International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 3574–3584

Identification Of Malicious Websites With HTML And URL Based Features Using Machine Learning

Shaik Irfan Babu[1*] and Dr.M.V.P.Chandra Sekhara Rao[2]

[1]*Research Scholar, Dept. of CSE, ANU College of Engineering, Acharya Nagarjuna University(ANU), Guntur, Andhra Pradesh, India*
[2]*Professor, Dept. of CSE, R.V.R.and J.C College of Engineering (A), Guntur, Andhra Pradesh, India*
[1]irfanbabushaik@gmail.com and [2]manukondach@gmail.com

Abstract

In recent years, with the rapid development of the Internet and the continuous growth of network services, the threat to user's privacy and security has increased. This is mainly because of malicious web pages. Malicious webpage detection technology as a core security technology to resist network attacks, can help users effectively avoid security threats caused by malicious webpage's and ensure network security. This paper aims to assess and identify malicious websites by building a malicious site identification model with the help of machine learning algorithms. The present work used the URL and HTML based features to identify malicious websites. It is found that both URL and HTML based features are effective in analyzing and classifying malicious URLs. Most of the samples in this study were taken from PhishTank and Alexa. Further, it is seen that there is huge improvement in classification precision using proposed approach and SVM (Support Vector Machine) ends up being the best classifier offering the accuracy of 91.8% with FPR and FNR as 0.90 and 0.82 respectively.

Keywords— *Web Page Classification, Artificial Neural Networks, Machine Learning, Logistic Regression, KNN, SVM Classifier, Naive Bayes.*

1. INTRODUCTION

In recent years, the internet, especially the mobile internet, has developed rapidly. Various applications such as e-commerce, social media, online banking and mobile networks have emerged one after another. People have begun to use the internet for online banking and transaction processing. However, as the e-commerce and online banking systems are booming, various network security issues are becoming more and more common. Hacking behaviors such as viruses, account theft, Trojan horses, and phishing have extremely bad effects on the internet environment. According to the network monitoring report of the Internet Security Organization, malicious website attacks have surpassed traditional malicious attacks and become the biggest threat to current network attacks.

In order to minimize user losses, various researchers have proposed many different methods of identifying malicious websites. The typical method is to use black and white list recognition technology, which can easily provide services in the form of browser plug-ins. However, due to the rapid increase in the number of malicious websites, it has become more and more difficult to establish a complete black and white list, and it has become almost impossible [1][2]. Therefore, some researchers began to use advanced machine learning technology to model from the perspective of URL abnormal features, web content features, Internet evaluation data, etc., and achieved good results [3][4][5]. However, before building a malicious website identification model, effective evaluation of malicious websites is very important for understanding the existence of malicious websites and building an effective identification model. At present, only a few studies analyzed and evaluated websites for malicious codes [6], but this evaluation method only analyzes website codes, which is not comprehensive enough and the analysis efficiency needs to be improved.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

This paper uses URL and HTML based features to find the evaluation of malicious websites. Based on the above features, malicious websites are identified. Machine learning algorithms are used to verify the effectiveness of the extracted main factors in identifying malicious websites.

2. Related works

In view of the serious damage done by malicious websites, the rapidity of spreading and the wide geographical scope, it is very important to identify malicious websites efficiently and accurately. Based on the previous literature, the existing research on malicious website recognition can be divided into recognition technology based on black and white lists and recognition technology based on machine learning. It is observed that black and white list-based technology is not effective in identifying malicious websites [2][7].

The recognition technology based on URL abnormal features uses URL features to construct a malicious website recognition model. Mohith Gowda et al. [1] proposed a malicious website URL detection method based on abnormal features by analyzing the structure and vocabulary characteristics of malicious website URL addresses. Ram B. Basnet et al. [7] extracted sensitive features from URLs, and constructed a malicious website classification recognizer based on number of publicly available features on URL alone , and achieved good detection results. According to URL characteristics, S. Carolin Jeeva [6] proposed a malicious website identification algorithm based on multi-tag rules, and generated new hidden knowledge (rules) that other algorithms could not find. Abdul Basit et al. [3] designed a malicious website detection system based on Artificial Intelligence (AI) techniques. The system mainly uses features extracted from URLs for identification. Ankit Kumar Jain et al. [5] identified malicious websites based on the hyper links of normal websites and malicious websites, and designed a lightweight phishing detection system. Suleiman Y. Yerima et al. [4] used URL characteristics to build a malicious website recognition model based on convolutional neural networks (CNN) algorithm and achieved good recognition results. However, website URL features are relatively easy to imitate and the number is limited, so there is a greater risk of identifying malicious websites only through URL features.

Recognition technology based on web content usually extracts effective features from web content such as titles, keywords, and description information to identify malicious websites. Ahmet Selman Bozkir et al. [2] used website Logo and Google image search function to identify normal and malicious websites. Vaghela S et al. [11] extract features such as titles and keywords to construct a classifier model to realize intelligent detection of malicious websites. G. Kalyani et al [15] proposed a decision tree based selection algorithm, which divides webpages into different types of visual regions, and proposed a webpage similarity evaluation method based on region matching to identify fake websites. Jyothi Mandala et al [16] proposed the idea of Particle Swarm Velocity Aided GWO to identify malicious websites. Different from the traditional page similarity comparison detection technology, this research no longer directly extracts URL feature maps, but compares partial pages to achieve the purpose of improving recognition accuracy. K. J. Patel et al [7] proposed a malicious website identification technology based on data mining. However, the above research focuses on the identification of malicious websites and ignores the evaluation of malicious websites themselves. URL based and HTML based feature technique will help in evaluating malicious websites more clearly, and then guide and improve the construction of malicious website identification models.

3. Dataset Used

This research arbitrarily gets 10,000 URLs of malicious and real sites from PhishTank (https://www.phishtank. com/) and Alexa (http:/www.alexa.com/). The assessment information of every website is gathered from Moz (https:/moz.com/), Majestic (https:/zh.majestic.com/), and other 4 notable websites. We have utilized 15 features (arranged into two groups: URL and HTML) for classifying web pages.

4. Proposed Methodology

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

This section discusses framework for proposed methodology, Features and evaluation parameters used in this work for detecting phishing websites. Further different classification methods used are also discussed.

4.1. Framework for malicious website evaluation and identification

In order to effectively evaluate and identify malicious websites, this paper proposes a machine learning based identification model, as shown in Figure 1. The model is generally divided into three parts: Data collection and processing, Feature Extraction and Classification.

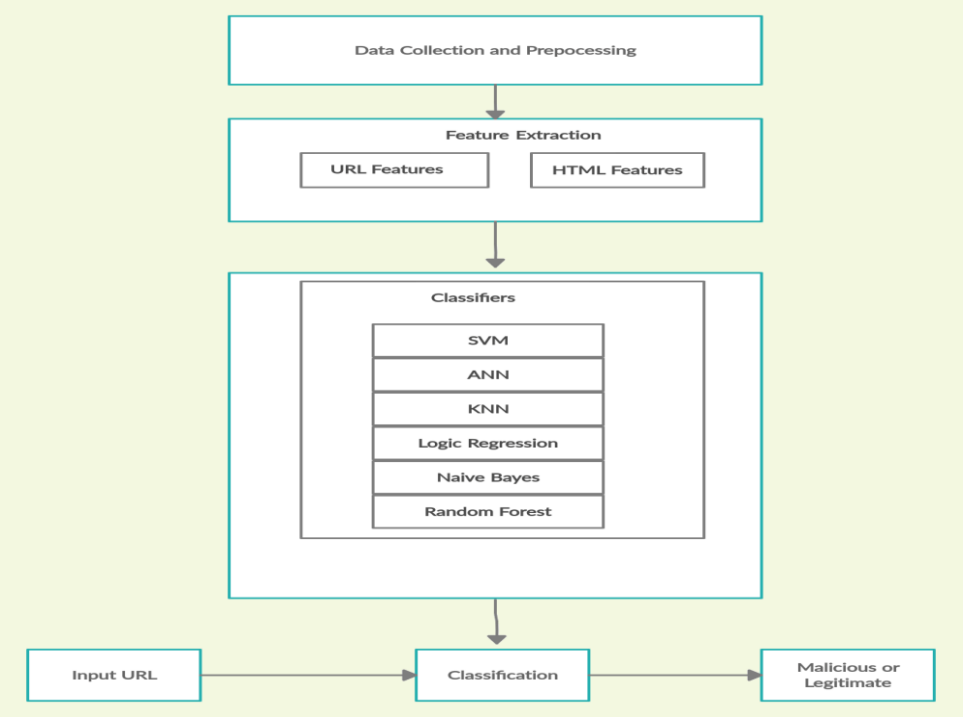

Figure 1: Model Framework

- **Data collection and Pre-processing:**

To build a classification model of malicious URLs, relevant data about malicious URLs and benign URLs are needed. This experiment conducts data collection through multiple channels. For malicious URLs, this work used authenticated malicious URLs from the well-known anti-phishing website PhishTank. For benign URLs, this work obtained URLs from the Alexa website. The URLs obtained from the above channels are consolidated to obtain the final data set used in the proposed model, which contains both malicious and benign.

- **Feature Extraction:**

URLs often have certain commonalities. Based on these commonalities, relevant features can be extracted, and then used for machine learning training. Here we take both URL and HTML feature as an example to illustrate the method of data analysis and feature selection.

For this work 15 features are used (categorized into two groups: URL features, HTML features) for classification of web pages. These features are given in Table 1. URL-based features are extracted after detailed analysis of text of URLs. Further these features can be classified into structural and statistical. The former are concerned with attributes like protocol, domain, subdomains, path, port, and top level domain and the later are concerned with the distribution of URL base elements, specific words, and characters in the text of URLs. Specifically URL based features are the number of dots, subdomains and length of words. HTML-based features are concerned with the number, status, and nature of hyperlinks (i.e., internal/external) used in HTML tags.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

Table 1: Evaluation Features

URL Based Features	HTML Based Features
Embedded Domains	Action Field Values
Age of Domain	Meaningless Tags Present in the URL
IP Address	Hyperlinks
Using "@"	Copied Cascading Style Sheet (CSS)
Using "//"	Redirecting a Webpage
Using "-"	IFrame Redirection
Using "HTTPS" Token in the Domain Part	
Number of Dots	
Position of Top Level Domain (TLD)	

- **Classification:** In recent years, the good performance of Machine Learning algorithms on large-scale data sets has made it the most popular method for detecting malicious websites. The basic idea of proposed learning algorithms is to train several classifiers first, and then use these classifiers so as to achieve the effect of improving the prediction accuracy. In this work, several classifiers algorithms were selected for training the proposed model, which includes Support Vector Machine (SVM), Logistic Regression (LR), Artificial Neural Network (ANN), K-Nearest Neighbour (KNN), Naive Bayes and Random Forest (RF).
- **Support Vector Machine**: It is a discriminative classifier defined by a classification hyper plane. For the current work the following hyper parameters are used: regularization C is considered to be 1000 and gamma value is 150 and kernel is RBF. Under these hyper parameters it is observed high accuracy.
- **Logistic Regression:** Logistic Regression is a machine learning method used to solve binary classification (0 or 1) problems, which is used to estimate the possibility of something. In the current work the cost function used is L2 norm, the regularization strength C is considered to be 10, solver is "lbfgs" and maximum iterations are 100. Under these hyper parameters it is achieved high accuracy.
- **Artificial Neural Networks:** The ANN algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the ANN algorithm is to reduce this error, until the ANN learns the training data. For this work no of input nodes used are 15, no of hidden layers are 2 of which each hidden layer is having 10 nodes for which activation function is relu and for output layers this work made use of sigmoid function.
- **K-Nearest Neighbour:** It makes use of a method of measuring the distance between different feature values for classification. The working principle of k-nearest neighbour algorithm (KNN) consists of the following i)Sample dataset, ii) label and iii) Classifier. For the current work higher accuracy is achieved when no of neighbour nodes (K) is equal to 5.
- **Naive Bayes:** The Naive Bayes classifier depends on Bayes' hypothesis and depends on the assumption that features are free of one another (accepting that there is a component in the

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

class that isn't identified with some other features). Regardless of whether these features are reliant or subject to the presence of different features, the Naive Bayes Classifier believes these features to be independent.

- **Random Forest:** This classifier is based on bootstrap aggregation. The random forest adds extra randomness to the model while the decision tree grows. For this work no of decision trees used are 100 and the quality of split parameter used is "entropy".

4.2 Evaluation Procedure

This work consists of balanced dataset consisting of 5000 malicious URLs from the well-known anti-phishing website PhishTank and 5000 Benign URLs from Alexa website. The dataset is divided in a proportion of 80:20 training and testing respectively. The machine used for implementation for this work is CPU in colab environment. The packages used are Tensorflow, Keras, Numpy and Matplotlib.

Evaluation Procedure consists of 6 basic steps as shown in Algorithm.

---

1. **Data Collection and Pre-processing:** Data is collected from various sources and the same is pre-processed
2. **Feature Extraction:** Feature Extraction is based on both URL and HTML features
3. **Model:** Mathematical Modelling of proposed methods
4. **Loss:** Evaluation of Loss functions of the respective proposed models.
5. **Train the model:** Proposed model will be trained using gradient descent function.
6. **Evaluate the model:** The proposed model will be evaluated against Precision, Recall and Accuracy.

---

5. Results and Discussion

This section provides proposed model classification results based on URL and HTML features for different classifiers.

5.1. Evaluation Parameters

The assessment boundaries utilized for contrasting different classifiers are False Positive Rate (FPR), False Negative Rate (FNR), Precision, F-Measure, and Accuracy alongside following extra parameters.

- **TP (True Positive):** In this both prediction and actual is positive. In other words, which class originally belonged to, but predicted to be that class, is called true positive.
- **FP (False Positive):** In this prediction is positive and the actual is negative. That is to say, one can predict that something is a certain category, but that something is actually not, or false positive.
- **FN (False Negative):** Here Prediction is negative and the actual is positive. That is, the forecast is not, but the actual is.
- **TN (True Negative):** Here both prediction and actual is negative.
- **FPR (False Positive Rate):** It is the rate of erroneously distinguished real webpage's.

$$\text{False Positive Rate} = \frac{False\ Positive}{False\ Positive + True\ Positive}$$

- **FNR (False Negative Rate):** It is the rate of incorrectly identified phishing webpage's.

$$\text{False Negative Rate} = \frac{False\ Negative}{True\ Positive + False\ Negative}$$

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

- **Precision:** It gauges the precision of a model. It is the likelihood for a genuine outcome to be classified effectively.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Recall:** It is the proportion of our model accurately distinguishing True Positives**.**

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- **F-Measure:** It is the symphonious mean of recall and precision. It lies somewhere in the range of 0 and 1, and gives a straightforward method to analyze classifiers.

$$\text{F-Measure} = \frac{2 * True\ Positive}{2 * True\ Positive + False\ Negative + False\ Positive}$$

- **Accuracy (%):** It is the level of accurately recognized website pages (both phish and real).

$$\text{Accuracy (\%)} = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Negative + True\ Negative + False\ Positive} * 100$$

- **Confusion Matrix:** In the field of AI, Confusion Matrix, otherwise called the chance of mistake. It is generally used to evaluate the performance of the given model. Each segment speaks to the anticipated worth, and each line speaks to the real class.

Table 2: Confusion Matrix

	Classified As		
Actual	**Class**	**Phishing**	**Legitimate**
	Phishing URLs	True Positive	False Negative
	Legitimate URLs	False Positive	True Negative

- **Receiver Operating Characteristic Curve (ROC):** It is the plot between the TPR (y-pivot) and FPR (x-hub). Since our model groups if the given site is genuine dependent on URL and HTML features, the probabilities created for each class, we can choose the limit of the probabilities too.
- **Area Under ROC Curve (AUS):** AUC is the area under the ROC curve, a performance index that measures the pros and cons of a learner.
- The ROC curve can easily detect the influence of any threshold on the generalization performance of the learner.
- It helps to choose the best threshold. The closer the ROC bend is to the upper left corner, the higher is the recall. The point on the ROC bend nearest to the upper left corner is the best edge with the least order mistake, and the overall number of positives and false negatives is the littlest.
- The performance of different learners can be compared. Draw the ROC curve of each learner into the same coordinate to visually identify the pros and cons. The ROC curve near the upper left corner represents the learner with the highest accuracy.

5.2. Classifiers used

For the present work, it used six classifiers, i.e., SVM, LR, ANN, KNN, Naive Bayes and RF as machine learning mechanism. Further detailed comparisons of performance of all these classifiers are given common indicators for evaluating the quality of a classification model are Precision, Recall, Accuracy and F1-score. Because malicious URLs often have serious threats, in the process of comparing the quality of each model, the recall rate is prioritized, and the rest of the indicators are

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

used as references. In addition, in order to compare the quality of each classification model more intuitively, an attempt is made to draw the Receiver Operating Characteristic Curve (ROC) of each classification model.

5.3. Results Analysis

- **Support Vector Machine:**

Proposed model is trained with SVM classifier. After training the proposed model with SVM, results obtained are shown in Figure 2.

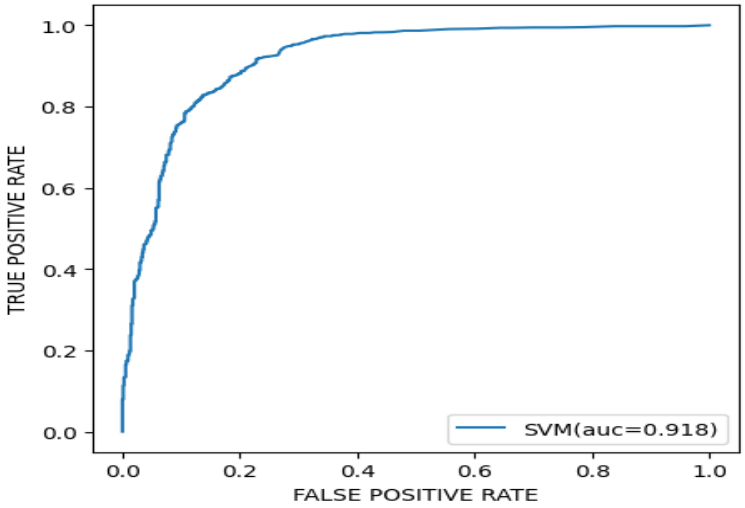

Figure 2: Evaluation results for Support Vector Machine

- **AUC Interpretation:** The model gave value of 0.918 as the AUC which is a pretty good score. In simplest terms, this means that the proposed model will be able to identify malicious websites with an accuracy of 91.8. When the threshold value is 0.23, true positive is almost near 1.
- **Logistic Regression:**

Results obtained after training the given model with Logistic Regression classifier is shown in Figure 3.

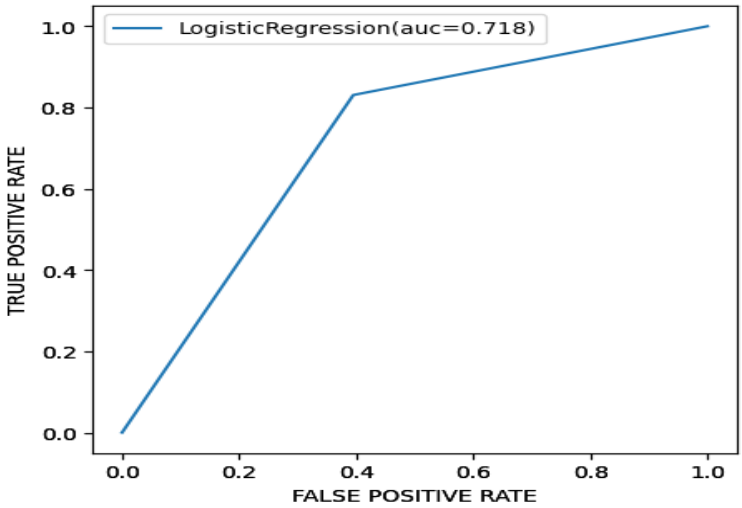

Figure 3: Evaluation results for Logistic Regression

- **AUC Interpretation**: The proposed model gave value of 0.718 as the AUC which is average score. In other words, this means that the proposed model will be able to identify malicious websites with an accuracy of 71.8.
- **Artificial Neural Network:**

Evaluation results after training the proposed model with Artificial Neural Network (ANN)

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

are shown in Figure 4 and the same is represented with ROC curve.

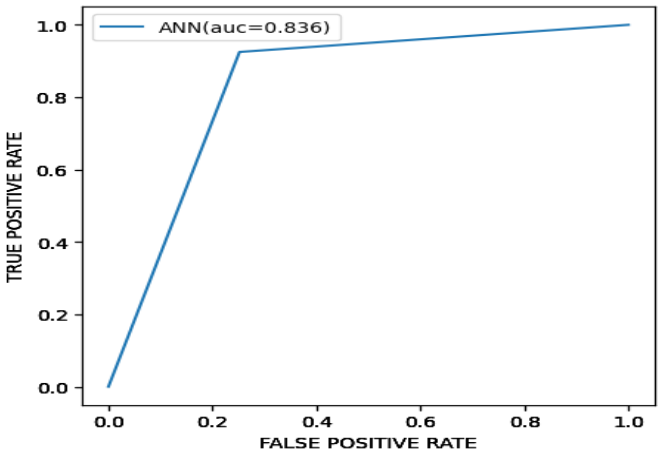

Figure 4: Evaluation results for Artificial Neural Network

- **AUC Interpretation:** The AUC score for the ANN classifier is 0.836 which is a decent value. This indicates that if we use this classifier for training, the machine learning algorithms prediction accuracy rate of identification of malicious websites is almost 83.6%.
- **K-Nearest Neighbour:**

After training the proposed model with KNN classifier, results are evaluated and are shown as ROC curve in Figure 5.

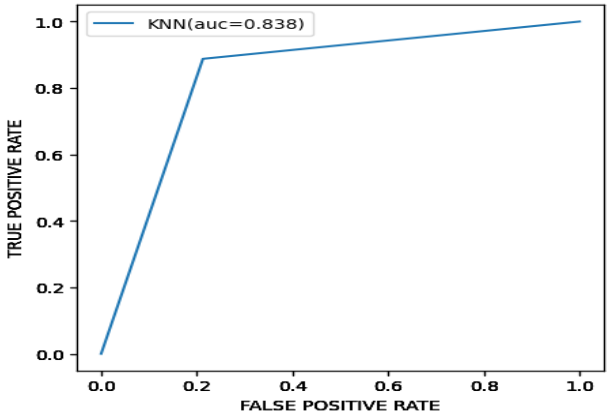

Figure 5: Evaluation results for K-Nearest Neighbour

- **AUC Interpretation:** The AUC score for the KNN classifier is 0.838 which is a good value. This indicates that if we use this classifier for training, the machine learning algorithms prediction accuracy rate of identification of malicious websites is almost 83.8%.

- **Naive Bayes:**

Proposed model is trained with Naive Bayes classification algorithm. After training the proposed, results obtained are shown in Figure 6 as ROC curve.

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

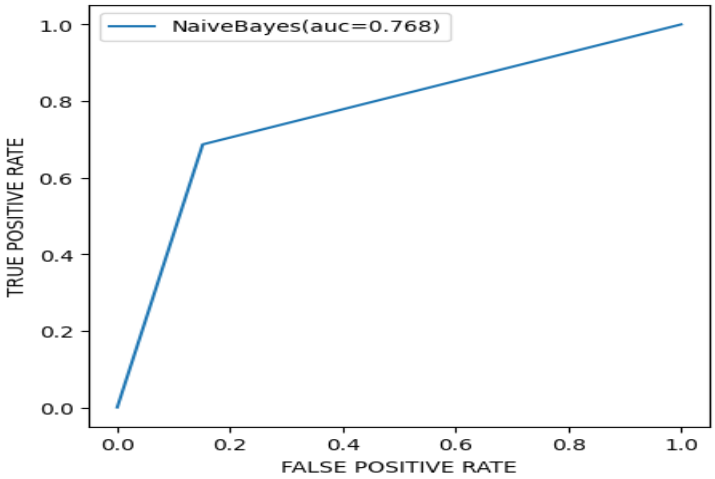

Figure 6: Evaluation results for Naive Bayes

- **AUC Interpretation:** Proposed model gave a value of 0.768 as AUC which is average score. In simplest terms, this means that the proposed model will be able to identify malicious websites with an accuracy of 76.8%. The prediction rate of this classifier is average.
- **Random Forest:**

Results obtained after training the given model with Random Forest classifier are given Figure 7 as the ROC curve.

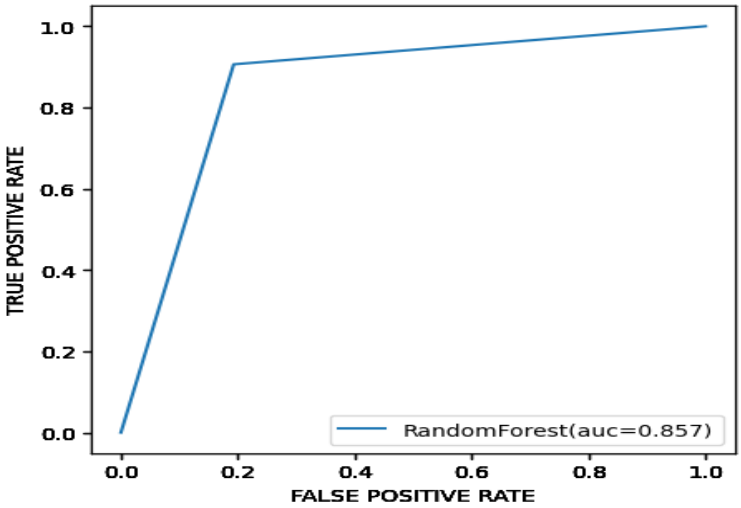

Figure 7: Evaluation results for Random Forest

- **AUC Interpretation:** Proposed model gave a value of 0.857 as AUC which is very good score. In simplest terms, this means that the proposed model will be able to identify malicious websites with an accuracy of 85.7\%. The prediction rate of this classifier is good.

Each of the six classifiers are utilized to characterize the phishing sites by considering URL and HTML features. Table 3 shows the weighted normal estimations of FPR, FNR, Precision, Recall, F-Measure, and Accuracy.

Table 3: Evaluation results for different classifiers

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

CLASSIFIER	FPR	FNR	Precision	Recall	F-Measure	Accuracy (%)
Support Vector Machine	0.855	0.072	0.2071	0.792	0.8227	91.66
Logistic Regression	0.169	0.395	0.67	0.60	0.64	74.9
Artificial Neural Network	0.079	0.237	0.847	0.76	0.802	86
K-Nearest Neighbors	0.112	0.212	0.81	0.79	0.8	85
Naive Bayes	0.313	0.151	0.62	0.85	0.72	75
Random Forest	0.093	0.192	0.84	0.81	0.82	87

The proposed model evaluated the performance of each classification method based on the feature set defined. Figure 2 gives the evaluation results of SVM classifier which is plotted by considering accuracy. The plot shows that the most elevated exactness for this class of features is accomplished by SVM (i.e., 91.8%). Among six classifiers SVM turns out to be the best classifier with classification accuracy 91.8% and with minimum FPR and FNR.

6. Conclusion

In order to effectively evaluate and identify malicious websites, this paper proposed a machine learning based model which is based on URL and HTML features extraction. The present work made use of machine learning algorithm to build a malicious website recognition model and an attempt is made to compare it with other common classification mechanisms. The proposed model improves the evaluation efficiency of malicious websites. The experimental results show that the website feature extraction based on URL and HTML have good interpretability, which is convenient for effective evaluation of malicious websites. Further, the experimental results show that there is a notable improvement in classification accuracy utilizing proposed model and SVM ends up being the best classifier offering the accuracy of 91.8% with FPR and FNR as 0.90 and 0.82 respectively. In future work to design a system which can also detect non-HTML websites with high accuracy.

References

[1] Mohith Gowda HR , Adithya MV, Gunesh Prasad S and Vinay S, “Development of anti-phishing browser based on random forest and rule of extraction framework”, HR et al. Cybersecurity,2020

[2] Ahmet Selman Bozkir, Murat Aydos, “LogoSENSE: A companion HOG based logo detection scheme for phishing web page and E-mail brand recognition”, Computers & Security, ELSEVIER, 2020

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC

[3] Abdul Basit, Maham Zafar, Xuan Liu, Abdul Rehman Javed, Zunera Jalil, Kashif Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques", Telecommunication Systems, Springer, 2020

[4] Suleiman Y. Yerima and Mohammed K. Alzaylaee, "High Accuracy Phishing Detection Based on Convolutional Neural Networks", International Conference on Computer Applications & Information Security (ICCAIS 2020), 19-21 March, 2020.

[5] Ankit Kumar Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information", Journal of Ambient Intelligence and Humanized Computing, 2018

[6] Carolin Jeeva and Elijah Blessing Rajsingh, "Intelligent phishing url detection using association rule mining", Human-centric Computing and Information Sciences, 2016.

[7] Ram B. Basnet, Andrew H. Sung and Quingzhong Liu, "LEARNING TO DETECT PHISHING URLs", IJRET: International Journal of Research in Engineering and Technology, Jun-2014

[8] K. J. Patel and K.J. Sarvakar, "Web Page Classification Using Data Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, No.7, July 2013

[9] J . Heaton, " Artificial Intelligence for Humans: Deep learning and neural networks", Vol 3, Heaton Research, Incorporated, 323 pages, 2015

[10] Leeladevi B., Sankar A.: Feature selection for web page classification using swarm optimization. Int. J. Comput. Control Quantum Inf. Eng. 9(1), 340–346, 2015

[11] Vaghela S., Chaudhary M.B., Chauhan D.: Web page classification using term frequency. Int. J. Technol. Res. Eng. 1(9), 949–954, 2014

[12] Kaur P., Kaur R.: An optimized approach for feature selection using membrane computing to classify web pages. Int. J. Curr. Eng. Technol. 4(5), 3579–3584, 2014

[13] Kaur P., Kaur R.: An optimized approach for feature selection using membrane computing to classify web pages. Int. J. Curr. Eng. Technol. 4(5), 3579–3584, 2014

[14] Liu J., Sun H., Ding Z.: An efficient webpage classification algorithm based on LSH. Intell. Comput. Big Data Era Commun. Comput. Inf. Sci. 503, 250–257, 2015

[15] G. Kalyani, Dr. M.V.P.Chandra Sekhar Rao "Decision Tree Based Data Reconstruction for Privacy Preserving Classification Rule Mining" Informatica 41 289-304. 2017

[16] Jyothi Mandala, Dr. M. V. P. Chandra Sekhara Rao, "PSV-GWO:Particle SwarmVelocity Aided GWO for Privacy Preservation of Data", Journal of Cyber Security and Mobility, Vol. 8 4, 439 - 466, 20 June 2019

ISSN: 2233-7857 IJFGCN
Copyright ⓒ2020 SERSC