10.1016/j.jkss.2015.07.002.

Arafah A.A., & Mukhlash I. (2015). The Application of Fuzzy Association Rule on Co-movement Analyze of Indonesian Stock Price. *Procedia Computer Science 59*, 235-243.

Asadi S., Hadavandi E., Mehmanpazir F., & Nakhostin M.M. (2012). Hybridization of evolutionary Levenberg-Marquardt neural networks and data pre-processing for stock market prediction. *Knowledge based systems 35*, 245-258.

Atsalakis G. S. , Dimitrakakis E. M. , & Zopounidis C. D. . (2011). Elliott Wave Theory and neuro-fuzzy systems, in stock market prediction: The WASP system. *Expert Systems with Applications 38* , 9196–9206.

Atsalakis G.S., & Valvanis K.P. (2009). Surveying stock market forecasting techniques- Part 2: Soft computing methods. *Expert Systems with Applications 36*, 5932-5941.

Babu C. N., & Reddy B. E. (2015). Prediction of selected Indian stock using a partitioning–interpolation based ARIMA–GARCH model. *Applied Computing and Informatics 11*, 130-143.

Babu C.N., & Reddy B.E. (2014). A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data . *Applied Soft Computing 23* , 27–38.

Babu C.N., & Reddy B.E. (2015). Prediction of selected Indian Stock using a partitioning-interpolation based ARIMA-GARCH model. *Applied Computing and Informatics 11*, 130-143.

Bahrammirzaee A. (2010). A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. *Neural Comput & Applic 19*, 1165–1195.

Bahrammirzaee, A. (2010). A Comparative survey of artificial intelligence applications in finance: Artificial neural networks, expert systems and hybrid intelligent systems. *Neural Comput & Applic (2010) 19.*, 1165-1195.

Ballings M., Poel D. Van den, Hespeels N., Gryp R. . (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications 42*, 7046–7056.

Ballini R., Luna I., Lima L.M., & Dasilveira R.L.F. (2010). *A comparative analysis of neurofuzzy, ANN and ARIMA models for Brazilian stock index forecasting.* Brazil: Department of Economic Theory, institute of Economics, University of Campinas.

Bola A.A., Adesola A.G., Olusayo O.E., & Adebisi A.A. (2013). Forecasting Movement of the Nigerian Stock Exchange All Share Index using Artificial Neural and Bayesian Networks. *Journal of Finance and Investment Analysis Vol 2 No 1*, 41-59.

Bollen J., Mao H., & Zeng X.-J. (2010). Twitter mood predicts the stock market.

Boyacioglu M. A., & Avci D. (2010). An Adaptive Network-Based

Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange. *Expert Systems with Applications 37* , 7908–7912.

Cai Q., Zhang D., Wu B., & Leung S.C.H. (2013). A novel forecasting model based on fuzzy timeseries and genetic algorithm. *Procedia computer science 18*, 1155-1162.

Chakravarty S., & Dash P.K. (2012). A PSO based integrated functional link net and interval type-2 fuzzy logic system for predicting stock market indices. *Applied Soft Computing 12*, 934-941.

Chang J.-R., Wei L.-Y., & Cheng C.-H. . (2011). A hybrid ANFIS model based on AR and volatility for TAIEX forecasting. *Applied Soft Computing 11*, 1388–1395.

Chang P.-C., & Fan C.-Y. (2011). A dynamic threshold decision system for stock trading signaldetection. *Applied soft computing 11*, 3998-4010.

Chang P.-C., & Liu C.-H. (2008). A TSK type fuzzy rule based system for stock price prediction. *Expert Systems with Applications 34*, 135-144.

Chen C.-I., Hsin P.-H., & Wu C.-S. . (2010). Forecasting Taiwan's major stock indices by the Nash nonlinear grey Bernoulli model. *Expert Systems with Applications 37*, 7557–7562.

Chen D., & Seneviratna D.M.K.N. (2014). Using Feed Forward BPNN for Forecasting All Share Price Index. *Journal of Data Analysis and Information Processing.*, 87-94.

Cheng C.-H., Chen T.-L., & Wei L.-Y. (2010). A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting. *Informational Sciences 180*, 1610-1629.

Cocianu C.-L., & Grigoryan H. (2015). An Artificial Neural Network for data forecasting purposes. *Informatica Economica Vol 19 No 2*.

Dai W., Yu J.-Y., & Lu C. -J. (2012). Combining nonlinear independent component analysis and neural network forthe prediction of Asian stock market indexes. *Expert Systems with Applications 39*, 4444-4452.

Das S.K., Kumar A., DasB., & Burnwal A.P. (2013). *On soft computing techniques in different areas.* India: Department of Computer Science and Engineering.

Dase R.K., & Pawar D.D. (2010). Application of Artificial Neural Network for stock market predictions: A review of literature. *International Journal of Machine Intelligence Vol 2, Issue 2.*, 14-17.

Dash R., & Dash P. (2016). Efficient stock price prediction using a Self Evolving Recurrent Neuro-Fuzzy Inference System optimized through a Modified technique. *Expert Systems With Applications 52*, 75-90.

Deng S., Mitsubuchi T., Shioda K. , Shimada T. , & Sakurai A. . (2011). Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction. *Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing* (pp. 800-807). IEEE Computer Society.

Desai J., Trivedi A., & Joshi N. A. . (2013). *Forecasting of Stock Market Indices Using Artificial Neural Network.* Ahmedabad: Shri Chimanbhai Patel Institutes.

Egrioglu E., Aslan Y., & Aladag C.H. (2014). A new fuzzy time series method based on Artificial Bee Colony Algorithm. *Turkish Journal of Fuzzy Systems Vol 5*, 59-77.

Enke D., Graver M., & Mehdiyev N. (2011). Stock Market prediction with multiple regression, fuzzy type-2 clustering and neural networks. *Procedia Computer Science 6*, 201-206.

Feng H.-M., & Chou H.-C. . (2011). Evolutional RBFNs prediction systems generation in the applications of financial time series data. *Expert Systems with Applications 38* , 8285–8292.

Fenghua W., Jihong X., Zhifang H., & Xu G. (2014). Stock price prediction basedon SSA and SVM. *Procedia Computer Science 31*, 625-631.

Guresen E., Kayakutlu G.,& Daim T. U. . (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications 38*, 10389–10397.

Guresen E., Kayakutlu G., & Daim T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications 38* , 10389–10397.

Hadavandi E., Shavandi H., & Ghanbari A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge Based Systems 23*, 800-808.

Hafezi R., Shahrabib J., & Hadavandi E. (2015). A bat-neural network multi-agent system (BNNMAS) for stock price prediction: Case study of DAX stock price. *Applied Soft Computing 29*, 196–210.

Hagenau M., Liebmann M., Hedwig M., & Neumann D. (2012). Automated news reading: Stock price prediction based on Financial news using content-specific features. *45th Hawaii International Conference on System Sciences.*

Hajizadeh E., Ardakani H.D., & Shahrabi J. (2010). Application of data mining techniques in stock markets: A survey. *Journal of Economics and International Finance Vol 2 (7).*, 109-118.

Hegazy O., Soliman O.S., & Toony A.A. (2014). Hybrid of neuro-fuzzy inference system and quantum genetic algorithmfor prediction in Stock Market. *Issues in Business Management and Economics Vol 2*, 094-102.

Hsieh T.-J., Hsiao H.-F., & Yeh W.-C. (2011). Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. *Applied Soft Computing 11*, 2510-2525.

Hsu C.-M. (2011). A hybrid procedure for stock price prediction by integrating self-organizing map and genetic programming. *Expert Systems with Applications 38*, 14026–14036.

Hsu S.-H., Hsieh J.J., Chih T.-C, & Hsu K.-C. (2009). A two-stage architecture for stock price

forecasting by integrating. *Expert Systems with Applications 36*, 7947–7951.

Huang C., Gong X., Chen X., & Wen F. (2013). Measuring and Forecasting Volatility in Chineses Stock Market Using HAR-CJ-M Model. *Hindawi Publishing Corporation Abstract and Applied Analysis*, Article ID 143194.

Huang C.-F. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing 12* , 807–818.

Huanhuan Y., Rongda C., & Guoping Z. (2014). A SVM Stock Selection Model within PCA. *Procedia Computer Science 31*, 406 – 412.

Isenah G.M., & Olubusoye O.E. (2014). Forecasting Nigerian Stock Market Returns using ARIMA and Artificial Neural Network Models. *CBN Journal of Applied Statistics Vol 5 No 2*.

Kara Y., Boyacioglu M.A., & Baykan O.K. (2011). Predicting direction of Stock index movement using artificial neural networks and support vector machines: The sample of the Instanbul Stock Exchange. *Expert Systems with Applications 38*, 5311-5319.

Kazem A., Sharifi E., Hussain F.K., & Saberic M. (2013). Support Vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing 13*, 947-958.

Khasei M., & Bijari M. (2010). An artificial neural network (p,d,q) model for time series forecasting. *Expert Systems with Applications 37*, 479-489.

Khashei M., & Bijari M. (2010). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing 2*, 2664-2675.

Khashei M., & Bijari M. (2012). A new class of hybrid models for time series forecasting. *Expert Systems with Applications 39*, 4344-4357.

Khashei M., Bijari M., & Ardali G.A.R. (2012). Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks (PNNs). *Computers & Industrial Engineering 63*, 37-45.

Lahmiri S. (2014). Wavelet low- and high-frequency components as features for predicting stock prices with backpropagation neural networks. *Computer and Information Sciences 26*, 218-227.

Li Y., & Ma W. (2010). Application opf Artificial Neural Networks in Financial Economics: A Survey. *International Symposium on Computational Intelligence and Design*, (pp. 211-214).

Liao Z., & Wang J. . (2010). Forecasting model of global stock index by stochastic time effective neural network. *Expert Systems with Applications 37* , 834–841.

Liu C.-F., Yeh C.-Y., & Lee S.-J. (2012). Application of type-2 neuro-fuzzy modeling stock price prediction. *Applied Soft Computing 12*, 1348-1358.

Lu C.-J. . (2010). Integrating independent component analysis-based denoising scheme with neural network for stock price prediction. *Expert Systems with*

Applications 37 , 7056–7064.

Luo L. , & Chen X. . (2013). Integrating piecewise linear representation and weighted support vector machine for stock trading signal prediction. *Applied Soft Computing 13* , 806–816.

Magaji A.S., & Adeboye K.R. (2014). An Intense Nigerian Stock Exchange Market Prediction Using Logistic with Back-propagation ANN model. *Science World Journal Vol 9 (No 2),* 8-13.

Magaji A.S., Isah A., Waziri V.O, & Adeboye K.R. (2013). A Conceptual Nigeria Stock Exchange Prediction: Implementation Using Support Vector Machines-SMO Model. *World of Computer Science and Information Technology Journal Vol 3. No 4.*, 85-90.

Majhi B., Rout M., & Baghel V. (2013). On the development and performance evaluation of a multiobjective GA-based RBF adaptive modelfor the prediction of stock indices. *ournal of King Saud University-Computer and Information Sciences 26*, 319-331.

Mathew O.O., Sola A.F.,Oladiran B.H., & Amos A.A. (2013). Prediction of Stock price using Autoregressive Integrated Moving Average Filter ((Arima(P,D,Q)). *Global Journal of Science Frontier Research Mathematics and Decision Sciences Vol 13 Issue 8*.

Merh N., Saxena V. P., & Pardasani K.R. (2010). A comparison between hybrid approaches of ANN and ARIMA for Indian stock rend forecasting. *Business Intelligegence Journal Vol. 3 No. 2*, 23-43.

Mostafa M. M. . (2010). Forecasting stock exchange movements using neural networks: Empirical evidence from Kuwait. *Expert Systems with Applications 37*, 6302–6309.

Neenwi S., Asagba P.O., & Kabari L.G. (2013). Predicting the Nigerian Stock Market using Artificial Neural Network. *European Journal of Computer Science and Information Vol 1 No 1*, 30-39.

Nguyen D.-H., & Le M.-T. (2014). A two-stage architecture for stock price forecasting by combining SOM and Fuzzy-SVM. *International Journal of Computer Science and Information Security Vol 12 No 8.*

Ni L.-P., Ni Z.-W., & Gao Y.-Z. . (2011). Stock trend prediction based on fractal feature selection and support vector machine. *Expert Systems with Applications 38* , 5569–5576.

Nikfarjam A., Emadzadeh E., & Muthaiyah S. (2010). Text mining approaches for stock price prediction. *http://www.researchgate.net/publication/224/32689*.

Oh C., & Sheng O.R. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. *32nd International Conference on Information Systems.* Shanghai.

Olatunji S.O., Al-Ahmadi M.S., Elshaferi M., & Fallatah Y.A. (2013). Forecasting the Saudi Arabia stock prices based on artificial neural networks model.

International Journal of Intelligent Information Systems, 77-86.

Park K., & Shin H. . (2013). Stock price prediction based on a complex interrelation network of economic factors . *Engineering Applications of Artificial Intelligence 26* , 1550–1561.

Pele D.T., & Mazurencu N. (2012). Modelling stock market crashes: the case of Bacharest Stock Exchange. *Procedia-Social and Behavioral Sciences 58*, 533-542.

Preethi G., & Santhi B. (2012). Stock Market Forecasting Techniques: A Survey. *Journal of Theoretical and Applied Information Technology Vol 46 No 1*.

Rounaghi M. M., Abbaszadeh M. R., & Arashi M. . (2015). Stock price forecasting for companies listed on Tehran stock exchange using multivariate adaptive regression splines model and semi-parametric splines technique. *Physica A 438* , 625–633.

Sakarya S., Yaruz M., Karaoglan A.D.,& Ozdemir. (2015). Stock Market Index Prediction with Neural Network During Financial Crisis: A review on BIST-100. *Financial Risk and Management Reviews*, 53-67.

Santosh K. D., Abhishek K., Bappaditya D., & Burnwal A.P. (2013). On soft computing techniques in various areas. *Computer Science & Information Technology (CS & IT)*, 59–68.

Schumaker R.P., Zhang Y., Huang C.-N.,& Chen H. (2012). *Evaluating Sentiments in Financila News Articles.* New Britain, Connecticut 06050, USA.: Department of Management Information Systems, Central Connecticut State University.

Shen W. , Guo X. , Wu C. , & Wu D. . (2011). Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm. *Knowledge-Based Systems 24*, 378–385.

Sheng Y., & Subhash K. (2012). *A Survey of Prediction Using Social Media.* Oklahoma: Department of Computer Science, Oklahoma State University.

Subhabrata C., Subhajyoti G., Arnab B., Kiran J. F., & Manoj K. T. (2014). A realtime clustering and SVM based price-volatility prediction for optimal trading strategy. *Neurocomputing 131*, 419–426.

Sureshkumar K.K., & Elango N.M. (2012). Performance Analysis using Artificial Neural Network. *Global Journal of Computer Science and Technology Volume 12 Issue 1*, 19-26.

Suwandi E., & Santica D.D. (2014). Prediction of Jakarta Composite Index Using Least Squares Support Vector Machines Approach. *Journal of Theoretical and Applied Information Technology Vol 63 No 2*.

Ticknor J. L. . (2013). A Bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications 40* , 5501–5506.

Tsai C.-F., Lin Y.-C., Yen D. C., & Chen Y.-M. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing 11*, 2452–2459.

Tsai C.-F., Hsiao Y.-C. . (2010).

Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems 50* , 258–269.

Vaisla K.S., & Bhatt A.K. (2010). An analysis of the performance of artificial neural network technique for stock market forecasting. *International Journal on Computer Science and Engineering Vol 02*, 2104-2109.

Wang J.-J., Wang J.-Z., Zhang Z.-G., &Guo S.-P. (2012). Stock Index Forecasting based on a hybrid model. *Omega 40*, 758-766.

Wang J.-Z., Wang J.-J., Zhang Z.-G., & Guo S.-P. (2011). Forecasting stock indices with back propagation neural network. *Expert Systems with Applications 38*, 14346–14355.

Wang L., & Wang Q. (2011). Stock market prediction using artificial neural networks based on HLP. *2011 Third International Conference on Intelligent Human-Machine Systems and Cybernetics* (pp. 116-119). IEEE Computer Society.

Wei K., & Cai-hong S. (2011). Building the model of artificial stock maret based on JASA. *Procedia Engineering 23*, 835-841.

Wei L.-Y., C. T.-L.-H. (2011). A hybrid model based on Adaptive network based fuzzy inference system to forecast Taiwan stock market. *Expert Systems with Applications 38*, 13625-13631.

Wei., L.Y. (2013). A hybrid model based on ANFIS and adaptive expectation genetic algorithm to forecast TAIEX. *Economic Modelling 33*, 893-899.

Wen Q. , Yang Z., Song Y., Jia P. . (2010). Automatic stock decision support system based on box theory and SVM algorithm. *Expert Systems with Applications 37*, 1015–1022.

Wu Y., Gaunt C., & Gray S. (2010). A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting & Economics 6*, 34-45.

Yeh C.-Y. , Huang C.-W., & Lee S.-J. . (2011). A multiple-kernel support vector regression approach for stock market price forecasting. *Expert Systems with Applications 38* , 2177–2186.

Yixin Z., & Zhang J. . (2010). Stock data analysis based on BP neural network. *2010 Second International Conference on Communication Software and Networks* (pp. 396-399). IEEE Computer Society.

Yu H., Chen R., & Zhang G. (2014). A SVM stock selection model within PCA. *Procedia Computer Science 31*, 406-412.

Yu S., & Kak S. (2012). *A survey ofprediction using social media.* Okhlahoma, USA 74078: Department of Computer Science, Oklahoma state university.

Yu T.H., & Huarng K.-H. (2010). A neural network based fuzzy time series model to improve forecasting. *Expert Systems with Applications 37*, 3366-3372.

Zahedi J., & Rounaghi M. M. . (2015). Application of artificial neural network models and principal component analysis method in

predicting stock prices on Tehran Stock Exchange. *Physica A 438* , 178–187.

Zarandi M. H. F., Hadavandi E.,Turksen I. B. (2012). A Hybrid Fuzzy Intelligent Agent-Based System for Stock Price Prediction. *International Journal of Intelligent Systems Vol 00*, 1-23.

ZheGao, & Yang J. (2014). Financial Time Series Forecasting with Grouped Predictors using Hierarchical Clustering and Support Vector Regression. *International Journal of Grid Distribution Computing Vol 7*, 53-64.

Covenant Journal of Informatics & Communication Technology. Vol. 5 No. 2, Dec, 2017

An Open Access Journal Available Online

Human Security for Sustainable Development in Nigeria: The Role of Information and Communication Technology (ICT)

Lukman Muhammad Bashar

Shehu Shagari College of Education, Sokoto, Nigeria
lukman.bashar@gmail.com

Abstract: Human security is needed in response to the complexity and the interrelatedness of both old and new security threats ranging from chronic and persistent poverty to ethnic violence, human trafficking, climate change, health pandemics, international terrorism, and sudden economic and financial downturns. The paper: Human Security for Sustainable Development in Nigeria: The Role of Information and Communication Technology (ICT) looks at the role of improved technology in human security and development, analyzes technology utilization in addressing human security and development issues in Nigeria. It recommends among others increasing the accessibility of information technology to reduce the threat to sustainable livelihood.

Keywords: Human Security, Sustainable Development, Information and Communication Technology (ICT)

1. Introduction

The development of any nation is usually measured by the degree and extent of the sociocultural, socio-economic, and economic improvements that are brought to bear through the enterprises of Science, Technology, and Mathematics. One of the many roles of the state is to provide peace, security, and a platform for development for its citizens.

Human security focuses primarily on protecting people while promoting peace and assuring sustainable

continuous development. It addresses issues such as organized crime and criminal violence, human rights and good governance, genocide and mass crimes, resources and environment. It emphasizes aiding individuals by using a people centered approach for resolving inequalities that affect security.

Information and communication technology (ICT) is a term that has several meanings across different sectors. It is used as an umbrella term to refer to the use of communication devices such as radio and cellular devices, satellite devices and channels, computers and utilities to manage the acquisition, dissemination, processing, storage, and retrieval of information (Ogu, Oyeyinka 2014).

The introduction of ICT into many aspects of everyday life has led to the development of the modern concept of the information society. This development offers great opportunities such as improving both access to education, and the quality of that education. Insecurity in Nigeria is at a level where it is necessary to deploy technological systems and professionals to help fight all crimes and corruption in the country.

Having a clear picture of issues, leads to a better understanding of events. In order to give this paper a better focus, basic concepts would be clarified. The major concepts of this paper are: human security, Information and communication technology (ICT), and sustainable development. It discusses the security challenges in Nigeria and the role (ICT) could play in ensuring human security and sustainable development of the Nigerian society.

2. Human Security

Security entails safety, it is protection against harm. It is the protection of a country, a building, or a person against attack or danger (Isaac, 2007). It has been defined as a state of well-being characterized by freedom from danger, risk, lack, uncertainties etc. (Nwankwo, 2013).

Human security is a concern for the well-being of human beings. It is the protection of human beings from threats and risks. It entails men and women having security at home, in the offices, and within the community. United Nations Development Programme (UNDP) report of 1994 defined human security as:

i. Safety from chronic threats such as hunger, disease and repression.
ii. Protection from sudden and hurtful disruptions in the pattern of daily life whether in jobs, in homes or communities.

Human development report highlights two major components of human security. These are: freedom from fear, and freedom from want. Human security therefore entails ability of individuals to live in peace and harmony free from such threats as disease, hunger, unemployment, political oppression, environmental degradation etc. justice, fair-play, tolerance, protection of human rights and a level playing ground for all citizens to participate

3. Information and Communication Technology (ICT)

Today's world is knowledge-based and technology-driven. Technology has widened the scope of interaction among people all over the world and it has led to emergence of new and innovative ways of doing things. This has turned the world into a global village resulting to a technological revolution that cannot be ignored. To progress in this world therefore, individuals and groups within societies must acquire, utilize and

communicate knowledge. Knowledge is power and a means of empowering all citizens. A knowledgeable citizenry is a productive one. As a result of this, nations across the globe are placing more and more emphasis on the acquisition of technological capabilities for all citizens.

Information and communication technology has been defined as modern equipment and tools that includes hardware, software, networks and media for collecting, storing, processing, transmitting, and presenting information (World Bank, 2002). Computer hardware and software network and other digital devices such as radios, videos, television etc. convert information in form of text, sound, or motion into digital forms. Mohammed (2007) defined ICT as an umbrella term that encompasses all technologies for the manipulation and communication of information. Wuru (2008) also defined ICT as a wide range of technologies that are enormous and powerful tools for development. It is a wide range of technologies that includes telephones (land and cellular), computer, satellite, telex fax, radios, television, videos etc. (Ikemelu, 2015). Nwabueze and Ozioka (2016) also defined ICT as a broad-based technology that supports the creation, storage, manipulation and communication of information.

Information and communication technology plays an important role in the use of the individual and society. It has brought tremendous awareness and new technologies are tremendously changing our world. They create jobs, transform education, healthcare delivery, and politics.

They help in the delivery of humanitarian assistance and contribute to security (Wuru, 2008).

4. Sustainable Development

Development is associated with progress, advancement, and the ability to provide for the material well-being of all citizens. It is the advancement in the social, economic, political, and spiritual well-being of all citizens. It involves social harmony and economic growth.

Sustainable development is the development that leads to the fulfilment of societal ideals. United Nations General Assembly in 1987 defined sustainable development as the development that meets the needs of the present generation without compromising the ability of future generation to meet their own needs. It is development that is regenerating and self-sustaining. It is a development that is needed to maximize the output of citizens. The satisfaction of physical, mental, spiritual needs and the mastery of environment are the parameters of development when applied to human society (Nbuezor & Ozioka 2015).

Sustainable development is only possible when the capacity of human beings is built through the process of human resource development which education is the regulator. Education is ideally and organically linked with the production process. Education is concerned with imparting knowledge and a means of enabling individuals to tackle personal and societal problems. It provides appropriate knowledge, skills, attitudes, abilities, and competence necessary for undertaking specific tasks and functions. Education improves human relationships, ensures economic growth, healthcare, effective citizenships, national consciousness, and national unity for enhanced human security.

A functional education promotes manpower development. It produces

competent men and women who can apply knowledge to solve personal and societal problems. That is, it provides training and skills for production of craftsmen/women, technicians, technologists and other skilled personnel who are enterprising and self-reliant.

However, for any education system to be relevant to the developmental needs of the society. It must develop the creative ability of individuals especially in the cultural and technological realms. It must foster in the individuals those values which make for good citizenship such as honesty, selflessness, tolerance, dedication, hard work and personal integrity. It must train the individual to relate and interact meaningfully with other individuals in the society and appreciate the importance of effective organization for his progress. It must also promote the culture of productivity by enabling every individual to discover the creative genius in him/her and apply it to the improvement of his existing skills.

The most important skills required for success in todays' world is information and communication technology (ICT). ICT equips individuals with critical wealth of skills, technical knowledge and diversity of understanding, values, and attitudes that are needed in order to live happily and contributes meaningfully to the development of the society. ICT provides limitless possibilities, allows opportunities in every field of human endeavor. Therefore, the ability to access and use information is no longer a luxury but a necessity, this underscores the need for basic knowledge of ICT for all citizens.

5. Security Challenges in Nigeria

Justice, equity, fair-play, and respect for the dignity of individuals are needed for responsible living within the society. In this respect, every citizen deserves the right to live in an environment that is free from social antagonism. That is, every citizen needs to be free from traumatic experiences, dysfunctional relationships and unsatisfactory conditions of life. Every citizen also deserves the right to resource information and the freedom of action to be able to fulfill social responsibilities. It is the responsibility of government to ensure that all citizens have equal right, obligations, and opportunities before the law.

In the last few years however, Nigerian citizens have suffered from social distinctions, and inequalities in the distribution of resources, social rights, privileges, and power thereby widening the gap between the haves and have not's. Poverty is now widespread and it is a leading factor to crime. Corruption, high rate of inflation, mismanagement land misappropriation of public funds due to poor governance have led to inability of government to protect and support its citizens physically, socially, and emotionally.

The Nigerian society is now witnessing serious political, religious, and ethnic disturbances, economic distress and high rate of youth unemployment which has made Nigerian youth to indulge in drug use and abuse, gangstarism, armed robbery, political thuggery, and other social vices leading to frequent crisis and violence. Absence of peace leads to insecurity and without security of human beings, sustainable development of the society is not possible.

6. Role of ICT in Sustainable Development

Economic security provides a durable foundation for peace and stability to prevail. Economic security however, depends on steady, regular, and

adequate income and gainful employment (Attah & Kyari, 2014). People must have a source of income to be responsible members of their families and the society at large. In a distressed economy, there is a high cost of basic necessities of life, high rate of inflation such that people cannot meet their basic needs of life. There is thus an unsatisfactory condition of life.

Unemployment and poverty are among the major symptoms of economic insecurity which deprive people from meeting their basic needs of life. Under such circumstances, children and youth are left uncartered for by their parents and the society. Many children and youth are forced out of school no matter their ability because their parents cannot afford the payment of school fees and other educational requirements. Those that are able to continue to graduation level hardly acquire more than the basic skills of reading, writing, and arithmetic. Many of them are now roaming the streets with certificates that cannot fetch them means of livelihood because they have not acquired relevant skills for gainful employment.

Education must prepare people for the future. To do this, people must learn with technology, and about technology. Amongst the yardsticks for measuring success at any education system is the marketability of the graduates of the system. Information and Communication Technology (ICT) provides unparalleled degree of communication collaboration, resources sharing, and unlimited access to information for more powerful and complete knowledge building. In ICT, skills are deployed and integrated with other types of knowledge and skills in a technological environmental context. That is technologically vibrant individuals.

The high rate of unemployment in Nigeria could be attributed to lack of equilibrium between skills needed in the labor market and the training received b youth. ICT trains the youth for economic and social responsibilities with which to convert poverty and promote peace. ICT provides relevant and comprehensive intellectual and vocational skills to meet the requirements of skilled manpower and improve access to jobs as an important requirement of the economy. Industrialization of the economy requires the production of competent engineers and technologists. ICT leads to emergence of new employment categories. Through the use of ICT, the world has become a global village where everyone can be reached.

Technological advances in transportation have made it possible for workers and traders to reach their destination in good time

ICT could be used in the laboratories to carry out experiments, monitor medical laboratory activities to provide efficient medical and veterinary doctors, pharmacists, nurses, radiologists, astronomers, engineers, architects, pilots, research scientists, and science educators.

ICT is used in saving lives, flying aircrafts, running nuclear power plants, processing orders, controlling production, making bookings, transferring vast amounts of money, controlling missile system and in the enhancement of educational practices. People can move about searching for new items in the world market, check for prices, and place order on what they need just in their living room.

7. Conclusion

It is evident from the discussion so far that ICT education for all individuals is not just a luxury, but rather a necessity. Information and Communication Technology (ICT) network is the basic facility through which information needs of industry, commerce and agriculture can be satisfied. Industrial development requires the coordination of a series of operation, including the acquisition of supplies, recruitment of labor, control stocks, processing of materials, and delivery of goods to buyers, as well as billing and record keeping. Information technology is vital to the effective development and control of many of these operations. Commerce is essentially on information processing activity; effective buying, selling and brokerage rely on the continual supply of up-to-date information regarding the availability of prices of goods and services. Farmers on the other hand, must not only grow food but they must sell effectively and buy seeds and fertilizer. They also need information on weather conditions, disease outbreaks and new agricultural techniques.

References

Agommuoh P.C (2015). Enhancing the Teaching of Physics Through the Use of ICT in Senior Secondary Schools. Proceedings of 56th Annual Conference of Science Teachers Association of Nigeria (STAN). Njoku Ed. University Nigeria Press Limited.

Attah S. C & Kyari Y.B (2014). Social Security as a Brand of Human Security: An Approach for Peace and Development in Nigeria. A Journal of Research Issues and Ideas. 13 (2). A Publication of Kashim Ibrahim College of Education Maiguguri, Nigeria. December.

Ikemelu C. R (2015). Towards Effective Application of ICT Education for Classroom Curriculum Delivery: Science Teacher Perspectives. Proceedings of 56th Annual Conference of Science Teachers Association of Nigeria (STAN). Njoku Ed University Nigeria Press Limited.

Mohammed A.A (2007). Role of University Education in the Economic Development of Nigeria. Journal of Educational Management and Planning (JEMP) 7(1) September.

Mohammed S. (2007). Educational Reform Through the Use of ICT to Improve Teaching and Learning in Nigeria Secondary Schools. Journal of Education Management and Planning (JEMP) 7(1) September.

Nwanko J. I (2013): Managing Education for National Security. Journal of Nigerian Association for Educational Administration and Planning (NAEAP). Uneage Publishing House Ibadan, Nigeria.

Shitu F. M (2014). Information and Communication Technology (ICT) in the English Language Class: An Overview of Issues, Problems, and Recommendations. Tambari Kano Journal of Education. A Journal of Federal College of Education, Kano. 9(13) December.

United Nations General Assembly (1987). Report of the World Commission on Environmental Development and International Cooperation.

Wuru MM (2008). The Teaching of Computer Science as a Prerequisite for E-Learning in the Current Education Reforms. Nigerian Journal of Professional Teachers. 1 (5). An International Journal of the Teachers Registration Council of Nigeria.

Covenant Journal of Informatics & Communication Technology. Vol. 5 No. 2, Dec, 2017

An Open Access Journal Available Online

Detecting Malicious and Compromised URLs in E-Mails Using Association Rule

Nureni Ayofe Azeez[1] & Emilia Anochirionye[2]

[1,2]Department of Computer Sciences,
Faculty of Science, University of Lagos, Nigeria
nazeez@unilag.edu.ng, emygreat1912@yahoo.com

Abstract: The rate of cybercrime is on the rise as more people embrace technology in their different spheres of live. Hackers are daily exploiting the anonymity and speed which the internet offers to lure unsuspecting victims into disclosing personal and confidential information through social engineering, phishing mails and sites and promises of great rewards which are never received. Thus resulting in great loss of property, finances, life, etc. and harm to their victims. This research work seeks to evaluate ways of protecting users from malicious Uniform Resource Locators (URLs) embedded in the emails they receive. The aim is to evaluate ways of identifying malicious URLs in emails by classifying them based on their lexical and hostname features. This study is conducted by extracting features from URLs sourced from phishing tank and DMOZ and adopting Association Rule of classification in building a URL classifier that analyzed extracted features of a URL and use it in predicting if it is malicious or not. 0.546 level of accuracy and an error rate of 0.484 was achieved as multiple URL features were employed in the classification process.

Keywords/Index Terms: Malicious, Association rule, URLs, Cybercrime, Hackers

1. Introduction

Information and Communication technology evolution has changed the way in which businesses are conducted all over the world. Prior to this era, messages where exchanged through

courier and postal services, businesses where confided within the walls of an organization and managing communication within and outside a business was pretty tedious. Contrarily, the advent of internet has increased the speed and automations with which businesses transact and communicate (Ayofe et. al., 2010). Irrespective of the institution, online visibility has become a key criteria for business survival and competitiveness within today's global business environment. It is however noted that online transactions are seriously being hampered across the globe because of insecurity (Azeez et. al., 2015). This is because, cybercriminals all over the globe are advancing on daily basis on their strategies to dupe and cause great financial loss on the internet users. This they achieve by sending fake and compromised URLs to internet users that perform online transactions (Azeez and Venter, 2013). Once such a user falls prey of their actions, they lose nearly if not all the money that is contained in their bank account. In an attempt to solve this challenge, researchers have adopted different data and machine learning algorithms to identify compromised URLs being used by phishers. Doing this has undoubtedly assisted internet users to identify which of the URLs sent to them is fake and phony. In this work, the authors proposed association rule for detecting malicious and compromised URLs in electronic mail. It our strong believe that this approach will supplement efforts made so far by researchers in this regard and specifically in the area of cybersecurity (Azeez and Ademolu, 2016).

1.2 Statement of problem

The benefits accruing from the internet evolution poses security challenges bordering around the preservation of intellectual, financial and personal information from fraudsters, who pose as legitimate internet users to steal valuable information from unsuspecting victims through vicious means such as malware, phishing mails, pharming, spams etc. Cybercrime has become a fast growing area of crime. Statistics from Forbe (Morgan, 2015) shows that cyber-attacks cost businesses as much as $400 to $500 billion a year excluding the cost of unreported cases. It is speculated that by 2017, the global cyber security market would skyrocket to $120.1 billion.(PWC, 2013)

To mitigate against the problem of users exposure to malicious links embedded in website or emails, the need to educate the populace on the risk of opening emails or attachments from unknown sender, clicking on links embedded in emails and ways of identifying a malicious link cannot be emphasized. However, this approach may not be sufficient as hackers rapidly adopts different ways of masking malicious URLs in legitimate emails and websites.

The alternative of matching suspected URLs against a blacklist which has been overly utilized, leaves the threat of a malicious URL going unidentified long before it is blacklisted. Currently, many researches have carried out researches and are still doing more in identifying malicious URLs using learning algorithms. This necessitates the need to evaluate the efficiency of this approach over the users' awareness campaigns and education on social engineering and malicious URL identification and adopting blacklist checks.

This research work centers on the "Identification of Malicious URL in Electronic Mails" using association rule and it aims to achieve the following:

- Implementation of a tool that extracts URLs from received electronic mails.
- Implementation of a URL analyzer that carries out analysis on URL based on some pre-defined features.
- Analyze the performance level in using each url features by using association rule

2. Literature Review

Previous works done in classifying URLs using machine learning algorithms adopted the methodology of classifying URLs based on the content of the website (Deri, 2015), which requires accessing and evaluating the contents of a website in order to ascertain its legitimacy.

Wardman and Warner used a technique that computes the similarity between the content files from potential and known phishing websites (Warner & Wardman, 2008). Ludl et al. classified phishing websites using features extracted from the main phishing webpage (Ludl, McAllister, Kirda, & Kruegel, 2007). This approach though more effective than traditional blacklist, poses a workload overhead on the classifier.

Work done by Soon and Jeffrey, evaluated the identification of a URL genuineness using Favicons (Soon & Jeffrey, 2014). This approach utilized google search-by-image API and semantic analysis in achieving 97.2% true positive in its classification.

Kan et. al. implemented URL classification by extracting features (Kan, 2005) in the URL and using them in the classifier. Initially features where extracted based on punctuation marks using tokens as the classifier's feature-set. Afterwards, researchers either used statistical (involved using statistical methods such as mean, variance calculation) to analyze the extracted information content (Kan, 2004) or brute-force approach to further segment URLs beyond punctuation marks. All possible sub-strings, n-grams in a URL are used as the classifier's feature-set in the brute-force approach (Iglesia, 2015).

Some of the URL features set utilized by researchers irrespective of their approach to URL classification using machine learning includes: lexical features (i.e URL features that URL string properties not considering host or page content) and External features (that is features that require querying a remote server) or a hybrid of the two features set to improve the performance of the classifier (Zhu, Lee, & Choi, 2001).

3. Proposed Framework

Process design is a technique used to describe the processes that transform data into useful information. It encompasses the flow of data through a system's process and /or logic, policies. This involves all the procedures to be implemented by a system's process. The flow chart below is used to explain the process flow of data as well as the policies that governed data flow and usage within the system.

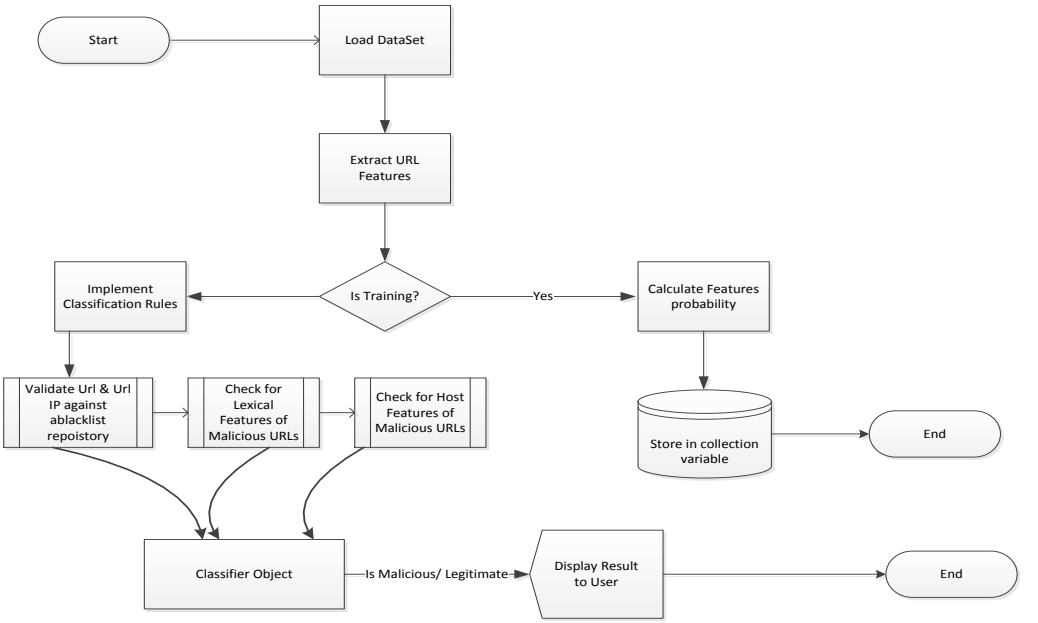

Figure 1: Process Flow Diagram

Algorithm 1: Association Algorithm

Algorithm: Association Rule

Rules:

Rule 1: if(URL IP is contained in IP blacklist)== true }class is malicious

Rule 2: If (URL is contained in blacklist database)==true } class is malicious

Rule 3: if (URL resolves to an IP Address) ==false } class is malicious

Rule 4: if (p(Malicious | Xo..Xn) > p(Legitimate | Xo..Xn)} class is malicious

Learning output Screenshot

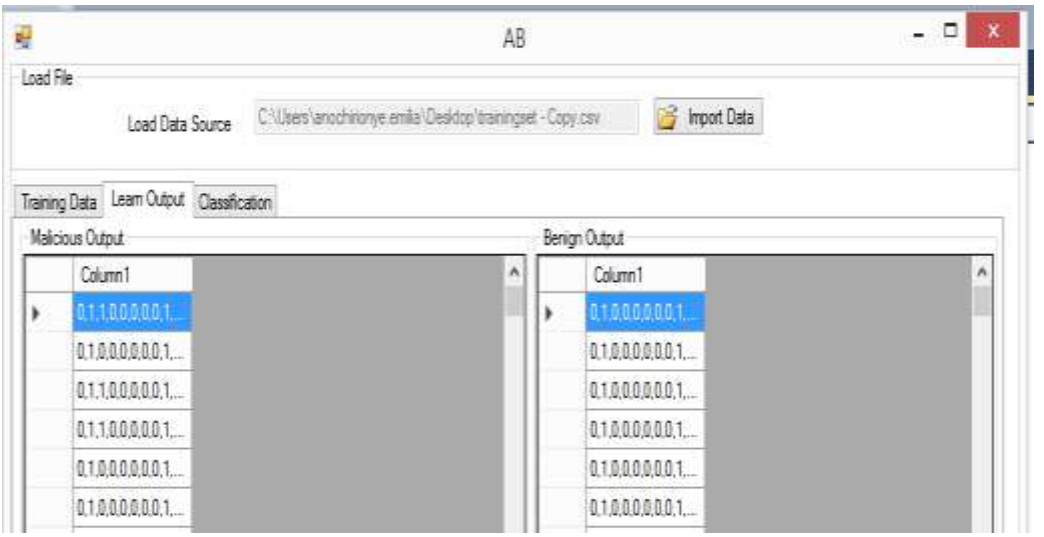

Figure 2: Training Data Result Screen

Classification Screenshot

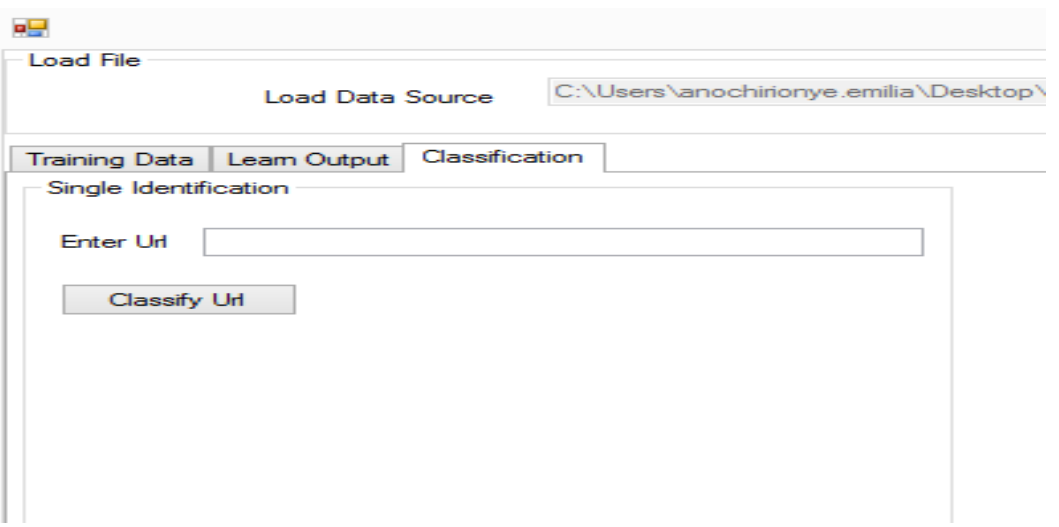

Figure 3: Single URL classification Screen

The screenshot for classifying URLs extracted from Emails is as shown below:

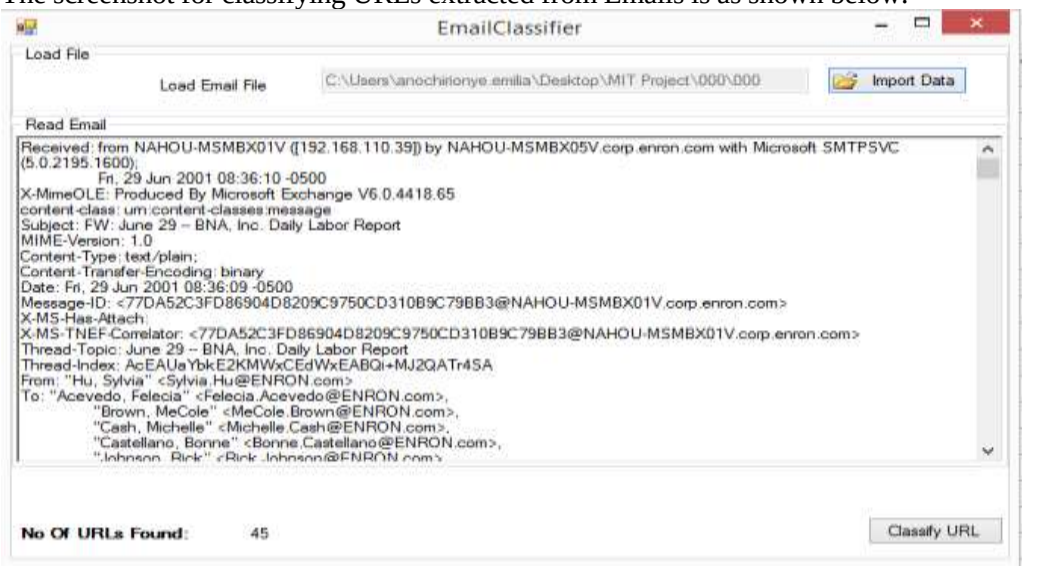

Figure 4: Classification Screen for urls in Emails

4. Output Design

This work has been designed to show users through screen outputs which can be exported out in excel format. The output screen are populated and presented to user mainly after features extraction or URL classification has been completed. The output report format adopted are "detailed reporting"; which one or more lines of output for each record processed is displayed. Each line of output printed is called a "Detail Line".

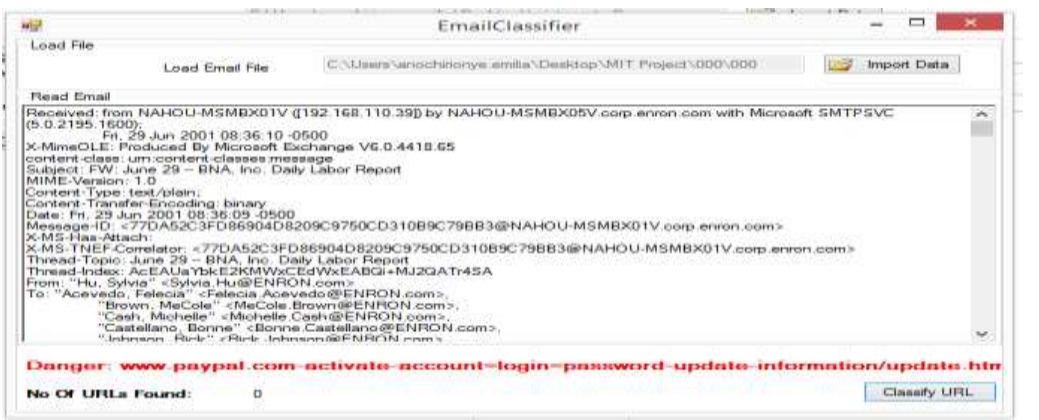

Figure 5: Email Test Data Result Screen

Figure 5 shows the screenshot obtained when some email dataset were used for the evaluation.

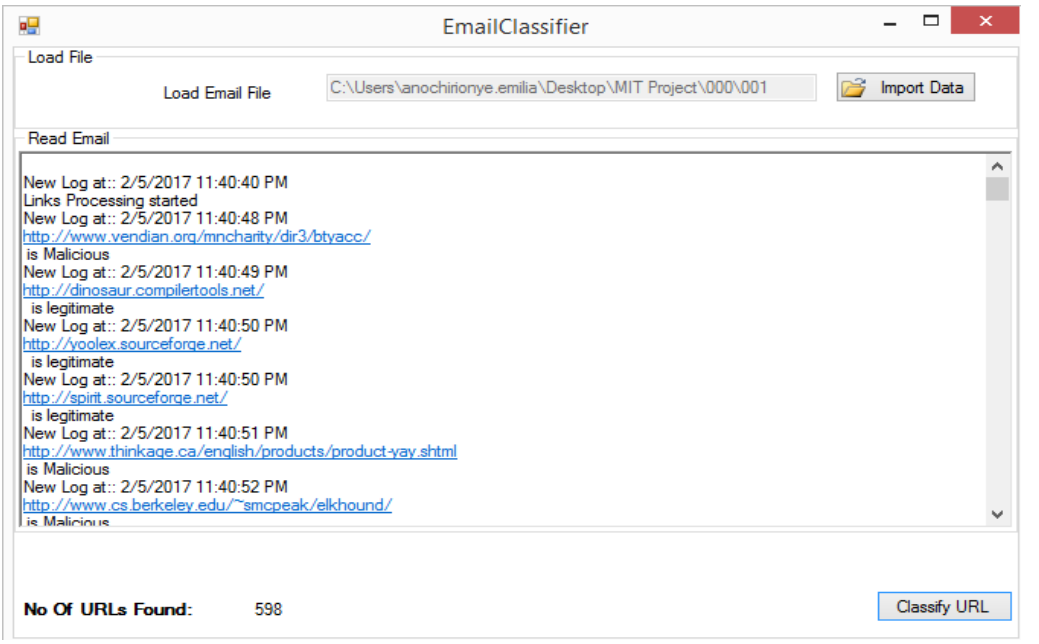

Figure 6: Email Test Data Result Screen

TABLE I: ANALYSIS OF FEATURES IN DATASETS

S/NO	Features	Frequency in Malicious dataset	Frequency in Legitimate dataset
1	Presence of URL Protocol (http/https) in the URL Path	4.0816%	0.0204%
2	Count of URL Resolved IP Address Not found in IP Blacklist	2.4490%	65.3061%
3	No of slashes in URL	42.8571	0.0204%
4	No of @ sign in domain	0.0204%	0.0204%
5	Count of IP in URL domain	0.0204%	0.0204%
6	Length of domain	0.0204%	0.0204%
7	Dots in domain	0.0408%	0.0204%
8	TLD found in TLDRepository	100%	97.9591%
9	No of Special Characters and keywords in URL	10.204%	0.0204%

Table I shows the corresponding values obtained for each of the features for the frequency of malicious and frequency of legitimate in the dataset. It should be recalled that there are nine (9) different features used for the evaluation. They are all listed in Table I.

5. Features Occurrence Evaluation

The probability of occurrence for each of the above mentioned features (See Table I) in the data classes for Malicious and Legitimate URLs was calculated in the training stage as shown in the Table I and results gotten from the analysis of these features in both Malicious and Legitimate datasets were used to build the predictive model used in detecting malicious URLs in the testing data based on the knowledge gained in the training stage on the rate of occurrences of the features in both the legitimate and malicious class.

i. Postulation of Decision Rule

From the review of the results obtained in the training stage, it was observed that most malicious URLs exhibited the presence of protocols (http/https) in their path, >5 slashes in the URL, unresolved IPs or IPs as their domain and the absence of TLD. Using association rule, which is a rule-based machine learning methodology that uses if/then statements that help uncover relationships between seemingly unrelated data in an information repository (Rouse, 2011), the Algorithm I was adopted in classifying the URLs in stage 3.

ii. Detecting Malicious Links within an Email

This is the last stage in the implementation of the URL classifier. It involves the extraction of over 40 URLs from a spam email dataset, extraction of the features predominant in malicious URLs and predicting the class (Malicious/ Legitimate) of each URL based on the knowledge acquired in the training stage. The association rules stated was used to evaluate the accuracy of the classifier.

6. Experimental Results

Results from the training stage on the probability of occurrence of each malicious URL features in the training dataset for the 2 classes (malicious and legitimate) is as shown in Figures 7,8,9 and 10. From the research, it is evident that validating URLs based on pre-defined features stated in the feature extraction stage will provide the following level of certainty for a dataset containing 50 URLs.

IP blacklist =24%

Number of slashes= 42.85%

Availability of special characters: 10.24%

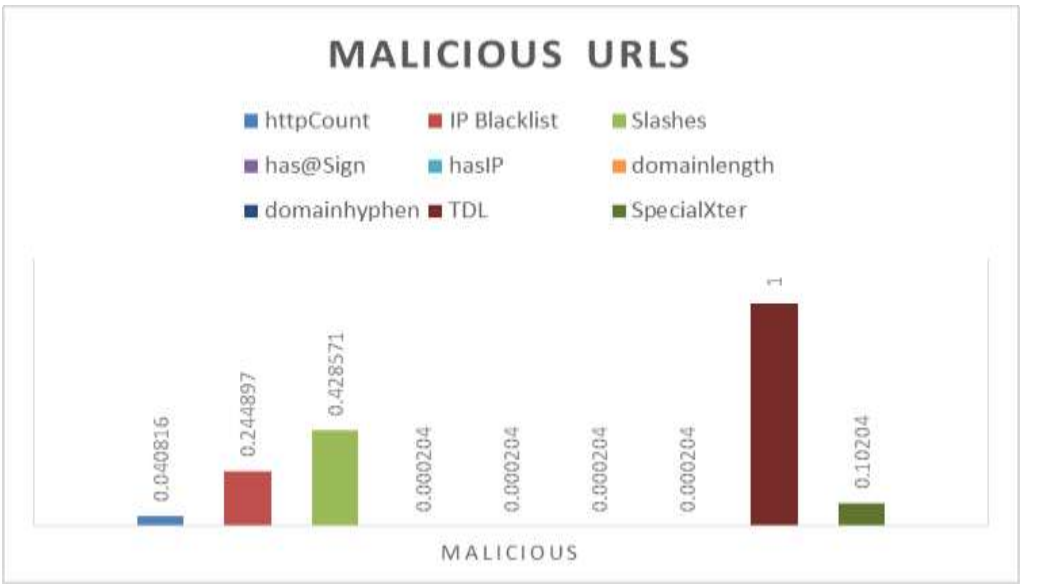

Figure 7: Malicious urls Feature Set Training Result for 50urls

Figure 7 shows the graphical representation of values obtained for malicious URLs with their corresponding features while Figure 8 depicts the graphical representation for legitimate URLs with their respective features.

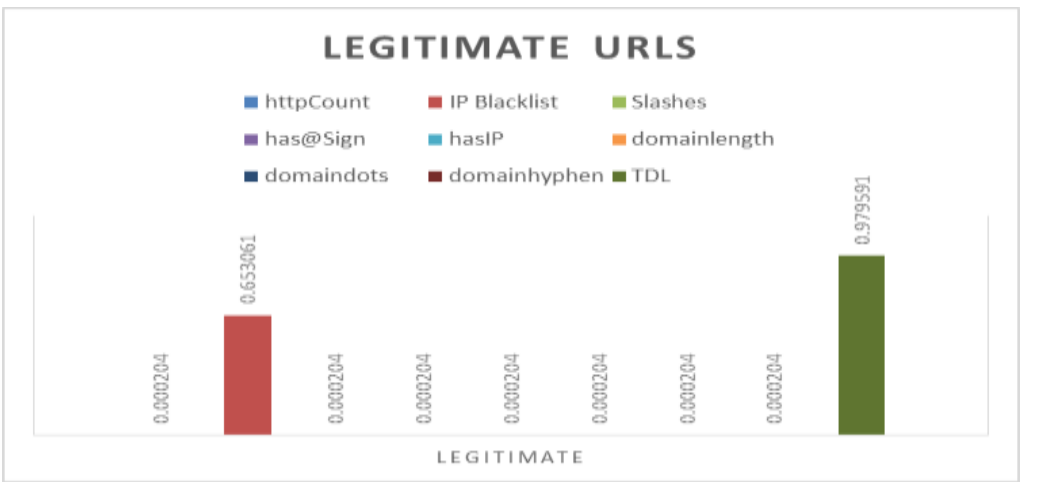

Figure 8: Legitimate Dataset Training Result for 50urls

While the presence of the Top Level Domain (TLD) in the repository, the valid TLDs gave the highest level of certainty.

Increasing the URL in the training dataset to 100, improved the certainty level for identifying malicious URLs using "domain length" and "domain dots" features from 0.000204 to 0.19 in legitimate URLs and 0.14 in malicious URLs.

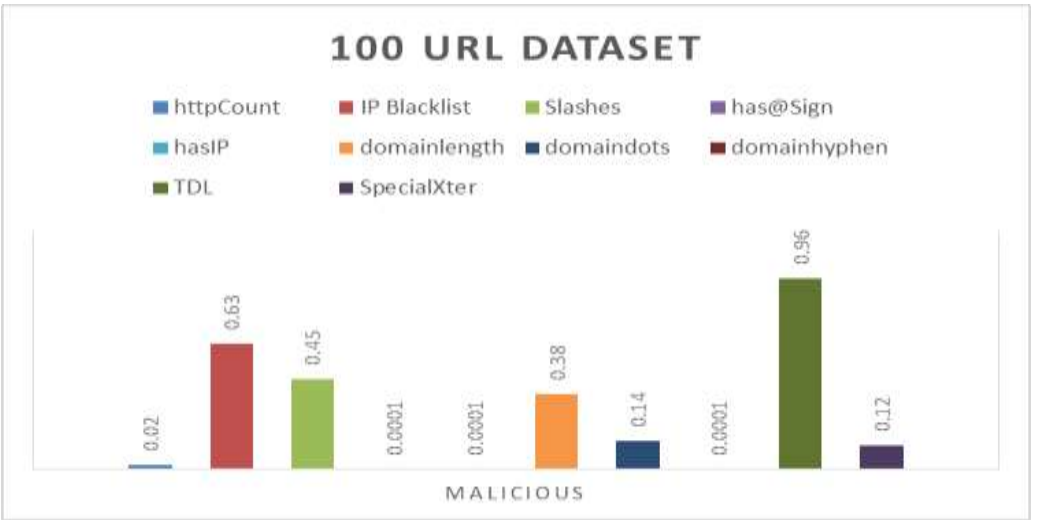

Figure 9: Malicious Dataset Training Result for 100urls

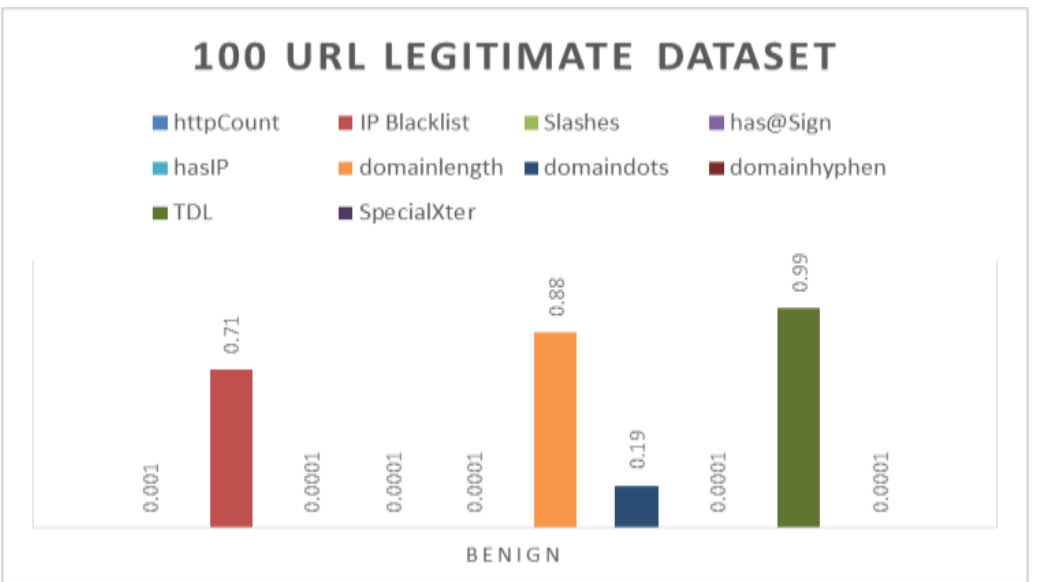

Figure 10: Legitimate urls Training Result for 100urls

URLs that failed detection based on Rules 1- 3 in the Association Rule Algorithm were moved to the classifier object and classification was done using the Bayes' rule by computing the following:

Posterior probability of a new URL(N) being legitimate = Prior Probability for Valid URLs * likelihood of N given a valid URL &

Posterior probability of a new URL(N) being malicious = Prior Probability for Malicious URLs * likelihood of N given a malicious URL.

Out of 100 malicious URLs and 100 legitimate URLs used in testing the classifier, the accuracy and error rates of the classifier were determined with the expression suggested by Damien (François, 2009):

Classification Accuracy = (TP + TN) / (TP + TN + FP + FN)

Error Rate = (FP + FN) / (TP + TN + FP + FN)

Where TP= True Positive, TN= True Negative, FP= False Positive and FN= False Negative.

The classifier developed in this research work, had 0.546 accuracy and an error rate of 0.484.

7. Conclusion

The rules implemented in the development of the URL classification in this research work can serve as a prototype for the development of a more robust solution which will help in identifying malicious URLS in email as well as most existing anti-virus solutions. Also, the continued advancement in the study and implementation of machine learning algorithms in the newly developed systems creates room for improved performance of the designed URL classifier in further works.

The major achievement of this project is the discovery of valid ways of:

Identifying malicious URLs in Electronic mails and

Analyzing URL features in classifying a URL as either legitimate or malicious.

Effort is ongoing to hybridize two different machine learning algorithms other than decision and association rule for the evaluation. Thereafter, a comparative assessment of the result with the association rule will be carried out.

Acknowledgement

The authors wish to acknowledge the efforts of anonymous referees for their valuable comments and helpful suggestions in shaping this paper into a publishable condition.

References

Ayofe, A.N, Adebayo, S.B, Ajetola, A.R, Abdulwahab, A.F (2010) "A framework for computer aided investigation of ATM fraud in Nigeria" International Journal of Soft Computing, Vol. 5, Issue 3 pp. 78-82

Azeez, N. A., & Ademolu, O. (2016). CyberProtector: Identifying Compromised URLs in Electronic Mails with Bayesian Classification. 2016 International Conference Computational Science and Computational Intelligence (CSCI) (pp. 959-965). Las Vegas, NV, USA: IEEE.

Azeez, N. A., & Iliyas, H. D. (2016). Implementation of a 4-tier cloud-based architecture for collaborative health care delivery. Nigerian Journal of Technological Development, 13(1), 17-25.

Azeez, N. A., & Venter, I. M. (2013). Towards ensuring scalability, interoperability and efficient

access control in a multi-domain grid-based environment. SAIEE Africa Research Journal, 104(2), 54-68.

Azeez, N. A., Iyamu, T., and Venter, I. M. (2011). Grid security loopholes with proposed countermeasures. In E. Gelenbe, R. Lent, and G. Sakellari (Ed.), 26th International Symposium on Computer and Information Sciences (pp. 411-418). London: Springer.

Azeez, N.A., and Lasisi, A. A. (2016). Empirical and Statistical Evaluation of the Effectiveness of Four Lossless Data Compression Algorithms. Nigerian Journal of Technological Development, Vol. 13, NO. 2, December 2016, 64-73.

Azeez, N.A, Olayinka, A.F, Fasina, E.P, Venter, I.M. (2015) "Evaluation of a flexible column-based access control security model for medical-based information" Journal of Computer Science and Its Application. Vol. 22, Issue 1, Pages 14-25

Azeez, N. A., and Babatope, A. B. (2016). AANtID: an alternative approach to network intrusion detection. The Journal of Computer Science and its Applications. An International Journal of the Nigeria Computer Society, 129-143.

Azeez, N.A and Otudor, A.E. (2016) "Modelling and Simulating Access Control in Wireless Ad-Hoc Networks" Fountain Journal of Natural and Applied Sciences. Vol 5(2), pp 18-30

Azeez, N.A Abidoye, A.P Adesina, A.O Agbele, K.K Venter, I.M Oyewole, A.S (2013) "Statistical Interpretations of the Turnaround Time Values for a scalable 3-tier grid-based Computing architecture" Computer Science & Telecommunications, Vol 39 (3), pp 67-75.

Zhu, H. L. (2001). Detecting Malicious Web Links and Identifying Their Attack Types. s.l., USENIX Conference on Web Application Development 2011 .

Brownlee, J., (2013). A Tour of Machine Learning Algorithms. [Online] Available at: http://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ [Accessed 01 February 2017].

Ludl, S. M, (2007). On the effectiveness of techniques to detect phishing sites.. Switzerland, Conference on Detection of Intrusions and Malware and Vulnerability Assessment (DIMVA).

Chandak, V., (2012). Parts of URL. [Online] Available at: https://www.virendrachandak.com/techtalk/parts-of-url/ [Accessed 01 February 2017].

François, D., (2009). Binary classification performances measure cheat sheet. Volume 10.

Grauschopf, S., (2016). A Simple Guide to Understanding URLs. [Online] Available at: https://www.thebalance.com/what-does-url-mean-897078 [Accessed 30 1 2017].

IBM, (2017). http://www.ibm.com. [Online]

Available at: http://www.ibm.com/support/knowledgecenter/SSGMCP_5.2.0/com.ibm.cics.ts.internet.doc/topics/dfhtl_uricomp.html [Accessed 30 January 2017].

IEEE, (2015). Large Scale Web-Content Classification. Programming and Systems (ISPS), 2015 12th International Symposium, Volume 10.1109/ISPS.2015.7244974, p. 15438528 .

Iglesia, T. A. (2015). URL-Based Web Page Classification:With n-Gram Language Models, s.l.: Springer International Publishing Switzerland 2015.

Kan, M., (2004). Web page classification without the web page. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters. s.l., ACM, pp. 262-263.

Kan, M. H. (2005). Fast webpage classification using url features.. The Proceedings of the 14th international conference on Information and knowledge, pp. 325-326.

Nureni , A. A., & Irwin, B. (2010). Cyber security: Challenges and the way forward. Computer Science & Telecommunications, 29, 56-69.

Morgan, S., (2015). The Business of Cybersecurity: 2015 Market Size, Cyber Crime, Employment, and Industry Statistics, s.l.: s.n.

PANDA, (2009). Malicious. [Online] Available at: http://www.pandasecurity.com/homeusers/security-info/213894/Malicious [Accessed 01 February 2017].

PWC, (2013). Cybercrime Event, s.l.: PWC Nigeria.

Rajsingh, S. C, (2016). Intelligent phishing url detection using association rule mining. s.l., Cross Mark.

Rouse, M., (2011). Definition: association rules (in data-mining). [Online] Available at: http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining [Accessed 4 February 2017].

Rouse, M., (2016). machine learning. [Online] Available at: http://whatis.techtarget.com/definition/machine-learning [Accessed 01 February 2017].

Soon Fatt Choo, J. e. a., 2014. Phisdentity: Leverage website Favicom to offset Polymorphic Phishing Website. s.l., Conference Publishig Services.

Warner, B. W, (2008). Automating phishing website identification through deep MD5 matching. eCrime Researchers Summit, 29 October, pp. 1-8.

WHUK, (2007). types-of-url-absolute-and-relative. [Online] Available at: https://www.webhosting.uk.com/blog/types-of-url-absolute-and-relative/ [Accessed 31 January 2017].

Covenant Journal of Informatics & Communication Technology. Vol. 5 No. 2, Dec, 2017

An Open Access Journal Available Online

Using Four Learning Algorithms for Evaluating Questionable Uniform Resource Locators (URLs)

Nureni Ayofe Azeez[1] & Opeyemi Imoru[2]

[1,2]Department of Computer Sciences,
University of Lagos, Nigeria.
[1]nazeez@Unilag.edu.ng;
[2]opeimoru@gmail.com

Abstract: Malicious Uniform Resource Locator (URL) is a common and serious threat to cyber security. Malicious URLs host unsolicited contents (spam, phishing, drive-by exploits, etc.) and lure unsuspecting internet users to become victims of scams such as monetary loss, theft, loss of information privacy and unexpected malware installation. This phenomenon has resulted in the increase of cybercrime on social media via transfer of malicious URLs. This situation prompted an efficient and reliable classification of a web-page based on the information contained in the URL to have a clear understanding of the nature and status of the site to be accessed. It is imperative to detect and act on URLs shared on social media platform in a timely manner. Though researchers have carried out similar researches in the past, there are however conflicting results regarding the conclusions drawn at the end of their experimentations. Against this backdrop, four machine learning algorithms:Naïve Bayes Algorithm, K-means Algorithm, Decision Tree Algorithm and Logistic Regression Algorithm were selected for classification of fake and vulnerable URLs. The implementation of algorithms was implemented with Java programming language. Through statistical analysis and comparison made on the four algorithms, Naïve Bayes algorithm is the most efficient and effective based on the metrics used.

Keywords: Malicious URLs; Pharming; Phishing, Attacks; Naïve Bayesian classifier; Decision tree, logistic Regression, k-means

1. Introduction

In the world today, online social networks have become powerful information diffusion platforms as they have attracted hundreds of millions of users. Online Social Networks (Guille et. al., 2013) (OSN) have changed the way people pursue social life and made it easy to connect with family members, classmates, friends and colleagues. In modern times, with increase in population the OSNs have become an easy and a much efficient platform in maintaining social relationships. Online Social Network sites like Facebook, YouTube, Badoo, Twitter, Linkedln, MySpace or Google+ have become popular sites on the Internet. They have attracted all ages from technicians to novice users. In the wide area sphere like research, working office, news media, organizations, entrepreneurship, industries, businesses, OSN have become a daily practice in use (Rao & Saleem, 2015). Most OSN are mainly used for information sharing and to express common interest views like political view, football discussion as well as fashion views etc. (Azeez et. al., 2014).

Its popular usage has been a major concern for the information technology society and experts and has alerted stakeholders to strengthen their defense against unauthorized entities such as malicious programs, Trojan horses, hackers, viruses etc. As online social networks sites have raised in popularity, cyber-criminals started to exploit these sites to spread malware and to carry out frauds (Rao & Saleem, 2015). Recent studies find that around 25% of all status messages in these systems contain URLs, amounting to millions of URLs shared per day. With this opportunity come challenges however from malicious users who seek to promote phishing, malware and other low quality content (Cao & Caverlee, 2015). The theft attacks such as phishing, pharming and spamming that are encountered by malicious e-mail URLs result in several loss to user and may lead to low usage of online services or e-commerce services. As a result of this negative occurrence and unfavorable experience, the authors propose a research work titled "investigating the performance of four learning algorithms for detecting fake and compromised urls". Classification of URLs was based on their lexical features and host-based features and the Naïve Bayes Algorithm, Decision tree model algorithm (ID3) (Azeez & Iliyas, 2016), K means and Logical Regression model Algorithm were used as a probabilistic model to detect if a URL is malicious or legitimate. Figure 1 is a sample phishing website.

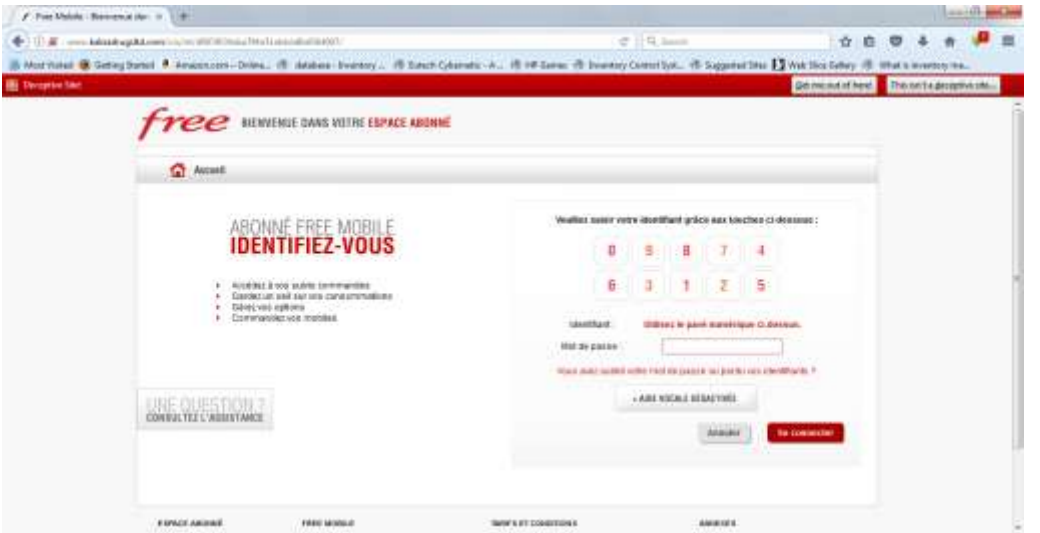

Figure 1: Sample Phishing Website

2. Background/Related Work

Online learning algorithms like Perceptron, Logistic Regression with Stochastic Gradient Descent, Passive Aggressive (PA) Algorithm and Confidence Weighted (CW) Algorithms can be used to detect malicious URLs. Online algorithms are not only used to process large numbers of URLs more efficiently than batch algorithms they can also adapt more quickly to new features in the continuously evolving distribution of malicious URLs as compared to batch learning algorithms. These features include lexical URL features, IP address properties, WHOIS properties, domain name properties, blacklist membership, geographic properties and connection speed. (Ma et. al., 2011) developed a real time system for gathering URL features and compared it with a real time feed of labeled URLs from a large Web mail provider. Using these features and labels, they were able to train an online classifier that detected malicious Websites with 99% accuracy over a balanced dataset. (Ma et. al., 2011) Presented a novel two stage classification model to detect malicious Web pages (Azeez & Venter 2013). They divided the detection process into two stages. In the first stage they have estimated the maliciousness of Web pages using static features.

In the second stage, they used the potential malicious webpages found in the first stage for final identification of malicious web pages by extracting run time features of these webpages (Azeez, 2013). They extracted the static features from contents or properties of webpages without rendering fully or executing the webpages. Potential run time features like foreign contents, script contents and exploit contents were extracted by rendering webpages fully and executing them on specific systems. They used scoring algorithm for the classification.

(Qi & Davison, 2009) evaluated their scoring algorithm on the dataset of 20000 benign webpages for training and 13,646 instances of benign and malicious Web pages for testing. Web based classification approach was conducted which was a survey on the

features and algorithms deployed for webpage classification.

The most common types of features used are the content features (text and HTML tags on the page), and Features of Neighbors (classification based on the class label of similar webpages). After the feature construction, standard classification techniques were applied, often with focus on multi-class classification and hierarchical classification (Azeez et. al., 2013). Like Spam detection, webpage classification also benefits significantly from text classification techniques.

(Gupta & McGrath, 2008) studied phishing infrastructure and the anatomy of phishing URLs. They pointed out the importance of features such as the length of the URL, age of linked/to domains, number of links present in the e/mails and the number of dots in the phishing URLs. (Sahoo et. al., 2017) Malicious URL Detection are broadly grouped into two major categories, (i) Blacklisting or Heuristics, and (ii) Machine Learning approaches.

- Blacklisting or Heuristic Approaches: Blacklisting approaches are a common and classical technique for detecting malicious URLs, which often maintains a list of URLs that are known to be malicious. Whenever a new URL is visited, a database lookup is performed. If the URL is present in the blacklist, it is considered to be malicious and then a warning will be generated; else it is assumed to be benign.
- Machine Learning: These approaches try to analyze the information of a URL and its corresponding websites or Webpages, by extracting good feature representations of URLs, and training a prediction model on training data of both malicious and benign URLs.

2.1 Url features

Phishing URLs can be examined based on two types of features: lexical features and host-based features of the URL. The lexical features analyse the format of the URL while the host based features identify the location, owner and how malicious sites are hosted and managed (Azeez & Ademolu 2016).

2.1.1 Lexical Features

According to (Azeez & Ademolu 2016), lexical features are the textual properties of the URL. It analyses the format of the URL not the content of the page it references. These properties include the length of the entire URL, presence of IP address in URL, the number of dots in the URL, presence of phishing keywords in URL, presence of suspicious characters such as @ symbol, hexadecimal characters and use of delimiters or special binary characters like “/”, “?”, “.”, “=”, “-”, “$”, “^” either in the host name or path (Dhanalakshmi & Chellappan, 2013).

a. Length of URL: Most phishing URLs use very large domain names to lure end-users so that the URL may appear legitimate. e.g. http://www.tsv1899benningen-ringen.de/chronik/update/alert/ibclogon.php.Thus, if the length of a URL is longer than 55 characters, the URL is flagged suspicious.
b. Use of IP address in URL : Some phishing websites contain an IP address in their URL instead of the domain name in order to hide the actual domain name which is malicious. When the URL in an email has its host name as an IP address. For example, in

http://65.222.204.76/co/, we flag the URL suspicious.

c. Using the hexadecimal character codes : A malicious URL can also be represented using hexadecimal base values with a '%' symbol to hide the actual letters and numbers in the URL. Thus, a URL that has hexadecimal character codes will be flagged suspicious.
d. Use of @ symbol in URL : The '@' character is used by phishers to make host names difficult to understand. A @ symbol in a URL will enable the string to the left of the '@' symbol which is the actual legitimate URL to be discarded while the string to the right which leads to the phishing site is treated as the actual website. For example, in the URL http://www.worldbank.com@phishingsite.com, "www.worldbank.com" will be discarded

2.1.2 Host-Based Features

Host-based features describe the location of malicious sites, that is, where they are being hosted, who these sites are managed by and how they are managed. Some of these features are age of domain, page rank, number of domains (Azeez & Ademolu 2016).

a. Age of domain : The age of the domain identifies when a website is hosted such that a website that has less age or is relatively new is flagged suspicious. Many phishing sites have registered domain names that exist only for a short period of time to evade detection. They may be recently registered and some domains may not even be available at the time of checking. The WHOIS lookups on the WHOIS server is used to retrieve the domain registration date, and if the domain registration entry is not found on the WHOIS server, the URL is considered suspicious.
b. Presence of Form Tag : One of the methods phishers use to collect information from users is the use of form tag in URL. For example, <FORM action=http://www.paypalsite.com/profile.php method=post, the PayPal URL contains a form tag which has the action attribute actually sending the information to http://www.paypalsite.com/profile.php and not to http://www.paypal.com. Thus, a URL that has the form tag is flagged suspicious.
c. Number of Domains: A phishing URL may contain two or more domain names which are used to forward address from one domain to the other. For example, "http://www.google.com/url?sa=t&ct=res&cd=3&url=http%3A%2F%2Fwww.antiphishing.org%2F&ei=0qHRbWHK4z6oQLTmBM&usg=uIZX_3aJvESkMveh4uItI5DDUzM=&sig2=AVrQFpFvihFnLjpnGHVsxQ" has two domain names where "google.com" forwards the click to "antiphishing.org" domain name. The number of domain names in the URL extracted from an e-mail is counted and if more than one, we flag the URL suspicious.

3 Algorithms Considered

Four supervised machine learning classifiers (Naïve Bayes, Decision Tree, K-means and Logical Regression), were used for verification of fake URLS. They are briefly described below:

3.1. Naive-Bayes Classification Algorithm

The Bayesian Classification represents a supervised learning method as well as a

statistical method for classification. 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 (Mihaela, 2010).

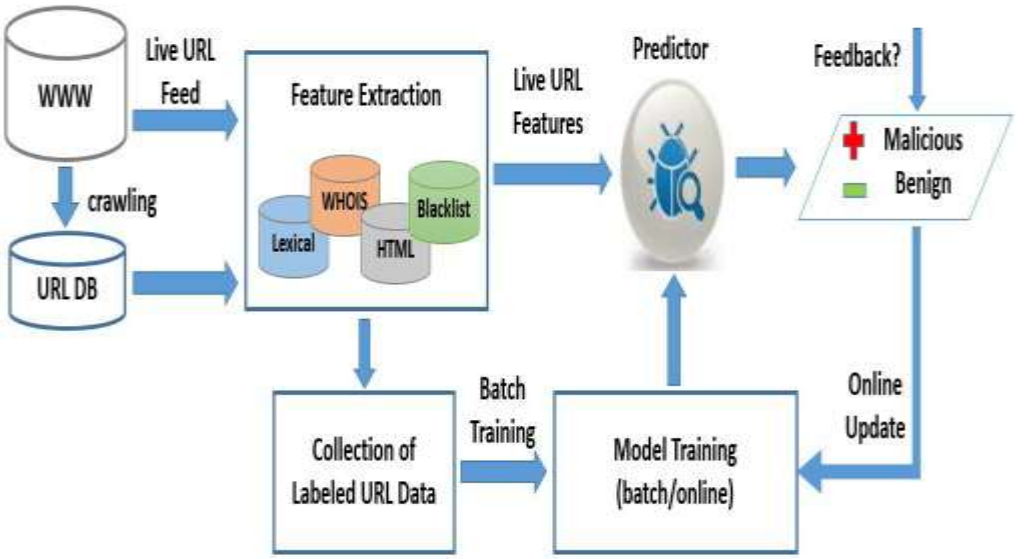

Figure 2: A framework for malicious url detection using machine learning (Sahoo et. al., 2017)

3.1.1 Uses of Naïve Bayes

1. Spam Filtering: It makes use of a naive Bayes classifier to identify spam e-mail. Bayesian spam filtering has become a popular mechanism to distinguish illegitimate spam email from legitimate email (sometimes called "ham" or "bacn").
2. Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering: Recommender Systems apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. It is proposed a unique switching hybrid recommendation approach by combining a Naïve Bayes classification approach with the collaborative filtering.
3. Naive Bayes text classification: The Bayesian classification is used as a probabilistic learning method (Naive Bayes text classification). Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents (Mihaela, 2010).

3.1.2 Naïve Bayes classifier

$p(c_j|d) =$ *Probability of class* c_j, *given that we have observed d*

$$p(c_j|d) = \frac{p(d|c_j)\,p(c_j)}{p(d)} \quad \ldots\ldots\ldots\ldots\ldots\ldots 1$$

$p(c_j|d)$ = Probability of instance d being in class c_j,

$p(d|c_j) =$ Probability of generating instance d given class c_j,

$p(c_j)$ = Probability of occurrences of class c_j,

$p(d)$ = Probaility of instance d occurring

Bayes classification for more features

To simplify the task naïve Bayesian classifiers assumes attributes have independent distribution and there by estimate

$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * \ldots * p(d_n|c_j)$ ……..2

$p(d|c_j)$ = Probability of class c_j generating instance d

$p(d_1|c_j)$ = The probability of class c_j generating the observed value for feature 1,

$p(d_2|c_j)$ = The probability of class c_j generating the observed value for feature 2,

$p(d_3|c_j)$ = The probability of class c_j generating the observed value for feature 3

3.2 Decision Tree Model Algorithm

The core algorithm for building decision trees called ID3 by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. ID3 uses Entropy and Information Gain to construct a decision tree.

Entropy is a measure of uncertainty associated with a random variable

For a discrete random variable Y taking m distinct values $\{y_1, \ldots, y_m\}$

$H(Y) = -\sum_{i=1}^{m} p_i \log(p_i)$, *where* $p_i = P(Y = y_i)$ .3

Conditional Entropy

$H(Y|X) = \sum_x p(x)\, H(Y|X = x)$……....4

Select the attribute with the highest information gain

Let pi be the probability that an arbitrary tuple in D belongs to class Ci, estimated by |Ci, D|/|D|

Expected information (entropy) needed to classify a tuple in D:

$info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$ …….…5

Information needed (after using A to split D into v partitions) to classify D:

$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times info(D_j)$…….6

Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

3.3 K-Means

This is the most commonly used algorithm for an iterative refinement technique. Due to its ubiquity, it is often called the k-means algorithm; it is also referred to as Lloyd's algorithm, particularly in the computer science community. Lloyd's algorithm is based on the simple observation that the optimal placement of a center is at the centroid of the associated cluster (Faber, 1994). The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments (the k-means++ algorithm addresses this problem by seeking to choose better starting clusters).

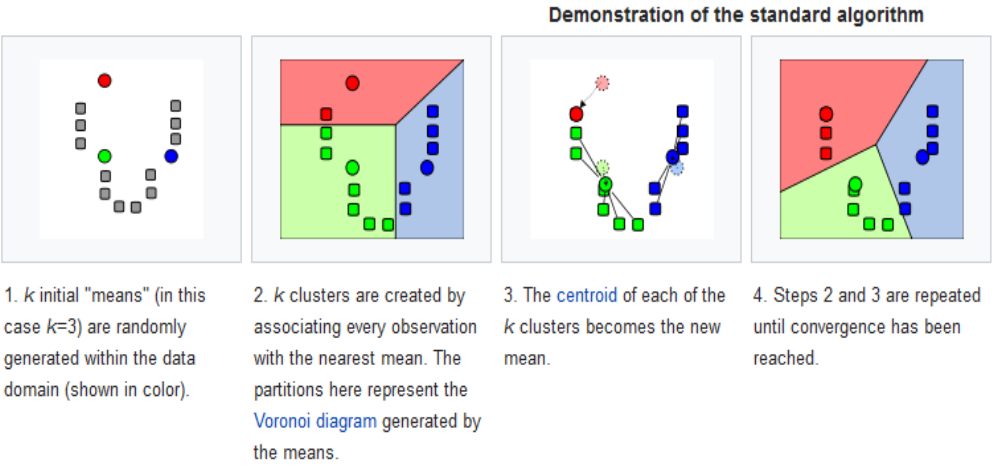

Figure 3: Demonstration of the Standard Algorithm (Guido, 2014).

3.3.1 K-means Algorithm

Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets (k ≤ n) S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS):

$$J = \sum_{j=1}^{K} \sum_{n \epsilon s_j} |x_n - \mu_j|^2 \quad \ldots\ldots\ldots\ldots\ldots 7$$

Where $|x_n - \mu_j|^2$ is a chosen distance measure between a data point and the cluster centre is an indicator of the distances of the n data points from their respective cluster centers.

Steps In k means algorithm

3.3.1.1 Assignment step: Assign each observation to the cluster with the closest mean

$$S_i^{(t)} = \{x_j : ||x_j - m_i^{(t)}|| \leq ||x_j - m_{i*}^{(t)}|| \; for\, all\, i^* = 1, \ldots\ldots, k\}$$

3.3.1.2 Update step: Calculate the new means to be the centroid of the observations in the cluster.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{xj \in S_i^{(t)}} X_j \quad \ldots\ldots\ldots\ldots\ldots 8$$

Complexity of k means algorithm is given by: Complexity is O (n * K * I * d) n = number of points, K = number of clusters, I = number of iterations, d = number of attributes.

3.4 Logistic Regression

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. The main advantage is to avoid confounding effects by analyzing the association of all variables together (Sperandei, 2014).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest.

Logistic regression models the probability of an event occurring depending on the values of the independent variables which can be categorical or numerical.

$P = {}^{outcome\ of\ interest}/_{all\ possible\ outcome}$

Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. If the probability of an event occurring is p, the probability of the event not occurring is (1-p).

$$odds = \frac{p(occuring)}{P(not\ occuring)} = \frac{p}{1-p} \ldots\ldots 9$$

3.4.1 Odd ratio in logistic regression

Odd ratio is the ratio of two odd, the odds ratio (OR) is a comparative measure of two odds relative to different events

$$odds\ ratio = \frac{odds_1}{odds_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_2}}$$

The dependent variable of logistic regression follows the Bernoulli distribution having an unknown probability. Bernoulli distribution is a special case of the binomial distribution where n =1 legitimate is "1" Malicious is "0".

$P(legitimate) = p$ and $P(Malicious) = 1-p$

In logistic regression we are estimating an unknown p for a given linear combination of the independent variable. We link together our independent variables to the Bernoulli distribution, the link is called Logit. The goal of logistic regression is to estimate p for a linear combination of the independent variables and, estimate of p is to tie together our linear combination of variables that could result unto the Bernoulli probability distribution with a domain from 0 to 1.

$$\text{logit}(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = a + \beta x \ldots\ldots 10$$

Where p is the probability of interested outcome and x is the explanatory variable. The parameters of the logistic regression are α and β. This is the simple logistic model. Taking the antilog of equation (1) on both sides, one can derive an equation for the prediction of the probability of the occurrence of interested outcome as

$$p = P(Y = interested\ outcome\ X = x, a\ specific\ value) = \frac{e^{a+\beta x}}{1+e^{a+\beta x}} = \frac{1}{1+e^{a+\beta x}} \ldots\ldots\ldots 11$$

Extending the logic of the simple logistic regression to multiple predictors,
one may construct a complex logistic regression as

$$logit(y) = \ln\left(\frac{p}{1-p}\right) = a + \beta_1 X_1 + \cdots + \beta_k X_k \ldots 12$$

Therefore

$$p = P(Y = interested\ outcome\ X = x_1 \ldots\ldots\ldots X = x_k) = \frac{e^{a+\beta_1 X_1 + \cdots + \beta_k X_k}}{1+e^{a+\beta_1 X_1 + \cdots + \beta_k X_k}} = \frac{1}{1+e^{a+\beta_1 X_1 + \cdots + \beta_k X_k}} \ldots..13$$

A simple logistic function is defined by the formula

$$y = \frac{e^x}{1+e^x} = \frac{1}{1-e^{-x}} \ldots\ldots\ldots\ldots\ldots 14$$

To provide flexibility, the logistic function can be extended to the form

$$y = \frac{e^{a+\beta x}}{1+e^{a+\beta x}} = \frac{1}{1+e^{a+\beta x}} \ldots\ldots\ldots..15$$

Where α and β determine the logistic intercept and slope. Logistic regression fits α and β, the regression coefficients.. The logistic or logit function is used to transform an 'S'-shaped curve into an approximately straight line and to change the range of the proportion from
0 – 1 to -∞ - +∞ as

$$\text{logit}(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = a + \beta x \ldots 16$$

Where p is the probability of interested outcome, α is the intercept parameter,β is a regression coefficient, and χ is a predictor.

4 Implementation, Findings and Results

The system is a web based application, it classifies a URL as malicious or legitimate based on lexical features and host based features. Four machine learning algorithms which are all supervised learning algorithms (Naïve Bayes algorithm, decision tree algorithm, k means algorithm and logistic regression algorithm) were used to classify the URL. Based on the trained features, the system classifies the URL as malicious else it is classified as legitimate. The collected features include both URL-based features and host-based features. The verification of fake urls using supervised learning algorithm based on repetitive and redundancy values have been implemented with java programming language in the Netbeans integrated Development Environment (IDE) and are tested against 200 URLs. This has been done to determine the algorithm that has the highest maximal level of effectiveness, accuracy and efficiency. Some of the collected features hold categorical values termed as ‘‘Legitimate ’and Malicious’’, these values have been replaced with numerical values 1, 0 and -1 instead of ‘‘Legitimate’’, ‘‘Malicious’’ and ‘‘Suspicious’’ respectively.

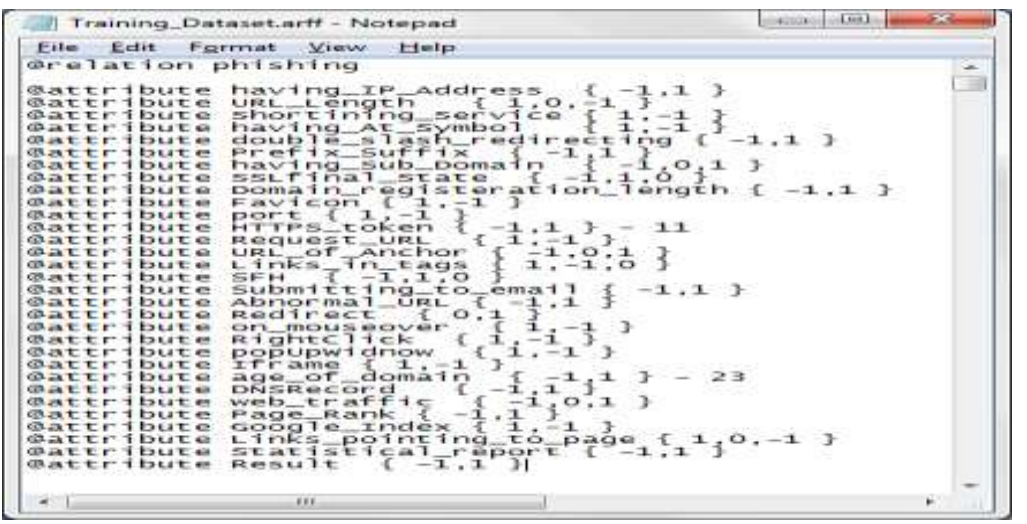

Figure 4: Lexical and Host-Based Features for url Classification

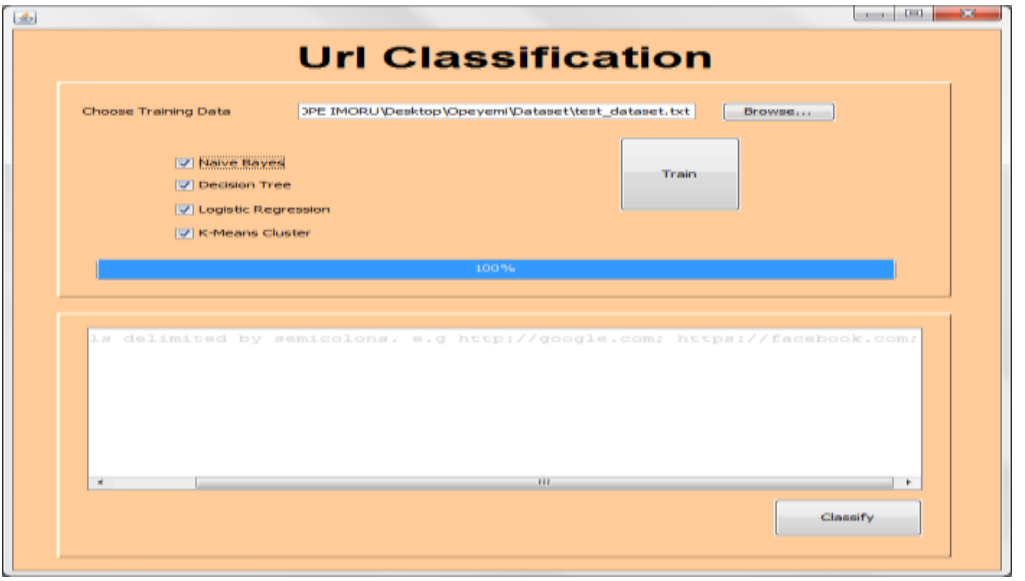

Figure 5: Initialization of the Program/ Training Dataset File

Sample of the url classification (2 url examples) for all features is shown in diagrams below:
http://www.Unilag.edu.ng;
http://www.unitedhealthgroup.com
http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/;
http://www.sunsofttec.com/eeeettt/2185266aadae98f002016e352372bba8/;
http://www.lagranherramienta.com/easy/ayo/ayo1/;
http://www.kabradrugsltd.com/css/nt/df6f3f034aba794e31abbdd8a0564007/;
http://ec2-54-200-151-255.us-west-2.compute.amazonaws.com/-/accord2/;
https://www.google.com/;
http://www.msn.com/en-us?cobrand=hp-notebook.msn.com&OCID=HPDHP&pc=HPNTDF;
http://www.folder365.world/yawa/aptgd/;
http://rooferexpert.com/css/8933617-dosar-nr-1817842015/394c-4735-82399c8f64a5248/botosani_firme/ec77154aef4d9311a65613d9a59cf370/;

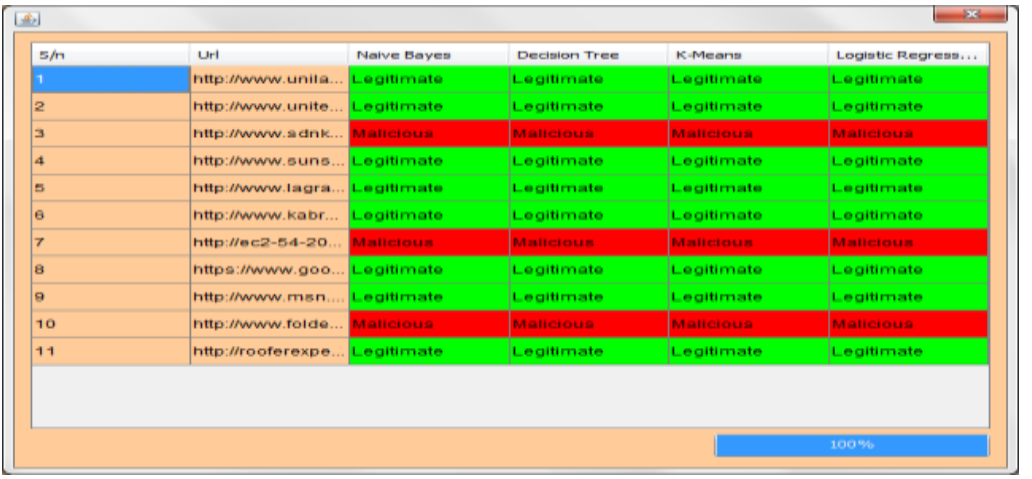

Figure 6: Classification of 10 Samples url

Table 1: Breakdown of Naïve Bayes Classifier for www.Unilag.Edu.Ng

URL FEATURES	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.855932	0.872549
LEGITIMATEURL	0.235294	0.145631
NORMALURL	0.79661	0.823529
NOATSYMBOL	0.974576	0.941176
NODOUBLESPLASH	0.813559	0.77451
NOPREFIXSUFIX	0.288136	0.009804
LEGITIMATEDOMAIN	0.352941	0.495146
MALICIOUSSSL	0.016807	0.242718
MALICIOUSREGISTRATIONLENGTH	0.220339	0.421569
NOHTTPSTOKENDMAIN	0.635593	0.627451
DOMAINAGEOLDERTHAN6MONTHS	0.677966	0.411765
HASDNSRECORD	0.542373	0.264706

Table1 shows the Naïve Bayes mathematical breakdown of the url features for www.Unilag.edu.ng. From Table 1, it was deduced that the Unilag url is a legitimate url based on the features.

4.1 Breakdown and Graphical Classification of Legitimate url

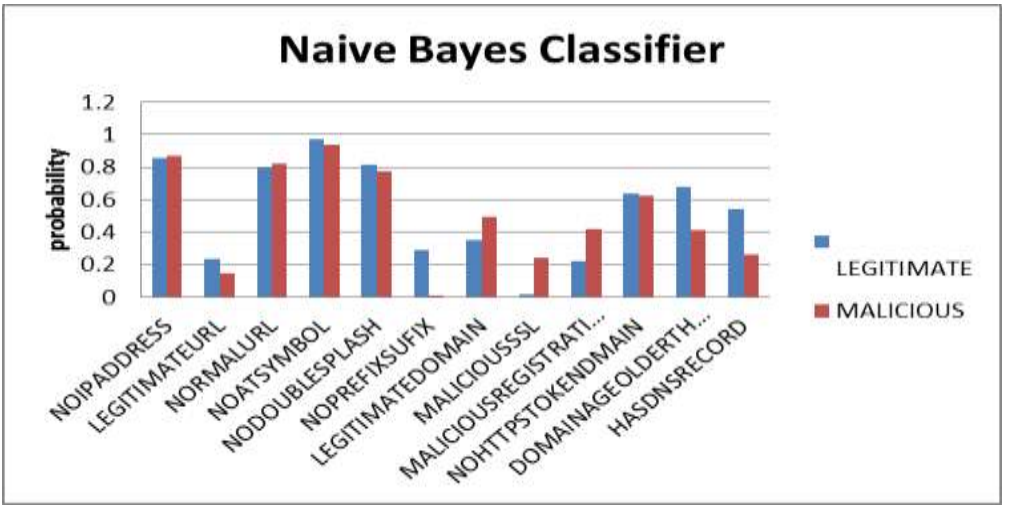

Figure 7: Graphical Figure of Naïve Bayes Classifier for www.unilag.edu.ng

Figure 7 is a graphical representation that shows the url features of legitimate and malicious breakdown as depicted in Table 1.

Table 2: Breakdown of Decision Tree www.unilag.edu.ng

URL FEATURES	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.14	0.13
LEGITIMATEURL	0.24	0.15
NORMALURL	0.8	0.82
NOATSYMBOL	0.97	0.94
NODOUBLESPLASH	0.81	0.77
NOPREFIXSUFIX	0.29	0.01
LEGITIMATEDOMAIN	0.35	0.5
MALICIOUSSSL	0.02	0.24
MALICIOUSREGISTRATIONLENGTH	0.22	0.42
NOHTTPSTOKENDMAIN	0.64	0.63
DOMAINAGEOLDERTHAN6MONTHS	0.32	0.59
HASDNSRECORD	0.54	0.26

Table 2 shows the mathematical breakdown of Decision Tree showing url features of Unilag website. From the table it was deduced that the Unilag url is a legitimate url based on the features as shown in Table 2.

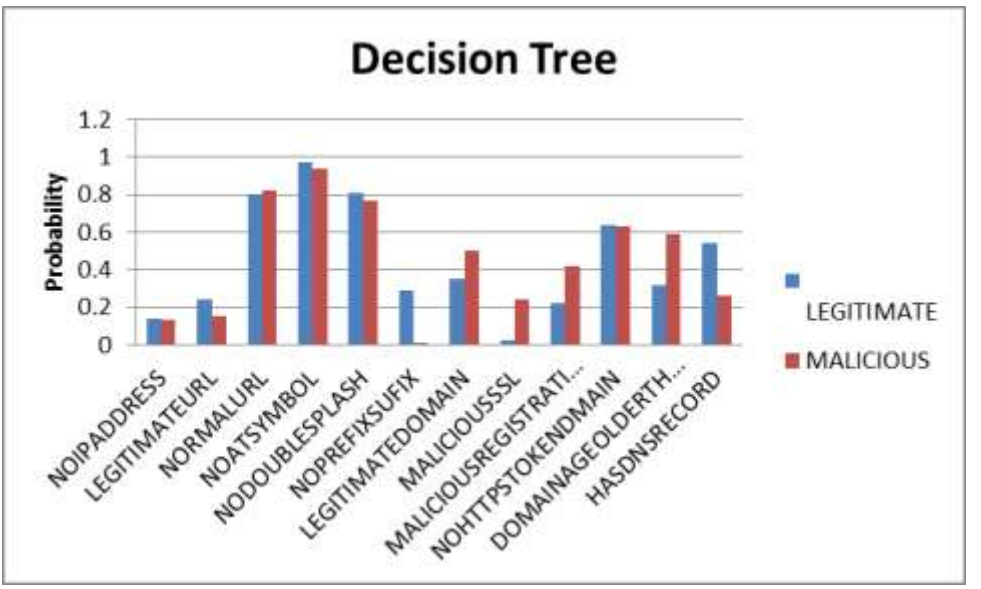

Figure 8: Graphical Figure of Decision Tree Classifier for www.unilag.edu.ng

Figure 8 is a graphical representation that shows the url feature of legitimate and malicious breakdown from Table 2.

Table 3: Breakdown of K-Means www.unilag.edu.ng

URL FEATURES	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.862069	0.88
LEGITIMATEURL	0.232759	0.14
NORMALURL	0.801724	0.83
NOATSYMBOL	0.982759	0.95
NODOUBLESPLASH	0.181034	0.22
NOPREFIXSUFIX	0.284483	0
LEGITIMATEDOMAIN	0.353448	0.5
MALICIOUSSSL	0.008621	0.24
MALICIOUS REGISTRATION LENGTH	0.215517	0.42
NOHTTPSTOKENDMAIN	0.637931	0.63
DOMAINAGEOLDERTHAN6MONTHS	0.681034	0.41
HASDNSRECORD	0.543103	0.26

Table 3 shows the numerical values of k-means url features of Unilag website. From the table, it was deduced that the Unilag url is a legitimate url based on the features depicted in Table 3.

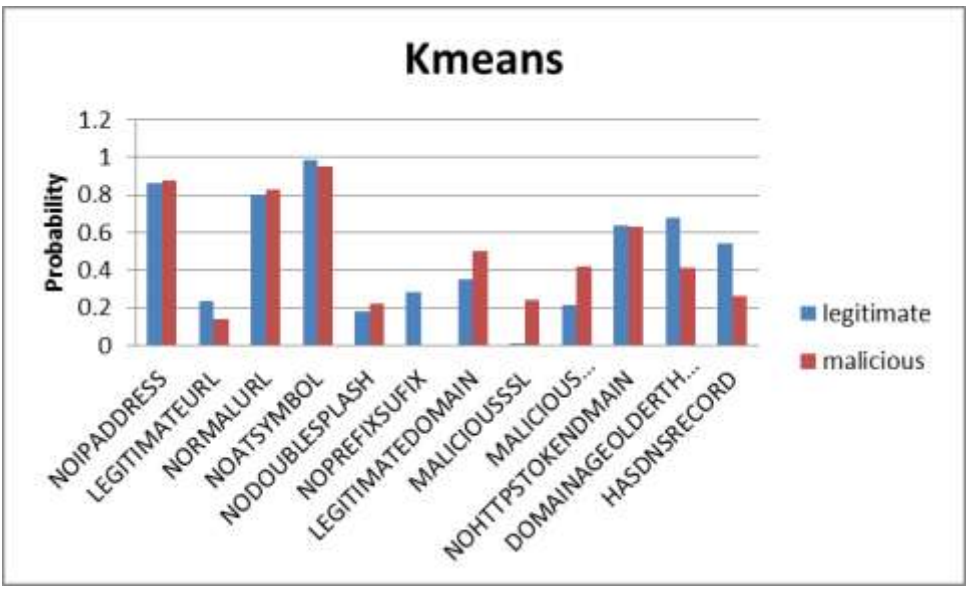

Figure 9: Graphical Figure of K Means Classifier for www.unilag.edu.ng

Figure 9 is a graphical representation that shows the k-means interpretation of url feature of legitimate and malicious breakdown as depicted in Table 3.

Table 4: Breakdown of Logistic Regression www.unilag.edu.ng

URL FEATURES	WEIGHT
NO IPADDRESS	-0.246116026
LEGITIMATE URL LENGTH	0.406965719
NORMAL URL	-0.023747636
NO ATSYMBOL	0.222137496
NO DOUBLESPLASH	0.159693131
NO PREFIXSUFIX	1.045097936
LEGITIMATE DOMAIN	-0.391132699
MALICIOUS SSL	-1.463416988
MALICIOUS REGISTRATION LENGTH	-0.084222514
NOHTTPSTOKENDMAIN	0.031575833
DOMAIN AGE OLDER THAN6MONTHS	0.526511068
HASDNSRECORD	0.580109969

Table 4 shows the numerical values obtained for Logistic Regression of url features of Unilag website. From the table it was deduced that the Unilag url is a legitimate url based on the features used.

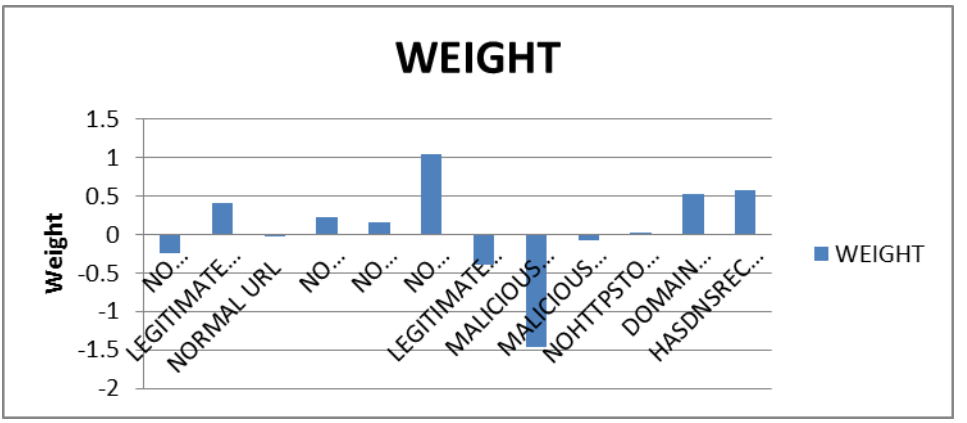

Figure 10: Graphical Figure of Logistic Regression Classifier for www.unilag.edu.ng

The above graphical representation shows the Logistic Regression interpretation of url feature of legitimate and malicious breakdown as depicted in Table 4.

4.2 Breakdown and Graphical Classification Of Malicious url

Table 5; Naïve Bayes Breakdown Details of http://www.sdnkasepuhan02btg.sch.id/cana/A2a7938099b2075bd8b9b69804524753/

URL FEATURES	Legitimate	Malicious
NOIPADDRESS	0.855932	0.872549
LEGITIMATEURL	0.058824	0.087379
NORMALURL	0.79661	0.823529
NOATSYMBOL	0.974576	0.941176
NODOUBLESPLASH	0.813559	0.77451
HASPREFIXSUFIX	0.711864	0.990196
MALICIOUSDOMAIN	0.369748	0.174757
MALICIOUSSSL	0.016807	0.242718
MALICIOUSREGISTRATIONLENGTH	0.220339	0.421569
NOHTTPSTOKENDMAIN	0.635593	0.627451
DOMAINAGEOLDERTHAN6MONTHS	0.322034	0.588235
HASDNSRECORD	0.457627	0.735294
	6.233513	7.279363

Table 5 shows the values obtained for Naïve Bayes of url features of Unilag website. From the table it was deduced that the url is malicious based on the url features in the table.

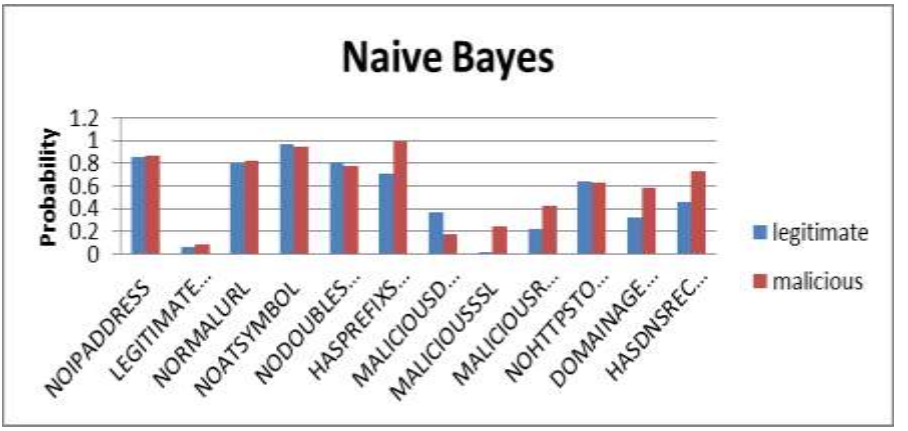

Figure 11: Graphical Figure of Naïve Bayes Classifier for http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

Figure 11 shows a graphical representation of the values obtained in Table 5 for the Naïve Bayes.

Table 6: Decision Tree Breakdown Details of http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

URL CLASSIFICATION	LEGITIMATE	MALICIOUS
NOIPADDRESS	**0.86**	**0.87**
LEGITIMATEURL	**0.24**	**0.15**
NORMALURL	**0.8**	**0.82**
NOATSYMBOL	**0.97**	**0.94**
NODOUBLESPLASH	**0.81**	**0.77**
HASPREFIXSUFIX	**0.71**	**0.99**
MALICIOUSDOMAIN	**0.37**	**0.17**
MALICIOUSSSL	**0.02**	**0.24**
MALICIOUSREGISTRATIONLENGTH	**0.22**	**0.42**
NOHTTPSTOKENDMAIN	**0.64**	**0.63**
DOMAINAGEOLDERTHAN6MONTHS	**0.68**	**0.41**
N0DNSRECORD	**0.46**	**0.76**

The features depicted in Table 6 shows the url features used for Decision Tree classification. It also shows the values obtained for the malicious url.

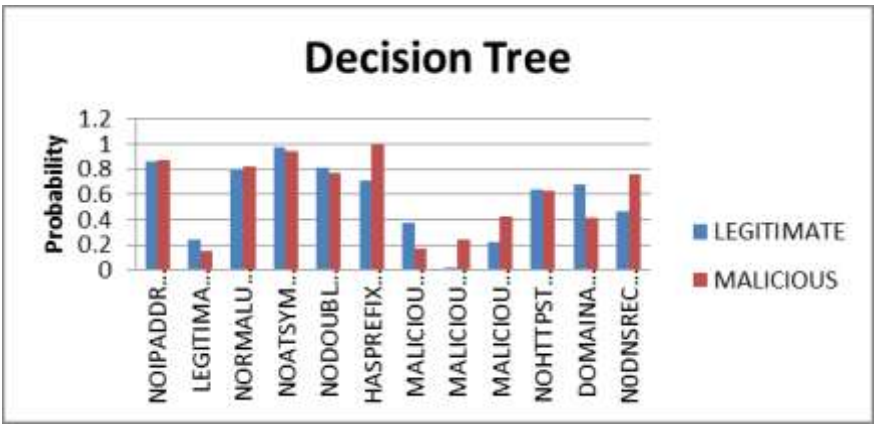

Figure 12: Graphical Figure of Decision Tree Classifier for http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

Figure 12 is a graphical representation that shows the interpretation of Decision Tree for url features of both legitimate and malicious as depicted in Table 6 above.

Table 7: K-Means Showing the Breakdown of http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

URL FEATURES	LEGITIMATE	MALICIOUS
NOIPADDRESS	0.862069	0.88
SUSPICIOUS URL LENGH	0.051724	0.08
NORMALURL	0.801724	0.83
NOATSYMBOL	0.982759	0.95
NODOUBLESPLASH REDIRECTING	0.181034	0.22
HASPREFIXSUFIX	0.715517	1
MALICIOUSDOMAIN	0.37069	0.17
MALICIOUSSSL	0.008621	0.24
MALICIOUSREGISTRATIONLENGTH	0.215517	0.42
NOHTTPSTOKENDMAIN	0.637931	0.63
DOMAINAGEOLDERTHAN6MONTHS	0.681034	0.41
N0DNSRECORD	0.456897	0.74

Table 7 shows the url features used for k-means classification. It shows the values obtained for both malicious and legitimate urls.

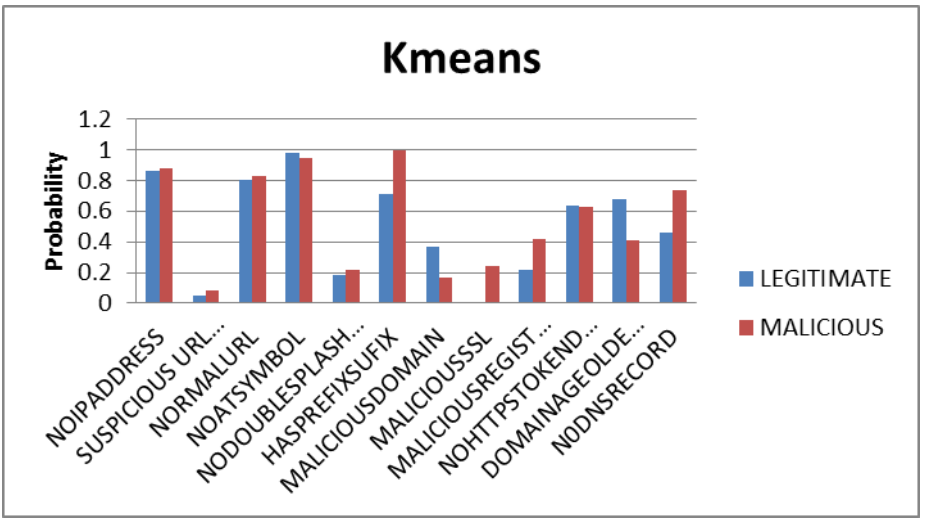

Figure 13: Graphical Figure of K-Means Classifier for
http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

Figure 13 is a graphical representation that shows the k-means interpretation of url features for legitimate and malicious as depicted in Table 7.

Table 8: Logical Regression Breakdown Details of
HTTP://WWW.SDNKASEPUHAN02BTG.SCH.ID/CANA/A2A7938099B2075BD8B9B69804524753/

FEATURES	WEIGHT
NOIPADDRESS	-0.246116027
SUSPICIOUS URL LENGH	-0.063673963
NORMALURL	-0.023749636
NOATSYMBOL	0.222137497
NODOUBLESPLASH REDIRECTING	0.159693132
HASPREFIXSUFIX	-1.018083616
MALICIOUSDOMAIN	0.006899523
MALICIOUSSSL	-1.463416988
MALICIOUSREGISTRATIONLENGTH	-0.084222514
NOHTTPSTOKENDMAIN	-0.004561513
DOMAINAGEOLDERTHAN6MONTHS	0.526510685
N0DNSRECORD	-0.553095649

Table 8 shows the url features used for Logistic Regression classification. It shows the corresponding values obtained.

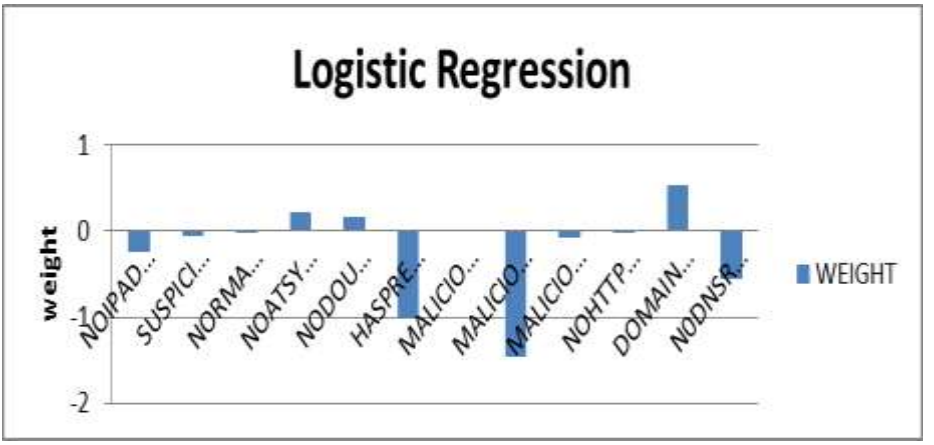

Figure 14: Graphical Figure of Logistic Regression Classifier for http://www.sdnkasepuhan02btg.sch.id/cana/a2a7938099b2075bd8b9b69804524753/

Figure 14 is a graphical representation that shows the logistic regression interpretation of url features for legitimate and malicious as contained in Table 8.

5. Conclusion

A study and evaluation of four Machine Learning Algorithms for evaluating legitimacy of urls has been successfully carried out. The algorithms were implemented and tested with different dataset. A comparison of all four algorithms was done to know their level of efficiency and effectiveness in detecting and evaluating both legitimate and malicious urls. It is of note that twelve different url features were considered and evaluated for each of the algorithms. With the available results, as observed in the numerical values and graphical representations for the experimentation, the Naïve Bayes Algorithm is considered to be the most effective and efficient of all the four machine learning algorithms evaluated. Naïve Bayes Algorithm yielded good results for detecting legitimate and malicious values when tested with the same url under the same features. Some future works, therefore for this admirable research work include the development of a new algorithm that can be more accurate than Naïve Bayes algorithm. This can be achieved by hybridizing two or more supervised learning algorithms in order to have a more accurate, efficient and reliable url legitimate evaluation.

References

Ayofe, A.N, Adebayo, S.B, Ajetola, A.R, Abdulwahab, A.F (2010) "A framework for computer aided investigation of ATM fraud in Nigeria" International Journal of Soft Computing, Vol. 5, Issue 3 pp. 78-82

Azeez, N.A., and Lasisi, A. A. (2016). Empirical and Statistical Evaluation of the Effectiveness of Four Lossless Data Compression Algorithms. Nigerian Journal of Technological Development, Vol. 13, NO. 2, December 2016, 64-73.

Azeez, N.A and Otudor, A.E. (2016) "Modelling and Simulating Access Control in Wireless Ad-Hoc Networks" Fountain Journal of Natural and Applied Sciences. Vol 5(2), pp 18-30

Azeez, N. A., & Ademolu, O. (2016). CyberProtector: Identifying Compromised URLs in Electronic Mails with Bayesian Classification. 2016 International Conference Computational Science and Computational Intelligence (CSCI) (pp. 959-965). Las Vegas, NV, USA: IEEE.

Azeez, N. A., & Iliyas, H. D. (2016). Implementation of a 4-tier cloud-based architecture for collaborative health care delivery. Nigerian Journal of Technological Development, 13(1), 17-25.

Azeez, N. A., & Venter, I. M. (2013). Towards ensuring scalability, interoperability and efficient access control in a multi-domain grid-based environment. SAIEE Africa Research Journal, 104(2), 54-68.

Azeez, N.A Abidoye, A.P Adesina, A.O Agbele, K.K Venter, I.M Oyewole, A.S (2013) "Statistical Interpretations of the Turnaround Time Values for a scalable 3-tier grid-based Computing architecture" Computer Science & Telecommunications, Vol 39 (3), pp 67-75.

Azeez, N. A., Iyamu, T., and Venter, I. M. (2011). Grid security loopholes with proposed countermeasures. In E. Gelenbe, R. Lent, and G. Sakellari (Ed.), 26th International Symposium on Computer and Information Sciences (pp. 411-418). London: Springer.

Azeez, N.A and Venter, I.M (2012). Towards achieving scalability and interoperability in a triple-domain grid-based environment (3DGBE)- Information Security for South Africa (ISSA), 2012, pp 1-10.

Azeez, N. A., and Babatope, A. B. (2016). AANtID: an alternative approach to network intrusion detection. The Journal of Computer Science and its Applications. An International Journal of the Nigeria Computer Society, 129-143.

Azeez, N. A. (2012). Towards Ensuring Scalability, Interoperability and Efficient Access Control In a Triple-Domain Grid-Based Environment. Cape Town: University of the Western Cape.

Cao, C. & Caverlee, J. 2015 Detecting Spam URLs in Social Media via Behavioral Analysis, Department of Computer Science and Engineering, Texas A&M University College Station, Texas, USA.

Dhanalakshmi, R., & Chellappan, C. (2013). Detecting Malicious URLs in E-mail - An Implementation. AASRI Procedia , 125-131.

Faber, V., 1994 Clustering and the Continuous k-means Algorithm, Los Alamos Science, vol. 22, pp. 138-144, 1994.

Gupta, M. & McGrath D. 2008, Behind phishing: an examination of phisher modi operandi, in proceedings of the USENIX Workshop on Large/scale Exploits and Emergent Threats (LEET), San Fransicso, CA, Apr 2008.

Guido, S., 2014 Kmeans Clustering With Scikit- Learn

Https://Www.Slideshare.Net/Sarahguido/Kmeans-Clustering-With-Scikitlearn.

Guille, A., Hacid,H., Favre, C., Zighed, D.A. 2013: Information Diffusion in online Social Networks: A Survey. ERIC Lab, Lyon 2 University, France, Bell Labs France, Alcatel-Lucent, France Institute of Human Science, Lyon 2 University, France. ACM. 2013 published in SIGMOD Record, Vol 42 ISS2, June 2013.

Ma, J., Saul, L., Savage, S. & Voelker,G. 2011. "Learning to Detect Malicious URLs",ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, no. 30, (2011),pp. 30:1-30:24.

Mihaela M. 2010 : Naïve-Bayes Classification Algorithm. 7 May 2010 http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf . .

Nureni, A. A., and Irwin, B. (2010). Cyber security: Challenges and the way forward. Computer Science & Telecommunications, 29, 56-69.

Qi, X. & Davison, B. 2009 Web Page Classification: Features and Algorithms ACM Comput. Surv. 41, 2, Article 12 (February 2009), 31 pages DOI = 10.1145/1459352.1459357 http://doi.acm.org/10.1145/1459352.1459357.

Rao, V. & Saleem, P.A., 2015 Twitter Adoption And Analysis Online Social Networks International Journal Of Global Innovations - Vol.2, Issue .I Paper Id: Sp-V2-I1-257 Issn, Dept Of Cse, Kkr And Ksr Institute Of Technology And Sciences (Kits) Guntur, A.P., India. 2015.

Sahoo, D. Liu,C & Steven C.H. Hoi 2017. Malicious URL Detection using Machine Learning: A Survey " arXiv:1701.07179v2 [cs.LG] 16 Mar 2017. https://arxiv.org/pdf/1701.07179.pdf

Sperandei, S. 2014 Understanding logistic regression analysis. Biochem Med (Zagreb). 2014 Feb; 24(1): 12–18.Published online 2014 Fe doi: 10.11613/BM.2014.003PMCID: PMC3936971

Srivastava, V.T. 2007 : Phishing and Pharming- The Deadly Duo, SANS Institute 2007 Accepted January 29, 2007. https://www.sans.org/reading-room/whitepapers/privacy/phishing-pharming-evil-twins-1731

Covenant Journal of Informatics & Communication Technology. Vol. 5 No. 2, Dec, 2017

An Open Access Journal Available Online

Hyper-Erlang Battery-Life Energy Scheme in IEEE 802.16e Networks

Ibrahim Saidu, Hamisu Musa, Muhammad Aminu Lawal & Ibrahim Lawal Kane

Department of Mathematics and Computer Science
Umaru Musa Yar'adua University (UMYU)
Katsina, Nigeria.
Saidu.ibrahim@umyu.edu.ng ,
hamisu.musa@umyu.edu.ng ,
muhammad.aminu@umyu.edu.ng
Ibrahim.lawal@umyu.edu.ng

Abstract: IEEE 802.16e networks is one of the broadband wireless technologies that support multimedia services while users are in mobility. Although these users use devices that have limited battery capacity, several energy schemes were proposed to improve the battery-life. However, these schemes inappropriately capture the traffic characteristics, which lead to waste of energy and high response delay. In this paper, a Hyper-Erlang Battery-Life Energy Scheme (HBLES) is proposed to enhance energy efficiency and reduce the delay. The scheme analytically modifies idle threshold, initial sleep window and final sleep window based on the remaining battery power and the traffic pattern. It also employs a Hyper-Erlang distribution to determine the real traffic characteristics. Several simulations are carried out to evaluate the performance of the HBLES scheme and the compared scheme. The results show that the HBLES scheme out performs the existing scheme in terms of energy consumption and response delay.

Keywords: Energy Saving Scheme; mobile WiMAX; Battery Lifetime.

1. Introduction

The IEEE 802.16e standard (IEEE, 2005) popularly known as mobile WiMAX is one of the broadband access technologies that provides mobility support, ubiquitous access and support to multimedia applications to a mobile subscriber station (MSS) (Saidu et al., 2015) . In order to support these applications while the MSS moves at vehicular speed, the MSS undergo frequent battery drain due to excessive power consumption than in traditional voice-centric technology. Thus, efficient energy schemes are highly needed.

The standard uses sleep mode operations to improve energy efficiency. The sleep modes are classified into three power saving classes: Type A, Type B, and Type C. The Type A employs for the delay insensitive traffics with an exponential increase sleep period. The Type B uses for the delay sensitive traffics with a constant sleep period. While the Type C designs for multicast and management operation with the sleep duration regulated by the base station (BS). These PSCs use idle threshold (Tit), initial sleep window (Tmin) and final sleep window (Tmax) parameters to reduce energy dissipation of the MSS. The Tit refers to a waiting time of the MSS before going to sleep. The Tmin represents a shortest duration of sleep interval. While Tmax is the longest period of sleep interval. The parameters are allocated to the MSS by the BS when it requests to sleep.

Several energy schemes were proposed based on these PSCs to extend the battery- life of the MSS (Xiao, 2005), (Xiao, 2006). Some schemes consider effects of sleep parameters on traffic arrival rate (Xiao, 2006), (Zhang & Fujise, 2006), (Xiao et al., 2006) and others, adjust the parameters with consideration to traffic load and delay requirements (Xiao et al., 2006), (Zhu & Wang, 2007), (Zhu et al., 2007), (Xue et al., 2008), (Sanghvi et al., 2008). While others (Kim et al., 2008), (Lin & Wang, 2013), (Ferng & Li, (2013), (Chou et al., 2013) dynamically adjust the parameters based on the load and the remaining energy. Among these schemes, only (Chou et al., 2013) employs all the parameters with significant energy efficiency but it wastes energy and increases response delay due to inappropriate capture of the traffic pattern.

This paper presents a HBLES to enhance energy efficiency of the MSS while reducing the response delay. The scheme modifies the sleep parameters based on a residual energy and traffic arrival pattern. It also uses a Hyper-Erlang distribution to capture the real traffic characteristics. Extensive simulations were used to evaluate the performance of the HBLES scheme and the compared scheme (Zhang, 2007). . The results show that the HBLES out performs the BLAPS scheme.

The rest of this paper is organized as follows. Section II presents related work. In Section III, the proposed HBLAPS is described. Section IV, presents performance evaluation. Finally, Section V presents conclusion of the paper.

2. Related Work

This section presents a review on some of the existing energy schemes in mobile WiMAX, which are as follows:

In (Xiao, 2005), an energy scheme is introduced to analytically model the sleep mode operation. The scheme examines the effects of Tmin and Tmax on arrival rate. It improves energy consumption with a smaller Tmin but increases the delay. The scheme also reduces delay and energy consumption with a smaller Tmax. However, it

considers only the downlink (DL) traffic.

In (Xiao, 2006), an enhanced model in (Xiao, 2005) is proposed which considers both uplink (UP) and DL traffics. Although the scheme reduces energy consumption in a shorter time, it consumes high energy in a longer duration due to incessant sleep-wake mode. While the scheme in (Zhang & Fujise, 2006) differentiates between the incoming and outgoing traffics. It reduces the energy consumption but waste energy due to excessive listening operations under low traffic arrival rate.

In (Xiao et al., 2006), an enhanced energy saving scheme is proposed to reduce the listening operations. The scheme considers the initial sleep interval in the next sleep operation as half of the previous sleep interval. The scheme also uses embedded Markov chain model for analysis and a closed-form expression to enhance energy. The scheme extends battery life but it has a higher response delay due larger sleep interval.

Heuristic Scheme in (Zhu & Wang, 2007) is proposed to enhance the battery-life. The scheme utilizes a heuristics algorithm to dynamically adjust Tmin based on the traffic load. The scheme extends the battery-life of MSS, however with small increase in the delay. It is enhanced in (Zhu et al., 2007), where the mechanism adjusts the sleep parameters based on the delay and traffic requirements. The scheme bounds the delay in a certain range.

Traffic Load Aware Scheme is proposed in (Xue et al., 2008) to improve energy consumption. The scheme analyses the sleep mode operation to determine relationships among traffic load, idle check time and power consumption. It employs dynamic technique to adjust the idle check time based on the traffic load measurement. The scheme enhances the battery life and reduces the mean delay but fails the standard conformity.

Adaptive Mechanism is proposed in (Sanghvi et al., 2008) to enhance battery performance. The mechanism dynamically regulates the sleep parameters based on the request period of each initiation of awakening (Tin). The scheme improves energy but it increases response delay.

Remaining Energy Aware Power Management mechanism (REAPM) (Kim et al., 2008) is proposed to prolong the battery. The REAPM employs smoothing technique with current inter arrival time to adjust Tmax. The mechanism also employs the remaining energy and the Tmax to adjust the Tmin. It reduces the response delay under sufficient energy but increases delay under insufficient energy.

In (Lin & Wang, 2013), an adaptive waiting time threshold estimation scheme is proposed to minimize energy consumption. The scheme predicts the threshold by dynamically adjusting the idle threshold based on the DL and UP traffics. It considers the time to be small under low-traffic arrival but large under heavy traffic arrival. The scheme minimizes the energy consumption but it leads to a small increase in the delay.

Predictive and Dynamic Energy-Efficient Mechanism Scheme (Ferng & Li, 2013) is proposed to improve energy efficiency and the delay. The scheme uses a prediction mechanism to determine when a MSS should wake up. It sets the maximum sleep interval when the probability of traffic arrival is slim in order to minimize energy wastage. In addition, it also sets the smaller sleep

interval when the traffic arrival is high to minimise the delay. It improves energy efficiency and minimizes delay but prediction error is not ignored.

In (Chou et al., 2013) , a battery aware scheme is proposes to prolong the battery life. The BLAPS dynamically adjusts the sleep parameters based on the residual energy and the traffic loads. The scheme enhances the battery lifetime, but it increases energy consumption and response delay due to failure to consider appropriate traffic pattern and distribution.

3. Proposed Hbles.

This section presents the proposed HBLES, which analytically modifies the sleep parameters as follows:

3.1 Idle Threshold

The T_{it} is adjusted based on residual battery lifetimes and the traffic arrival. Firstly, T_{it} is computed as:

$$T_{it} \begin{cases} T_{it\_min} & if\ T_{it\_min} < T_{it\_max} \\ T_{it\_max} & otherwise \end{cases} \quad (1)$$

Where, T_{it_min} and T_{it_max} is minimum and maximum idle threshold, respectively. The T_{it_min} is derived from (Saidu et al., 2015) as follows:

$$T_{it\_min} = \frac{E_w}{E_i} \quad (2)$$

Where E_w is sleep-wake energy and E_i is idle state energy.

Then the next T_{it} is predicted as:

$$T_{it} = T_{it} + \left(\frac{E_{residual}}{E_{total}}\right) * (\lambda_{new-weight} - \lambda_{old-weight}) \quad (3)$$

$$\lambda_{new-weight} = (1-\gamma)\,\lambda_{current-arrival} + \gamma * \lambda_{old-weight}, 0 < \gamma < 1, \quad (4)$$

where γ is the proportionality constant, $\lambda_{new-weight}$ and $\lambda_{old-weight}$ is the new and old arrival rate, respectively.. $\lambda_{current-arrival}$ is current arrival rate, $E_{residual}$ and E_{total} is residual and total energy, respectively.

3.2 Initial Sleep Window

The $T_{\min}$ is dynamically updated according to the residual energy and the new weight arrival rate as follows:

$$T_{\min} = max\left(\left\lceil \frac{E_{residual}}{Etotal}\,\lambda_{new-weight} \right\rceil, 1\right) \quad (5)$$

3.3 Final Sleep Window

The $T_{\max}$ is computed when the $T_{\min}$ is determined and the frame response delay is given. Firstly, let the frame arrival follows a Hyper-Erlang distribution with pdf (Zhang, 2007) as.

$$p_{h_d}(t) = \sum_{i=1}^{I} \beta_i \frac{(\omega_i \lambda_i)^{\omega_i} t^{\omega_i - 1}}{(\omega_i - 1)!} e^{-\omega_i \lambda_i t} \quad (6)$$

Where λ_i and β_i are constants with $\lambda_i \geq 0$, $0 \leq \beta_i \leq 1$ and ω_i are positive integers.

Integrate Equation (6) from $0 < x \leq t$ to derive the CDF as

$$C_{h_d}(t) = \int_0^t f_{t_d}(x)dx \quad (7)$$

Next, suppose $h_{d,r}$ is the remaining packets inter-arrival time with pdf $p_{h_{d,r}}$. $p_{h_{d,r}}$ is computed using Equations (6) and (7) as:

$$p_{h_{d,r}} = \lambda \sum_{i=1}^{I} \beta_i \sum_{j=0}^{\omega_i - 1} \frac{(\omega_i \lambda_i t)^j}{j!} e^{-\omega_i \lambda_i t} \quad (8)$$

Integrate Equation (8) from $0 < p_{h_{d,r}} \leq t$, to obtain the CDF $C_{h_{d,r}}$ as

$$C_{h_{d,r}} = 1 - \lambda \sum_{i=1}^{I} \frac{\beta_i}{\omega_i \lambda_i} \sum_{j=0}^{\omega_i - 1} \sum_{k=0}^{j} \frac{(\omega_i \lambda_i t)^k}{k!} e^{-\omega_i \lambda_i t} \quad (9)$$

Finally, S_h denotes the summation of 1st, 2nd, . . ., hth derived as:

$$S_h = \sum_{j=1}^{h} (T_j + L), h \geq 1 \quad (10)$$

Assume T_h represents the length of the h^{th} sleep window as

$$T_h = \begin{cases} 2^{h-1} T_{\min} \text{ if h} < Y \\ T_{\max} \text{ otherwise} \end{cases} \quad (11)$$

CASE I

The probability of frame arrival at the idle state is computed as:

$$P_h(N = T_{it}) = 1 - \lambda \sum_{i=1}^{I} \frac{\beta_i}{\omega_i \lambda_i} \sum_{j=0}^{\omega_i - 1} \sum_{k=0}^{j} \frac{(\omega_i \lambda_i t)^k}{k!} e^{-\omega_i \lambda_i t} \quad (12)$$

CASE II

The probability of frame arrival during the h^{th} sleep interval is :

$$P_h(N = h) = \lambda \sum_{i=1}^{I} \frac{\beta_i}{\omega_i \lambda_i} \sum_{j=0}^{\omega_i - 1} \sum_{k=0}^{j} \frac{(\omega_i \lambda_i)^k}{k!} e^{-\omega_i \lambda_i L} \cdot [(S_{h-1} - L)^k e^{-\omega_i \lambda_i c_{h-1}} - (S_{h-1} - L)^k e^{-\omega_i \lambda_i c_n}] \quad (13)$$

if $h < Y$

Replacing Equations (10) and (11) into Equation (13),we have

$$P_h = A \times \left[(X - L)^k e^{-\omega_i \lambda_i X} - (Z - L)^k e^{-\omega_i \lambda_i Z} \right] \quad (14)$$

Where $X = ((2^{h-1} T_{min} + hL) - (T_{min} + L)$,

$Z = (2^h T_{min} + hL) - T_{min}$ and

$$A = \lambda \sum_{i=1}^{I} \frac{\beta_i}{\omega_i \lambda_i} \sum_{j=0}^{\omega_i - 1} \sum_{k=0}^{j} \frac{(\omega_i \lambda_i)^k}{k!} e^{-\omega_i \lambda_i L}$$

if $h \geq Y$

Similarly, Replacing Equations (10) and (11) into Equation (13),we have

$$P_h = A \times \left[(P - L)^k e^{-\omega_i \lambda_i P} - (Q - L)^k e^{-\omega_i \lambda_i Q} \right] \quad (15)$$

Where

$P = ((2^{Y-1} T_{min} + YL) - (T_{min} + L)$ and

$Q = (2^Y T_{min} + YL) - T_{min}$

Using Equations (10), (11), (14) ,(15) and (16), the average response delay is obtained as:

$$A[R] = \frac{L}{2} + D + A \frac{T_{\max}}{2} \left[(P - L)^k e^{-\omega_i \lambda_i P} - (Q - L)^k e^{-\omega_i \lambda_i Q} \right] \quad (16)$$

Finally, from Equation (16), we obtain

$$T_{max} = \frac{2A[R] - L + 2D}{A\left[(P-L)^k e^{-m_i \lambda_i P} - (Q-L)^k e^{-\omega_i \lambda_i Q} \right]} \quad (17)$$

4. Performance Evaluation

To evaluate the performance of the HBLES and the BLAP, simulation is used. The discrete event simulation is developed using C++ Programming Language. The metrics used in the simulation are as follows:

The average energy consumption is as

$$A(E) = T_{it} + E_i \sum_{k=0}^{\infty} P_k \sum_{l=0}^{k} (T_l E_S + L E_L + E_W) \quad (18)$$

Where E_S *and* E_L is the energy consumed during the sleeping and listening mode respectively.

The Equation (16) provides response delay.

The simulation parameters and the topology used are in (Xiao, 2005). The topology consists of one BS station and several MSSs. The BS transmits only non-real-time traffics to the MSSs, which is a downlink transmission. The MSSs operate on a 300 mAh initial battery capacity and use a simple linear power model. Both the MSSs and the BS use the power consumption

parameters adopted in the SQN1130 System-on-Chip (SOC) (Sequans Communications). . In addition, the MSS uses 1 frame duration for the listening window because no benefit setting it higher than 1 frame.

Figures 1and 2 illustrate the average energy consumption and average response delay of the proposed HBLES and the BLAPS in terms of mean arrival rate, respectively .The HBLES achieves superior performance than the compared scheme because of the accurate capture of the traffic characteristics.

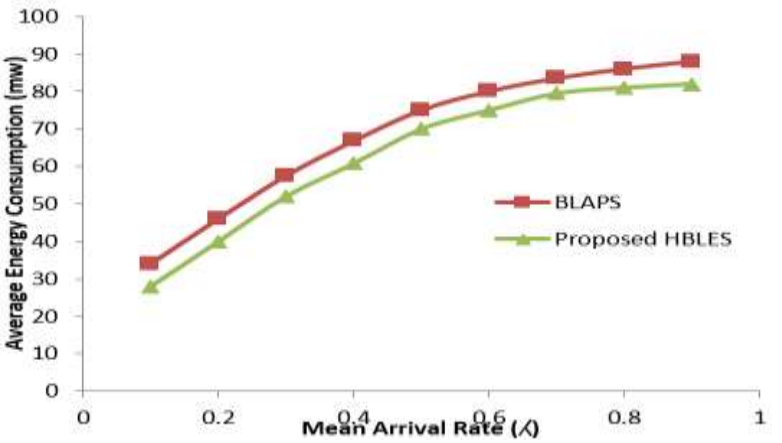

Figure 14: Graphical Figure of Logistic Regression Classifier for

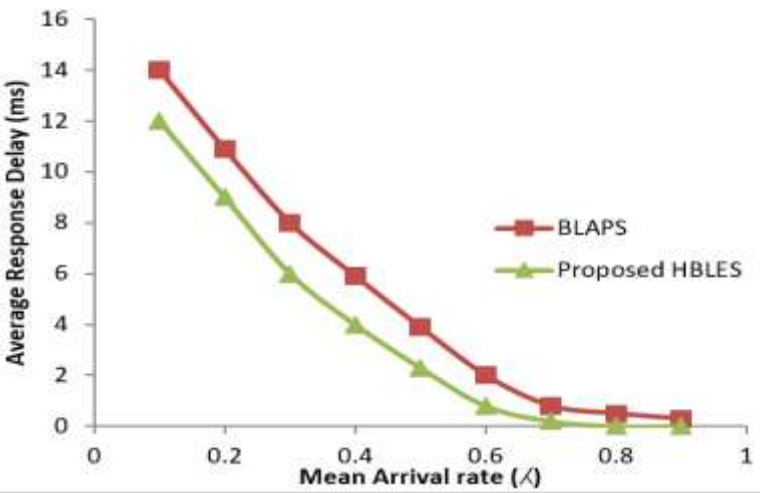

Figure 2.Response Delay vs Arrival Rate.

8. Conclusion

In this paper, an HBLES is proposed to improve energy efficiency while reducing the response delay. The scheme modifies the sleep parameters based on the remaining battery power

and downlink traffic arrival pattern. It also uses Hyper-Erlang distribution to determine the actual traffic arrival characteristics. Several simulations are conducted for the evaluation of the HBLES and the BLAPS. The results demonstrate that the HBLES achieves superior performance than the existing scheme in terms of the metrics used.

Acknowledgment

The authors acknowledge the support received from the Umaru Musa Yar'adua University, Katsina, and Katsina State, Nigeria.

References

IEEE 802.16e WG (2005). IEEE Standard for Information Technology - Telecommunications and Information Exchange between Systems - LAN/MAN Specific requirements, Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems

Saidu, I., et al. (2015). An efficient battery lifetime aware power saving (EBLAPS) mechanism in IEEE 802.16 e networks. *Wireless Personal Communications"* vol. 80, pp. 29-49.

Xiao, Y. (2005). Energy saving mechanism in the IEEE 802.16e wireless MAN, IEEE Communications Letters, vol. 9, pp. 595–597.

Xiao, Y. (2006). Performance Analysis of an Energy Saving Mechanism in IEEE 802.16 e wireless MAN, in 3rd IEEE Conference on Consumer Communications and Networking CCNC'06, vol. 1, pp. 406–410.

Zhang, Y., & Fujise, M. (2006). Energy management in the IEEE 802.16 e MAC, IEEE Communications Letters, vol. 10, pp. 311–313.

Xiao, J., Zou, S., Ren, B., & Cheng, S. (2006). WLC17-6: An enhanced energy saving mechanism in IEEE 802.16 e, in IEEE Global Telecommunications Conference GLOBECOM'06, pp.1–5.

Zhu, S., & Wang, T. (2007). Enhanced power efficient sleep mode operation for IEEE 802.16 e based WiMAX, in IEEE Mobile WiMAX Symposium, pp. 43–47.

Zhu, S., Ma, X., & Wang, L. (2007) A delay-aware auto sleep mode operation for power saving WiMAX, in Proceedings of 16th International Conference on Computer Communications and Networks ICCCN'07 , pp. 997–1001.

Xue, J., Yuan, Z., & Zhang, Q. (2008). Traffic load-aware power-saving mechanism for IEEE 802.16 e sleep mode, in 4th IEEE International Conference on Wireless Communications, Networking and Mobile Computing WiCOM'08, pp. 1–4.

Sanghvi, K., Jain, P. K., Lele, A., & Das, D. (2008). Adaptive waiting time threshold estimation algorithm for power saving in sleep mode of IEEE 802.16e, in 3rd IEEE International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE '08), pp. 334–340.

Kim, M.G., Kang, M., & Choi, J. Y. (2008). Remaining Energy-aware power management mechanism in the 802.16e mac, in 5th IEEE Consumer Communications and Networking Conference CCNC'08, pp. 222–226.

Lin, Y.W., & Wang, J.S. (2013). An Adaptive QoS Power Saving Scheme for Mobile WiMAX, Wireless Personal Communications , vol. 69, no. 4, pp. 1435–1462.

Ferng, H.W., & Li, H.Y. (2013). Design of predictive and dynamic energy efficient mechanisms for IEEE 802.16e, Wireless Personal Communications, vol. 68,no. 4, pp. 1807–1835.

Chou, L.D., Li, D. C., & Hong, W.Y. (2013). Improving energy-efficient communications with a battery lifetime-aware mechanism in IEEE 802. 16e wireless networks, Concurrency and Computation: Practice and Experience, vol. 25, pp. 94–111.

Zhang, Y. (2007). Performance modeling of energy management mechanism in IEEE 802.16e mobile WiMAX. In *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE*, pp. 3205-3209.

Sequans Communications, "Datasheet: SQN1130 System-on-Chip (SOC) for WiMAX Mobile Stations."