Detecting the Phishing Website with the Highest Accuracy

Hesham Abusaimeh [1], Yusra Alshareef [1]

[1] Computer Science Department, Middle East University, Amman, 11831, Jordan

Abstract **– Phishing attacks are increasing and it becomes necessary to use appropriate response methods and to respond effectively to phishing attacks. This paper aims to uncover phishing attack sites by analyzing a three-module set to prevent damage and reconsider the awareness of phishing attacks. Based on the analyzed content, a countermeasure was proposed for each type of phishing attack by using website features. These features will be classified in order to determine the effectiveness of the countermeasure. Finally, the proposed method enhanced the site security as anti-phishing technology. The phishing detection used three classification algorithms, which are the decision tree; the supporting vector machine and the random forest were combined into one system that was proposed in this paper for the purpose of obtaining the highest accuracy in detecting phishing sites. The results of the proposed algorithm showed 98.52% higher accuracy than others.**

Keywords **– Decision tree, Supporting vector machine, Random forest.**

1. Introduction

The Internet has made way in this technological age and has become an important part of our lives, in which the internet provides many methods of rest, including the communication, education, entertainment, shopping and so on.

They found criminals with the advancement of the Internet+ the opportunity to transfer their crimes in a virtual environment. The Internet not only provides amenities, there is also a negative side to using the Internet, examples of this are the spread of multiple types of crimes at the present time that have been conducted through the Internet, due to the lack of disclosure of identity of the users that the Internet provides to them. Therefore, the main focus will be on phishing because it is considered a type of electronic crime [1]. The process of fraud takes place when the prank deceives and deceives targeted people into obtaining sensitive information such as passwords, identification information, and a personal social security number. The fraud takes place when the criminal lures and deceives the target into obtaining sensitive information such as passwords, identification information, and a personal social security number. Phishing attack is carried out in four steps [2]. First, the criminal creates a completely similar website to a legitimate website. Second, the prankster pretends to be a legitimate organization or company and therefore sends a link uniform resource locator (URL) website to the victims he wants to target. Third, the prankster works to persuade the victims to visit the fake website. Fourth, after the persuasion process, the victims will enter the website and enter the required personal information that serves the prankster, and through that information the prankster begins to carry out fraud activities with the victim. To avoid suspicions of victims and users, phishing attacks are not professionally performed [3]. The amount of damages between the years 2013 to 2016 reached a minimum of 2.3 billion dollars, as a result of phishing scams, according to a report from the federal bureau of investigation [4]. A survey was conducted earlier on the topic "Why phishing works" This study showed that the 23% of the participants was determining the legitimacy of the page through their reliance on the content of the web page. In addition, many people cannot differentiate the contents of the page, the lock icon as a favorite icon and the lock icon in the browser [5]. An automated approach should be considered for the purpose of detecting phishing sites in order to address some challenges, including the user's ability to determine address (URL) if it is legitimate or a phishing site

DOI: 10.18421/TEM102-58
https://doi.org/10.18421/TEM102-58

Corresponding author: Hesham Abusaimeh,
Middle East University, Amman, Jordan.
Email: habusaimeh@meu.edu.jo

Received: 22 September 2020.
Revised: 22 April 2021.
Accepted: 29 April 2021.
Published: 27 May 2021.

[6]. This paper presents an algorithm consisting of a three ensemble classification. The main objective from this algorithm is to discover the phishing website with high accuracy.

The rest of the paper is organized as follows. The section 2 presents a review of the literature in phishing. The section 3 explains the components of the proposed methodology. Section 4 presents the performance evaluation and results analysis. Section 5 explains the confusion matrix comparison between models. Finally, section 6 presents the conclusion.

2. Literature Review

Phishing uses e-mail to reach the target, the prankster sends a message by e-mail to users who may represent a commercial activity such as banking or financial institution or a company for that phishing has become very harmful and it has become necessary to detect phishing sites. Phishing in cyberspace stimulates researchers to find solutions to make websites safer [7]. The researchers described the pros and cons of automated education technologies and the extent to which these technologies can be applied in order to detect phishing. To obtain appropriate tools to counter phishing, various types of automated education techniques were verified in order to reach appropriate options. In order to show the actual performance of the models of automated education techniques and to discover the defects and advantages of those models, therefore, they compared a large number of automated education techniques with regard to the various measures in the real data of phishing data. The results showed that the best anti-phishing solution is the coverage approach model; this is because of the good rate of phishing detection and their effective and simple cognitive bases [8].

A. C. Bahnsen et al. [9], suggested the use of Uniform Resource Locator (URL) address in order to predict phishing sites and considered (URL) an introduction to machine learning. Compared new method based on neural networks with a random forest classifier followed by a feature engineering approach. And it turned out that the neural network approach outperforms the random forest method with 5% percentage and provides an accuracy rate of 98.7%. This means that the system is fast-acting, scalable, and does not require content analysis.

I. Qabajeh et al. [10] dealt with a review of anti-phishing models and an analysis of those models in a smart, educational, training and legal way. It also highlighted smart and traditional methods of combating phishing, In addition, the negative and positive aspects and the expected performance of the user, similarities and differences in the curricula were revealed.

N. Abdelhamid et al. [11], proposed a new algorithm, a type Associative Classification (AC) algorithm called Multi-label Classifier (MCAC) was used to detect phishing sites. The results showed that the (MCAC) algorithm had higher susceptibility and accuracy to detect phishing sites than other algorithms. Therefore, the (MCAC) algorithm participated in improving the predictive performance more than the rest of the algorithms, because (MCAC) algorithm generates knowledge of hidden rules that the rest of the algorithms cannot find.

F. Aburub et al. [12] proposed a new type of Associative Classification (AC) algorithm called a Fast Associative Classification Algorithm (FACA). A comparison was made between the proposed algorithm and four other algorithms of the type of Associative Classification (ECAR, MCAR, CMAR and CBA). The comparison was made between algorithms in terms of (F1) results. The results of the comparison showed that FACA had better results than the other algorithms in both F1.

Alyssa Anne Ubing et al. [13] focused on improving the accuracy of detecting phishing sites in addition to evaluating web sites if they were phishing sites or legitimate sites. Therefore, the collective learning methodology was combined with the feature selection algorithm was compared with other classification models such as prediction model, logistic regression, and random forest. The results showed that the proposed algorithm can produce an accuracy rate that may reach 95% which is higher than other algorithms, and this means that it has a promising accuracy rate to detect phishing sites.

3. Proposed Methodology

A meta-algorithm has been proposed for the purpose of improving predictions and reducing variance; meta-algorithm is the combination of a number of automated learning techniques in a single predictive model. In the proposed methodology, three algorithms were used (decision tree, random forest, and support vector machine (SVM)) and that were combined into one system as shown in the Table 1. The purpose of the proposed methodology is to obtain a high accuracy in detecting phishing sites that occur on the websites.

Table 1. The Parameters for the Proposed Methodology

No.	Parameter	Value
1.	No. of attributes	30
2.	Threshold ranking	-1.7977
3.	No. of cross-validation folds	10-fold
4.	The rate of momentum for backpropagation algorithm	0.2
5.	The rate of Learning for backpropagation algorithm	0.3

6.	No. of consecutive increases of error allowed upon validation testing before completion of the training	20
7.	The percentage of validation set size that was used to end the training	0
8.	No. of periods to train	500
9.	Threshold of confidence for pruning	0.25
10.	The value of seed for random number generator	1 -num-slots
11.	Gamma	auto
12.	The min number of cases per leaf	2
13.	The size of each bag	100
14.	No. of bag error	100
15.	Min variance for split	0.001
16.	No. of attributes	0
17.	Min number of instances	1
18.	No. of execution slots	1
19.	Seed for random number generator	1
20.	Set the max number of iterations	-1
21.	The exponent for the polynomial kernel	1
22.	Complexity constant	250007
23.	Sets the epsilon for round-off error.	1.0E-12

4. Performance Evaluation and Result Analyses

According to what he worked on [14], the features can be classified into four categories in order to determine the main characteristics of the phishing site on the Internet. The first category, are the features in the title bar, as the title bar can display a sneaky or suspicious website. This category shows the sub-types related to the title bar, such as (the phisher can use the "@" in the (URL) address, Scammer uses a long URL to hide the suspicious part, redirect using the "//" shortening, and use the IP address in the address bar and a lot of features may appear on the address bar. The second category, which is known as abnormal features or anomalies of multiple types such as (server form handler, URL of anchor, sending information to email, request URL address, abnormal URL address and links in <link>, <script> and <meta> tags. The third category is JavaScript and HTML based website redirects such as iframe redirection, right-click disabling, status bar customization, and popup window. The fourth category, which is one of the features that are based on the domain where phishing sites can be determined according to the website traffic, according to the age of the domain, a Google index, page rank, (DNS) records and other similar characteristics.

Table 2 shows a comparison between the results of the models (decision tree, random forest and support vector machine) individually and the results of the proposed model. The results showed that the proposed model has a higher accuracy than the support vector machine by (3.0996%), higher than the random forest by (1.2%) and a decision tree by (2.584) if the models are used separately. Consequently, the proposed model is highly effective and more reliable in detecting phishing sites compared to models (decision tree, random forest and support vector machine). Figure 1 shows an analysis of a comparison between the results of the proposed model and the current algorithms in terms of the most common factors that have been considered and included by researchers. The commonly used assessment measures can be distinct and cannot be used without identifying the corresponding levels of chance and a clear understanding of biases, as well as determining the basic state of the statistic and from these measures are (F-measure, Accuracy, Rand Accuracy and Recall). When using these measures, the model can demonstrate better performance.

Table 2. Comparison between the Results of the Proposed Model and Other Models

Detecting Method	Random Forest	SVM	Decision Tree	Proposed Model
Correctly Classified Instances	97.2592 %	95.3596 %	95.8752 %	98.5256%
Difference compared to the Model	1.2664%	3.1660%	2.6504%	0%

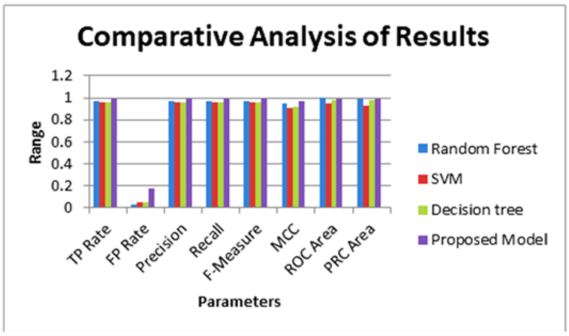

Figure 1. Comparative Analysis of Results

Figure 1 shows an analysis of the results of the comparison between the proposed model, which consists of three models and with other techniques (decision tree, support vector machine and random forest). The performance was compared in terms of several different parameters. The results showed that the proposed model obtained the highest accuracy rate compared to other techniques in the basic situation. The models (decision tree and support vector machine) got the lowest accuracy by (95.35% and 95.87%), while the proposed model obtained the highest accuracy by (98.52%) and the random forest model obtained the second degree with precision (97.25%), the proposed improved model got a higher accuracy than the random forest by (1.27%). The reason for the proposed improved results may be due to the lines that address the proposed titles due to additional alternatives. This paper aims to enhance the classification of phishing sites, and thus the results achieved the goal of this study. In Figure 1 the confusion matrix and the Receiver Operating Characteristics curve (ROC) are displayed for the trained random forest using the set of features provided. The results showed that the random forest has a low rate in terms of False Negatives FN rate and False Positives FP rate, while it has a high rate in terms of True Negatives TN rate and True Positives TP rate. It is clear that their Receiver Operating Characteristics curve (ROC) is worse than ours, with a smaller Area Under Curve (AUC).

A. Correctly and Incorrectly Classified Instances

Figure 2 shows that the proposed model has the highest performance compared to other current methods in the Attribute-Relation File Format ARFF data set. As the highest rate of classification of proposals is due to the proposed model, which reaches (98.52%), while the support vector machine model gives the lowest correctly classified rate, which reaches (95.35%). Classification accuracy (%) is used for paradoxical algorithms that are derived from phishing data. Accuracy indicates the ability of the algorithm to correctly predict the name of a class in case of unknown class designation. Accuracy is used in comparison and evaluation of basic recipes. The equation below shows the method for measuring accuracy [15].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....................} \quad (1)$$

Figure 2 shows the difference between the proposed model and other models in terms of correctly classified cases. The result showed that the proposed model has the highest accuracy rate of (98.52%) which is higher than the decision tree by (2.65%), and higher than the random forest by (1.26%) and by (3.16%) of support vector machine. While the figure 3 show the difference between the proposed model and other models in terms of incorrectly categorized cases. The results showed that the proposed model has the lowest percentage of (1.47%) and is the lowest by (2.65%) of the decision tree and of the random forest by (1.27%) and by (3.16%) of support vector machine.

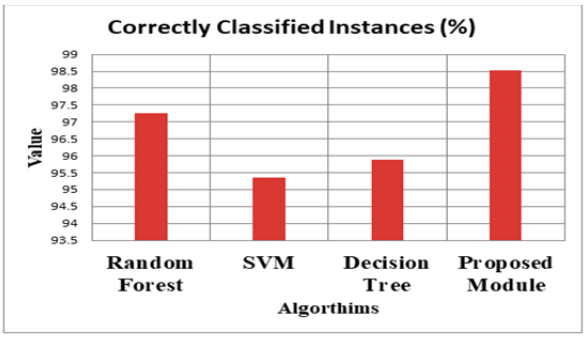

Figure 2. Correctly Classified Instances

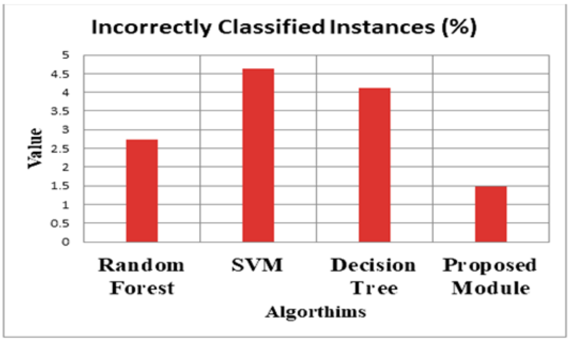

Figure 3. Incorrectly Classified Instances

B. Kappa Statistics

Figure 4 shows the results of the Kappa Statistics. The Kappa statistic is used to measure the inter-reliability of the categorical elements, as well as to measure the reliability between the evaluators [16]. The results showed that the proposed algorithm has the highest percentage which is (97.01%) while other models have a lower percentage such as random forest by (94.44%), the decision tree by (91.62%) and Support Vector Machine has the lowest percentage (90.58%). The equation 2 is used to calculate the Kappa statistic, where (Pe) represents the hypothetical probability of chance agreement while (Po) represents the relative observed agreement between the evaluators.

$$k = 1 - \frac{1-P_o}{1-P_e} \quad \text{........................} \quad (2)$$

To calculate the odds, the observed data is used, and each observer says each category randomly.

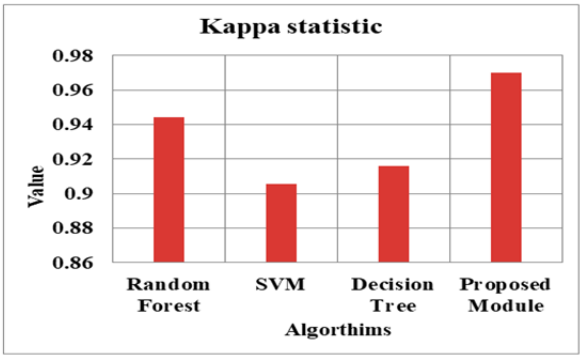

Figure 4. Kappa Statistics

The value of (K=1) is when the evaluators are completely in agreement, and the value of (K<=0) when there is no agreement between the evaluators and contrary to what can be expected by chance as shown in (Pe) and when the result is equal to what is expected by chance then the value (K=0) [17].

C. Mean Absolute Error (MAE)

Mean Absolute Error is a quantity used to measure the predictions of the end results or how close the forecasts are. Equation 3 is used to calculate the Mean Absolute Error, which is a rate of absolute errors. (Yi) represent the real value, while (Fi) represents the value of prediction [18].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad \text{........} \ (3)$$

Figure 5 shows the difference in results between the proposed model and the other models. The results showed that the proposed model has the lowest percentage of (3.75%) and is considered the best of the rest of the models, while the decision tree gave the worst result by (15.67%).

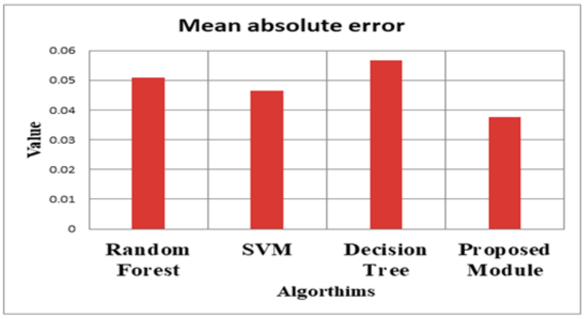

Figure 5. Mean Absolute Error (MAE)

D. Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is a measure used for the differences between the population or sample values predicted by a model or estimator and the observed values. RMSE represents a model for the standard deviation due to the differences between the observed values and the expected values. It is also used to compare prediction errors for different models for a given variable, so it is a good measure of accuracy. Root Mean Squared Error has high predictive power, as it aggregates the magnitudes errors in forecasts for different times in single measure. The equation 4 is used to calculate the RMSE, where (Y) represents the variable of the regression, and (t) represents times while (n) represents different predictions such as the square root [19].

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y)^2}{n}} \quad \text{................} \ (4)$$

Figure 6 shows the RMSE ratio of the proposed model and the other three models, where the results showed that the proposed model has the lowest ratio of (11.53%) and is better than the results of the rest of the other models, while the support vector machine gave the worst result which is (21.54%).

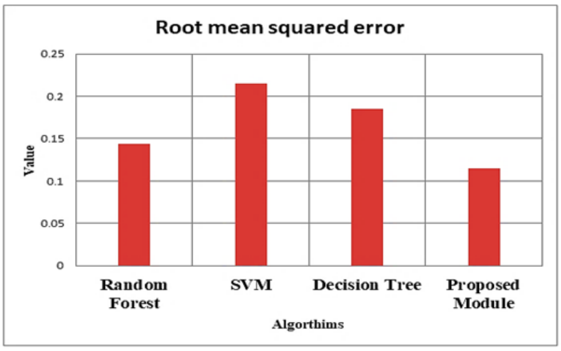

Figure 6. Root Mean Squared Error (RMSE)

E. Relative Absolute Error (RAE)

Relative Absolute Error represents the percentage of one's result deviation from the real value, and is measured in percentage. The Figure 7 shows the ratio of relative absolute error and the difference between the result of the proposed model and the other three models, where the results showed that the proposed model has the lowest percentage (7.59%) which is the best compared with results of other models and also the results showed that the decision tree has the worst result and is (11.48%).

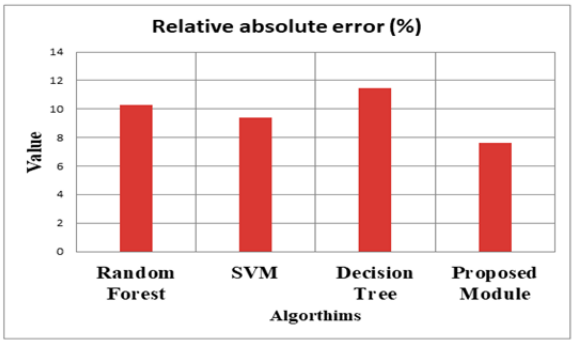

Figure 7. Relative Absolute Error (RAE)

F. Root Relative Squared Error (RRSE)

The Root Relative Squared Error takes the set of the squared error and then normalizes it by dividing the squared error squared by the sum of the squared error of the simple predictor, and the error can be reduced by taking the square root of the relative squared error.

Figure 8 shows the value of the Root Relative Squared Error, where the results showed that the proposed model is the best of the results of other models as it has the lowest ratio which is (23.2056%) while a support vector machine model has worst result by (43.3655%).

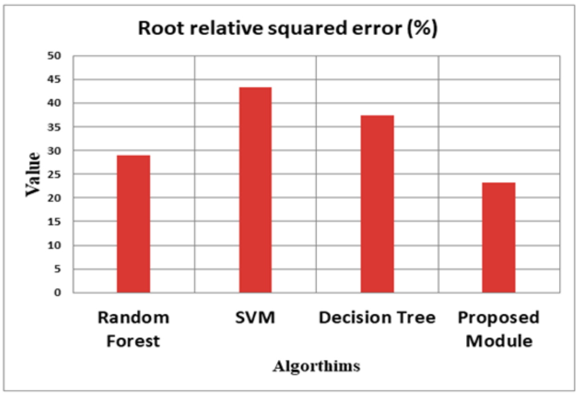

Figure 8. Root Relative Squared Error (RRSE)

5. Confusion Matrix Comparison between Models

Comparing the confusion matrix between the proposed model and the other three models does not require any changes to the current authentication systems that the website uses. Rather, this method only requires minimal user training. A set of reference data is used for the purpose of assessing the accuracy of phishing detection patterns. A summary of the standardized lines of the column shows the incorrect and distinct perception rates for each actual category. The following parameters have been used to measure the accuracy of phishing detection charts: The first is the number of true positives (TP), which represent sites that are correctly classified as phishing sites. Second, the true negatives number (TN), which represents the sites that were correctly classified as illegal sites. Third, the number of false positives (FP) representing sites incorrectly classified as hunting sites. Finally, the number of false negatives (FN), which represents sites incorrectly categorized as legitimate sites. Weighted average is the amount of your result deviation from the real and expected values. Table 3 and the Figure 9 shows the weighted average that was used to measure confusion, where the results showed that the proposed model has the best result compared to the results of other models where the proposed model obtained the highest value, which is (4782).

Table 3. Weighted Average of Confusion Metric Comparison between the Models

No.	Classification	TP	FN	FP	TN
1	Random Forest	4705	193	110	6047
2	SVM	4591	307	206	5951
3	Decision Tree	4615	283	173	5984
4	proposed Model	4782	116	110	6047

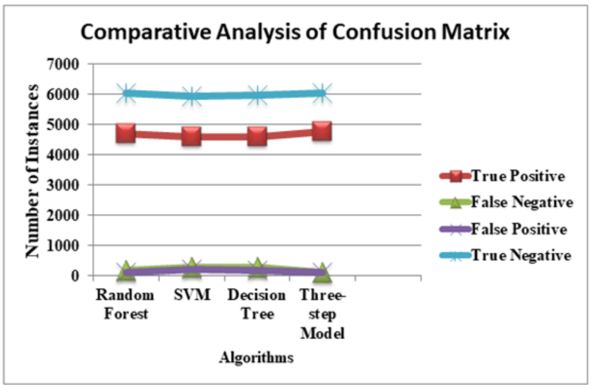

Figure 9. Weighted Average of Confusion Metric Comparison between the Models

The sections relate to the expected category and compare these lines with the actual separation. Oblique cells and cells from corner to corner inaccurately and very effectively compare perceptions required individually.

6. Conclusion

In this paper, we proposed to use three detection models that are combined with each other, namely (decision tree, random forest and support vector machine), to investigate the problem of phishing on sites in addition to using the forms separately for the purpose of comparison with the proposed model, and the proposal was implemented and evaluated using the data set. The results showed that the three models recorded a slight difference in their results, but all of them had less accuracy than the proposed model in detecting phishing sites. For the purpose of classification in this research, support vector machine multi-class classifier was used. The results showed that the percentage of improvement of accuracy in the proposed model compared to the detector reaches (1.2) via a data set attribute-relation file format (ARFF). A comparison in terms of the accuracy of detection of phishing sites was made between the proposed model and the other models that were used individually. The results showed that the proposed model outperforms the decision tree model by (2.584%), and from the support vector machine model by (3.0996%) and finally the accuracy of the proposed model exceeds the random forest model by (1.2%). Consequently, the proposed model has proven to be extremely effective in detecting phishing sites. As shown in the results of each individual model, which is the random forest model (97.259%), support vector machine model by (95.3597%), the decision tree model by (95.8752%), and finally the proposed model scored the highest accuracy by (98.5256%). It can be concluded that the proposed model has proved its validity by using the three detection forms together. In addition, we have demonstrated in this study the disadvantages of using

Uniform Resource Locator address (URL) features to detect phishing sites. An example is the lengths of addresses (URL) that can give accuracy in detecting phishing sites, but in the future they may not do so. This study is potentially very effective even with severe phishing which is specially designed for the purpose of deceiving experienced users.

References

[1]. Chaudhry, J. A., Chaudhry, S. A., & Rittenhouse, R. G. (2016). Phishing attacks and defenses. *International Journal of Security and Its Applications*, *10*(1), 247-256.

[2]. Khan, A. & Sharma R.,(2018). A Survey Paper on Detection of Phishing Website by URL Technique. *International Journal of Computer Science and Mobile Applications, 6*, 33-37.

[3]. Sakunthala R. & Shankar S.,(2018). Various Methods for Phishing Detection. *EAI Endorsed Transactions on Energy Web and Information Technologies, 5*(20), 3-11.

[4]. Abusaimeh, H. (2020). Security Attacks in Cloud Computing and Corresponding Defending Mechanisims. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(3).

[5]. Dhamija, R., Tygar, J. D., & Hearst, M. (2006, April). Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (pp. 581-590).

[6]. Cui, Q., Jourdan, G. V., Bochmann, G. V., Couturier, R., & Onut, I. V. (2017, April). Tracking phishing attacks over time. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 667-676).

[7]. MahaLakshmi, A., Goud, N. S., & Murthy, G. V. (2018). A Survey on Phishing And It's Detection Techniques Based on Support Vector Method (SVM) and Software Defined Networking (SDN). *International Journal of Engineering and Advanced Technology*, *8*(2).

[8]. Abdelhamid, N., Thabtah, F., & Abdel-jaber, H. (2017, July). Phishing detection: A recent intelligent machine learning comparison based on models content and features. In *2017 IEEE international conference on intelligence and security informatics (ISI)* (pp. 72-77). IEEE.
doi: 10.1109/ISI.2017.8004877.

[9]. Bahnsen, A. C., Bohorquez, E. C., Villegas, S., Vargas, J., & González, F. A. (2017, April). Classifying phishing URLs using recurrent neural networks. In *2017 APWG symposium on electronic crime research (eCrime)* (pp. 1-8). IEEE.
doi: 10.1109/ECRIME.2017.7945048.

[10]. Qabajeh, I., Thabtah, F., & Chiclana, F. (2018). A recent review of conventional vs. automated cybersecurity anti-phishing techniques. *Computer Science Review*, *29*, 44-55.

[11]. Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based associative classification data mining. *Expert Systems with Applications*, *41*(13), 5948-5959.

[12]. Hadi, W. E., Aburub, F., & Alhawari, S. (2016). A new fast associative classification algorithm for detecting phishing websites. *Applied Soft Computing*, *48*, 729-734.

[13]. Ubing, A. A., Jasmi, S. K. B., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Phishing website detection: An improved accuracy through feature selection and ensemble learning. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *10*(1).

[14]. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, *25*(2), 443-458.

[15]. Ezziyyani, M., Bahaj, M., & Khoukhi, F. (Eds.). (2017). *Advanced Information Technology, Services and Systems: Proceedings of the International Conference on Advanced Information Technology, Services and Systems (AIT2S-17) Held on April 14/15, 2017 in Tangier* (Vol. 25). Springer.

[16]. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276-282.

[17]. Armitage, P., & Colton, T. (2000). *Encyclopedia of epidemiologic methods*. John Wiley & Sons.

[18]. Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

[19]. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688.