Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection

Article *in* International Journal of Advanced Computer Science and Applications · January 2017

DOI: 10.14569/IJACSA.2017.080910

Waleed Ali

29 PUBLICATIONS **577** CITATIONS

SEE PROFILE

Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection

Waleed Ali
Department of Information Technology
Faculty of Computing and Information Technology, King Abdulaziz University
Rabigh, Kingdom of Saudi Arabia

***Abstract*—The problem of Web phishing attacks has grown considerably in recent years and phishing is considered as one of the most dangerous Web crimes, which may cause tremendous and negative effects on online business. In a Web phishing attack, the phisher creates a forged or phishing website to deceive Web users in order to obtain their sensitive financial and personal information. Several conventional techniques for detecting phishing website have been suggested to cope with this problem. However, detecting phishing websites is a challenging task, as most of these techniques are not able to make an accurate decision dynamically as to whether the new website is phishing or legitimate. This paper presents a methodology for phishing website detection based on machine learning classifiers with a wrapper features selection method. In this paper, some common supervised machine learning techniques are applied with effective and significant features selected using the wrapper features selection approach to accurately detect phishing websites. The experimental results demonstrated that the performance of the machine learning classifiers was improved by using the wrapper-based features selection. Moreover, the machine learning classifiers with the wrapper-based features selection outperformed the machine learning classifiers with other features selection methods.**

Keywords—Phishing website; machine learning; wrapper features selection

I. Introduction

In recent years, the Web has evolved explosively due to the availability of numerous services such as online banking, entertainment, education, software downloading and social networking. Accordingly, a huge volume of information is downloaded and uploaded constantly to the Web. This gives opportunities for criminals to hack important personal or financial information, such as usernames, passwords, account numbers and national insurance numbers. This is called a Web phishing attack, which is considered as one of the major problems in Web security [1], [2].

In a Web phishing attack, phishing websites are created by the attacker, which are similar to the legitimate websites to deceive Web users in order to obtain their sensitive financial and personal information. The phishing attack is initially performed through clicking a link received within emails. Victims receive an email containing a link to update or validate their information. If this link is clicked by the target victims, the Web browser will redirect them to a phishing website that appears similar to the original website. The attackers can then steal the important information of the web users, since they are asked to input the sensitive information on the phishing website. Eventually, the attackers can carry out financial theft after phishing occurs [3]-[5].

Due to the inevitability of phishing websites targeting online businesses, banks, Web users, and government, it is essential to prevent Web phishing attacks in the early stages. However, detection of a phishing website is a challenging task, due to the many innovative methods used by phishing attackers to deceive web users [6]-[8].

The success of phishing website detection techniques mainly depends on recognizing phishing websites accurately and within an acceptable timescale [2], [4]. Many conventional techniques based on fixed black and white listing databases have been suggested to detect phishing websites. However, these techniques are not efficient enough, since a new website can be launched within few seconds. Therefore, most of these techniques are not able to make an accurate decision dynamically on whether the new website is phishing or not. Hence, many new phishing websites may be classified as legitimate websites [1], [2], [6]-[8].

As alternative solutions to the conventional phishing website detection techniques, some intelligent phishing detection methods have been developed and suggested in order to effectively predict phishing websites. In recent years, the intelligent phishing website detection solutions based on supervised machine learning techniques have become common, which are smart and more adaptive to the Web environment compared to the conventional phishing website detection methods.

He et al. [6] proposed a phishing pages detection scheme using a support vector machine based on 12 features. Barraclough et al. [7] utilized a Neuro-Fuzzy scheme with five inputs (Legitimate site rules, User-behavior profile, PhishTank, User-specific sites, Pop-Ups from emails) to detect phishing websites with high accuracy in real-time. Mohammad et al. [9] suggested rule-based data mining classification techniques with 17 different features to distinguish phishing from legitimate websites. Mohammad et al. [4] proposed an intelligent model for predicting phishing attacks based on self-structuring neural networks. Abdelhamid et al. [1] developed an approach called Multi-Label Classifier based Associative Classification (MCAC) to detect phishing websites. In addition, neural network (NN), support vector machine, (SVM), naïve Bayes (NB), decision tree, random forest and other classification techniques have been employed in detection of phishing websites [5], [8], [10]-[13].

In these intelligent approaches, the discriminating features, which play an important role in enhancing the performance of the classifier, are selected manually [14] or using statistical methods [1], [15] to help in distinguishing the phishing websites from legitimate ones. As these approaches do not take into consideration any classifier to evaluate the significance of features, some features may be useful in an inductive classifier but not significant in other classifiers.

Unlike the previous studies, the most influential features are selected in this paper using the wrapper-based features selection method, which uses the classifier for evaluating the significance of features to be utilized in precisely predicting website phishing. More significantly, the most common supervised machine learning techniques are validated and evaluated in order to investigate the most effective intelligent machine learning techniques that can be used to detect phishing websites. Furthermore, the performance of each of these intelligent phishing website detection techniques with the wrapper-based features selection method is comprehensively discussed and compared in this paper.

The remaining parts of this paper are organized as follows. Section II introduces the background and related works to phishing websites detection. Wrapper features selection is presented in Section III, while Section IV describes briefly the machine learning techniques used in this study. In Section V, a methodology for phishing website detection based on supervised machine learning classifiers with wrapper features selection is illustrated and explained in details. The results of phishing website detection based on supervised machine learning classifiers with wrapper features selection are presented and discussed in Section VI. Finally, the works presented in this paper are concluded and summarized in Section VII.

II. Phishing Websites Detection

A. Phishing Websites

The number of phishing attacks has been growing considerably in recent years and is considered as one of the most dangerous modern internet crimes, which may lead individuals to lose confidence in e-commerce. Consequently, it has a tremendous negative effect on online commerce, marketing efforts, organizations' incomes, relationships, customers, and overall business operations [1], [2], [6]-[8].

In order to steal the user identities and credentials, the phisher usually develops a fake replica of the original website, which is similar in appearance to the original website. Subsequently, the phisher sends a forged email to victims in order to criminally perform fraudulent financial transactions on behalf of the web users.

Basically, the phisher constantly sends emails to many Web users including hyperlinks to the forged website in as attempt to deceive Web users. As most of Web users are not specialists in Internet security, they follow the link in the phishing email and log in to the fake website. Thus, they would simply fall into the phishing website trap and credentials information such as account information, passwords, and credit card numbers would fall under the control of the phisher. Fig. 1 illustrates the steps of the phishing process [3]-[5].

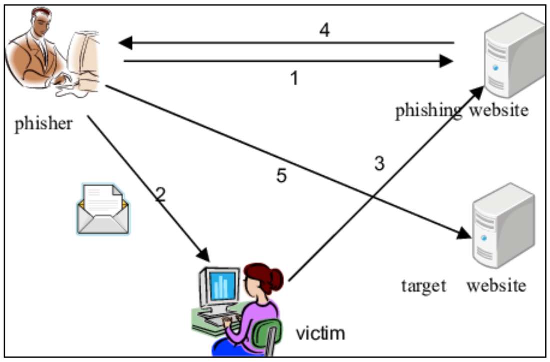

Fig. 1. Steps of Web phishing process.

B. Techniques for Phishing Websites Detection

It is a vital step to detect the phishing websites early, in order to warn the users against sending their sensitive information through these fake websites. The effectiveness and accuracy of phishing websites detection techniques are crucial for the success of the phishing detection mechanisms [2], [4].

Several conventional techniques for detecting phishing websites have been suggested in the literature to cope with the Web phishing problems. However, the decision regarding the phishing websites in these techniques was predicted imprecisely [1], [2], [6]-[8]. This led to most of the legitimate websites being classified as phishing. In general, two popular approaches are used to detect the phishing websites:

- *Blacklist and whitelist based approach*: This approach is based on the blacklist or whitelist to verify if the currently visited website is either a phishing or legitimate website respectively. The main drawback of the blacklist and whitelist based approach is that it cannot distinguish the newly created phishing websites from legitimate websites.

- *Intelligent heuristics-based approach*: In this approach, some features of websites are collected and evaluated to select the most influential website features, which play an important role in detecting the phishing websites. The selected significant features of many websites can be utilized as training dataset. Then, the machine learning techniques are trained based on the prepared training dataset in order to effectively classify the websites as either phishing or legitimate. After verification of the performance, the trained classifiers have the generalization ability to correctly detect the new phishing websites in the real implementation, which may have been unseen in the training phase. Therefore, unlike the blacklist and whitelist based approach, the intelligent heuristics-based approaches are able to effectively detect newly created phishing websites [5], [8], [10]-[13].

III. Wrapper Features Selection

It is impractical to use all the available features to train machine learning classifiers. In machine learning, the selection of discriminating features can play an important role in enhancing the performance of the classifier. In addition to highlighting the importance of features, the features selection establishes a trade-off between the adequacy of the learned model and the number of selected features [16]-[17].

In features selection, there are two main categories used for features evaluation: wrapper-based evaluation and filter-based evaluation [18]-[19]. In the filter-based evaluation techniques, the significant features are selected based on statistical measures to evaluate and weigh the features without classification information. In the filter-based evaluation techniques, the high dependency on target class and less inter-correlation are used to select the important features in order to be utilized later in a classification or a regression model. Information gain (IG) is one of the most common filter-based techniques, which measures how common a feature is in a class compared to all other classes.

Unlike filter-based evaluation, wrapper-based strategies use an inductive classifier to evaluate the significance of the features subset. The inductive classifier is separately trained with many subsets to eliminate the redundant and irrelevant features. The score for each subset is then given based on the classification error rate of the classifier model. In the wrapper-based evaluation, a search algorithm is used to search through the space of possible features and evaluate each subset by running a model on the subset. The wrapper-based evaluation techniques are usually computationally intensive for large dataset, since they train a new classifier for each subset. However, the wrapper-based techniques usually provide the most influential features set and achieve the best performance for that particular type of classifier [18]-[19]. Therefore, the wrapper-based evaluation is used in this study to enhance the performance of machine learning classifiers.

IV. Supervised Machine Learning

Machine learning concentrates on developing the computational algorithms that reason and induce patterns and rules from externally supplied instances and priori data in order to produce general models, which are able to make predictions about future instances. The machine learning is called supervised if known labels are given with instances in the training phase, whereas instances are unlabeled in unsupervised machine learning. Many supervised learning algorithms have been successfully employed in different real applications [19]-[20]. However, this section focuses on some popular machine learning techniques such as back-propagation neural network (BPNN), radial basis function network (RBFN), support vector machine (SVM), naïve Bayes classifier (NB), decision tree (C4.5), random forest (RF), and k-Nearest neighbor (kNN).

A. Back-Propagation Neural Network (BPNN)

Back-propagation neural networks (BPNNs) are the most well known algorithms in neural network models, which are effectively applied in many real classification and prediction problems. The learning in BPNNs is carried out in two phases: the forward pass and backward pass phases. In the forward pass phase, a training input pattern is presented to the input layer of the network. The input pattern is propagated from layer to layer in the network until the output is produced. In the backward pass phase, the output is compared with the desired output of pattern in order to compute an error. Accordingly, the error is propagated backward through the network from the output to the input layers and the weights are adjusted to minimize the error.

B. Radial Basis Function Network (RBFN)

A radial basis function network (RBFN) is a specific type of neural networks that uses radial basis functions as activation functions. The architecture of RBFN consists of a three-layer feedback network: an input layer, a hidden layer and an output layer. In RBFN, a radial activation function is executed in each hidden unit, while a weighted sum of the outputs of hidden units is implemented for each output unit. The learning of RBFN is usually carried out through two stages. In the first stage, clustering algorithms are utilized to determine the centers and widths of the hidden layer. In the second stage, Least Mean Squared (LMS) or Singular Value Decomposition (SVD) algorithms are used to optimize the weights connecting the hidden layer with the output layer.

C. Support Vector Machine (SVM)

The support vector machine (SVM) is one of the most well-known and robust supervised machine learning techniques, which has been utilized effectively in many science and engineering applications. SVM is based on maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances to reduce an upper bound on the expected generalization error. Some instances of the training dataset called support vectors, which are close to the separating hyperplane and provide the most useful information for classification, are utilized in SVM training. In addition, an appropriate kernel function is used to transform the data into a high-dimension to use linear discriminate functions.

D. Naïve Bayes Classifier (NB)

Naive Bayes network (NB) is a very simple Bayesian network, which includes directed acyclic graphs with just a single parent (representing the class label) and some children (corresponding to features). NB ignores any correlation among the attributes and assumes that all the attributes are conditionally independent given the class label. In order to assign a class to an observed instance, NB is based on probability estimations, called a posterior probability. The classification decision is expressed as estimating the class posterior probabilities given a test example. The most probable class is assigned to that test example.

E. Decision Tree (C4.5) and Random Forest (RF)

One of the most broadly utilized and practical strategies for inductive induction are the decision tree. In the decision tree, the instances are classified by sorting them based on evaluation of feature values. A node in the tree corresponds to a feature in an instance to be classified. Each branch of the tree represents a value that the node can predict. The C4.5 algorithm [21] is the most common algorithm among the other decision trees. In the C4.5 decision tree, the tree can also be represented as set of if-then rules to improve readability and interpretation.

Random Forest (RF) is another popular decision tree, which can be used for both classification and regression. RF is an ensemble of a number of decision trees independently trained on selected training datasets. The classification information is then determined by voting among all the trained decision trees. Therefore, Random Forest usually achieves a better classification accuracy compared to a single tree.

F. K-Nearest Neighbour (kNN)

K-Nearest Neighbour (kNN) is a non-parametric supervised machine method, which has been employed successfully in many real classification and regression issues. kNN supposes that the instances within a training dataset are usually available in closeness to other instances that have similar features. In other words, the class of the k closest neighbour instances is utilized to detect the classification decision of any instance.

V. METHODOLOGY

Fig. 2 illustrates the methodology of phishing website detection based on supervised machine learning classifiers with wrapper features selection.

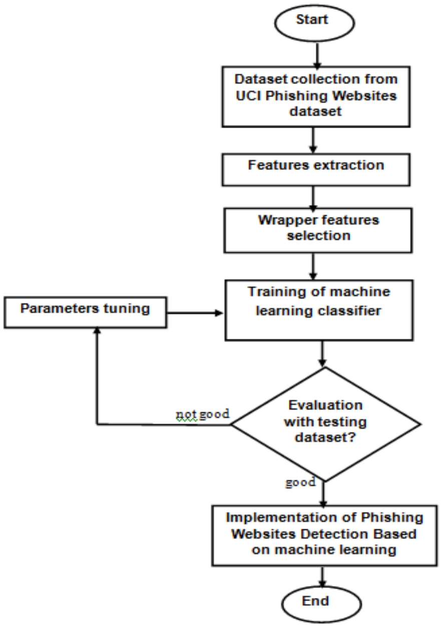
Start
Dataset collection from UCI Phishing Websites dataset
Features extraction
Wrapper features selection
Training of machine learning classifier
Parameters tuning
Evaluation with testing dataset?
not good
good
Implementation of Phishing Websites Detection Based on machine learning
End

Fig. 2. A methodology of phishing website detection based on machine learning classifiers with wrapper features selection.

As shown in Fig. 2, five steps are required to be accomplished in order to detect the phishing website: dataset collection, features extraction, features selection, training of machine learning classifiers, and evaluation of machine learning classifiers.

A. Data Collection

The dataset of phishing and legitimate websites were collected from the UCI Machine Learning Repository [22], which is freely available for use. This dataset consists of 4898 phishing websites and 6157 legitimate websites which were used to extract several website features. The phishing websites dataset was collected essentially from Phishtank archive, MillerSmiles archive, and Googleâ€™s searching operators.

B. Features Extraction

Several features can be extracted from a website to distinguish phishing websites from legitimate ones. The extracted features' goodness is crucial for the success of the phishing website detection mechanisms.

In the phishing websites dataset available in the UCI Machine Learning Repository [22], 30 key features of websites that have been proven in [14] to be efficient and influential in predicting the phishing and legitimate websites. Table 1 summarizes the key features that can contribute in the effective prediction of phishing websites. More details about these features and their meaning are given in [14].

TABLE I. THE KEY FEATURES THAT CAN CONTRIBUTE IN THE EFFECTIVE PREDICTION OF THE PHISHING WEBSITES

Feature Group	**Features Names**
Address bar -based features	Using the IP Address, Long URL to Hide the Suspicious Part, Using URL Shortening Services "TinyURL", URL's having "@" Symbol, Redirecting using "//",Adding Prefix or Suffix Separated by (-) to the Domain, Sub Domain and Multi Sub Domains, HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer), Domain Registration Length, Favicon, Using Non-Standard Port , and The Existence of "HTTPS" Token in the Domain Part of the URL
Abnormal-based features	URL of Anchor, Links in <Meta>, <Script> and <Link> tags, Server Form Handler (SFH), Submitting Information to Email and Abnormal URL
HTML and JavaScript-based features	Website Forwarding, Status Bar Customization, Disabling Right Click, Using Pop-up Window, and IFrame Redirection,
Domain-based features	Age of Domain, DNS Record, Website Traffic, Page Rank, Google Index, Number of Links Pointing to Page, and Statistical-Reports Based Feature

C. Wrapper Features Selection

The features selection step aims to select a subset of significant features from the phishing websites dataset that can efficiently describe the website dataset, and decrease the computation time, as well as reducing the noise and irrelevant features, which may negatively affect the performance of machine learning techniques.

As mentioned in Section III, the wrapper-based features selection usually produces the best performing features set for that particular kind of classifier. Therefore, in this paper, the wrapper-based features selection is used to select the most influential features, which can be utilized to distinguish phishing from legitimate websites.

In the wrapper-based features selection, the machine learning classifier is considered the main part used to evaluate the goodness of all the selected features subsets, as shown in Fig. 3. The wrapper method conducts a search in space of all the possible features subsets and utilizes a machine learning classifier as an evaluation function of the features subsets. The best features subset is decided based on the highest evaluation to be used in the training of the machine learning classifier.

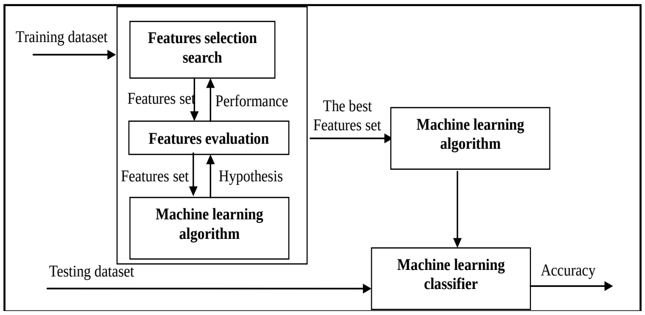

Fig. 3. The wrapper features selection approach used for predicting the phishing websites.

D. Training of Machine Learning Classifiers

The training in supervised machine learning is also known as inductive learning or classification. It is the task of inferring a function (classifier) from a supervised (labeled) training phishing websites dataset. A supervised learning algorithm analyzes the training phishing websites dataset and produces a classifier, which can predict the correct class for unseen dataset and effectively detect the newly created phishing websites.

Once the significant features are selected properly using the wrapper approach, the machine learning techniques can be trained in order to correctly classify the website, as either a phishing or legitimate website.

As shown in Fig. 4, the selected significant features are used as inputs of the machine learning algorithm, which analyzes and processes them to produce an output representing the class of the website, either a phishing or a legitimate website. If the output is different from the desired output, an error will be calculated and then the machine learning classifier will be iteratively retrained till the actual output becomes closer to the target output. The goal of the training phase is to correctly map inputs to outputs in order to minimize the error between the actual output and the target output.

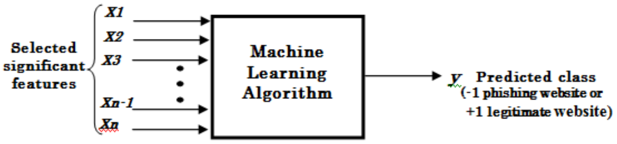
Selected significant features
X1
X2
X3
Xn-1
Xn
Machine Learning Algorithm
Y Predicted class (-1 phishing website or +1 legitimate website)

Fig. 4. Inputs and output of the machine learning classifiers used for predicting the phishing website.

E. Evaluation of Machine Learning Classifiers

In the training phase, a learning algorithm uses the training data to generate a classification model (classifier). In testing phase, the learned classifier is evaluated using the testing dataset to get the correct classification accuracy. If the correct classification accuracy for the testing dataset is acceptable, the trained classifier can be used in real-world applications. Otherwise, some further procedures can be carried out to improve the classification accuracy; for example, parameters tuning or more processing of the data. If the accuracy cannot be improved, another machine learning algorithm can be implemented in order to select the most efficient machine learning algorithm.

In this study, n-fold cross-validation was used to evaluate the machine learning classifiers used for predicting the phishing websites. In n-fold cross-validation, the dataset are divided into n equal-size disjoint datasets. Each dataset is then used as the testing dataset, while the remaining n-1 datasets are combined and used as the training dataset to train a classifier. This process is then run n times. The accuracy is computed for each run. Thus, the final accuracy of learning from this dataset is the average of the n accuracies for all runs.

In addition to the correct classification rate (CCR), other important measures extracted from a confusion matrix (see Table 2) can be calculated in order to accurately evaluate the machine learning classifiers. As described in Table 3, the performance of machine learning classifiers used in phishing website detection can also be evaluated using additional accurate measures such as sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), and geometric mean (GM).

TABLE II. CONFUSION MATRIX FOR A TWO-CLASS PROBLEM

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (*TP*)	False Negative (*FN*)
Actual Negative	False Positive (*FP*)	True Negative (*TN*)

TABLE III. THE MEASURES USED FOR EVALUATING PERFORMANCE OF MACHINE LEARNING CLASSIFIERS

Measure name	**Formula**
Correct Classification Rate	$CCR = \frac{TP+TN}{TP+FP+FN+TN}$ (%)
True Positive Rate	$TPR = \frac{TP}{TP+FN}$
True Negative Rate	$TNR = \frac{TN}{TN+FP}$
Geometric Mean	$GM = \sqrt{TPR*TNR}$

VI. RESULTS AND DISCUSSION

The phishing websites dataset was obtained from UCI Machine Learning Repository [22] to evaluate the supervised machine learning classifiers used in phishing websites detection. In the phishing websites dataset, 4898 phishing websites and 6157 legitimate websites were gathered and used for training and evaluating the supervised machine learning classifiers used in phishing websites detection.

Table 4 provides the important information about the phishing websites dataset including the number of attributes, number of instances (websites), and class distribution. For each website, a website pattern vector was extracted and formed to be used as an instance in the training dataset, which has 30 important features for that website. The website pattern vector corresponding to the legitimate website is assigned to a class with label +1 and the phishing website is assigned to a class with label -1.

TABLE IV. THE DESCRIPTION OF THE PHISHING WEBSITES DATASET

Description	**Value**
#Attributes	30
# Instances(Websites)	11055
# Phishing Websites	4898
Phishing Websites Percentage (%)	44 %
# Legitimate Websites	6157
Legitimate Websites Percentage (%)	56 %

The performances in terms of correct classification rate (CCR), true positive rate (TPR), true negative rate (TNR), and geometric mean (GM) of BPNN, RBFN, SVM, NB, C4.5,kNN and RF were compared together and discussed before and after the wrapper-based features selection.

In this study, five-fold cross validation was implemented using WEKA software in order to evaluate the performances of machine learning classifiers with the wrapper-based features selection in phishing websites detection. In addition, their performances were compared with two other popular features selection methods: Information Gain (IG) that was used in [15] and Principal Component Analysis (PCA).

Fig. 5 shows a comparison of the CCRs of BPNN, RBFN, SVM, NB, C4.5, kNN and RF before and after the features selection methods were applied for the phishing websites dataset in the testing phase using five-fold cross-validations.

As can be seen in Fig. 5, BPNN, kNN and RF achieved the best CCR while RBFN and NB achieved the worst CCR for detecting the phishing websites. Fig. 5 compares the performance in terms of CCR obtained by the machine learning classifiers with the wrapper-based features selection against their performances with PCA and IG features selection methods. It is clear from Fig. 5 that the CCRs of most of the machine learning classifiers were improved by using the wrapper-based features selection. Although the wrapper-based features selection has low impact on NB, C4.5, kNN and RF, the machine learning classifiers with wrapper-based features selection were able to maintain the CCRs using only fewer features. The experimental results in Fig. 5 also demonstrate that the machine learning classifiers with the wrapper-based features selection outperformed the machine learning classifiers with PCA and IG features selection methods.

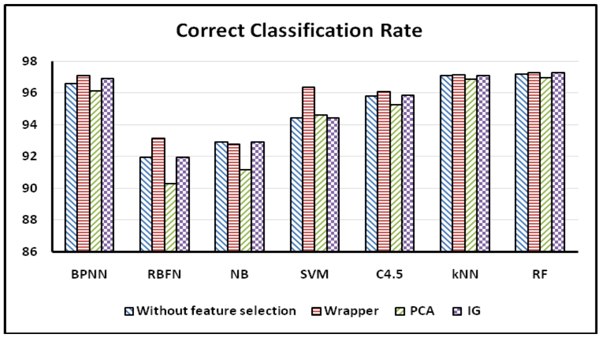

Fig. 5. A comparison of CCR between the machine learning classifiers with features selection methods.

In addition to the CCR measure, Table 5 shows the performance in terms of TPR, TNR, and GM of the supervised machine learning classifiers with the wrapper-based features selection used to detect the phishing websites. In Table 5, the best and the worst values of the measures are highlighted in bold font and underline font, respectively.

TABLE V. PERFORMANCE MEASURES OF THE MACHINE LEARNING CLASSIFIERS WITH FEATURES SELECTION METHODS

	Measures	**Without features selection**	**With features selection**		
			Wrapper	***PCA***	***IG***
BPNN	***TPR***	0.966	**0.971**	0.961	0.969
	TNR	0.963	**0.969**	0.958	0.967
	GM	0.964	**0.970**	0.959	0.968
RBFN	***TPR***	0.919	**0.931**	0.903	0.919
	TNR	0.917	**0.926**	0.902	0.917
	GM	0.918	**0.928**	0.902	0.918
NB	***TPR***	**0.929**	0.927	0.911	**0.929**
	TNR	**0.924**	0.922	0.907	**0.924**
	GM	**0.926**	0.924	0.909	**0.926**
SVM	***TPR***	0.944	**0.964**	0.946	0.944
	TNR	0.94	**0.962**	0.942	0.94
	GM	0.942	**0.963**	0.944	0.942
C4.5	***TPR***	0.958	**0.961**	0.952	0.959
	TNR	0.955	**0.958**	0.949	0.956
	GM	0.956	**0.959**	0.950	0.957
kNN	***TPR***	**0.971**	**0.971**	0.969	**0.971**
	TNR	0.969	**0.97**	0.966	0.969
	GM	**0.970**	**0.970**	0.967	**0.970**
RF	***TPR***	0.972	**0.973**	0.969	**0.973**
	TNR	0.969	**0.97**	0.967	**0.97**
	GM	0.970	**0.971**	0.968	**0.971**

Table 5 obviously shows that most of the supervised machine learning classifiers with the wrapper-based features selection accomplished a better performance compared to the others. In particular, the supervised machine learning classifiers with the wrapper-based features selection achieved the best TPR, TNR, and GM. This was due to the fact that the wrapper-based features selection utilizes a machine learning classifier as evaluation function to evaluate the goodness of all the selected features subsets.

On the other hand, the supervised machine learning classifiers with the PCA features selection method achieved the worst TPR, TNR, and GM for the phishing websites dataset. Table 5 also shows that the classifiers with IG features selection method had a somewhat better performance when compared to the classifiers with the PCA features selection method.

VII. CONCLUSION AND FUTURE WORKS

In this paper, the wrapper-based features selection method was used for selecting the most significant features to be utilized in predicting the phishing websites accurately. Accordingly, BPNN, RBFN, SVM, NB, C4.5, kNN and RF were applied with these significant features selected using the wrapper features selection in order to detect the phishing websites. The experimental results showed that BPNN, kNN and RF achieved the best CCR while RBFN and NB achieved the worst CCR for detecting the phishing websites. More significantly, the machine learning classifiers using wrapper-based features selection outperformed the machine learning classifiers with PCA and IG features selection methods. The machine learning classifiers based on wrapper-based features selection accomplished the best performance while these classifiers with PCA features selection method achieved the worst performance in terms of CCR, TPR, TNR, and GM.

Although the wrapper-based features selection method may consume more time and require extra computational overhead with some classifiers, the wrapper-based features selection method is usually used once in order to provide the most influential features. The machine learning classifiers should then be retrained with these selected features regularly in the update process in order to improve the efficiency and adaptability of the intelligent phishing websites detection approaches. Furthermore, the wrapper-based features selection can be used with ensemble learning to improve the performance of the intelligent phishing website detection techniques.

References

[1] N. Abdelhamid, A. Ayesh, F. Thabtah, “Phishing detection based associative classification data mining,” Expert Systems with Applications, vol. 41(13), pp. 5948-5959, 2014.

[2] R. M. Mohammad, F. Thabtah, L. McCluskey, “Tutorial and critical analysis of phishing websites methods,” Computer Science Review, vol. 17, pp. 1-24, 2015.

[3] H. Huang, S. Zhong, J. Tan, “Browser-side countermeasures for deceptive phishing attack,” Fifth International Conference on Information Assurance and Security IAS'09, vol. 1, pp. 352-355, IEEE, 2009.

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

[5] M. A. U. H. Tahir, S. Asghar, A. Zafar, S. Gillani, “A Hybrid Model to Detect Phishing-Sites Using Supervised Learning Algorithms,” International Conference on Computational Science and Computational Intelligence (CSCI), pp. 1126-1133, IEEE, 2016.

[6] M. He, S.J. Horng, P. Fan, M.K. Khan, R.S. Run, J.L. Lai, R.J. Chen, Sutanto, “An Efficient Phishing Webpage Detector,” Expert Systems with Applications, vol. 38(10), pp. 12018-12027, 2011.

[7] P. A. Barraclough, M. A. Hossain, M. A. Tahir, G. Sexton, N. Aslam, “Intelligent Phishing Detection and Protection Scheme for Online Transactions,” Expert Systems with Applications, vol. 40(11), pp. 4697-4706, 2013

[8] H. H. Nguyen, D. T. “Nguyen, Machine learning based phishing web sites detection,” In AETA 2015: Recent Advances in Electrical Engineering and Related Sciences , pp. 123-131, Springer International Publishing, 2016.

[9] R. M. Mohammad, F. Thabtah, L. McCluskey, “Intelligent Rule-based Phishing Websites Classification, “ IET Information Security, vol. 8(3), pp. 153-160, 2014.

[10] V. S. Lakshmi, M. S. Vijaya, “Efficient prediction of Phishing Websites Using Supervised Learning Algorithms,” Procedia Engineering, vol. 30, pp. 798-805, 2012.

[11] J. James, L. Sandhya, C. Thomas, “Detection of Phishing URLs Using Machine Learning Techniques,” International Conference on Control Communication and Computing (ICCC), pp. 304-309, IEEE, 2013.

[12] M. Al-diabat, “Detection and Prediction of Phishing Websites using Classification Mining Techniques”, International Journal of Computer Applications, vol. 147(5), pp. 5-11, 2016.

[13] A. Hodzic, J. Kevric, A. Karadag, “Comparison of Machine Learning Techniques in Phishing Website Classification,” 2016.

[14] R. M. Mohammad, F. Thabtah, L. McCluskey, “An Assessment of Features Related to Phishing Websites Using An Automated Technique,” International Conference for Internet Technology and Secured Transactions, pp. 492-497, IEEE, 2012.

[15] I. Qabajeh, F. Thabtah, “An Experimental Study for Assessing Email Classification Attributes Using Feature Selection Methods,” 3rd International Conference on Advanced Computer Science Applications and Technologies (ACSAT), pp. 125-132, IEEE, 2014.

[16] H. Liu, J. Li, L. Wong, “A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns,” Genome informatics, vol. 13, pp. 51-60, 2002.

[17] Z. S. J. Hoare, "Feature selection and classification of non-traditional data: examples from veterinary medicine," University of Wales, 2007.

[18] R. Kohavi, G. H. John, “Wrappers for feature subset selection,” Artificial intelligence, vol. 97(1-2), pp. 273-324, 1997.

[19] G. Chandrashekar, F. Sahin, “A survey on feature selection methods,” Computers & Electrical Engineering, vol. 40(1), pp. 16-28, 2014.

[20] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and P. S. Yu, "Top 10 algorithms in data mining", Knowledge and Information Systems, vol. 14(1), pp. 1-37, 2008.

[21] Quinlan J.R., “C4.5: Programming for machine learning”, Morgan Kauffmann, 1993.

[22] UCI Machine Learning Repository: Phishing Websites Data Set. Retrieved May 9, 2016, from https://archive.ics.uci.edu/ml/datasets/Phishing+Websites