Phishing Detection Based on Machine Learning and Feature Selection Methods

https://doi.org/10.3991/ijim.v13i12.11411

Mohammed Almseidin (✉)
University of Miskolc, Miskolc, Hungary
alsaudi@iit.uni-miskolc.hu

AlMaha Abu Zuraiq, Mouhammd Al-kasassbeh
Princess Sumaya University for Technology, Amman, Jordan

Nidal Alnidami
National Information Technology Center, Amman, Jordan

Abstract—With increasing technology developments, the Internet has become everywhere and accessible by everyone. There are a considerable number of web-pages with different benefits. Despite this enormous number, not all of these sites are legitimate. There are so-called phishing sites that deceive users into serving their interests. This paper dealt with this problem using machine learning algorithms in addition to employing a novel dataset that related to phishing detection, which contains 5000 legitimate web-pages and 5000 phishing ones. In order to obtain the best results, various machine learning algorithms were tested. Then J48, Random forest, and Multilayer perceptron were chosen. Different feature selection tools were employed to the dataset in order to improve the efficiency of the models. The best result of the experiment achieved by utilizing 20 features out of 48 features and applying it to Random forest algorithm. The accuracy was 98.11%.

Keywords—Phishing Detection, Machine Learning, Feature Selection, Random Forest, Multilayer Perceptron.

1 Introduction

The Internet is everywhere today, and the society uses web services for a range of activities such as sharing knowledge, social communication, and performing various financial activities, which include buying, selling and money transferring and more other things. Malicious websites are a severe threat to the Internet's users, and unaware users can become victims of malicious URLs that host undesirable content such as spam, phishing, drive-by-download, and drive-by-exploits. Phishing is a conventional attack on the Internet, and it is defined as the social engineering process of luring users into fraudulent websites to obtain their personal or sensitive information such as their user names, passwords, addresses, credit card details, social security numbers, or any other valuable information. According to the Anti-Phishing Working Group (APWG) report [1], the number of different phishing incidents reported to the organization over the last quarter of the year 2016 was 211,032 and they increased up by 12% in last quarter of 2018 which received 239,910 reports.

Furthermore, a recent Microsoft security intelligence (volume 24) report [2] found that phishing attacks were on the top of the discovered web attacks of 2018, and it is expected to continue increasing. The major challenge when detecting phishing attacks lies in discovering the techniques utilized. Phishers continuously enhance their strategies and can create web pages that are able to protect themselves against many forms of detection. Accordingly, developing robust, effective and up to date phishing detection methods is very necessary to oppose the adaptive techniques employed by the phishers [3].

Surveying the literature on phishing detection techniques, it can be categorized to the following approaches: Blacklist based, Content-based, Heuristic-based, and Fuzzy rule-based approaches. Each of these approaches has its own characteristics and limitations. The blacklist approach maintains a list of Suspicious or malicious URL's that are collected using different approaches like Google safe browsing, Phish Tank, and users voting. So, when a web page is initiated, the browser searches the blacklist for it and alerts the user if the webpage was found. Finally, the blacklist can be stored on the user's machine or in a server [4]. Blacklists are often used to classify websites as malicious or legitimate. But while these techniques have low false-positive rates, they lack the ability to classify newly-produced malicious URLs [5]. The content based approach deploys an in-depth analysis of the pages content. Building classifiers and extract features from page contents and third-party services such as search engines and DNS servers. Yet, these methods are ineffective because of a massive number of training features and the reliance on third-party servers which assault user's privacy by uncovering his browsing history [3].

A Heuristic Based Approach, the detection technique is based on employing various discriminative features extracted by understanding and analyzing the structure of phishing web pages. The method used in processing these features plays a considerable role in classifying web pages effectively and accurately [6]. Since Fuzzy logic permits the intermediate level among values, the fuzzy rule-based approach is utilized to classify web-pages based on the level of phishness that appeared in the pages by implementing and employing a specific group of metrics and predefined rules [7]. Using fuzzy approach allows processing of ambiguous variables. Fuzzy logic integrates human experts to clarify those variables and relations between them. Also, fuzzy logic approaches using linguistic variables to explain phishing features and the phishing web page likelihood [8].

The aim of this paper is to present a study of existing methods used in the detection of phishing web-pages that employed the machine learning algorithms and focus on the most common feature selection methods that are used for dealing with various problems and enhance the performance and effectiveness of phishing dataset. Moreover, we will apply feature selection to an existing novel phishing data set to enhance the effectiveness of the data set and decrease the time taken to build the models, then compares between different machine learning algorithms to find which one is more efficient.

2 Related Works

In this section, recent works that used phishing detection approaches that utilized with machine learning algorithms will be discussed.

According to content-based approach, in [9], a novel method that utilizes a logo image to determine the identity of the web page by matching real and fake web-pages. The proposed approach is composed of two phases, which are logo extraction and identity verification. In the first phase, machine learning algorithms are used to detect the right logo image. While in the second phase, image search offered by Google is used to return the fake identity, then it will be utilized for the verification. Because the relation among the logo and domain name is unique, the domain name is treated as the identity of the logo. So, a comparison among the domain name retrieved by Google with the one from web page query will permit us to distinguish between phishing and legitimate web pages. The experimental results notice that logo extraction phase enhanced phishing detection accuracy, and it is more useful than extraction phases based on textual features. The system has been evaluated by using two different datasets that made of 1140 phishing obtained from Phish-Tank and legitimate web-pages obtained from Alexa. They only selected the most sensitive eight features out of 23 features. They justify utilizing feature selection because using all the 23 features would consuming the time. The accuracy of the proposed system is 93.4%.

On the other hand, some studies combined a heuristic based with a machine learning algorithm to enhance a classification process of web pages. Machine learning algorithms are utilized a clarify features and effective algorithm to produce an accurate classifier model to distinguish between phishing and legitimate web-pages. In the work of [10], they suggested heuristic based phishing detection method that used to recognize the phishing site. In the beginning, the system extracts and utilize URL-based features. Then, these features are applied to machine learning algorithms, and it will recognize if the web page is phished or legitimate. The system used 10 features on the input URL's dataset. It implements features extraction from URL inputs using .NET Script. The output results are categorized as either Legitimate or Phishing. Support Vector Machine algorithm is used on extracted features result and find the value for FP, TP, FN and TN and also have calculated the value of F1-measure and the accuracy that presented 96%. Dataset of URLs are collected from Phish-Tank and yahoo directory, which contains 200 Legitimate and phishing web pages URLs.

Likewise, in [11], they implemented a heuristic based phishing detection approach besides machine learning algorithms features of URL. The proposed method elicited URL features of web pages requested by the user and applied them to decide if a requested web page is phishing or not. To choose a classifier that most effectiveness for employing URL-based features, five machine learning techniques are utilized: support vector machine (SVM), naive Bayes, decision tree, k-nearest neighbour (KNN), random tree, and random forest. To evaluating and training a classifier a dataset that collected 3,000 phishing web-pages from Phish-Tank and 3,000 legitimate webpages from DMOZ. 26 URL-based features are extracted and utilized. The experiment results show that machine learning classifier that achieved the best performance is Random Forest (FR) with 98.23% of accuracy.

Additionally, in [12], authors also proposed a heuristic based method to detect phishing URLs by utilizing URLs features. The system is evaluated using data sets that consist of more than 16,000 phishing and 31,000 non-phishing URLs is employed. They used a set of 138 features in detecting phishing URLs. Features are categorized into four groups, which are Lexical based features, Keyword based features, Reputation-based features, and Search engine-based features. Furthermore, seven different classifiers are implemented which are Support Vector Machines (SVM with RBF kernel), SVM with linear kernel, Multilayer Perceptron (MLP), Random Forest (RF), Nave Bayes (NB), Logistic Regression (LR) and C4.5. According to experiment results, Random Forest (RF) achieved a higher accuracy rate and lower error rate.

In the previous works, a heuristic based approach is implemented with a machine learning algorithms, each of them has its own data sets, employing different features and applying several machine learning algorithms, but in both Random Forest algorithm is achieved the most effective classification rate of web-pages, likewise, in our work, we use different dataset, different features and applying in different machine learning algorithms in addition to employing different feature selection techniques but also the random forest shows the best results. Next two studies will demonstrate a hybrid machine learning approaches that get a benefit from strengthens of each algorithm and overlooked about the weaknesses, because more effective techniques are needed to limit the fast evolution of phishing attacks.

The study of [4], they proposed a method that combines two algorithms, K-nearest neighbors (KNN) algorithm which is effective against noisy data and Support Vector Machine (SVM) algorithm, which is a robust classifier, a combination is done in two phases. At first, applying KNN then SVM is employing as a classification tool. The dataset used for the experiment is taken from related work, the dataset contains more than 1353 sample gathered from various sources, each sample record composed of nine features and the class label which is Phishing, Legitimate or Suspicious web page. Consequently, the clearness of KNN is integrated with the effectiveness of SVM, regardless of their own disadvantages when they used individually. The accuracy of the proposed method is 90.04%. In [13], authors proposed a fast and accurate phishing detection method that combined both Naive Bays (NB) and Support Vector Machine (SVM), utilizing features of URLs and web-page contents. NB is used in detecting web pages. As long as the web pages are not detected efficiently and still suspicious, SVM will be employed to reclassifying the web pages. The used learning dataset is generated from Phish Tank which is 600 phishing web pages, and 400 are legitimate ones, 100 legitimate and 100 phishing web pages are occupied as the training set, and the rest are carried as testing dataset. Experimental results exhibit that this proposed approach achieved high detection accuracy and lower detection time.

3 Phishing Website Dataset

Data set used in this study is offered by Chiew et al [14] which composed of 48 features taken out from 5000 phishing web-pages and 5000 legitimate web-pages. Phishing webpages are collected from Phish-Tank and Open-Phish, while legitimate web-pages are collected from Alexa and Common Crawl. These web-pages are downloaded on two distinct sessions, from January to May 2015 and through May to June 2017. Browser automation framework is employed to improve the feature extraction method, which is more accurate and robust in contrast with parsing technique based on regular expressions. Features in this dataset are classified into three groups, which are Address bar-based, Abnormal-based, and HTML/JavaScript-based features. Address bar-based are the features in the URL of the web page like URL's length and port number, abnormal-based are features of abnormal actions on the web page like downloading objects from external domains, and HTML/JavaScript-based are features of HTML and JavaScript methods placed in the source code of the web page [15]. In this work, we chose this dataset because it is the most recent dataset in this field.

4 Machine Learning Techniques

Different experiments have been done on different machine learning classifiers such as Bayes net, Naive Bayes, J48, Logistic, Random forest, Bagging, and Multilayer perceptron. Then we chose three algorithms which obtained the best accuracy rates and the most commonly used classifiers based on the literature, which are J48, Random forest and Multilayer perceptron.

4.1 J48 Algorithm

J48 is a type of C4.5 decision tree algorithm deployed for classification purposes; it employs a set of training data that composed of classified samples. Every sample demonstrates the feature value of that sample. The decision tree is constructed by the algorithm using the training data set. Each node in the tree is recognized by the feature that effectively divides its set of samples into new subsets using the value of the information gain (Fig 1). The significant characteristics of decision trees are their clarity to illustrate, explain and consider the relationships and interactions of the features. While decision trees are requiring reconstructing the tree if new samples exist [16], [17].

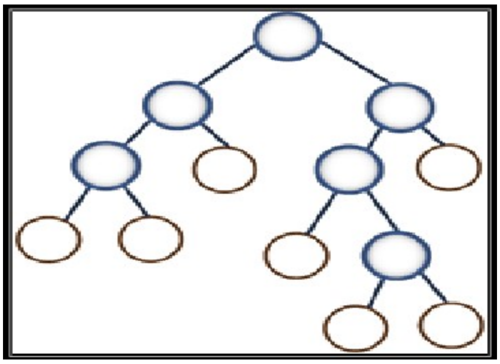


Fig. 1. Decision Tree Structure

4.2 Random Forest Algorithm

Random forest is a classification method based on the decision tree algorithm. It is appropriate for enormous datasets for the reason that it can hold a considerable number of variables in the dataset; at the training phase, it builds a group of different decision trees (Fig.2). Where each tree runs on a set of predefined attributes that selected randomly. The classification process is done by majority vote the outcomes from every single tree. Random Forest is trained on several portions of the training data set. Characteristic of using the random forest is that it solved the over-fitting problem that is commonly occurred when using individual decision trees. However, reproducibility process is absent because the operation of building the forest is random [18], [19].

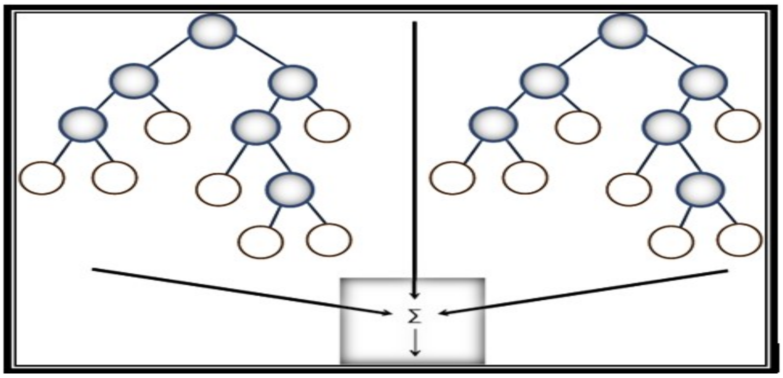


Fig. 2. Random Forest Structure

4.3 Multilayer Perceptron Algorithm

A Multilayer Perceptron (MLP) is the most popular and frequently used artificial neural network. Like a neural network, MLP consists of multi interconnected components. They are constructed of three different layers which are an input layer, hidden layer, and output layer each has its own functionality (Fig. 3), an input layer is used to obtain the signal, an output layer turns out a decision about the input, and there is at least one hidden layer that is the computational engine of the MLP. It is usually utilized to supervised learning problems: it is trained on a group of input-output pairs and learns the correlation and dependencies among them [20].

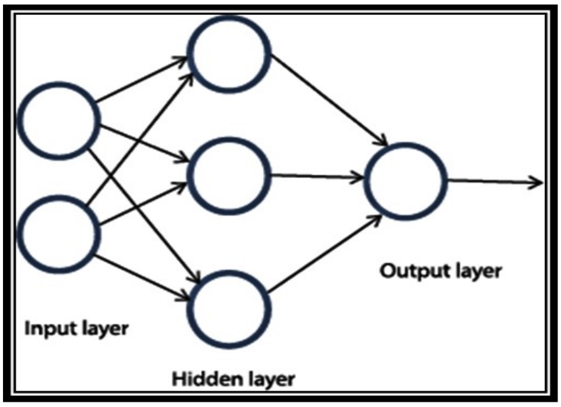


Fig. 3. Multilayer Perceptron Structure

5 Feature Selection

Feature selection is employed to decrease the size of the data to enhance the model's performance and reducing the computation time. Simply, the feature selection keeps the most important fields and eliminates unimportant ones. However, it also gives useful and robust results. In this work, different feature selection methods will be utilized to enhance the phishing detection method by increasing the accuracy rate and decreasing the time that taken to build the model.

5.1 Feature Selection Methods

Feature selection methods are classified based on the evaluation criteria into three categories, which are filters, wrappers, and hybrid methods. In filter methods, the features are chosen based on the performance measure with the independence of the used data modeling algorithm or any utilized predictor. Then, after picking out the best features, the modeling algorithm can employ them. In wrapper methods, the features subsets are considered based on the quality of the performance on modeling the algorithm. This method is significantly slower than the filter method in finding excellent features subsets because it depends on the modeling algorithm. Whereas, the wrapper method is more efficient in acquiring features subsets than the filter method because the subsets are assessed using an actual modeling algorithm. In the hybrid method, the best characteristics of filter and wrapper methods are combined. Primarily, a filter method is employed to decrease the feature space. After that, a wrapper is used to asset the best subsets. Hybrid methods obtained high accuracy and high-efficiency rates [21].

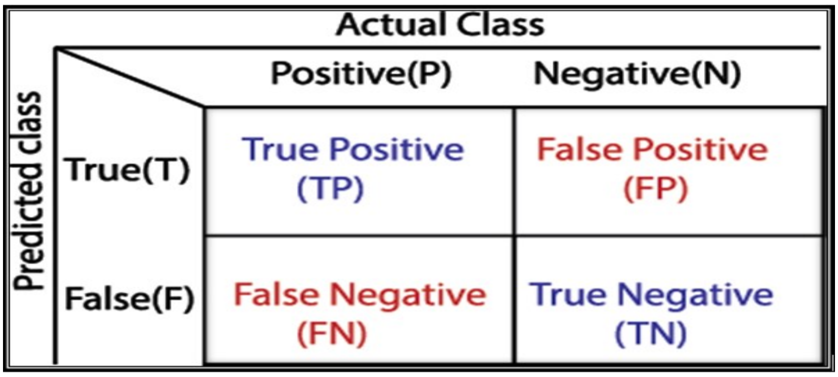
The aim of this study is to assess different feature selection techniques in term of accuracy and computational execution. Out of the overall 48 features used in phishing detection, some features will be optional in detecting phishing web pages. Therefore, the essential features are taken away from the original dataset that is particularly effective in phishing detection, which will be debated in the results section. Different experiments had been done on different filters methods of feature selection techniques such as InfoGain, ReliefF, PCA, and attribute. However, InfoGain and ReliefF had been chosen in our work because they attain the best accuracy rates than the remnant techniques.

- **InfoGain:** It shows the significance of the features and determines which one of them is the most helpful for distinguishing among the classes. The value of InfoGain is calculated in the training data set. It is used in decision tree algorithms because it can help in deciding the best split; which high value indicates that split is excellent and low value indicates that the split is not good enough. The equation (1) used to estimate the value of an attribute by calculating the information gain according to the class [17].
- **ReliefF:** As a filter-based feature selection method, Relief used to evaluate the quality of every feature according to the context of other features and the relevance of the feature to given target notion [22]. The produced value of the algorithm is between - 1 and 1 for every feature in addition with positive numbers designating more significance or weighted attributes. The weight of an attribute is reduplicative upgraded, and it has a probabilistic description. The fundamental principle of relief is that important attributes are equivalent to instances of the same class.

$$\text{InfoGain(Class,Attribute)} = \text{H(Class)} - \text{H(Class|Attribute)} \tag{1}$$

6 Model Evaluation

To evaluate the models, there are many assessment tools. But we attend to evaluate our model using the accuracy equation because the utilized dataset is Binary and Balanced data set. So, calculating the accuracy rates will be enough, efficient and accurate. To apply the accuracy formula, we should mention that there are two kinds of classification methods in accordance with the number of classes which are binary classification and multi-class classification. Where in binary classification there are only two classes whereas in multi-class classification the number of classes is more than two. In binary classes (Fig. 4), assume we have two classes, P for the positive class and N for negative class [23].



		Actual Class	
		Positive(P)	Negative(N)
Predicted class	True(T)	True Positive (TP)	False Positive (FP)
	False(F)	False Negative (FN)	True Negative (TN)

Fig. 4. An Example of The 2 X 2 Confusion Matrix

- True Positive (TP): the true prediction rate of the positive samples. The predicted value is positive, and the actual value is also positive
- False Positive (FP): negative value incorrectly classified as positive
- True Negative (TN): the true prediction of negative samples. The predicted value is negative, and the actual value is also negative
- False Negative (FN): positive value incorrectly classified as negative

Accuracy refers to the ratio of correctly classified instances. It is the most used evaluation metric for the performance of binary classification problems. Also, it is determining the accuracy of the classification model. Accuracy is calculated using the following equation (2).

$$Accuracy = \frac{\sum true\ positive + \sum true\ negative}{\sum total\ population} \tag{2}$$

7 Experiments And Results

In this study, the dataset mentioned in section 3 was employed, which contains 48 different features. For analysis and comparing between used classifiers, Weka 3.8.3 has been utilized. Weka is a set of machine learning algorithms used for different data mining functions such as data preparation, classification, regression, clustering, association rules mining, and visualization. Two different feature selection algorithms have been used in this study: InfoGain and ReliefF. The details of the top 15 extracted features from both algorithms are described in Table 1.

Table 1. The Top 15 Extracted Features

Method	Top 15 Features
InfoGain	27, 28, 48, 34, 14, 35, 47, 5, 39, 1, 30, 3, 22, 25, 23.
ReleifF	48, 30, 35, 34, 47, 40, 27, 39, 28, 29, 45, 31, 3, 26, 14.

For the experiments, the 10-fold cross-validation technique is utilized in testing the models for the reason that it minimizes the estimation variance. By using this technique, the training dataset should be divided into 10 subsets, then each of these subsets must be tested in the remaining nine subsets. Every test subset is employed once a time in all 10 repetitions. Table 2,3 and 4 show the performance of the three selected algorithms (J48, RF, and MLP) using infoGain and reliefF feature selection methods with top 5, top 10 and top 15 features.

Table 2. The Performance of J48 Algorithm.

Algorithm	Accuracy	Taken Time (seconds)
J48	97.31	1.2
J48+infogain+top5	95.31	0.12
J48+infogain+top10	96.17	0.21
J48+infogain+top15	96.96	0.35
J48+ reliefF +top5	89.59	0.08
J48+ reliefF +top10	97.08	0.16
J48+ reliefF +top15	97.28	0.29

Table 3. The Performance Of Random Forest (RF) Algorithm

Algorithm	Accuracy	Taken Time (seconds)
Random Forest (RF)	98.37	4.18
RF+infogain+top5	95.96	2.24
RF+infogain+top10	96.87	2.87
RF+infogain+top15	97.91	2.68
RF+ reliefF +top5	89.75	1.25
RF+ reliefF +top10	97.7	2.29
RF+ reliefF +top15	97.87	2.48

Table 4. The Performance of Multilayer Perceptron (MLP) Algorithm

Algorithm	Accuracy	Taken Time (seconds)
MLP	96.59	117.79
MLP +infogain+top5	91.89	6.04
MLP +infogain+top10	93.45	12.02
MLP +infogain+top15	95.74	18.92
MLP + reliefF +top5	88.22	4.75
MLP + reliefF +top10	95.63	9.74
MLP + reliefF +top15	96.19	14.93

Furthermore, other two experiment were performed to get the best accuracy and the least time to build the model. First one is the intersect of top 15 features using infoGain and reliefF - that present 10 features which are 27, 28, 48, 34, 14, 35, 47, 39, 30, 3. As it has seen in Table 5. The second experiment results in 20 features which are the Union of top 15 features using infoGain and reliefF. These features are 27, 28, 48, 34, 14, 35, 47, 39, 30,3 ,5, 1, 22, 25, 23.See Table 5.

Table 5. Intersect Of Info Gain and Relief Using 10 Features

Algorithm	Accuracy	TakenTime (seconds)
Intersect of infoGain and relief using J48	96.65	0.56
Intersect of infoGain and relief using RF	97.49	2.44
Intersect of infoGain and relief using MLP	95.57	9.69

The experiments results show that using the 20 features that result from the Union of top 15 features using infoGain and reliefF is presents very close accuracy rates of using the whole 48 feature. In addition, it takes much less time to build the model.

Table 6. Union Of Infogain And Relief Using 15 Features.

Algorithm	Accuracy	Taken Time (seconds)
Union of infoGain and relief using J48	97.03	0.4
Union of infoGain and relief using RF	98.11	2.61
Union of infoGain and relief using MLP	96.64	23.91

8 Conclusion

Nowadays there is an enormous number of web pages, phishing web-pages take a significant part of them. Phishing web-pages are trying to lure users to get the benefits from them. This paper proposed a method of phishing detection using machine learning algorithms and employing a dataset of 5000 legitimate web-pages and 5000 phishing ones. Best results are acquired by utilizing feature selection tools that eliminate the number of features from 48 to only 20. The time taken to construct the model was 2.44 seconds and performed an accuracy rate of 98.11 by employing 20 features to the Random Forest algorithm.

9 References

[1] Anti-Phishing Working Group . phishing activity trends report 4 th quarter. https://docs.apwg.org, 2018.

[2] Microsoft Security Intelligence Report . volume 24. https://www.microsoft.com/security, 2019.

[3] Hossein Shirazi. *Unbiased phishing detection using domain name based features*. PhD thesis, Colorado State University. Libraries.

[4] Altyeb Altaher. Phishing websites classification using hybrid svm and knn approach. *International Journal of Advanced Computer Science and Applications*, 8(6):90–95, 2017. https://doi.org/10.14569/ijacsa.2017.080611

[5] Yi-Shin Chen, Yi-Hsuan Yu, Huei-Sin Liu, and Pang-Chieh Wang. Detect phishing by checking content consistency. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*, pages 109–119. IEEE, 2014. https://doi.org/10.1109/iri.2014.7051880

[6] Neda Abdelhamid, Aladdin Ayesh, and Fadi Thabtah. Phishing detection based associative classification data mining. *Expert Systems with Applications*, 41(13):5948–5959, 2014. https://doi.org/10.1016/j.eswa.2014.03.019

[7] Mahmood Moghimi and Ali Yazdian Varjani. New rule-based phishing detection method. *Expert systems with applications*, 53:231–242, 2016. https://doi.org/10.1016/j.eswa.2016.01.028

[8] Maher Aburrous, M Alamgir Hossain, Keshav Dahal, and Fadi Thabtah. Intelligent phishing detection system for e-banking using fuzzy data mining. *Expert systems with applications*, 37(12):7913–7921, 2010. https://doi.org/10.1016/j.eswa.2010.04.044

[9] Kang Leng Chiew, Ee Hung Chang, Wei King Tiong, et al. Utilisation of website logo for phishing detection. *Computers & Security*, 54:16–26, 2015. https://doi.org/10.1016/j.cose.2015.07.006

[10] Jaydeep Solanki and Rupesh G Vaishnav. Website phishing detection using heuristic based approach. In Proceedings of the third international conference on advances in computing, electronics and electrical technology, 2015.

[11] Jin-Lee Lee, Dong-Hyun Kim, and Lee Chang-Hoon. Heuristic-based approach for phishing site detection using url features. In *Proc. of the Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology-CEET*, pages 131–135, 2015. https://doi.org/10.15224/978-1-63248-056-9-84

[12] Ram B Basnet and Tenzin Doleck. Towards developing a tool to detect phishing urls: a machine learning approach. In *2015 IEEE International Conference on Computational Intelligence & Communication Technology*, pages 220–223. IEEE, 2015. https://doi.org/10.1109/cict.2015.63

[13] Xiaoqing Gu, Hongyuan Wang, and Tongguang Ni. An efficient approach to detecting phishing web. *Journal of Computational Information Systems*, 9(14):5553–5560, 2013.

[14] Kang Leng Chiew, Choon Lin Tan, KokSheik Wong, Kelvin SC Yong, and Wei King Tiong. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, 484:153–166, 2019. https://doi.org/10.1016/j.ins.2019.01.064

[15] Mahdieh Zabihimayvan and Derek Doran. Fuzzy rough set feature selection to enhance phishing attack detection. *arXiv preprint arXiv:1903.05675*, 2019. https://doi.org/10.1109/fuzz-ieee.2019.8858884

[16] Adwan Yasin and Abdelmunem Abuhasan. An intelligent classification model for phishing email detection. *arXiv preprint arXiv:1608.02196*, 2016.

[17] Mohammad Almseidin, Maen Alzubi, Szilveszter Kovacs, and Mouhammd Alkasassbeh. Evaluation of machine learning algorithms for intrusion detection system. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*, pages 000277–000282. IEEE, 2017. https://doi.org/10.1109/sisy.2017.8080566

[18] Mouhammad Alkasassbeh and Mohammad Almseidin. Machine learning methods for network intrusion detection. *Icccnt 2018 - The 20TH International Conference On Computing, Communication And Networking Technologies*, 2018.

[19] Ibrahim Obeidat, Nabhan Hamadneh, Mouhammd Alkasassbeh, Mohammad Almseidin, and Mazen AlZubi. Intensive pre-processing of kdd cup 99 for network intrusion classification using machine learning techniques. 2019. https://doi.org/10.3991/ijim.v13i01.9679

[20] Mouhammd Alkasassbeh, Ghazi Al-Naymat, AB Hassanat, and Mohammad Almseidin. Detecting distributed denial of service attacks using data mining techniques. *International Journal of Advanced Computer Science and Applications*, 7(1):436–445, 2016. https://doi.org/10.14569/ijacsa.2016.070159

[21] Mouhammad Alkasassbeh. An empirical evaluation for the intrusion detection features based on machine learning and feature selection methods. *Journal of Theoretical and Applied Information Technology*, 95(22), 2017.

[22] Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: introduction and review. *Journal of biomedical informatics*, 2018. https://doi.org/10.1016/j.jbi.2018.07.014

[23] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2018.

10 Authors

Mohammed Almseidin works at the Department of Information Technology in, University of Miskolc at Miskolc in Hungary alsaudi@iit.uni-miskolc.hu

AlMaha Abu Zuraiq, Alm20178050@std.psut.edu.jo Princess Sumaya University for Technology, Amman, Jordan

Mouhammd Al-kasassbeh works at Princess Sumaya University for Technology in Amman at Jordan m.alkasassbeh@psut.edu.jo

Nidal Alnidami works in National Information Technology Center in Amman situated at Jordan Nidal.n@nitc.gov.jo

Article submitted 2019-07-31. Resubmitted 2019-09-15. Final acceptance 2019-09-21. Final version published as submitted by the authors.