ISSN 2303-4521
Periodicals of Engineering and Natural Sciences *Case Study*
Vol. 9, No. 4, October 2021, pp.622-631

An adaptive approach for internet phishing detection based on log data

Ahmed J. Obaid [1], Kareem K. Ibrahim [2], Azmi Shawkat Abdulbaqi [3], Salwa Mohammed Nejrs [4]
[1] Faculty of Computer Science and Mathematics, University of Kufa, Iraq
[2] Faculty of Computer Science and Mathematics, University of Kufa, Iraq
[3] College of Computer Science and Information Technology, University of Anbar, Iraq
[4] Directorate of Private University Education, Ministry of Higher Education, Baghdad, Iraq

ABSTRACT

The Internet has become one of the most important daily socials, financial and other activities. the number of customers who use the Internet to conduct their business and purchases is very large. This results in billions of dollars being transferred every day online. Such a large amount of money attracts the attention of cybercriminals to carry out their illegal activities. "Fraud" is one of the most dangerous of these methods, especially phishing, where attackers try to steal user credentials using fraudulent emails, fake websites, or both. The proposed system for this paper includes efficient data extraction from the web file through data collection and preprocessing. and web usage mining procedure to extract features that demonstrate user behavior. and feature-extracting URL analysis to detect website phishing addresses. After that, the features from the above two parts are combined to make the number of features sixty-three. Finally, a classification algorithm (Random Forests) is applied to determine if website addresses are phishing or legitimate. Suggested algorithms performance is determined by using a confusion matrix and a number of metrics that shows the robustness of the proposed system.

Keywords: Fraud, Phishing, Legitimate weblog, Phishing log data

Corresponding Author:

Ahmed J. Obaid
Faculty of Computer Science and Mathematics
University of Kufa, Iraq
Email: ahmedj.aljanaby@uokufa.edu.iq

1. Introduction

The development in the field of communications and information technology (IT) in recent years has led to a very large growth in services provided on the web such as shopping, banking, e-commerce, games, forums, and file sharing [1]. Internet users are exposed to several types of phishing. through the use of fraudulent emails or a fake website, attackers try to obtain sensitive information from users such as user credentials, passwords, etc. [2].A phishing attacker uses social engineering techniques to simulate legitimate websites and lure users to phishing web pages in various ways, etc. [3]. A common method asks to enter the malicious link on the page to reset your sensitive information and this directs the user to a phishing website [4]. Phishing attacks are among the most serious threats to web-based services including financial institutions, e-commerce, and individuals [2][5]. According to a report by the Anti-Phishing Working Group (APWG). In the first quarter of 2021. The number of phishing attacks doubled during 2020. Then it peaked in January 2021 [6]. In general, phishing attack detection techniques fall into two main categories: blacklisting and On the basis of the heuristic. The first technique compares the requested URL with the one in the phishing list. Recent studies have proven the ineffectiveness of the blacklist against the number of sites hosted daily [7][4]. Conversely, other heuristic technology uses machine learning algorithms to extract features from web pages such as features extracted from URLs or web usages such as detecting user behavior. Depending on these features a web page is classified as legitimate or phishing. The second method is considered more effective, fast and reliable, due to its ability to detect a new phishing website [7].

2. Literature review

The researchers describe the advantages and disadvantages of machine learning and why it is important to apply these techniques in order to identify and detect phishing. To get the right anti-phishing tools [8].

2.1. Review related concepts

Phishing is a fake web page created similar to a legitimate page, and most often they take advantage of well-known pages, to increase the user's confidence and access to this page. The aim is to steal the sensitive and personal information of users [9]. Phishing attacks are divided into two groups:

A-Social engineering

Social engineering means an act that influences a person to achieve desired goals. This includes obtaining information from the target to take a particular action.

B- Technical Subterfuge Attacks

These common methods of scams, where fraudsters send some malicious code which is attached either to fraudulent emails or fraudulent websites that are through (XSS-based programming, session hijacking, phishing software) [10].

Security experts and researchers have taken advanced steps to solve the problem of phishing by multiple techniques that can be categorized into (user training, blacklist, and heuristic-based), heuristic-based two common methods, URL parsing, and page contents analysis such as knowing user behavior. URL analysis extracts features from a web page link, analyzes and detects either a phishing web page or a legitimate [12].

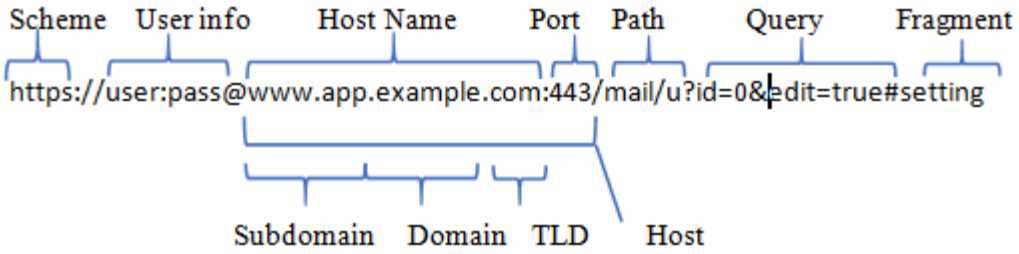

Figure 1. Typical URL syntax [13]

2.2. Related works

Detection of the phishing content received a critical attention in recent years, due the explosion growth of transmission content over the internet through wide range of social media applications. However, detecting legal and authenticated content still faces many challenges because of the complexity of detecting the fraud or phishing contents, in which this content may change over time and cannot be presented in formal manner. In this section, we illustrate most related works focused on detected phishing content. V.Preethi et al. (2016) presented a study of an algorithm which is called PrePhish, which is a machine learning technique to analyze whether URLs are fraudulent or not. The URL features used in machine learning are based on an actual data set. With a range value and limit value set for each feature. Three of the basic classifiers, Random Forest, Naive Bayes, and SVM are used to increase security. The results had a high predictive level with an accuracy of 97.83% and an error of 1.82% [14]. Pratik Patel et al. (2016) presented a study focusing on tackling phishing by clarifying phishing methods and the methods used to detect phishing. In addition to phishing prevention methods, it is also provided an effective model for detecting and preventing malicious attacks [15]. Nandhini.S et al. (2017) conducted a study aimed at identifying the important and effective features in the performance of classification for the detection of fraud sites. The results, after applying a number of algorithms to these features, were that the random forest algorithm gives the highest percentage of correctly classified cases[13]. Alejandro Correa Bansen et al. (2017) presented a study that used URLs as input for machine learning through a method based on recurrent neural networks (Long/Short Term Memory Neural Network (LSTM)) and compared their proposed method with (Random Forest Classifier RF), and their results were the best accuracy (98.7% Vs 93.5% to RF) and takes of memory (581 KB Vs 288.7 MB to RF) [16]. R. Kumar et al. (2018) proposed a multi-layer model, where they use 4 algorithms as a filter to identify malicious URLs. In the last two layers, the Naive Bayesian classifier and the CART Decision Tree classifier are used respectively. This component model achieved a high level of accuracy [17]. MahaLakshmi et al. (2018) discussed the types of phishing, what are their harms, and explain that social engineering phishing is an act that affects a person in several ways such as malicious email or malicious websites to obtain sensitive

information such as passwords, credit card details, and usernames. Phishing is also countered by countermeasures from anti-phishing techniques [8]. Alyssa Anne Ubing et al. (2019) used a method in which the feature selection algorithm was combined with the collective learning methodology. Where by the results of the current phishing identification has a good accuracy rate of between 70% and 92.52% the accuracy rate may reach experimental results in the proposed system to 95%, which is better than Many current techniques in detecting phishing sites [18]. Shisrut Rawat et al. (2019) used classic machine learning techniques with deep neural networks and unsupervised learning techniques. They also used a comparative analysis of some models of deep learning versus machine learning. The results had obtained an accuracy of 93.82% with a reduction of training time by 98.8% [19]. Eint Sandi Aung et al. 2019 introduced a systematic survey of phishing techniques based on URL features. Focuses on deception detection by discussing commonly used algorithms and features. They proposed a model that classifies a fraud attack, based on feature extraction criteria. They also emphasized that it is necessary for the user to check the URL before entering any website[20]. Hesham Abusaimeh et al. (2021) suggested using three combined algorithms (random forest, decision tree, and support vector machine) to detect phishing sites in addition to using these models separately for comparison with the proposed model. The results that emerged was that the three models combined had a higher accuracy of detecting phishing sites than using them alone, where the percentage was (98.52%) [21]. P. Kalaharsha et al. 92021) discussed different types of phishing attacks and phishing website detection techniques. Technologies include list-based, visual measurement, machine learning, and heuristics. and different performance methods for data sets. Knowing this information is very important to help end-users in combating phishing sites [22].

3. Research methodology

In this section the model used to detect phishing sites is described as well as the data set, algorithms, and metrics used in the evaluation of the model.

3.1. Why heuristic based phishing detection

This technique depends on the characteristics of phishing sites or the behavior of the attackers; Although these techniques have high accuracy of results, it is not always possible to guarantee the presence or selection of important characteristics in phishing detection. If the method (technique, features) chosen is effective in identifying phishing, phishing attacks can be detected at zero hour. This technique is against blacklist technology. They are very quick to respond when compared to visual similarity technical because it does not require any initial legitimate image database and does not include no comparison of images with image database. Thus, the calculation cost is lower as compared to visual similarity assessment technique. It is useful from blacklist or whitelist approach in a phishing attack is detected.

3.2. Phishing data set

Machine learning technology was used to develop the proposed model for phishing attack detection by selecting data for training and for validation. To develop a new phishing attack detection model, a phishing training dataset was collected from the *Aalto University, Finland (AU)* repository dataset of approximately 96012 entries Preprocessing and cleaning of outlier data, and 102 records were found an outlier and used data to train and test the model. A Random Forest algorithm was chosen for classification and is one of the most popular algorithms in identifying and discovering websites that are phishing or legitimate.

Table 1. Dataset information

Source	Type	Size of Dataset samples	No. Features considered	No. Instances in Dataset
Aalto University (Finland)	Dataset	11.6 MB	63 F	96012

3.3. Adaptive random forest algorithm

Random forest is a supervised learning algorithm which is used for classification and regression tasks. The "forest" it builds, or A classifier is a collection of multiple decision trees. Randomness is added to the model to generate decision trees. It defines a random subset of features to split nodes. this that measures a feature's importance by looking at how much the tree nodes that use that feature reduce impurity across all trees in the

forest. Based on the prediction of each decision tree, each tree performs a unit vote for the most popular category in the input data. It computes this score automatically for each feature after training.

3.4. Design flowchart of the phishing attack detection model

The following figure describes the proposed system design for detecting phishing attacks. Which starts from entering the weblog and conducting analyzes and even detecting phishing sites.

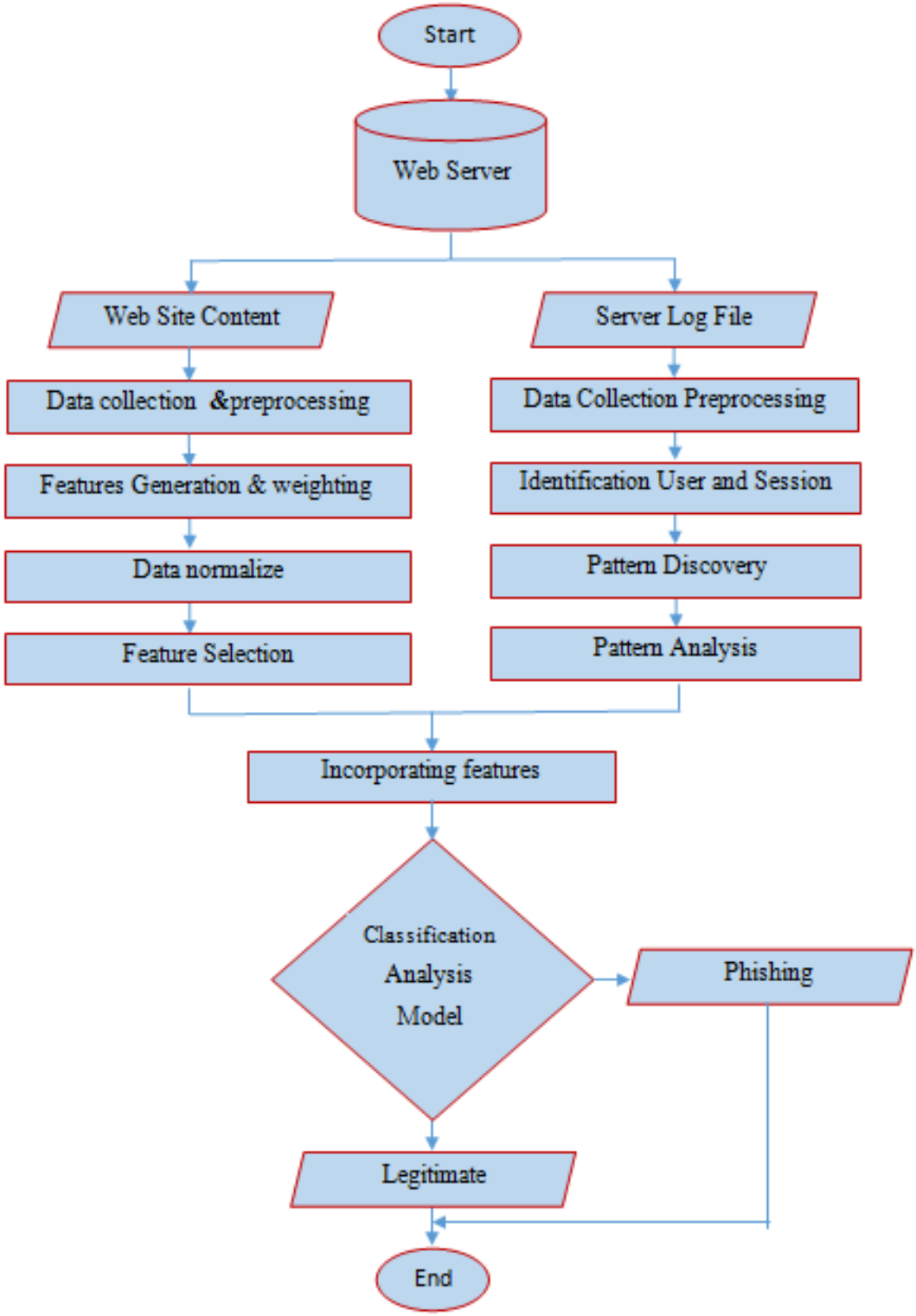

Figure 1. Proposed system flowchart

Algorithm (3.4) Adaptive Random Forest Algorithm (ARFA)

Input data: *Dataset (DS), Training Data (TD, Iteration (I, J), No. of Classes (L, P)*
Output Parameters: *Root Node (RN), Total Features (F), Selected Features (sF), Tree Node (Tn),*
STEP1: *Generate sF (Selected Features = M) From (F)*
For I= 1 To M do
From TD, Select sf Where sF << M.
Find RN that has the best Split Point, Create TN that Not containing in RN
Call Build Tree (Tn)
End for
Step 2: *Build Tree (Tn) by split nodes into Couple Nodes*
If Tn consists of instances of only one class
Return
Else
Randomly select x% of the possible split features in sF
select the feature F with the Highest information gain to split on
Create f child nodes of N, N1, Nf where F has f possible values (F1, Ff)
For J=1 To f do
Set the consists of Tn to RN where RN is all instances in Tn that match RN
Call Build Tree (Tn)
End for
End if
Step 3: End All Iterations.
Step 4: Classified All Instances in TD

After collecting data from the webserver, perform preprocessing, web usage mining, and analysis of URLs to extract features. 63 influential features were obtained in the process of phishing detection.
[(Domain, Port, Host Type, Query, Having IP, Having Subdomain, URL Length, URL Length Threshold, URL Depth, Redirections, SSL Type, Shortening Services, Prefix & Suffix, URL Have Sign (. , -, _ ,/ ,? , = ,& ,! , , , , ~ ,+ , * ,# ,$,% , @) , Domain Have Sign (. , -, _ ,/ ,? , = ,& ,! , , ~ ,+ , * ,# ,$,% , @),Path (. , -, _ ,/ ,? ,= ,& ,! , , ~ ,+ , * ,# ,$,% , @)].
The proposed system is implemented in the C# programming language in stages. The next stage represents the program interface. It contains parameters implemented on the previously selected data set as shown in the following figure.

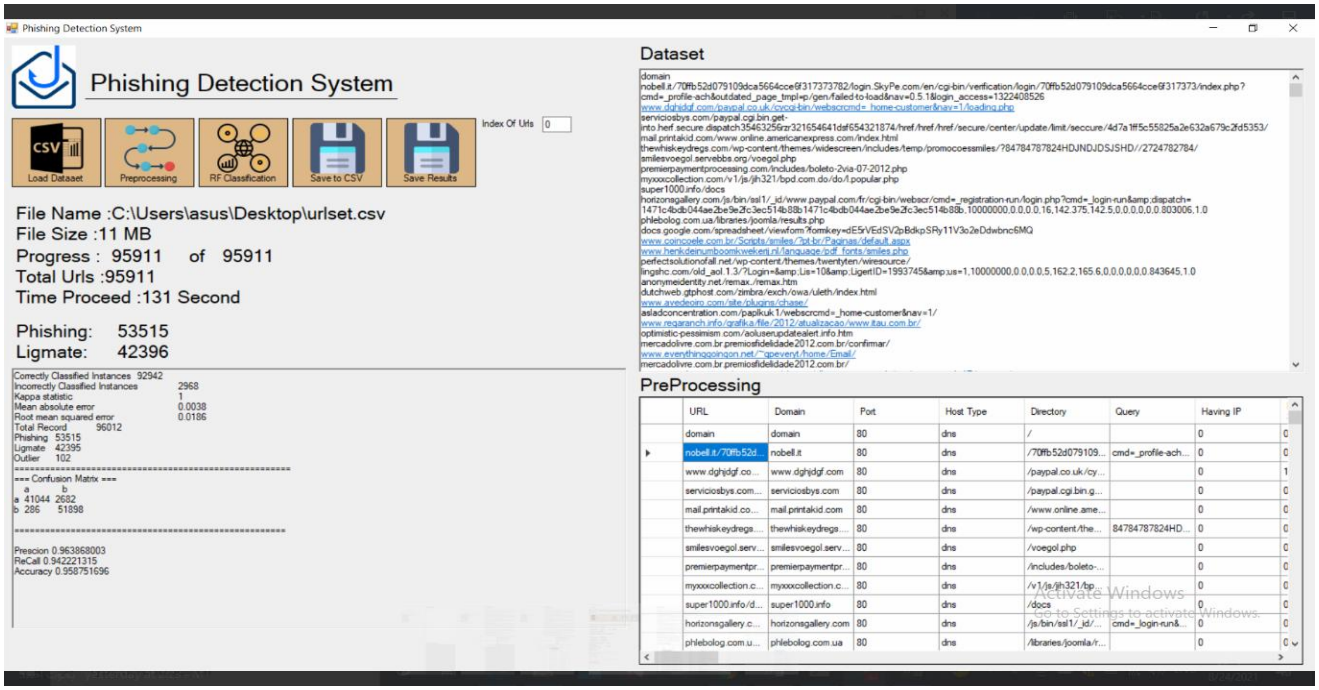

Figure 2. Main screen of proposed system platofrom

Exploratory techniques are effective in detecting fraudulent websites, based on the characteristics of phishing websites and user behavior features. These techniques have a high detection accuracy of fraudulent websites. If the chosen method (technique and features) is effective in identifying phishing, phishing attacks can be detected at zero hour. The proposed model was built with these techniques, and using a data set consisting of 96,012 records, the results shown in the following figure were obtained.

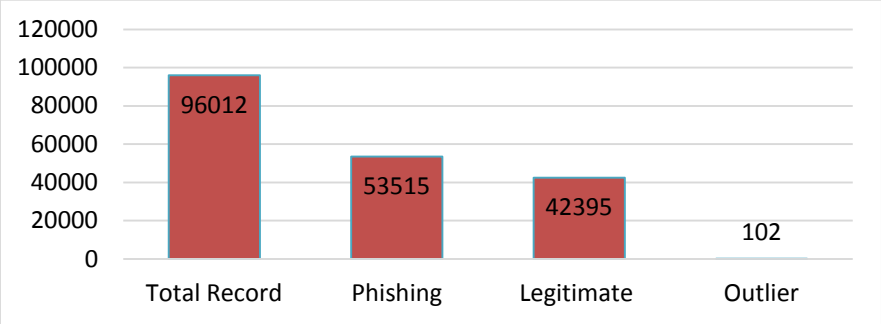

Figure 3. Results of AU dataset using proposed system

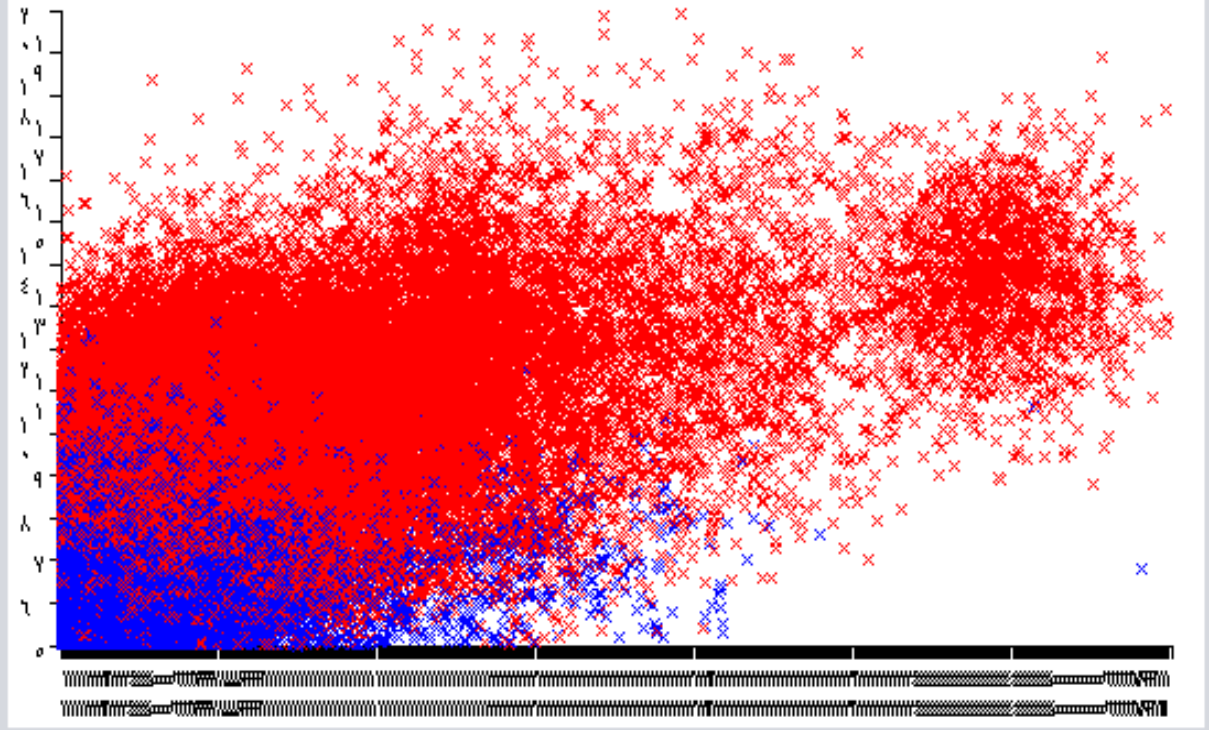

Figure 3. Ploting of AU data point

3.4.1. Time and space complexity

Time complexity is the amount of time it takes a computer to run a given algorithm. As a function of input length, it measures the time it takes to execute each statement of code in an algorithm. Space complexity represents the total amount of memory an algorithm or process uses to run (with input values into the algorithm) to execute and produce the result.

Table 2. Time and space complexity

Detecting Model	Preprocessing time	Execution Time	Memory	Space
Proposed Model	215 Second	167 Second	305 MB	22 MB

4. Performance evaluation

These are automatic algorithms for quality assessment that could analyses data and report their quality without human involvement.

4.1. Confusion matrix

When it comes to classification problems, the confusion matrix is a widely used measure.[22].

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+FP+FN+TN}} \qquad (1)$$

Table 3. Confusion matrix of tested dataset

Classification	Test Result			
Proposed Model	True Positive	False Positive	True Negative	False Negative
	0.54111%	0.00298%	0.42794%	0.02796%

4.2. Precision and recall (P/R)

Precision indicates how well the model predicts positive values. The recall is a useful metric for determining a model's ability to predict positive outcomes. The following are the formulas for measuring precision and recall[23].

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \qquad (2)$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad (3)$$

4.3. F-measure (F)

F-measure, also known as F-value, A-weighted harmonic mean of precision and recall[24].

$$\boldsymbol{F1} = \frac{\boldsymbol{2 * Precision * Recall}}{\boldsymbol{Precision+Recall}} \qquad (4)$$

4.4. Kappa statistic (KS)

The Kappa statistic is used to measure interference Among categorical items. [21].

$$(ks) = \boldsymbol{1} - \frac{\boldsymbol{1-Po}}{\boldsymbol{1-Pe}} \qquad (5)$$

Po is the relative observed agreement among raters, Pe is the hypothetical probability of chance agreement.

4.5. Mean absolute error (MAE)

The mean absolute error is quantity is used to measure Expectations in the end results. [21].

$$Mean\ Absolute\ Error = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \qquad (6)$$

where fi is the prediction value , yi is the true value.

4.6. Root mean square error (RMSE)

The root mean square error (RMSE) is a measure of the file User for differences between the Number of sample values estimated or predicted by a model and observed values.[21].

$$\boldsymbol{RMSE} = \sqrt{(\sum_{i=1}^{M}\sum_{j=1}^{N}(\boldsymbol{I(i,j)} - \boldsymbol{C(i,j)})^2 / \boldsymbol{M * N})} \qquad (7)$$

After applying the previous measures to the specified data set, the following results were obtained.

4.7. Experimental parameters

To Comparing the proposed model with fraud detection methods, similar inputs were tested on each one of the four detectors which are SVM, Naïve Bays, KNN, and Decision Tree in addition to the proposed system individually. The results of the other algorithms differed, and each recorded a lower accuracy than the proposed system. As shown in the following table.

Table 4. Comparative analysis between existing and proposed system

Classifier	Accuracy	P	R	KS	MAE	RMSE	F
SVM	95.36%	0.93324	0.91132	0.9058	0.047%	0.231	0.92214
Naïve Bays	92.91%	0.9034	0.86421	0.81	10.51%	0.2314	0.883370
DT	95.88%	0.9223	0.9134	0.9162	5.67%	0.4356	0.917828
KNN	95.69%	0.9532	0.90342	0.8943	6.56%	0.3251	0.927642
ARFA	96.91%	0.9638	0.94222	1.0	0.0038%	0.0186	0.952921

4.8. Correctly and incorrectly classified instances

As we can see in Figure 4.1 shows which cases are correctly classified and the cases are incorrectly classified. This indicates the performance of the proposed model and its high ability to detect malicious websites.

Table 5. Comparison of correctly and incorrectly classified instances

Detecting Method	SVM	DT	Naïve Bays	KNN	ARFA
Correctly Classified Instances	95.36%	95.88%	92.91%	95.69%	96.91%
Difference Compared to Proposed Model	3.17%	2.65%	7.00%	3.82%	3.10%

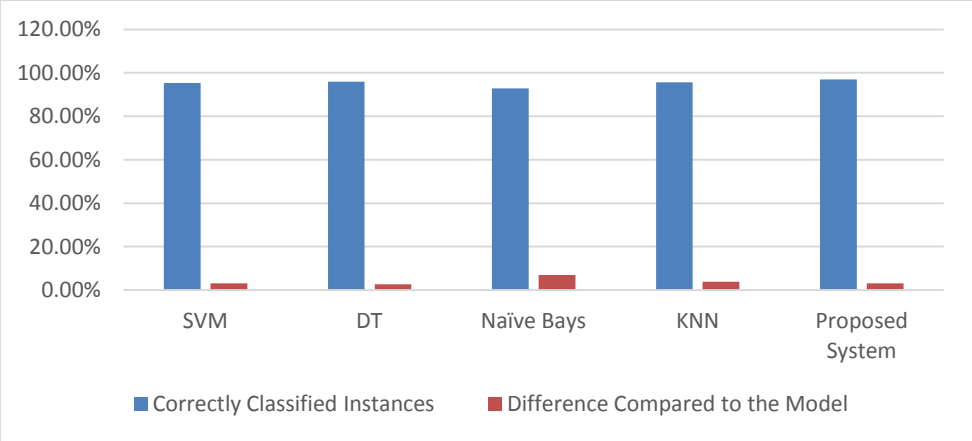

Figure 5. Analysis of correctly / incoreectly classified instances of proposed system

5. Conclusion and recommendations

The phishing attack is one of the most sophisticated web attacks and it is considered a serious threat to website users, this paper proposed a model based on the Random Forest algorithm for the purpose of classifying and detecting phishing sites based on 63 important features in identifying phishing by URL, domain, or path characteristics. The performance of the classification algorithm with feature selection based on classifier attributes evaluator was evaluated using a phishing dataset consisting of the combination of URL, Domain, and Path-based features. The result of the evaluation shows that the proposed model has a high accuracy of 96.91% and low error rates of 0.03% compare to other existing machine learning-based models. For future

work, it is hoped that more feature selection most relevant features and further improves the performance of the phishing attack detection model.

Reference

[1] G. Varshney, M. Misra, and P. K. Atrey, "A survey and classification of web phishing detection schemes," *Secur. Commun. Networks*, vol. 9, no. 18, pp. 6266–6284, Dec. 2016, doi: 10.1002/SEC.1674.

[2] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu, "A stacking model using URL and HTML features for phishing webpage detection," *Futur. Gener. Comput. Syst.*, vol. 94, pp. 27–39, May 2019, doi: 10.1016/J.FUTURE.2018.11.004.

[3] R. S. Rao and A. R. Pais, "Jail-Phish: An improved search engine based phishing detection system," *Comput. Secur.*, vol. 83, pp. 246–267, Jun. 2019, doi: 10.1016/J.COSE.2019.02.011.

[4] R. Verma and A. Das, "What's in a URL: Fast feature extraction and malicious URL detection," *IWSPA 2017 - Proc. 3rd ACM Int. Work. Secur. Priv. Anal. co-located with CODASPY 2017*, pp. 55–63, Mar. 2017, doi: 10.1145/3041008.3041016.

[5] B. E. and T. K., "Phishing URL Detection: A Machine Learning and Web Mining-based Approach," *Int. J. Comput. Appl.*, vol. 123, no. 13, pp. 46–50, Aug. 2015, doi: 10.5120/IJCA2015905665.

[6] "APWG | Phishing Activity Trends Reports." https://apwg.org/trendsreports/ (accessed Aug. 11, 2021).

[7] H. Liu, X. Pan, and Z. Qu, "Learning based Malicious Web Sites Detection using Suspicious URLs," *34th Int. Conf. Softw. Eng.*, pp. 3–5, 2016, [Online]. Available: http://www.dtic.mil/cgi-.

[8] A. Mahalakshmi, N. S. Goud, and G. V. Murthy, "A survey on phishing and it's detection techniques based on support vector method (Svm) and software defined networking(sdn)," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 2, pp. 498–503, 2018.

[9] M. Aydin and N. Baykal, "Feature extraction and classification phishing websites based on URL," *2015 IEEE Conf. Commun. NetworkSecurity, CNS 2015*, pp. 769–770, Dec. 2015, doi: 10.1109/CNS.2015.7346927.

[10] M. J. Hamid Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 208–215, 2018, doi: 10.14569/IJACSA.2018.090630.

[11] F. Aburub and S. Alhawari, "A new fast associative classification algorithm for detecting phishing websites," *Appl. Soft Comput. J.*, vol. 48, pp. 729–734, 2016, doi: 10.1016/j.asoc.2016.08.005.

[12] WangRong, ZhuYan, TanJiefan, and ZhouBinbin, "Detection of malicious web pages based on hybrid analysis," *J. Inf. Secur. Appl.*, vol. 35, pp. 68–74, Aug. 2017, doi: 10.1016/J.JISA.2017.05.008.

[13] S. Nandhini and V. Vasanthi, "Extraction of Features and Classification on Phishing Websites using Web Mining Techniques," vol. 5, no. 4, pp. 1215–1225, 2017.

[14] I. V. I. P. A. V Preethi and G. Velmayil, "Automated Phishing Website Detection Using URL Features and Machine Learning Technique."

[15] P. Patil and P. P. R. Devale, "A Literature Survey of Phishing Attack Technique," vol. 5, no. 4, pp. 198–200, 2016, doi: 10.17148/IJARCCE.2016.5450.

[16] A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. Gonz, "Classifying Phishing URLs Using Recurrent Neural Networks," 2017.

[17] R. Kumar, X. Zhang, H. A. Tariq, and R. U. Khan, "Malicious URL Detection Using Multi-Layer Filtering Model," no. November, 2018.

[18] A. A. Ubing, S. Kamilia, B. Jasmi, A. Abdullah, N. Z. Jhanjhi, and M. Supramaniam, "Phishing Website Detection : An Improved Accuracy through Feature Selection and Ensemble Learning," vol. 10, no. 1, pp. 252–257, 2019.

[19] S. Rawat, A. Srinivasan, and R. Vinayakumar, "Intrusion detection systems using classical machine learning techniques versus integrated unsupervised feature learning and deep neural network," p. 9, 2019.

[20] E. Sandi, "A Survey of URL-based Phishing Detection," pp. 1–8, 2019.

[21] H. Abusaimeh and Y. Alshareef, "Detecting the Phishing Website with the Highest Accuracy," vol. 10, no. 2, pp. 947–953, 2021, doi: 10.18421/TEM102.

[22] B. M. P. K. Mehtre, "Detecting Phishing Sites - An Overview," pp. 1–13, 2021.

[23] "Classification: Precision and Recall | Machine Learning Crash Course." https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

(accessed Aug. 01, 2021).
[24] D. J. Hand, P. Christen, and N. Kirielle, "F*: an interpretable transformation of the F-measure," *Mach. Learn. 2021 1103*, vol. 110, no. 3, pp. 451–456, Mar. 2021, doi: 10.1007/S10994-021-05964-1.