Int. J. Nonlinear Anal. Appl. 12 (2021) No. 1, 2179-2188
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2021.5367

Fraud usage detection in Internet users based on log data

Kareem K. Ibrahim[a,*], Ahmed J. Obaid[a]

[a]*Faculty of Computer Science and Mathematics, University of Kufa, Iraq.*

(Communicated by Madjid Eshaghi Gordji)

Abstract

The Internet has become one of the most important daily social, 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 in 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 that shows the robustness of the proposed system.

Keywords: Fraud, phishing, Legitimate Weblog, phishing Log Data.
2010 MSC: Please write mathematics subject classification of your paper here.

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 [23]. Internet users are exposed to several types of phishing. Through

*Corresponding author
Email addresses: karimk.aljabri@student.uokufa.edu.iq (Kareem K. Ibrahim), ahmedj.aljanaby@uokufa.edu.iq (Ahmed J. Obaid)

Received: February 2021 *Accepted:* June 2021

the use of fraudulent emails or a fake website, attackers try to obtain sensitive information from users such as user credentials, passwords, etc. [13]. A phishing attacker uses social engineering techniques to simulate legitimate websites and lure users to phishing web pages in various ways, etc. [18]. 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 [24]. Phishing attacks are among the most serious threats to web-based services including financial institutions, e-commerce, and individuals [13, 6]. 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 [3]. 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 [24, 14]. 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 [14].

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 [15].

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 [5]. 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)[8].

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)[1]. 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 [20].

2.2. Review of Related Works

R. Kumar et al. 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 [12].

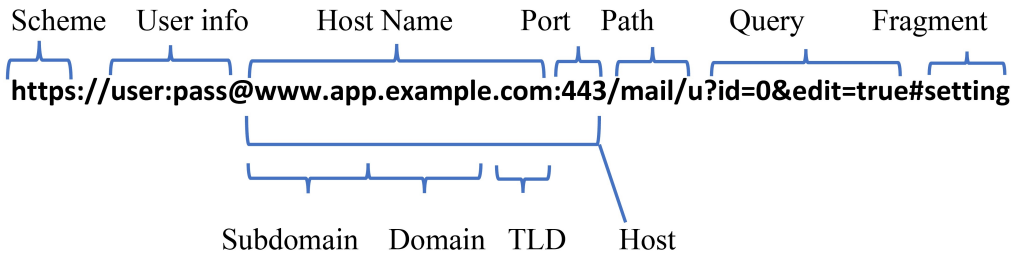

Figure 1: Typical URL syntax [16]

S. Jagadeesan et al. [10] discuss the metadata of URLs used by certain attributes and how to determine if a URL is a fraud. By applying specific algorithms such as Random Forest and linear and non-linear SVM to the web file. The results indicate that the RF algorithm is more accurate than others.

Bayu Adhi Tama et al. Presented a study comparing individual classification algorithms such as decision tree with the algorithms of different ensembles, such as alternating forest and random forest. Then they evaluated the performance using (AUC) and the experimental results indicated that the random forest is the best in accuracy and performance [21].

Issa Qabajeh et al. presented a comparative study between technological methods for controlling phishing sites and traditional methods. Technological solutions such as using predictive machine learning algorithms, traditional methods are to enforce cybercrime laws and sue the creators of malicious websites. And raise awareness of several tips for users of phishing sites. Their study focused on raising awareness and educating users about phishing [17].

MahaLakshmi et al. discuss the types of phishing, what are their harms, and explained 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 [15].

Alyssa Anne Ubing et al. discussed the use of 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 [22].

Shisrut Rawat et al. explained the 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. 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 [4].

Hesham Abusaimeh et al. 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%) [2].

P. Kalaharsha et al. 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 [11].

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. 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) repository database of approximately 11,055 entries and used 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.

3.2. Design Flowchart of the Phishing Attack Detection Model

The following figure 2 describes the proposed system design for detecting phishing attacks. Which starts from entering the weblog and conducting analyzes and even detecting phishing sites.

.

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.

.

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 $(., -, \_, /, ?, =, \&, !, , \sim, +, *, \#, \$, \%, @)$, Domain Have Sign $(., -, \_, /, ?, =, \&, !, , \sim, +, *, \#, \$, \%, @)$, Path $(., -, \_, /, ?, =, \&, !, , \sim, +, *, \#, \$, \%, @)$)]

4. Performance Evaluation Metrics

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 [11].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.1}$$

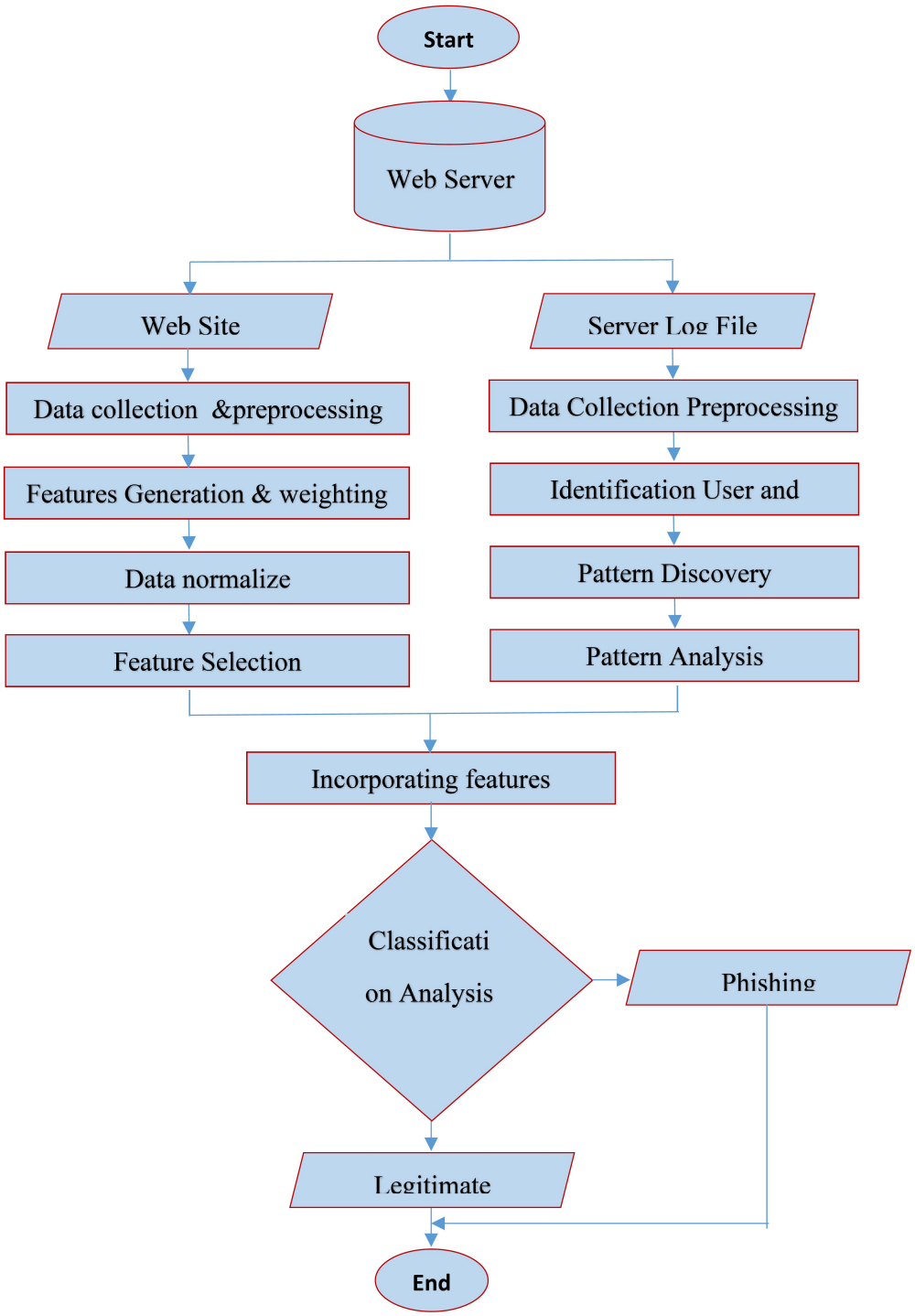

Figure 2: Flowchart of the Proposed System

Algorithm (3.4) Adaptive Random Forest Algorithm

Generate m Classifier
For e= 1 To m do
From training data G take a random sample, with substitution to produce Ge Create root node, Ne containing Ge
Call Build Tree (Ne)
End for
Build Tree (N):
If N consists of instances of only one class
Return
Else
Randomly select x% of the possible splitting features in N
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 e=1 To f do
Set the consists of Ne to Ge where Ge is all instances in N that match Fe
Call Build Tree(Ne)
End for
End if
END

Table 1: Experiment Parameters for Adaptive Random Forest

No.	Parameter	Value
1	Size of each bag	100
2	No. of Iteration	100
3	No. of execute slots	1
4	No. of Attributes to investigate randomly	0
5	Minimum number of cases per sheet	1
6	Seed for random number generator	1
7	No. of cross-validation folds	10-fold

Table 2: shows the Confusion matrix for binary classification

Predicted	Actual Values		
		Positive (1)	**Negative (0)**
	Positive (1)	**TP**	**FN**
	Negative (0)	**FP**	**TN**

4.2. Precision and Recall

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 [7].

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

4.3. F-measure

F-measure, also known as F-value, A-weighted harmonic mean of precision and recall [9].

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4.4}$$

4.4. Kappa Statistic

The Kappa statistic is used to measure interference Among categorical items [2].

$$(ks) = 1 - \frac{1 - Po}{1 - Pe} \tag{4.5}$$

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

4.5. Mean Absolute Error

The mean absolute error is quantity is used to measure Expectations in the end results [2].

$$MeanAbsoluteError = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{4.6}$$

where f_i is the prediction value, y_i 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 [2].

$$RMSE = \sqrt{\left(\sum_{i=1}^{M}\sum_{j=1}^{N}(I(i,j) - C(i,j))^2/M * N\right)} \tag{4.7}$$

After applying the previous measures to the specified data set, the following results 3 were obtained.

4.7. Correctly and Incorrectly Classified Instances

As we can see in Figure 3 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.

The results of the proposed system are also determined and its ability to detect real phishing addresses. As shown in Table 4.

The accuracy is calculated which indicates the true algorithm's ability to predict the correct class naming of the unknown class naming states.

Table 3: Experiment Parameters for The Proposed Module

No.	Parameter	Value
1	No. of attributes	63
2	Correctly Classified Instances	10599
3	Incorrectly Classified Instances	456
4	KS	0.9162
5	MAE	0.0567
6	RMSE	0.1835
7	Relative absolute error	11.49%
8	Root relative squared error	37.30%
9	Total Number of Instances	11055
10	No. of cross-validation folds	10-fold
11	Seed for random number generator	1
12	No. of Decimal Places	2
13	Batch Size	100
14	No. of execution slots	1
15	Size of each bag	100
16	Minimum division difference	0.001
17	No. of Attrition	100

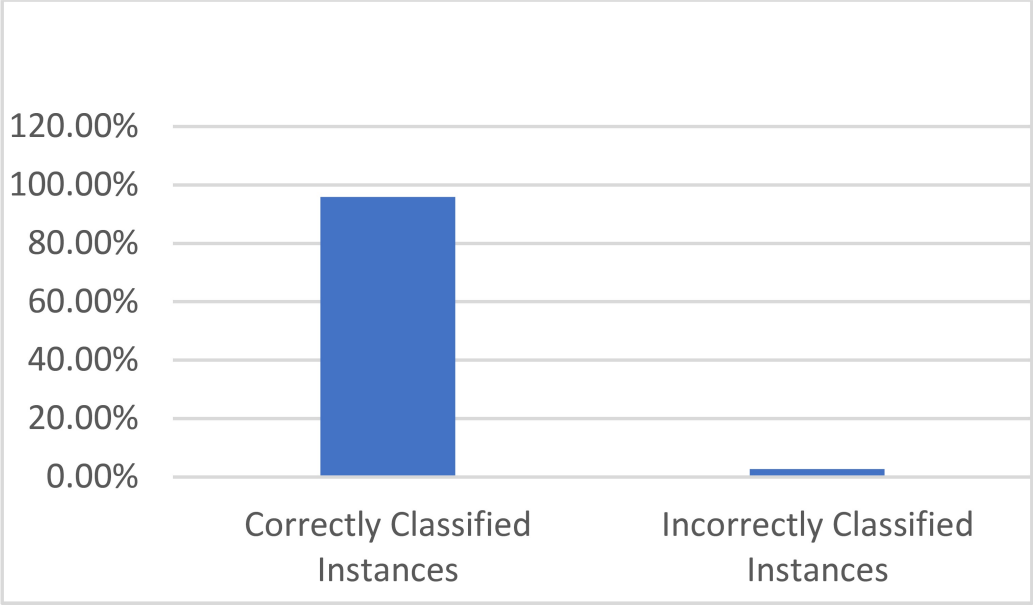

Figure 3: Correctly and Incorrectly Classified Instances

Table 4: Accuracy and retrieval

	percentage	percentage
Actual Phishing	0.417458164	0.025599276
Actual Legitimate	0.015649028	0.541293532

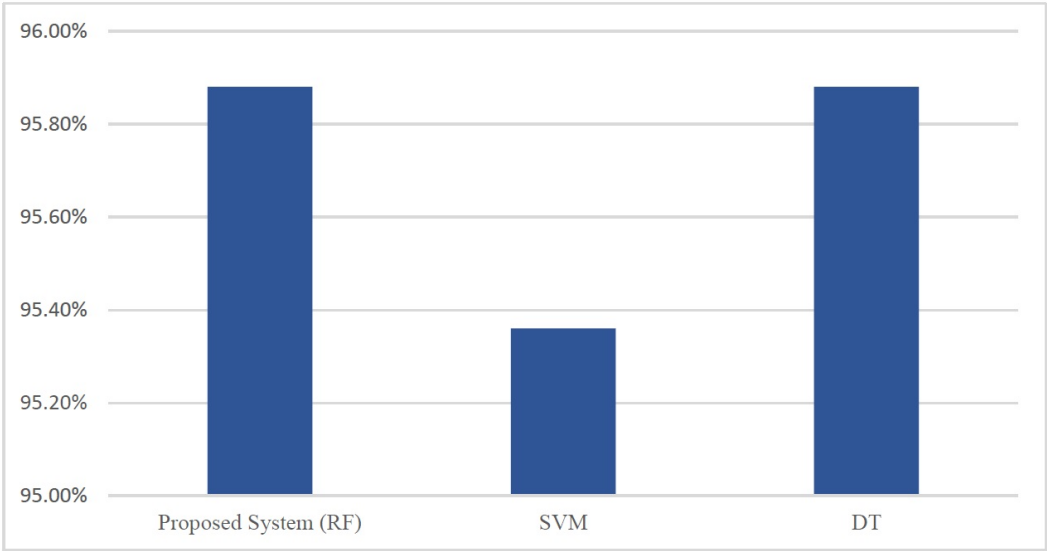

Figure 4: Correctly Classified Instances

5. Conclusions and Future Works

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 95.88% 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.

References

[1] F. Aburub and S. Alhawari, *A new fast associative classification algorithm for detecting phishing websites*, Appl. Soft Comput. J. 48 (2016) 729–734.
[2] H. Abusaimeh and Y. Alshareef, *Detecting the phishing website with the highest accuracy*, TEM J. 10(2) (2021) 947–953.
[3] Anti-Phishing Working Group, Inc. (APWG), *Phishing Activity Trends Reports*, 1st Quarter 2021.
[4] E.S. Aung, C.T. Zan, and H. Yamana, *A survey of URL-based phishing detection*, DEIM Forum (2019) 1–8.
[5] M. Aydin and N. Baykal, *Feature extraction and classification phishing websites based on URL*, 2015 IEEE Conf. Commun. Network Sec. CNS 2015 (2015) 769–770.
[6] E. Bhagyashree and K. Tanuja, *Phishing URL detection: a machine learning and web mining-based approach*, Int. J. Comput. Appl. 123(13) (2015) 46–50.
[7] *Classification: Precision and Recall — Machine Learning Crash Course*, 2021.
[8] M.J. Hamid Mughal, *Data mining: web data mining techniques, tools and algorithms: an overview*, Int. J. Adv. Comput. Sci. Appl. 9(6) (2018) 208–215.
[9] D.J. Hand, P. Christen, and N. Kirielle, *F*: an interpretable transformation of the F-measure*, Mach. Learn. 110(3) (2021) 451–456.
[10] S. Jagadeesan, *URL phishing analysis using random forest*, International Journal of Pure and Applied Mathematics, 118(20) (2018) 4159–4163.
[11] P. Kalaharsha, and B.M. Mehtre, *Detecting phishing sites - an overview*, arXiv preprint arXiv:2103.12739, (2021) 1–13.

[12] R. Kumar, X. Zhang, H.A. Tariq, and R.U. Khan, *Malicious URL detection using multi-layer filtering model*, 2017 14th Int. Comput. Conf. Wavelet Active Media Tech. Inf. Proc. (2017) 97–100.
[13] 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. 94 (2019) 27–39.
[14] H. Liu, X. Pan, and Z. Qu, *Learning based malicious web sites detection using suspicious URLs*, 34th Int. Conf. Softw. Eng. (2016) 3–5.
[15] 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. Tech. 8(2) (2018) 498–503.
[16] S. Nandhini, and V. Vasanthi, *Extraction of features and classification on phishing websites using web mining techniques*, Int. J. Engin. Dev. Res. 5(4) (2017) 1215–1225.
[17] I. Qabajeh, F. Thabtah, and F. Chiclana, *A recent review of conventional vs. automated*, Comput. Sci. Rev. 29 (2018) 44–55.
[18] R.S. Rao and A.R. Pais, *Jail-Phish: An improved search engine based phishing detection system*, Comput. Secur. 83 (2019) 246–267.
[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*, arXiv:1910.01114v1, CoRR, (2019) 1–9.
[20] W. Rong, Z. Yan, T. Jiefan and Z. Binbin, *Detection of malicious web pages based on hybrid analysis*, J. Inf. Secur. Appl. 35 (2017) 68–74.
[21] B.A. Tama and K. Rhee, *A comparative study of phishing websites classification based on classifier ensembles*, J. Korea Mult. Soc. 5(2) (2018) 99–104.
[22] 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*, Int. J. Adv. Comput. Sci. Appl. 10(1) (2019) 252–257.
[23] G. Varshney, M. Misra and P.K. Atrey, *A survey and classification of web phishing detection schemes*, Secur. Commun. Networks, 9(18) (2016) 6266–6284.
[24] R. Verma and A. Das, *What's in a URL: Fast feature extraction and malicious URL detection*, IWSPA 2017 - Proc. 3rd ACM Int. Work. Sec. Priv. Anal. (2017).