*Menoufia J. of Electronic Engineering Research (MJEER), Vol. 30, No. 1, Jan.2021*

# A Smart Model for Web Phishing Detection Based on New Proposed Feature Selection Technique

Mohamed A. El-Rashidy
*Computer Science and Engineering Department*
*Faculty of Electronic Engineering*
*Menoufia University*
*Egypt*
E-mail: mohamed.elrashedy@el-eng.menofia.edu.eg

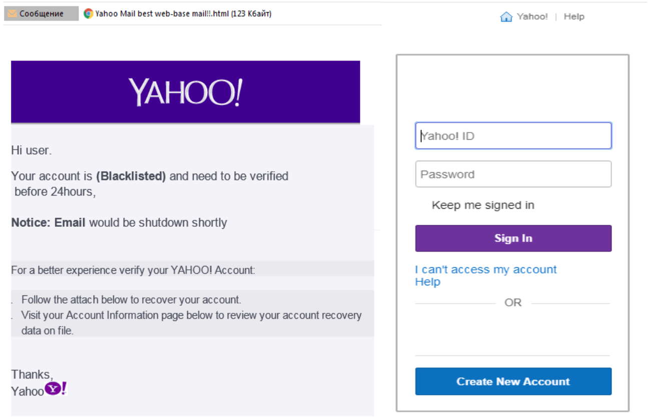
***Abstract*—Web-phishing attacks are one of the most serious cybercrime. It enables hackers to access the devices of many users and spy on their personal data such as passwords and credit card details. Hackers use a lot of tricks through the internet, which make users to share data, download files or open links that attack a computer. This research proposes meta-heuristic based approach to protect the internet users from the web-phishing. It consists of three phases, the first phase uses a new proposed method for evaluating and ranking the features of URL, HTML and JavaScript code, text, images and domain name of the web page. The second phase extracts the effective subset of the ranked features that achieves the highest classification accuracy of the web-phishing. The third phase constructs the Random forest classifier training by data features of the extracted subset. The new proposed method of the feature selection achieved the highest classification accuracy compared to the correlation feature selection, information gain, principle component analysis, and Relief feature selection algorithms. The proposed methodology of the web-phishing detection was also evaluated, it obtained the highest classification accuracy at the least possible time compared to the adaptive Neuro-fuzzy inference system.**

***Keywords*—*web-phishing, feature selection, classification, Random forest***

## I. Introduction

The importance and influential role of modern electronic devices and technologies has emerged in our lives, it has become an inseparable part of anyone's life, and associated with him in all his places from home to work or place of study. Many people cannot abandon the use of their personal phones whatever their surrounded circumstances. However, this excessive attachment of the electronic devices has penetrated the privacy of individuals, which makes it permissible and visible without any restrictions to preserve it. Therefore, while using electronic devices the privacy care has become an urgent necessity, while working in the internet environment. When most people realize that their privacy is becoming more vulnerable to hacking, manipulation or circumvention, they are keen to take appropriate means and ways to protect their data and information. Many internet users have faced threats from hacking into their computers and electronic accounts, which enters them into a cycle of conflicts to restore their stolen privacy.

Web-phishing is one of the most common hacker ways to get data of the internet users, because many internet users are unaware of the web-phishing nature and its undisclosed goals, which are hidden by criminals, such as sending an email randomly to get the account information, or to cite a name or address to win a cash prize requesting the financial statements of the recipient, or by relying on Pop-Up advertising that suddenly appear in front of the internet users, which his curiosity leads him to a website that downloads malicious software to copy all his personal information during the web browsing. Web-phishing has different several techniques that deceive the internet users, such as the existence of “@” symbol within the URL that causes the browser to block the preceded address of the “@” symbol and the authentic address follows the “@” symbol, using “//” symbol in the URL path that redirects the user to another website, adding letters separated by (-) to the beginning or the ending of the domain name. For example, http://www.yah-oo.com/, which the internet user feels that he is browsing with the real web site as shown in Fig. 1, script language of the spoofed web pages may be used mail () function to submit user information to a server for processing. Iframe of HTML tag may be used to show an extra web page onto the displayed page at that time, which makes the browser to show a visual delineation [1].



Сообщение Yahoo Mail best web-base mail!!.html (123 Кбайт)

Hi user.

Your account is **(Blacklisted)** and need to be verified before 24hours,

**Notice: Email** would be shutdown shortly

For a better experience verify your YAHOO! Account:

. Follow the attach below to recover your account.
. Visit your Account Information page below to review your account recovery data on file.

Thanks,
Yahoo

Fig. 1. Spoofed E-mail to get the account information.

According to the report of the Anti-Phishing Working Group (APWG), there are 1,220,523 phishing attacks recorded in 2016, which increased 65% over 2015 [2], these techniques newly generated at all the time. In 2019 attacks increased 226% over Q3 of 2019 and 476% over 2018 [3]. Therefore, many different solution types of anti-phishing models have been proposed. There are three approaches to detect the web-phishing [4], the first approach is using a blacklist of URLs which reported from the internet users and organizations. It compares the required URL with the URLs on the blacklist to detect whether it is a Web-phishing or real web page [5]. The disadvantage of this approach that cannot involve all web-phishing pages, because the hackers of the web pages are constantly renewed. The second approach is using the similarity metric to identify the web-phishing pages, which searches about the similarity between the real and spoofed page. The similarity between the web pages is calculated according to how far the content in the web pages is similar. This approach achieves highest detection accuracy to detect the web-phishing, but it is not able to prevent it, because it takes too long to open the spoofed web page and search about the similarity between the real and spoofed web page, this time allows to download malicious software while opening the spoofed web page. The third approach is called meta-heuristic based approach, it analyses various web features such as address bar, HTML and JavaScript code, text, images and domain name features; to discover whether it is a spoofed or a real web page.

Most of the recent researches [6-8] that based on meta-heuristic approach has tended to increase the number of the web page features, which used in the analysis process to identify the web-phishing page. Also, the purpose of this increment is to enhance the accuracy of the spoofed web page detection. But at the same time, this increment came at the cost of the classifier approach time for detecting the spoofed pages, where the relationship between the number of the analyzed web features and the classification time to detect the phishing is a direct relationship. The more number of the web features leads to increase the classification time to detect the phishing, which allows for the opportunity of downloading malware during the process of analyzing the features of the web page. The least number of analyzed features leading to reduce the required time of detecting the phishing and prevent the possibility of downloading the malicious programs. Therefore, the other recent researches in [9-11] tended to search about new features to enhance the accuracy of phishing detection.

In this paper, the proposed methodology was developed to detect and prevent the possibility of the web-phishing page. It is based on a new proposed method of the feature selection and random forest classification algorithm, the new feature selection method is proposed to find the least number of the effective web page features for detecting and preventing the web-phishing pages with the highest detection accuracy. The new feature selection method of the proposed methodology leads to shrink the classification time of the random forest classifier for detecting the spoofed web pages, which prevents the possibility of the web-phishing by reducing the detection time. At the same time, it is taken into consideration the enhancement of the web-phishing detection accuracy. It consists of three phases, the first phase uses a new proposed method for evaluating and ranking the features of URL, HTML and JavaScript code, text, images and the web page domain name. The second phase extracts the effective subset of the ranked features that achieves the highest classification accuracy of the web-phishing. The third phase constructs the Random forest classifier training by data features of the extracted subset.

The proposed methodology was evaluated using Two benchmark datasets of the University of California Irvine [12] and University of Huddersfield [13]. The new proposed method of the feature selection achieved the highest classification accuracy compared to correlation feature selection, information gain, principle component analysis, and Relief feature selection algorithms. The proposed methodology was also evaluated and compared to Adaptive Neuro-Fuzzy Inference System (ANFIS) [6], which analyzed more features of text, frame and images of the web page to enhance its accuracy of the web-phishing detection, although the proposed methodology reduced the analyzed web features to shrink the required time of phishing detection, the proposed methodology achieved the highest accuracy of web-phishing detection compared to ANFIS using the least number of features.

The rest of the paper is structured as follows. In section 2, the literature review is presented. In section 3, the details of the proposed methodology phases will be explained. In section 4, the experimental results will be discussed. This research will be concluded in section 5.

## II. Literature Review

### A. *Blacklist Based Approach*

This approach has been developed as anti-phishing software, the blacklist of this software has been updated to known the web-phishing sites [14]. There are various researches investigated the problem based on this approach, Anti-Phishing Working Group (APWG) and Phish Tank proposed a model to record each reported URL after verifying it into the blacklist [15]. A web survey of the Net craft server records the visiting times, hosted country, hosted organization name, and risk rating of the web pages. The Net Craft Toolbar approach used this survey to detect the phishing attack [16]. Another proposed approach using Google's page rank value of each site as a list to detect the phishing pages [17].

### B. *Content Similarity Based Approach*

This approach identifies web-phishing page by inspecting the content similarity between the real and spoofed web pages, the content similarity between the two web pages is calculated according to how far the content in the web pages is similar [14]. Therefore, this approach opens the spoofed web page to compare the content similarity, which allows to download the malicious programs. There are various researches developed based on this approach, Term Frequency/Inverse document Frequency (TF-IDF) technique is developed to retrieve information to identify the spoofed website, it suffers from the required time of querying Google to classify the website [9]. Visual similarity based approach between two web pages was proposed, it depended on the content similarity of block level, web page layout, overall style, and frame [18]. Earth Mover’s Distance (EMD) algorithm is used to calculate the similarity of the web pages through measuring the distances of web images [19]. An anti-phishing approach was developed based on a visual cryptography technique. It generates two shares of images using cryptography approach, the first share of the image was stored during the user registration on a website, and the other part is uploaded to the site, it verifies the real website by comparing the location of the image of both shares during each login the website [20]. Another new model that is dependent on the website favicon was developed for identifying the web page, it depends on the search engine of Google search by image API [21]. A new approach dependent on fuzzy logic integrated with a data-mining system was developed to detect the spoofed web pages of the E-banking systems, it implemented six models to select text features of the web pages to analyze the content similarity and classify it [10]. A model of Neuro-fuzzy and fuzzy rules was proposed to identify real and spoofed web pages; it depends on text-based features to analyze the content similarity [22]. Cascading style sheet (CSS) method is used to measure the visual similarity of web pages in [23] to detect the web-phishing pages.

### *C. Meta-Heuristic Based Approach*

Meta-heuristic based approach uses a URL, HTML and JavaScript code, text, images and the web page domain name to detect the phishing. There are various researches developed based on this approach. Spoof Gurad was proposed as an anti-phishing browser plug-ins, it computes the spoof index value to classify it as phishing page or legitimate page, if the calculated value is more than a pre-defined threshold value, the page is identified as spoofed page [11]. A genetic algorithm is used to detect the phishing; it is based on a system of rules that is used to match the hyperlinks [24]. Various machine learning approaches were analyzed using phishing data set to detect the phishing, Random forest achieved the highest accuracy of 93%. Link Guard model was developed to analyze the features of the web page and URL similarity for detecting and preventing the phishing [8].Various classification techniques were suggested to classify the phishing E-mail [25]. Blend of artificial neural networks was introduced to analyze the text features of the news for discovering the fake news [26]. New custom rule-based algorithm was developed to analyze the sentiment score, user and text post features for detecting the fake tweets [27]. Multi-label Classifier based Associative Classification (MCAC) was proposed to generate the hidden knowledge rules between the web features, which contributed to improve the detection accuracy of the web-phishing [28]. Several deep learning systems in [29-32] were presented to detect the spoofed web pages basing on the multidimensional web features, the experimental results of these researches showed that the detection accuracy was enhanced. An adaptive neuro-fuzzy inference system was developed to classify the web-phishing, it is the first work introduced the best integration of the text, image and frame features to enhance the detection accuracy, which based on increasing the number of features for the enhancement [6].

## III. Proposed Methodology

The proposed methodology has been developed for detecting and preventing the web-phishing. It depends on analyzingthe features of URL, HTML and Java Script code,

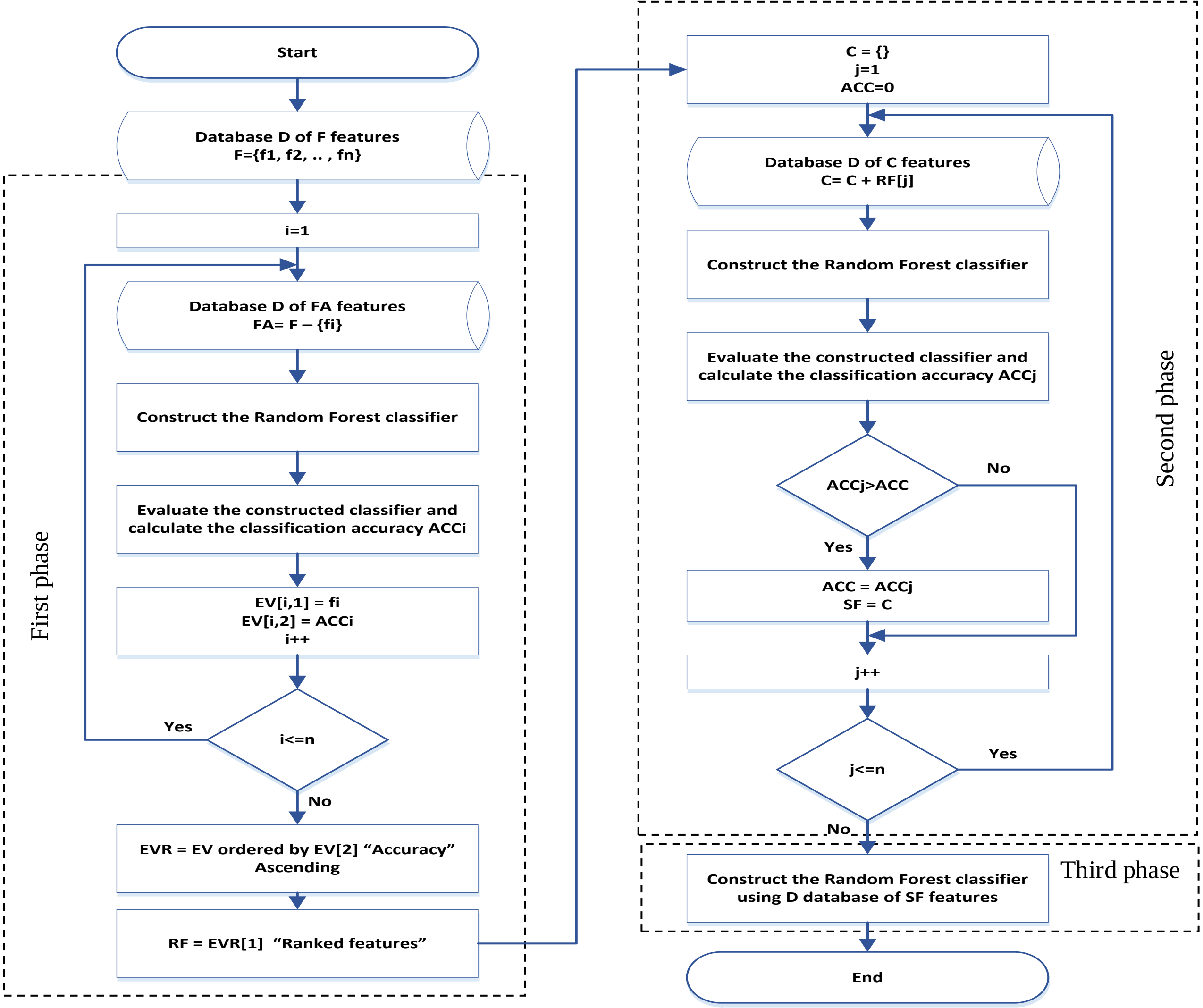

Start
Database D of F features
F={f1, f2, .. , fn}
i=1
Database D of FA features
FA= F – {fi}
Construct the Random Forest classifier
Evaluate the constructed classifier and calculate the classification accuracy ACCi
EV[i,1] = fi
EV[i,2] = ACCi
i++
Yes
i<=n
No
EVR = EV ordered by EV[2] "Accuracy" Ascending
RF = EVR[1] "Ranked features"
First phase
C = {}
j=1
ACC=0
Database D of C features
C= C + RF[j]
Construct the Random Forest classifier
Evaluate the constructed classifier and calculate the classification accuracy ACCj
ACCj>ACC
No
Yes
ACC = ACCj
SF = C
j++
j<=n
Yes
No
Second phase
Construct the Random Forest classifier using D database of SF features
Third phase
End

Fig. 2 Flowchart of the proposed methodology.

text, images and the web page domain name. It extracts the most significant of these features that contributes to improve the detection accuracy of the web-phishing.The feature selection process also excludes the other features that have a negative impact on the efficiency of preventing the web-phishing, or increase the required time of the web-phishing detection, which helps to increase the probability of downloading malware from the web-phishing pages. The main goal of the proposed methodology is to extract the least effective number of the features of URL, HTML and JavaScript code, text, images and domain name that can be used to detect the web-phishing with the highest accuracy and lowest possible time, which leads to reduce the probability of downloading malicious software from the web-phishing pages. It consists of three phases as shown in Fig. 2, the first phase uses a new proposed method for evaluating and ranking the features, the second phase extracts the effective subset of the ranked features that achieves the highest classification accuracy of the web-phishing, and the third phase constructs the Random forest classifier training by data features of the extracted subset.

The first phase of the proposed methodology is based on a new proposed method of the feature selection, it takes into its consideration the study and analysis of the absence impact for each database feature on the classification accuracy of the web-phishing, if the classification accuracy increased during the absence of the one of the database features, this indicates that this feature has a negative impact on the classification accuracy and that its absence has a positive effect on the web-phishing detection. Therefore, the features are evaluated based on the impact of their absence on the classification accuracy and ranked by the classification accuracy values during the absence of each feature from the smallest to the largest value. The features of the highest effective on the web-phishing detection are recognized by the lowest values of the classification accuracy during their absence in the classification process. The first phase calculates the classification accuracy, as in (1),*n* times as shown in Fig. 3.

$$ACC=(TP+TN)/(TP+TN+FP+FN) \qquad (1)$$

Where P is the positive case “spoofed web page”, N is the negative case“real web page”, TPis the number of true classified positive cases, TN is the number of true classified negative cases, FP is the number of false classified positive cases, FN is the number of false classified negative cases and *n* is the number of features. At each time, the Random forest classifier constructed by the database *D* of URL, webpage and images features, within the absence of*fi* attribute to analyze the impact of its absence on the classification accuracy. The features of the database *D* are ordered by the classification accuracy calculating in the absence of each feature from the smallest to the largest value in the *RF* array, which the feature with the lowest value of the classification accuracy during its absence has the highest effect in its presence on the accuracy of the web-phishing detection.

The second phase of the proposed methodology works to select the most effective subset of the ordered features achieving the highest classification accuracy of the web-phishing detection, it calculates the classification accuracy of the Random forest classifier training by the data of the first feature of the *RF* array. The calculation process of the classification accuracy is repeated *n* times. At each time, the next feature of the *RF* array is added to the training data of the classifier as shown in Fig. 4. The subset of data features that achieves the highest classification accuracy during the repetition process is selected to be the best subset of the features.

```
First phase: Features Evaluation
Input:Database D of F features F={$f_1, f_2, .., f_n$}.
Output:Ranked features  RF={$RF_1,RF_2, …………, RF_n$}.
begin
     1.  Feature evaluation array EV[n][2];
     2.  i=1;
     3.  Features set of the training data FA={$f_1, f_2, .., f_n$} - {$f_i$};
     4.  Train the Random Forestclassifier using Database D of FA
         Features set;
     5.  Compute the classification accuracy $ACC_i$ of the trained
         classifier;
     6.  EV[i][1]=$f_{i;}$
     7.  EV[i][2]=$ACC_{i;}$
     8.  i=i+1;
     9.  Go to step 3 until stopping criteria (i>n);
     10. Ascending order the EV[n][2] by the second column →
         EVR[n][2];
     11. RF=the first column of the EVR[n][2];
end
```

Fig. 2. Pseudo code of the first phase for the features evaluation process.

```
Second phase: Features Selection
Input:Database D of F features F={$f_1, f_2, .., f_n$}and Ranked features
RF={$RF_1,RF_2, …………, RF_n$}.
Output:Selected features SF={$SF_1,SF_2, …………, SF_m$}.
begin
     1.  ACC=0;
     2.  j=1;
     3.  C={};
     4.  Features set of the training data C=C+RF[j];
     5.  Train the Random Forestclassifier using Database D of C
         Features set.
     6.  Compute the classification accuracy $ACC_j$ of the trained
         classifier.
     7.  If $ACC_j$>ACC then
         7.1   ACC = ACCj;
         7.2   SF{}=C{};
     8.  End if;
     9.  j=j+1;
     10. Go to step 4 until stopping criteria (j>n).
end
```

Fig. 3. Pseudo code of the second phase for the features selection process.

The third phase of the proposed methodology is used to construct the Random forest classifier training by the data of the selected subset features as shown in Fig. 5.

```
Third phase: Classifier construction of the web-phishing detection
Input:Database D of F features F={$f_1, f_2, .., f_n$}and Selected features
SF={$SF_1,SF_2, …………, SF_m$}.
Output:CMWFD Classifier model of the web-phishing detection.
begin
     1.  Training database TD = Database D of SF features;
     2.  Train the Random Forest classifier using training database
         TDto obtain the classifier model CMWFD;
end
```

Fig. 4. Pseudo code of the Third phase for constructing the classifier of the web-phishing detection.

## IV. Results and Discussion

Two benchmark datasets of the University of California Irvine [12] and University of Huddersfield [13] were used and analyzed to evaluate the proposed methodology. The first database has thirty-five features of address bar, HTML and JavaScript code, images, text and domain with a total number of 11,056 records, the second database has the same features with a total number of 2,700 records, the detailed description of these databases available on UCI machine learning repository [12 and 13], the total number of the database records is 13,756.

Several experiments were conducted on the database to evaluate the proposed methodology using precision (positive predictive rate), recall (sensitivity), F Score average, Area Under the ROC Curve (AUC), the Matthew's Correlation Coefficient (MCC) and classification accuracy [33].

The database was partitioned into 10 independent training and testing datasets using 10-fold cross validation technique.Table 1 shows the comparison between different classification techniques to detect the web-phishing by analyzing the database features, these algorithms were trained by all the features of the database included address bar, HTML and JavaScript code, text, images and domain features, the results show that the Random forest algorithm achieved the highest classification accuracy compared to support vector machine, Naïve Bayes, K-nearest neighbors, regression, artificial neural network algorithms. Therefore, in this research, the proposed methodology is based on Random forest technique to construct the classification model of the web-phishing detection, and developed to enhance the classification time and accuracy of the Random forest classification technique, by reducing the number of database features using a new proposed feature selection model.

TABLE I. COMPARISON between different classification techniques for web-phishing detection.

| Classification Technique | Precision (%) | Recall (%) | F Score average | AUC | MCC | Accuracy (%) |
|---|---|---|---|---|---|---|
| Naïve Bayes | 92.18 % | 92.10 % | 0.921 | 0.980 | 0.854 | 92.15 % |
| K-Nearest Neighbors | 96.18 % | 96.14 % | 0.962 | 0.984 | 0.925 | 96.16 % |
| Support Vector Machine | 92.80 % | 92.76 % | 0.928 | 0.937 | 0.873 | 92.79 % |
| Random Forest | 96.38 % | 96.29 % | 0.963 | 0.987 | 0.926 | 96.33 % |
| Artificial Neural Network | 95.98 % | 94.96 % | 0.949 | 0.982 | 0.915 | 94.98 % |

The first phase of the proposed methodology was applied using a new proposed method of the feature selection, it analyses the absence impact of each feature of address bar, HTML and JavaScript code, text, images and domain features on the web-phishing detection accuracy. The classification accuracy was calculated thirty-fivetimes, at each time, the Random forest classifier constructed by the database features within the absence of one feature to analyze the impact of its absence on the classification accuracy. As shown in Fig. 6, when the feature of the “Https in URL” absented in the constructing process of the classifier, the classification accuracy of the classifier achieved 96.63 %,it is more than the achieved accuracy “96.33 %” of the random forest classifier constructing by all the features as shown in Table 1, this indicates to the “ Https in URL” feature has a negative impact on the classification accuracy and its absence has a positive impact on the phishing detection. Therefore, the achieved accuracy values during the absence of each feature is ranked from the smallest to the largest value as shown in Fig. 6, The features of the highest effective on the web-phishing detection were achieved the lowest values of the classification accuracy during their absence in the classification process.

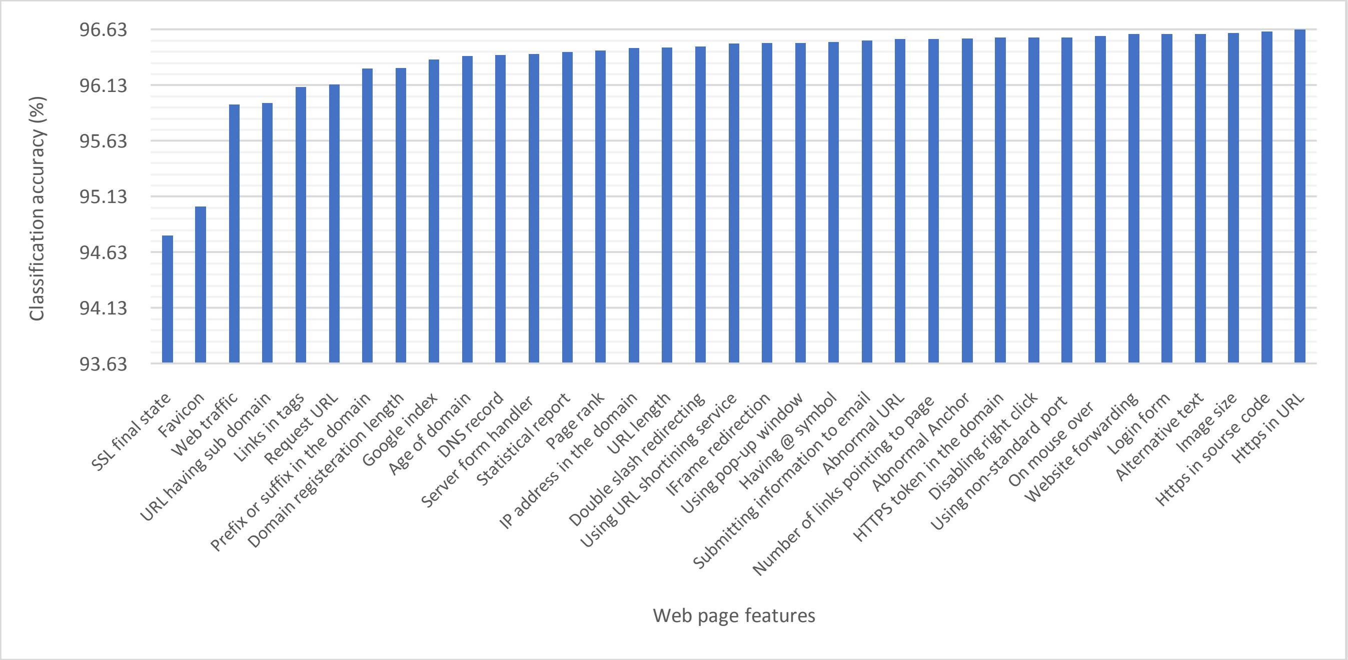


Fig. 5. Classification accuracy of Web-phishing detection for each absence of one feature from the all features.

The second phase of the proposed methodology was applied to select the most effective subset of the ordered features, it calculated the classification accuracy of the Random forest classifier training by the data of set 1 including the first feature of the ranked features. The calculation process of the classification accuracy was repeated thirty-fivetimes. At each time, the next feature of the ranked features was added to the subset. As shown in Fig. 7, The subset No. 28 that includes all the ordered features except the last seven features, achieved the highest classification accuracy during the repetition process. Therefore, it was selected to be the best subset of the selected features.

The third phase of the proposed methodology was applied to construct the Random forest classifier training by the data of the selected subset of the features. The classification accuracy of the selected subset of the features was compared with different feature selection techniques to evaluate it. As shown in Table 2, the new proposed feature selection model achieved the highest classification accuracy compared to Correlation Feature Selection (CFS), Information Gain (IG), Principle Component Analysis (PCA), and Relief feature selection algorithms.

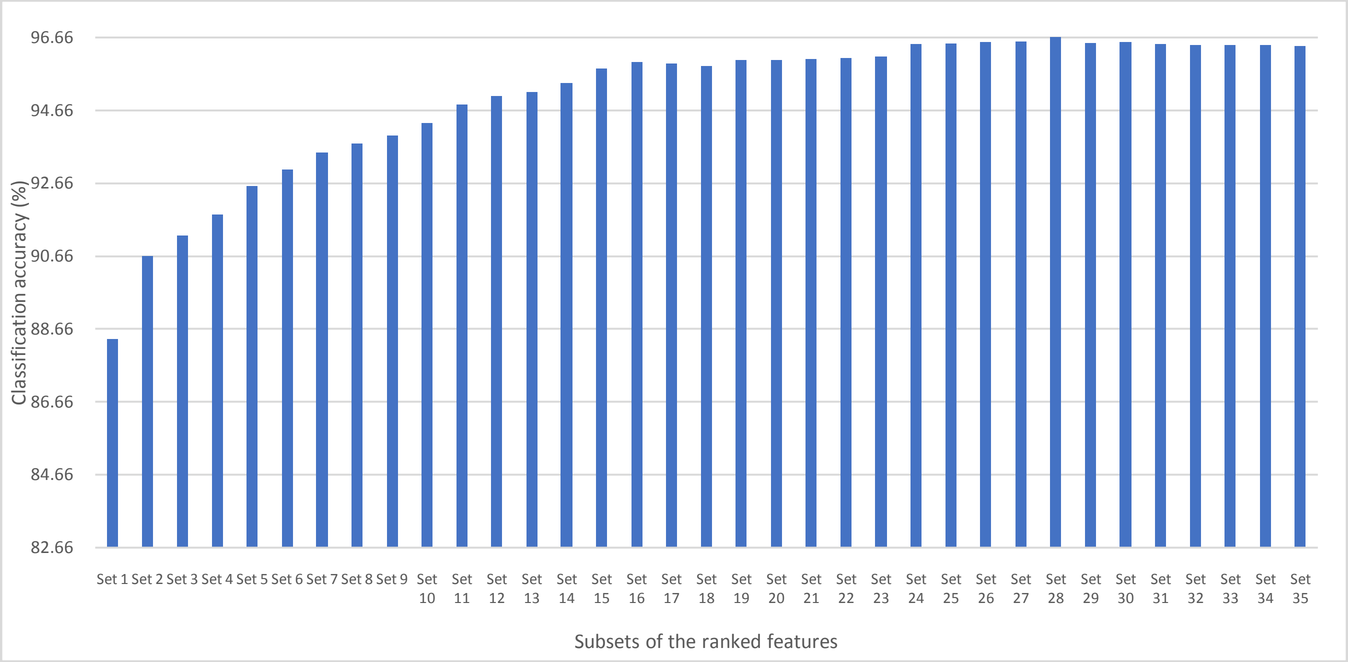


Fig. 6. Classification accuracy of the different subsets of the ranked features.

TABLE II. Comparison between different feature selection techniques and the proposed model.

| Algorithm | No. of features | Precision (%) | Recall (%) | F Score average | AUC | MCC | Accuracy (%) |
|---|---|---|---|---|---|---|---|
| Random Forest + CFS | 26 | 96.34 | 96.26 | 0.963 | 0.987 | 0.925 | 96.30 |
| Random Forest + IG | 12 | 94.89 | 94.84 | 0.949 | 0.982 | 0.914 | 94.88 |
| Random Forest + PCA | 25 | 96.14 | 96.07 | 0.961 | 0.984 | 0.921 | 96.12 |
| Random Forest + Relief | 26 | 96.05 | 96.02 | 0.960 | 0.983 | 0.917 | 96.04 |
| Proposed methodology | 28 | 96.68 | 96.62 | 0.967 | 0.991 | 0.942 | 96.66 |

The proposed methodology was also compared with recent algorithms to evaluate it. The proposed methodology achieved the highest classification accuracy compared to Adaptive Neuro-Fuzzy Inference System (ANFIS) models [6] and the developed approach MCAC in [28] using 10-fold cross validation technique as shown in Fig. 8.

The web-phishing detection time of the meta-heuristic based approaches depends on the computational time of the classification process O(*snc*), where *s* is the number of database samples, *n* is the number of features, and *c* is the computationtime of the operational method of the classifier. Number of trained samples “phishing attack techniques” are increased at all the time, according to the report of the APWG. Therefore, the number of features and the complexity time of the classification method have a significant role in the time of the spoofed web page detection. Most of the recent researches in [29-32] that based on artificial neural network and deep learning classification techniques were developed to enhance the detection accuracy of the web-phishing, but these researches ignored the importance of the computational time of the classifier, where the artificial neural network and deep learning algorithms have complex architectures and consume extremely expensive time to construct the classifiers.

The main contribution of this research is to detect the spoofed web pages at the least possible time for decreasing the probability of downloading malware from the web-phishing pages during the detection process. Therefore, the proposed methodology is based on new proposed feature selection method to select the least number of the effective web page features, which contributes to decrease the computational time of the web-phishing detection time. It also based on the Random Forest classification algorithm to construct the classifier of the web pages, which it'scomputational time is much less than artificial neural network and deep learning algorithms. As shown in the Table 3, the computational time of the proposed methodology to detect the spoofed web pages is very low compared to ANFIS by significant rate. The proposed methodology is based on 28 features to detect the Web-phishing and ANFIS is depended on 35 features of images, text and frame. ANFIS is also based onartificialneur-

-al fuzzy network algorithm, which is more complex in the operational method than Random Forest classifier. The detection time of the proposed methodology significantly contributes to prevent the web-phishing. At the same time, this significant improvement of the detection time did not have a negative impact on the accuracy of the spoofed web page detection, but it also has a positive effect on the detection accuracy as shown in Fig. 8.

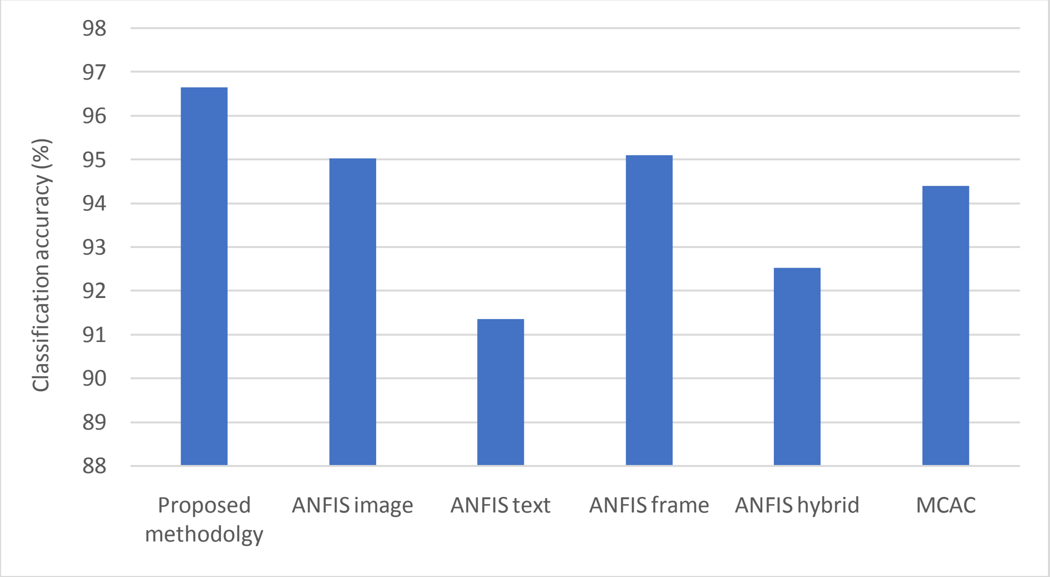


Fig. 7. Comparison between the proposed methodology and recent methods for web-phishing detection using 10-fold cross validation testing technique.

TABLE III. Table 3. Comparison between the classifier features of the ANFIS approach and the proposed methodology.

| Algorithm | No. of features | Classifier | Computational time |
|---|---|---|---|
| ANFIS approach [6] | 35 | Neural Fuzzy Network | 26.72 sec |
| Proposed methodology | 28 | Random Forest | 00.67 sec |

## V. Conclusion

In this paper, the proposed methodology introduced to detect and prevent the web-phishing, it depends on new proposed method to analyze the features of URL, HTML and JavaScript code, text, images and the web page domain name, it extracts the most significant of these features that contributes to improve the efficiency of the web-phishing detection. The proposed methodology was evaluated using two benchmark datasets, the new proposed method of the feature selection achieved the highest classification accuracy 96.66% compared to various feature selection algorithms. It also obtained the highest accuracy of web-phishing detection compared to ANFIS using the least number of the most effective features, which leads to reduce the probability of downloading malicious software from the web-phishing pages to a computer.

## References

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

[2] J. Maoa, J. Biana, W. Tiana, Sh. Zhua, T. Weic, A. Lid and Z. Liange, “Detecting Phishing Websites via Aggregation Analysis of Page Layouts”, In Proceedings of the International Conference on Identification, Information and Knowledge in the Internet of Things, China, 19-21 October, 2017.

[3] https://www.wombatsecurity.com/blog/the-latest-in-phishing-first-of-2019. (Accessed on: 2020)

[4] A. Jain and B. Gupta, “PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning”, Cyber Security, Advances in Intelligent Systems and Computing, vol. 729, pp. 467-474, 2018.

[5] N. Sanglerdsinlapachai and A. Rungsawang,“Using Domain Top-page Similarity Feature in Machine Learning-based Web Phishing Detection”, In Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Thailand, 09-10 Jan, 2010.

[6] M. Adebowale, K. Lwin, E. Sánchez and M. Hossain, “Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text”, Expert Systems with Applications, vol. 15, pp. 300-313, 2019.

[7] I. Hamid, A. Rahmi and A. Jemal “Phishing e-mail feature selection approach 2011.” In Proceedings of the International Joint Conference of IEEE, Taiwan, 25-27 May, 2011.

[8] N. Shekokar, C. Shah, M. Mahajan, and S. Rachh, “An ideal approach for detection and prevention of phishing attacks”, Procedia Computer Science, vol. 49, pp. 82-91, 2015.

[9] Y. Zhang, I. Hong, and F. Cranor, “Cantina: a content-based approach to detecting phishing web sites”, In Proceedings of the 16th

international conference on World Wide Web, ACM, Canada, 08-12 May, 2007.

[10] M. Aburrous, A. Hossain, K. Dahal, and F. Thabtah, “Intelligent phishing detection system for e-banking using fuzzy data mining”, Expert Systems with Applications, vol. 37(12), pp. 7913-7921, 2010.

[11] A. Barraclough, A. Hossain, A. Tahir, G. Sexton, and N. Aslam, “Intelligent phishing detection and protection scheme for online transactions. (Re- port)”, Expert Systems with Applications, vol. 40 (11), pp. 4697-4706, 2013.

[12] UCI Machine Learning Repository available at: https://archive.ics.uci.edu/ml/machine-learning-databases/00327/ Training%20Dataset.arff. (Accessed on: 2020)

[13] Phishing websites Database available at: http://eprints.hud.ac.uk/24330/9/Mohammad14JulyDS_1.arff. (Accessed on: 2020)

[14] A. Ahmed and N. Abdullah, “Real Time Detection of Phishing Websites”, In Proceedings of the IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference, Canada, 13-15 October, 2016.

[15] L. Cranor, S. Egelman, I. Hong, and Y. Zhang, “Phishing Phish: An Evaluation of Anti-Phishing Toolbars”, In Proceedings of the Network and Distributed System Security Symposium Conference, NDSS, USA, 28th February – 02nd March,2007.

[16] B. Osareh, "Intrusion Detection in Computer Networks based on Machine Learning Algorithms", International Journal of Computer Science and Network Security, vol. 8(11), pp. 15-23, 2008.

[17] H. Shahriar and M. Zulkernine, “Information Source-based Classification of Automatic Phishing Website Detectors”, IEEE/IPSJ International Symposium on Applications and the Internet, Munich, pp. 190-195, 2011.

[18] L. Wenyin1, G. Huang1, L. Xiaoyue, Z. Min, and X. Deng, “Detection of phishing webpages based on visual similarity”, In Proceedings of the 14th international conference on World Wide Web, Japan, 10-14 May, 2005.

[19] A. Fu, L. Wenyin, and X. Deng, “Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD)”, IEEE Transactions on Dependable and Secure Computing, vol. 3(4), pp. 301 - 311, 2006.

[20] V. Kumar and R. Kumar, “Detection of a phishing attack using visual cryptography in ad-hoc network”, In Proceedings of the IEEE International Conference on Communications and Signal Processing (ICCSP), INDIA, 02-04 April, 2015.

[21] S. Fatt, K. Leng, and S. Nah, “Phishdentity: Leverage Website Favicon to Offset Polymorphic Phishing Website”, In Proceedings of the IEEE Ninth International Conference on Availability, Reliability and Security (ARES), Switzerland, 08-12 September, 2014.

[22] A. Barraclough, A. Hossain, A. Tahir, G. Sexton, and N. Aslam, “Intelligent phishing detection and protection scheme for online transactions”, Expert Systems with Applications, vol. 40(11), pp. 4697-4706, 2013.

[23] M. Jian, T. Wenqian, L. Pei, W. Tao and L. Zhenkai, “Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity”, IEEE Access, vol. 5, pp. 17020-17030, 2017.

[24] V. Shreeram, M. Suban, P. Shanthi, and K. Manjula, “Anti-phishing detection of phishing attacks using genetic algorithm”, In Proceedings of the IEEE International Conference on Communication Control and Computing Technologies (ICCCCT), India, 7-9 October, 2010.

[25] A. Yasin and A. Abuhasan, “An intelligent model for phishing email detection”, International Journal of Network Security & Its Applications(IJNSA), vol. 8(4), pp. 55-72, 2016.

[26] A. Agarwal, M. Mittal, A. Pathak and L. Goyal, “Fake News Detection Using a Blend of Neural Networks: An Application of Deep Learning”, SN Computer Science, 1:134, pp. 1-9, 2020.

[27] C. Monica and N. Nagarathna, “Detection of Fake Tweets Using Sentiment Analysis”, SN Computer Science, 1:89, pp. 1-7, 2020.

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

[29] Y. Ping, G. Yuxiang, Z. Futai, Y. Yao, W. Wei and Z. Ting, “Web Phishing Detection Using a Deep Learning Framework”, Wireless Communications and Mobile Computing, pp. 1-9, 2018.

[30] Y. Peng, Z. Guangzhen and Z. Peng, “Phishing Website Detection based on Multidimensional Features driven by Deep Learning”, IEEE Access, vol. 7, pp. 15196-15209, 2019.

[31] Z. Erzhou, Ch. Yuyang, Y. Chengcheng, L. Xuejun and L. Feng, “OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network”, IEEE Access, vol. 7, pp. 73271-73284, 2019.

[32] R. Mohammad, F. Thabtah and T. Mccluskey, “Predicting phishing websites based on self-structuring neural network”, Neural Computing and Applications, vol. 25(2), pp. 443-458, 2013.

[33] A. Tharwat, “Classification assessment methods”, Applied Computing and Informatics, pp. 1-13, 2018.