Machine Learning Approach for Detection of Spam Url: Performance Evaluation of Selected Algorithms

Ahiaba Moses Okpanachi & Morufu Olalere
Department of Cyber Security Science
Federal University of Technology
Minna, Nigeria
E-mails: ahiabamoses@yahoo.com; lerejide @futminna.edu.ng

ABSTRACT

The Internet, web consumers and computing systems have become more vulnerable to cyber-attacks. Spam which exist in different form has recently becomes one of the techniques attackers use to get confidential information from their victims. Whatever is the form of spam, Uniform Resource Locator (URL) serves as a key driver for spam. Hence, detection of spam URL has attracted attention of many researchers. Machine learning approach is one of the approaches researchers have used in this area of study. Meanwhile, no researcher has reported 100% accuracy with any machine leaning algorithm and not all machine learning algorithms has been explored in this area of research. Consequently, this study presents performance evaluation of some selected algorithms with the aim of identifying best algorithm in terms of accuracy, precision, sensitivity, specificity, mean Squared Error. WEKA data mining tool was used carry out experiment on the selected algorithms. The results of our experiment revealed that K-NN out performed other algorithms with highest values in accuracy, precision, sensitivity and with lowest values in specificity and mean squared error.

Index Terms— Spam URL, machine learning, Naïve Bayes, J48, Multilayer perceptron, K-NN..

iSTEAMS Multidisciplinary Conference Proceedings Reference Format

Ahiaba Moses Okpanachi & Morufu Olalere (2019): Machine Learning Approach for Detection of Spam Url: Performance Evaluation of Selected Algorithms. Proceedings of the 22nd iSTEAMS Multidisciplinary SPRING Conference. Aurora Conference centre, Osogbo, Nigeria. 17th – 19th December, 2019. Pp 243-252 www.isteams.net/spring2019. DOI - https://doi.org/ 10.22624/AIMS/iSTEAMS-2019/V22N1P23

1. INTRODUCTION

Spam can be defined as unsolicited, unwanted email that is sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient [1] or an unsolicited commercial mail usually sent to a large group of recipients at the same time by service providers such as internet service providers (ISPs) to market their products [1] and services, current and new ones. Spam is an ongoing issue that has no perfect solution and there is no complete solution technique about spam problem [2]. According to reports produced by MacAfee laboratory the average global spam rate was two trillion a day MacAfee (2011). Spam exists in different forms such as Usenet spam, SMS spam or IM spam sent by instant messaging services and web log spam among others. URL is the abbreviation of Uniform Resource Locator, which is the global address of documents and other resources on the World Wide Web. A URL has two main components: (i) protocol identifier (indicates what protocol to use) (ii) resource name (specifies the IP address or the domain name where the resource is located) [3].

The URLs could direct users to websites that contain malicious content, drive-by download attacks, phishing, spam, and scams; this type of URL is known as Malicious URL [4]. Compromised URLs that are used for cyber attacks are termed as malicious URLs [3]. Malicious URL is a URL created with the intention of stealing data, money or personal information through the fake website [5]. Through malicious URLs, users are made to fill in their email addresses which have expose legitimate users to receiving spam emails that exhaust their computer resources, steal their personal details and scam these users. Machine learning algorithms have been one of the powerful techniques in detecting spam URL attack. In this study, comparatively analyze machine learning classification algorithms that can be employed for employed for spam URL attacks detection, employing the following machine learning classification algorithms Decision tree (J48), CS Forest, K-nearest neighbor, Naive Bayes and Multilayer Perception.

2. RELATED LITERATURE

The majority of studies in spam URL attack aim to find the most predictive features that they can acquire and the best algorithm to develop a classifier model.

[6] proposed effectiveness of anomaly detection applied to spam filtering by using Artificial Immune System (Negative Selection Algorithm). They achieved an actual detection rate of spam and ham of 98.5%.

[7] proposed detection of spam URL attack and image spam filtering using machine learning algorithm that is Support Vector Machine (SVM). They achieved accuracy in detection of the spam URL and image spamming, but could not measure FP, FN, TP and other parameters.

According to [8] proposed the detection of social spamming on facebook platform using KNN, classification decision, SVM, Naïve Bayes, Ramdom Forest and J48 as machine learning algorithm. They achieved accuracy of 96.3% with the machine learning algorithms. But there is not robustness and applicability of the end classifier.

[4] tried to detect suspicious URLs in online social networks using Random Forest classification Model; the model achieved a recall of 92% however the accuracy was not recorded.

[9] designed a machine learning to detect spam tweet using hybrid technique of Naïve Bayes and support vector machine, the system was able to categorize the message into spam and non spam, however, there was no record of any evaluation metrics.
[
[10] proposed support vector machine to detect spam URL, accuracy of 81% F1 score of 74% was achieved, however, the technique could not close the space for feasible attacks.

1) Spam

Literally, It is a trademark for a canned meat product made mainly form ham. Generally, spam (Specialized Automated Mail) is the electronic version of “junk mail, unrequested e-mail messages that advertise products to consumers that have Internet e-mail boxes [11]. Spam is abuse of electronic messaging system to send unsolicited bulk messages. Today large volumes of spam emails are causing serious problem for the users, and internet services. Such as, it degrades user search experience, It assists propagation of virus in network, It increase load on the network traffic, It wastes the resources such as bandwidth, storage, and computation power, It also wastes the user time and energy [11]. General advices to avoid spam’s are use the spam filter, Never reply the spam, Don’t post your email address on your web site, and Never buy anything from spam [11].

2) Effect of spam

Spam has severe negative effects on e-mail users. It consumes computer, storage and network resources as well as human time and attention to unwanted messages. Moreover, it has various indirect effects which are very difficult to account for - the spectrum reaches from measurable costs like spam filter software and administration to not measurable costs like a lost e-mail (expensive for a business, not that expensive for a private person) [5].

3) Universal Resource Locator

URL's, also known as Universal Resource Locator, as the name indicates are used to find a particular resource on the Internet, they are also known as web address. A URL finds by providing, to the browser for example, an abstract of the location of the resource, when this resource is found the system can execute a great diversity of operations [12]. A URL is formed by the protocol used to access the resource, the location of the server to be accessed, which may be on the form of the domain or on the form of the IP address and the path where the resource is located [12]. Some of the most popular types of malware or malicious URLs attacks are Drive-by Download, Phishing and Social Engineering, and Spam [13] and [14].

4) Phishing URL

[15] define a phishing web page as "any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewers would only trust a true agent of a the third party." This definition, which is similar to the definition of web forgery", covers a wide range of phishing pages from typical ones – displaying graphics relating to a financial company and requesting a viewer‟s personal credentials – to sites which claim to be able to perform actions through a third party once provided with the viewer‟s login credentials. Thus, a phishing URL is a URL that leads user to a phishing web page.

5) Malicious URL

To define it, a malicious URL is a link created with the purpose of promoting scams, attacks and frauds. By clicking on an infected URL, you can download a malware or a Trojan that can take your devices, or you can be persuaded to provide sensitive information on a fake website [15]. Malware propagators are users who tweet malicious links, which on clicking leads to the downloading of malwares. Malware's are malicious software [16].

6) Detecting Spam URL

An identity theft that occurs when a malicious web site masquerades a legitimate one is called spam. Such a theft occurs in order to procure sensitive information such as passwords, bank account details, or credit card numbers. Spam URL makes use of spoofed emails which look exactly like an authentic email. These emails are sent to a bulk of users and appear to be coming from legitimate sources like banks, e-commerce sites, payment gateways etc.

7) Principles of Detecting Malicious URLs

Several other methods have been endeavored to tackle the problem of Malicious URL Detection. According to the fundamental principles, these methods can be broadly grouped into two major categories: (i) Blacklisting or Heuristics, and (ii) Machine Learning approaches [17].

8) Machine Learning Approach

Machine Learning is a field of science that deals with the design of computer programs and systems that can learn rules from data, adapt to changes, and improve accuracy or performance with experience. For this study, out of the many machine learning algorithms the author have selected these five algorithms, CS Forest, K-Nearest Neighbor, naive Bayes, Multi Layer Perceptron and J48 to study and identify the best classifier for detection of spam URL attack used.

Naive Bayes

Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem (or Bayes's rule) with strong independence (naive) assumptions. Parameter estimation for Naïve Bayes models uses the maximum likelihood estimation. It takes only one pass over the training set and is computationally very fast [18].

Decision Trees (J48)

Decision tree learners are a nonparametric supervised method used for classification and regression. In a decision tree, classifier is represented as a tree whose internal nodes represent the condition of the variable and final nodes or leaves are the final decision of the algorithm. In the processof classification, a well-formed decision tree can efficiently classify a document by running a query from the root not until it reaches a certain node. The main advantage of using decision tree is that it is simple and easy to understand and interpret for naïve users. The risk associated with decision tree is over fitting which occurs when a tree is fully grown, and it may lose some generalization capabilities. Some common reasons of over fitting are the presence of noise, lack of representation instance, and multiple comparison procedures. Over fitting can be avoided by severing approaches such as pre-pruning and post-pruning [19].

K-Nearest Neighbour

The K-Nearest Neighbour Algorithm (KNN) is the simplest machine learning algorithm. To determine the category of the test data, K-NN performs a test to check the degree of similarity between documents and training data to store a certain amount of classified data. Since K-NN classifies instances, in our research, it will be malicious and benign code instances nearest to the training space. The classification of unknown instances is performed by measuring the distance between the training instance and unknown instance. Since instances are classified based upon the majority vote of neighbour, the most common neighbour is measured by a distance function. If, then the instance is assigned for the class of its nearest neighbor [19].

Multi - Layer Perceptron

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. MLP utilizes a supervised learning technique called back propagation for training the network. Neural networks are different paradigm for computing and it is an inspiration from neuroscience. Neural networks are particularly effective for predicting events when the networks have a large database. This network imitates the human brain. Artificial neurons or processing elements are highly simplified models of biological neurons. As in biological neurons, artificial neurons have a number of inputs, cell body and output that can be connected to a number of other artificial neurons. This network is densely interconnected together by learning rule which to adjust the strength of the connection between the units in response to externally supplied data [2].

3. METHODOLOGY

Spam URL dataset was collected from University of New Brunswick (UNB) dataset, the dataset is made of both spam URL attacks and non spam URL attack, the downloaded dataset is comma-separated values (CVS) format. Attack and normal instances are combined and labeled with "benign" for normal traffic and "spam" for URL attack traffic, the dataset is made of 79 attributes and 14,479 instances. The collected dataset was preprocessed by removing the instances that are non numeric in nature; also the data was manually cleaned up by removing blank spaces.

Feature selection is very necessary in machine learning algorithm, because it helps in removing useless features from the dataset. For this study, the feature selection process was done by dropping the features with least importance and the related features. This dataset is made of 79 attributes, but for this study, only 6 attributes was selected for the experiment because the features are lexical in form, so 73 related features were dropped. These 6 selected attributes have high impact on our prediction.

The intended approach and the selected metrics for carrying out the performance analysis are outlined as follows:

Sensitivity: sensitivity is the percentage of the test set that the model predicts; it is also referred to as True Positive Rate (TPR). It is denoted as:

$$Sensitivit\ y(TPR) = \frac{TP}{TN + FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Specificity: this can be defined as the percentage of the test set that is projected as correct, it can also be called True Negative Rate (TNR) which is designated as:

$$Specificity(TNR) = \frac{TP}{TN + FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

False Positive Rate: this can also be called the false alarm rate, it is the percentage of the test set that the model predicts falsely as positive when it was actually negative. It is denoted as:

$$(FPR) = \qquad \frac{FP}{TN + FP} = 1 - Specifty \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

False Negative Rate: this can be referred to the percentage of the test set that the model predicts falsely as negative when it is actually positive. It is denoted as:

$$(FNR) = \frac{FN}{TR + FN} = 1 - Sentivity \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Accuracy is a term that describes the average value of sensitivity and specificity, defined by a formula:

$$Accuracy = \frac{Sensitivity + Specificity}{2} \qquad \text{........(5)}$$

Precision: it is the percentage of the test set that the model predicts correctly. It is denoted as:

$$\Pr ecision = \frac{TP}{TP + FP} \qquad \text{...............................(6)}$$

5. EXPERIMENTAL RESULTS

After our experiment, the results obtained are presented in this section. Figure 1 presents accuracy of each of the machine learning algorithms. As shown in Figure 1, k-NN performed best with accuracy of 99.54%. Meanwhile, Naïve Bayes has lowest accuracy of 95.34%. This implies that K-NN has ability to detect highest number of spam URL and Benign URL correctly compare to other algorithms.

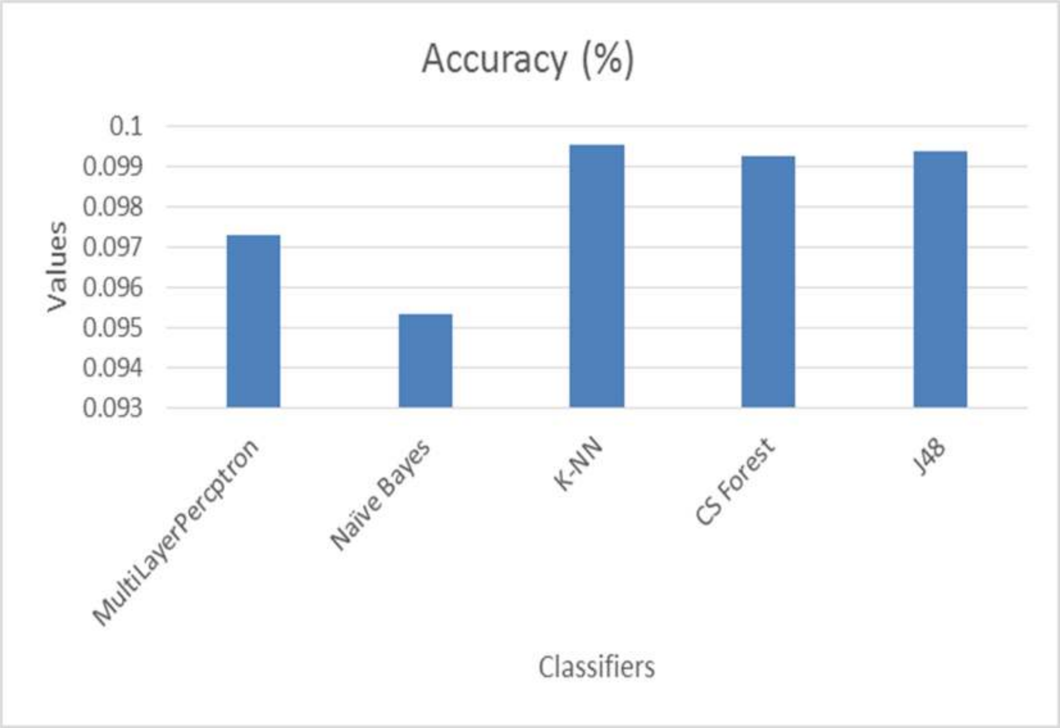

Figure 1. Accuracy of each of the machine learning algorithms

In Figure 2, the precision of each of the algorithms is presented. As it is shown in Figure 2, K-Nearest Neighbor has the highest precision values. The higher the precision of classification algorithm, the better is the performance of the algorithm. Therefore, K-NN performed best in term of precision. However, Naïve Bayes has lowest precision value. This implies that Naïve Bayes is a least performed algorithm for the job of detection spam URL.

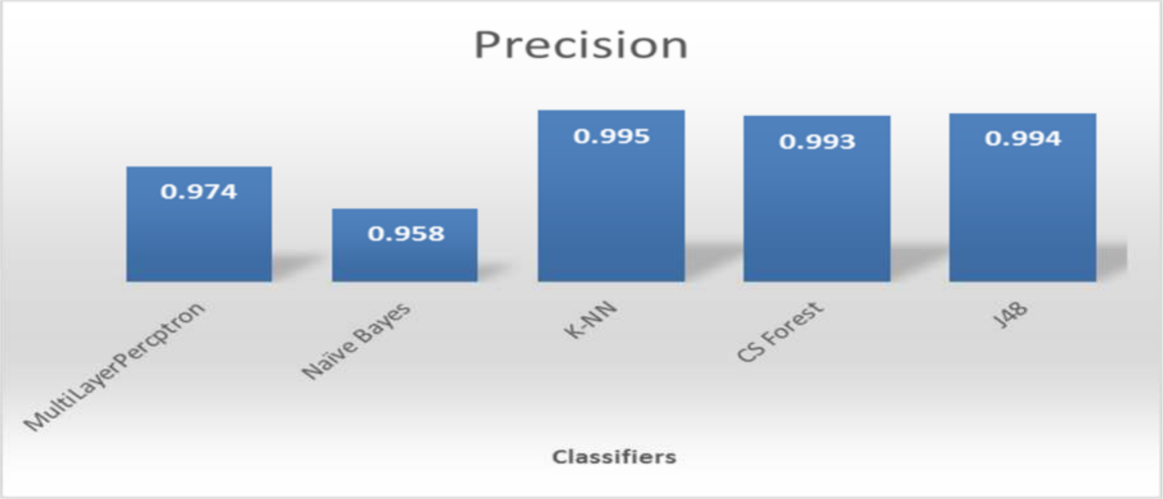

Figure 2: Precision of each of the machine learning algorithms

Figure 3 presents sensitivity of each of the algorithms. In Figure 3, it can be seen that K-NN has highest has highest sensitivity. This is followed by CS Forest, J48 and Multilayer Perceptron. Naïve Bayes has lowest precision. The higher the sensitivity of any classification algorithm, the better the algorithm.

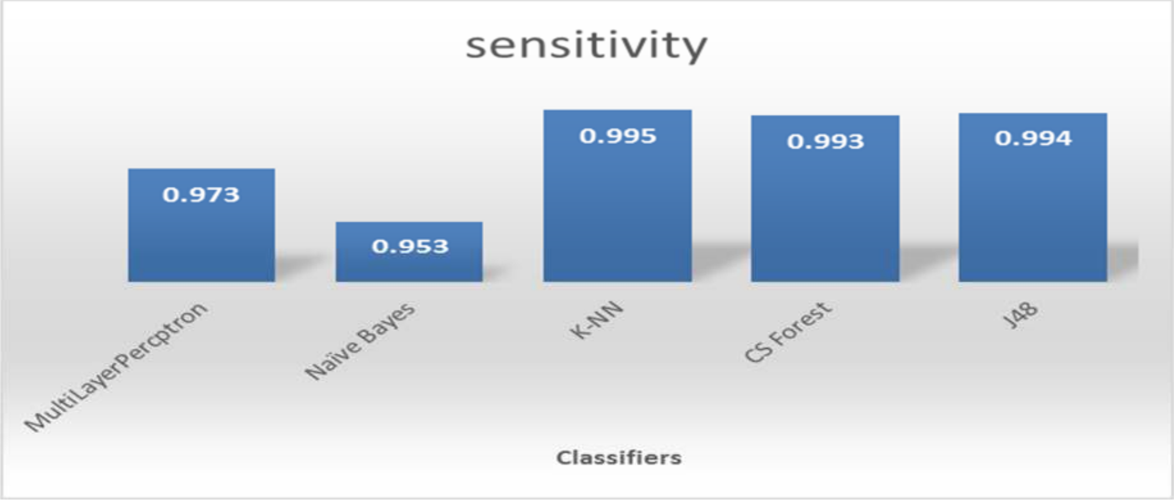

Figure 3. sensitivity of each of the machine learning algorithms

Figure 4 and Figure 5 present Specificity and Mean Squared error of each of the algorithm respectively. In Figure 4, K-NN has lowest specificity while Naïve Bayes has highest specificity. The lower the specificity of any classification algorithm, the better the algorithm. Therefore, K-NN with lowest specificity performed best compare with other algorithms. Also in In Figure 5, K-NN has lowest Mean squared error while Naïve Bayes has highest Mean squared error. The lower the Mean squared error of any classification algorithm, the better the algorithm. Therefore, K-NN with lowest Mean squared error performed best compare with other algorithms.

.

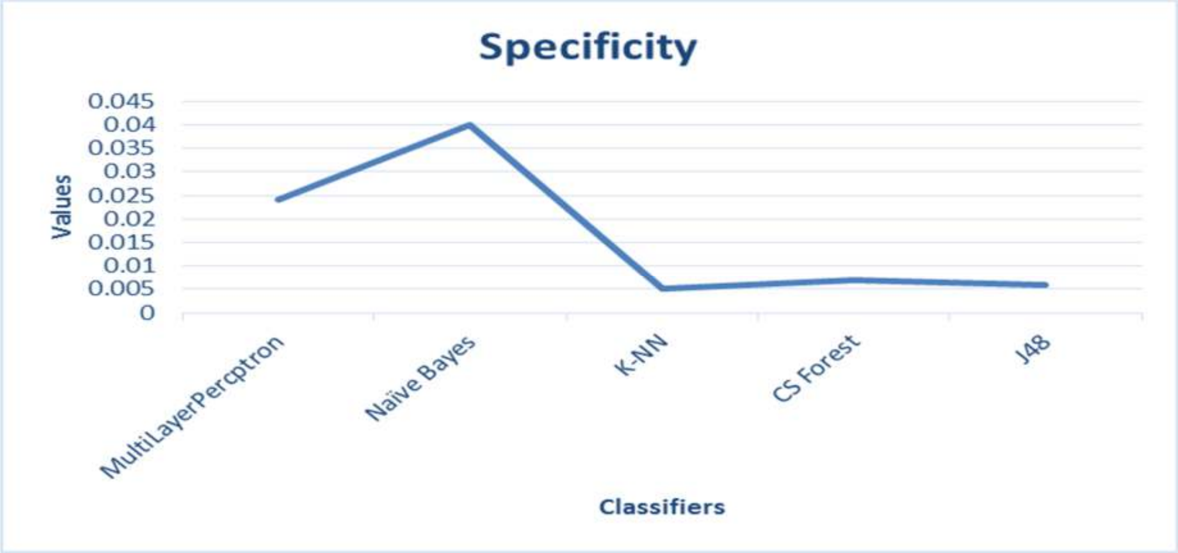

Figure 4. Specificity of each of the machine learning algorithms

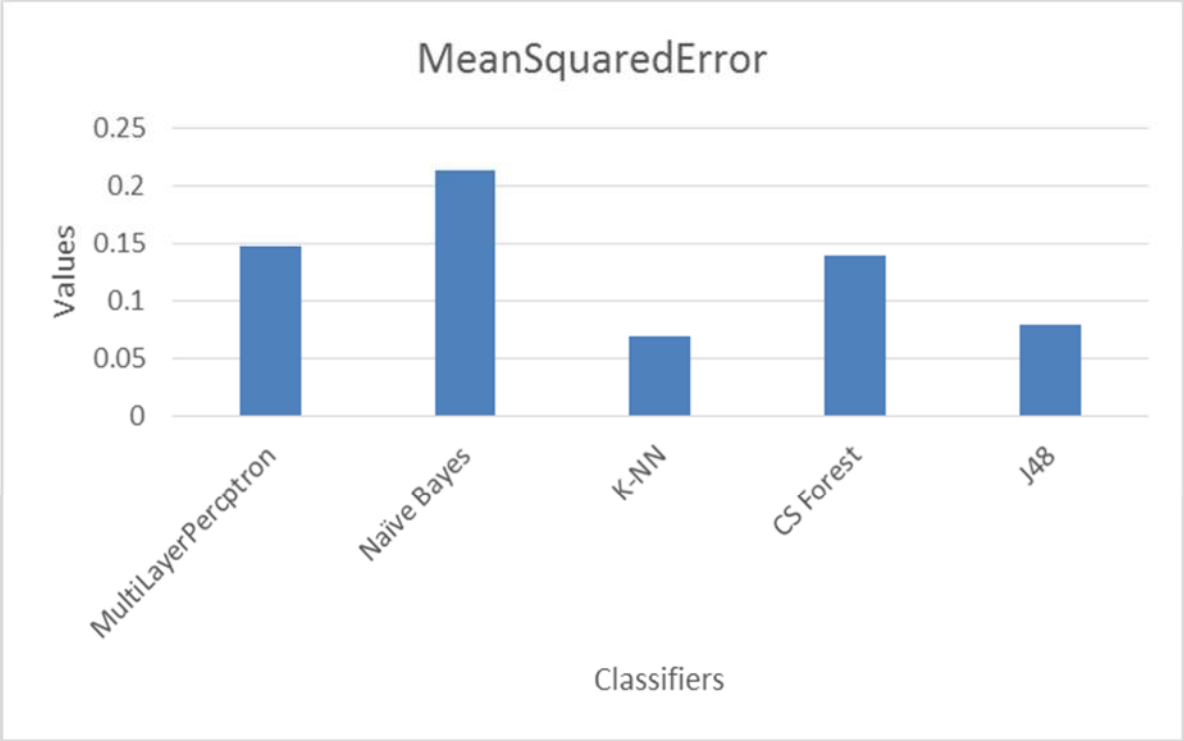

Figure 5. Mean Squared Error of each of the machine learning algorithms

6. CONCLUSION AND RECOMMENDATION

In this study, we have experimented; evaluated and made comparisons of some selected machine learning algorithm for the detection of spam URL attack. The results obtained clearly shows that K-Nearest Neighbour has the best performance than Multilayer Perceptron, CS Forest, and J 48 with accuracy of 99.54% and specificity of 0.005 with mean squared error of 0.0692. Naïve Bayes has the worst performance of 95.34% accuracy and 0.040 specificity with mean squared error of 0.2139.K-NN algorithm was the best for spam URL detection, which is effective than other selected machine learning classification algorithms.
K-NN algorithm is useful and accurate tool for classifying spam URL.

REFERENCES

[1] Ndumiyana D., Munyaradzi M, & Lucy S. (2013). Spam Detection using a Neural Network Classifier. Online Journal of Physical and Environmental Science Research, 2, (2), pp. 28-37.

[2] Torabi Z. S., Mohammad H. N.S & Akbar N. (2015). Efficient Support Vector Machines for Spam Detection. (IJCSIS) International Journal of Computer Science and Information Security,13,(1), pp 11 – 28.

[3] Sahoo D., Chenghao L, & Steven C.H. (2019). Malicious URL Detection using Machine Learning. 1,(1). Pp 1- 35. https://doi.org/10.1145 .

[4] Al-Janabi M. Ed de Q. & Peter A. (2017) . Using supervised machine learning algorithms to detect suspicious URLs in online social networks. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp 1104 – 1110.

[5] Selvaganapathy S., Mathappan N.& Hema P. (2018). Natarajan Deep belief network based detection and categorization of malicious URLs. Information Security Journal: A Global Perspective, pp 1 – 18. DOI: 10.1080/19393555.2018.1456577.

[6] Saleh Abdul Jabbar, A., Bharanidharan S., Sami A., Krishnan K., Jonkman M. & Friso D. B. (2019). An Intelligent Detection Model Based on Artificial Immune System.

[7] Parekh Parth, Kajal Parmar & Pournima Awate (2018). Spam URL Detection and Image Spam Filtering using Machine Learning, International Research Journal of Engineering and Technology. 05, (07), 2370 – 2374.

[8] Tartu Barigou, Naouel Barigou, and Baghdad Atmani (2019). Detection of social spamming on facebook platform. International Journal o Mechanical Engineering and Technology 8,(12), pp. 536–54.

[9] Kailas S. T & Kshirsagar.D.B. (2016). Design of Machine Learning Approach For Spam Tweet Detection. International Journal of Advance Research and Innovative Ideas in Education,2,(5), 2395-4396.

[10] R.Jeena, G.preethi, A.Praveena, & A.preethi (2019). Malicious URL Detection Using Machine Learning Techniques. International Journal of Innovative Research in Science, Engineering and Technology. 8, (3), pp 1751- 1754. DOI:10.15680/IJIRSET.2019.0802010.

[11] Teli S. P., Biradar Santoshkumar , (2017). Effective Email Classification for Spam and Non-Spam. 4th International Conference on Advances in Electrical Engineering.

[12] Marcelo F. (2019). Malicious URL Detection using Machine Learning Algorithms. Proceedings of the Digital Privacy and Security Conference, pp 114 - 122.

[13] Douksieh Abdi1 F. & Wenjuan L.(2016). Malicious URL Detection Using Convolutional Neural Network. International Journal of Computer Science, Engineering and Information Technology , 7(6),

[14] Lekshmi A. R & Seena T. (2019).The Kozinec-SVM Model for Detecting Malicious URLs. International Journal of Engineering Research & Technology (IJERT), 8 (07), pp 135 – 139.

[15] Ram B. B, Andrew H. S., Quingzhong L. (2014). Learning to Detect Phishing URLs. International Journal of Research in Engineering and Technology. 03 (06), pp 11 -24.

[16] Chorey S. A., Sawade R..N., Badnera R., Priyanka C., Mamanka P.V. (2016). Detecting Spam Classification on Twitter Using URL Analysis, Natural Language Processing and Machine Learning. International Journal of Innovative and Emerging Research in Engineering, 3, (1), 141-145.

[17] Sridevi M. & Sunitha K.V.N. (2017). Malicious URL detection and prevention at browser level. .International Journal of Mechanical Engineering and Technology. 8,(12), pp. 536–54.

[18] James P., Abhilash R., Praveen K. R. (2013). URL attacks: Classification of URLs via Analysis and Learning. International Journal of Electrical and Computer Engineering, 6,(3), 980 ~ 985,. DOI: 10.11591/ijece.v6i3.7208 ρ 980.

[19] N. Khan, J. Abdullah, and A. S. Khan, “Learning Classifiers,” vol. 2017, 2017.

[20] Baharuddin M.F., Tengku A. T & Mohd S.M. (2018).Malicious URL Classification System Using Multi-Layer Perceptron Technique. Journal of Theoretical and Applied Information Technology, 96,(19), pp 6454 – 6462.