Bulletin of Electrical Engineering and Informatics
Vol. 8, No. 4, December 2019, pp. 1545~1555
ISSN: 2302-9285, DOI: 10.11591/eei.v8i4.1344

Review of the machine learning methods in the classification of phishing attack

John Arthur Jupin[1], Tole Sutikno[2], Mohd Arfian Ismail[3], Mohd Saberi Mohamad[4], Shahreen Kasim[5], Deris Stiawan[6]

[1,3]Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, Pahang, Malaysia
[2]Department of Electrical and Computer Engineering, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[4]Institute for Artificial Intelligence and Big Data, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100 Kota Bharu, Kelantan, Malaysia
[4]Faculty of Bioengineering and Technology, Universiti Malaysia Kelantan, Jeli Campus, Lock Bag 100, 17600 Jeli, Kelantan, Malaysia
[5]Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Johor, Malaysia
[6]Department of Computer Engineering, Universitas Sriwijaya, Palembang, Indonesia

Article Info

Article history:

Received Jun 18, 2019
Revised Sep 24, 2019
Accepted Oct 6,2019

Keywords:

Classification
Machine learning
Phishing

ABSTRACT

The development of computer networks today has increased rapidly. This can be seen based on the trend of computer users around the world, whereby they need to connect their computer to the Internet. This shows that the use of Internet networks is very important, whether for work purposes or access to social media accounts. However, in widely using this computer network, the privacy of computer users is in danger, especially for computer users who do not install security systems in their computer. This problem will allow hackers to hack and commit network attacks. This is very dangerous, especially for Internet users because hackers can steal confidential information such as bank login account or social media login account. The attacks that can be made include phishing attacks. The goal of this study is to review the types of phishing attacks and current methods used in preventing them. Based on the literature, the machine learning method is widely used to prevent phishing attacks. There are several algorithms that can be used in the machine learning method to prevent these attacks. This study focused on an algorithm that was thoroughly made and the methods in implementing this algorithm are discussed in detail.

Corresponding Author:

Mohd Arfian Ismail,
Faculty of Computer Systems & Software Engineering,
Universiti Malaysia Pahang, Malaysia.
Email: arfian@ump.edu.my

1. INTRODUCTION

Network security is an important and critical issue that needs to be considered and emphasized in the network, especially in an organization. The authorization process of an account of a person can be considered as a network security, where the process commonly uses the username and password which inhibit and monitor the unauthorized access to a particular account [1]. An authentication process is needed to protect sensitive and important data from being exposed and stolen by unauthorized users or hackers. In certain situations, although a network security has already been implemented, there are still some chances that sensitive and important data can be stolen. One way to steal sensitive and important data can be done through a phishing attack.

Phishing attack can be considered as an act of imitating genuine websites to collect sensitive information from a victim and using them for committing crimes, such as illegal financial gains [2]. This attack typically starts when the attacker or hacker sends an email that seems original to the victim and persuades them to update and verify their information by clicking on a Uniform Resource Locator (URL) link in the email [3]. Usually, the phishing email will redirect users to the infected website and ask them to provide their information, such as their personal details and bank account information, which will be used by the attacker or hacker to steal whatever important information that the users have entered [4]. Phishing attack is always related to spam emails received by the victims. Some of those emails may contain the link that will redirect the victims to the phishing websites. Phishing attack is usually difficult to identify because the email sent by the attacker or hacker looks like a legit email. In addition, the attacker or hacker can hide the location of their server and sometimes disguise the URL of the phishing website to work like the legitimate website. Moreover, even a good security software is unable to detect phishing websites because they do not depend on the malware infection of the computer [5].

Currently, many works have been proposed for phishing attack detection in the literature and commercial products. There are four features that can be used in detecting a phishing attack. The features are given in Figure 1. The URL-based feature works based on URL. A phishing attack works based on a URL that redirects a user to a certain page that has been duplicated by the attacker from the official page. The URL and the duplicated page can be recognized from a malicious URL. The malicious URL can be detected based on the total length of URL, the count digit in URL, the correct spelling of URL, and whether the URL includes a legitimate brand name or not. The domain-based feature works by detecting the domain name of the URL, where the domain name will determine whether the URL can be classified as a phishing attack or not. The URL can be considered as phishing based on the status of the domain name; whether the domain name is in the list of blacklists of well-known reputation services, the age of domain name, and the owner of the domain name. The third feature, which is page-based works, based on the information from the pages where the information will determine the reputation ranking services. The reputation will determine the reliability of the pages. Normally, the reputation ranking is determined by the Global page rank, Country PageRank, and position indexed by Alexa. Usually, the ranking services will give information regarding the user activities in the site including an estimated number of visitors of a page in terms of daily, weekly or monthly; the average visit of the page, web traffic, category of the domain, and similar websites with the page. Meanwhile, the content-based feature works based on the scanning process of the domain. The items being scanned are usually the page title, meta tags, hidden text, text in the body, and images in the page. The scanning process is to determine whether the page requires the login process, the category of the page, and the user of the page.

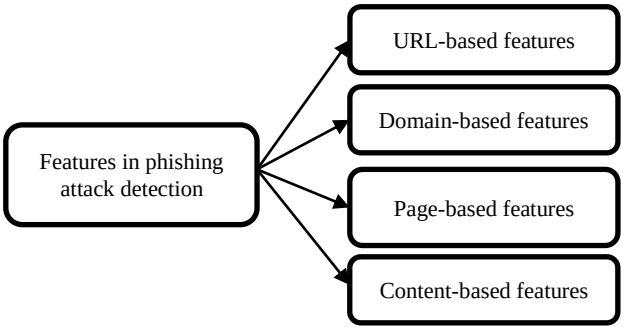

Figure 1. Features in phishing attack detection

All of the features discussed are widely used in the identification of phishing attacks. In some cases, the mentioned features may not be effective to detect the phishing attack due to the limitations of these features. Consider a situation where the content-based feature is used to develop a fast mechanism in analyzing the phishing of many pages. It will take time to scan a huge number of pages. Hence, the feature that will be chosen depends on the objective of the detection mechanism and should be selected carefully.

2. RESEARCH METHOD

Detecting phishing attack is a challenging task due to its mechanism where the attack exploits human vulnerabilities, not on the system error. Detecting phishing attack can be considered as a classification problem, which means that the attack needs to be labeled, whether the page is a phishing attack or legitimate. For this purpose, a good method is needed. Using machine learning (ML) methods is appropriate to be

applied in phishing attack detection because ML is able to transform the problem in phishing attack detection into a classification task.

ML is a subfield of artificial intelligence where the goal is to enable the computer to learn about something on its own or learn from experience. ML works based on the idea that computational method is able to learn information directly from data, identify patterns in the observed data, and make decisions without relying on a predetermined equation as a model. ML uses a technique where it trains a model on known input and output to predict the class of data. This technique seems suitable for the detection of phishing attacks because it can convert the detection problem into a classification task.

In detecting a phishing attack, the ML method will train a classification method with some features or rules to declare whether the attack is classified as phishing or not. Usually, the ML method works by extracting the features from a URL or the content of a web page and train a prediction model based on the features that have been discussion in the previous section before deciding whether the web page is legitimate or fake. There are many ML methods in detecting phishing attacks that are currently and widely used, which include the Artificial Neural Network (ANN) algorithm, Decision Tree (DT) algorithm, k-means clustering algorithm, Naïve Bayes (NB) algorithm, Random Forest (RF) algorithm, and Support Vector Machine (SVM) algorithm. These methods were chosen because of their performance and high accuracy in detecting phishing attacks. Following this is the description of these methods.

2.1. Artificial Neural Network

The ANN is a model influenced by the structure of the human brain to stimulate human behavior in processes. The learning process takes place in these networks through a set of simple processing, called artificial neurons [6, 7]. ANN is an information-processing model which stimulates how the biological nervous system processes information [8-10]. ANN is assembled and composed in layers, where each layer has artificial neurons that are connected with each other. The basic elements of an ANN can be shown in Figure 2.

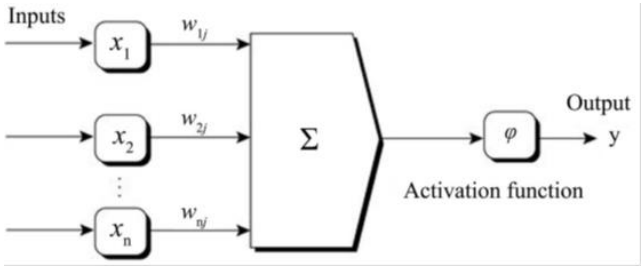

Figure 2. Basic elements of an ANN [6]

Based on Figure 2, the data which is the input vector of the neuron $(x_1, x_2, \dots, x_n)$ and the neurons of the input layer $(w_{1j}, w_{2j}, \dots, w_{nj})$ with their respective heights are observed. The additive junction, or known as sum, is represented by the letter sigma (Σ), whereas the activation function and output are denoted by φ and y, respectively.

A work that used this approach was proposed in [6]. Their aim was to classify the websites with the phishing characteristics. Based on the results obtained, it showed that the ANN correctly classified 87.61% in the training of 1000 records obtained from Phishing Websites Data Set of the University of California's Machine Learning and Intelligent Systems Learning Center. By comparing the other methods, such as Dynamic evolving neural network based on reinforcement learning, it had a slightly increased accuracy percentage where of 98.63%, which was a 0.40% difference in the accuracy percentage. This may be due to the order of the attributes used during the implementation. The study suggested that the order of the attributes should be changed to find the better groups to be processed by ANN. Besides that, [11] also proposed the ANN approach in detecting phishing in emails. The neural network models were tested with 17 and 12 features. The neural network with 17 features was tested by using 8,801 vectors, 587 train vectors, and 8,214 test vectors, whereas the neural network with 12 features was tested with 8697 vectors, 282 train vectors, and 8,415 test vectors. Based on the results obtained, the accuracy obtained was the same with 4, 5, and 6 hidden neurons. As the number of neurons in the hidden layer increased, the time taken to train the neural also increased. Thus, the number of hidden nodes taken was 4, which showed the best result. The number of hidden neurons taken was 3 since it showed the best accuracy among all of them. The results also showed that as the number of nodes increased above 3 neurons, the accuracy dropped down to 50%. Next, in [12] proposed the phishing URL detection by using ANN with Particle Swarm Optimization (PSO). Their aim is to show that the ANN-PSO can achieve better accuracy compared to Back Propagation Neural Network

(BPNN). The result of this work shows that the NN-PSO model achieved better accuracy, where all the accuracy reading for three different learning ratios in ANN-PSO model is above 97%, whereas the highest accuracy reading for the BPNN is 96.81%.

2.2. Decision tree algorithm

The DT algorithm is an algorithm that belongs to the supervised classification algorithm. This algorithm is used in solving regression and classification problems by creating a training model which will predict the class or value of target variables summarized from the training data. There are two types of DT algorithm, which include Iterative Dichotomiser 3 (ID3) algorithm and C4.5 algorithm. The ID3 utilizes the process for creating a decision tree in the "top-down" form. It has been proven a very useful method, but still has a huge number of constraints, which will deter the application of this algorithm in many real-world situations [13]. The C4.5 algorithm was developed to overcome these problems and has been considered as a good solution when using a large size, missing, and continuous variables data. The DT algorithm can be expressed in the (1-5) as follows:

$$Info(D) = -\Sigma_{i=1}^{m} p_i log_2(p_i) \quad (1)$$

$$Info_A(D) = \Sigma_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (2)$$

$$SplitInfo_A(D) = -\Sigma_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2\left(\frac{|D_j|}{|D|}\right) \quad (3)$$

$$Gain(A) = Info(D) - Info_A(D) \quad (4)$$

$$GainRatio(A) = \frac{Gain(\mathrm{A})}{SplitInfo_A(D)} \quad (5)$$

where D is the training set of class-labeled tuple, D_j is denoted as a subset of D, and C_i is the class label of tupple (for $i = 1,2,\ldots,m$). Meanwhile, p_i refers to the probability of a tuple in D and belongs to class C_i and $|D|$ is the number of tuples in D.

In a work proposed by [14], they used C4.5 algorithm in WEKA for the detection of phishing websites. Their work used a dataset that contained 300 websites. Based on the results obtained by their work, it was found that 200 websites were detected as phishing websites. The success rate and error rate obtained were 0.826 and 0.173, respectively, after the prediction confusion matrix was generated. Thus, the accuracy of the classifier model that trained with 750 instances was 82.6%. Proposed a phishing detection system using the ID3 algorithm [15]. The objective of their system was to distinguish whether the URL was legitimate or a phishing URL. The proposed system had four main steps, which included step 1: data preparation; step 2: feature extraction of URL; step 3: implementation of ID algorithm where ID3 will perform the classification process; and finally step 4: a model of their method. Another work was performed using the ID3 algorithm by [16]. In their work, they used the ID3 algorithm as a classifier and the info gain feature selection technique was applied by using different top selected feature subsets for determining the phishing websites. They used data from UCI repository in conducting their experiments where the data consist of 30 features, 11055 instances, and one class which can be classified as phishing websites and legitimate websites. From the results obtained, their method performed well compared to other classification methods.

2.3. K-means clustering approach

K-means clustering is the algorithm used to cluster data points into different clusters where the minimization of the distance between the elements of the cluster and its centroid is made [17]. Basically, K-means algorithm is used for partitioning of the n observation into k clusters, in which each of the observation belongs to the cluster with the nearest mean [18]. According to [19], the algorithm of this approach is iterated between two steps, which are the data assignment step and centroid update step. The algorithm is as follows:

$$argmin_{c_i \in C} dist(c_i, x)^2 \quad (6)$$

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_1} x_i \quad (7)$$

where dist(-) is the standard Euclidean distance. Let the set of data point assignments for each i^th cluster centroid be S_i. Both (6) and (7) in k-means clustering algorithm will be iterated until a stopping criterion is met.

A work that used this approach was conducted by [20]. They proposed the Kernel K-means clustering, which is the extension of the k-means algorithm. The vectors were mapped from the vector space to a higher dimensional feature space through kernel function and then k-means was applied in the feature space. Based on the results obtained, it showed that the proposed method (K-means clustering approach) was better compared to the ensemble clustering algorithm. Based on the accuracy result, it showed that the accuracy increased by 13.25% when testing 500 phishing websites and 1.48% when testing 2,000 phishing websites. The work that used k-means clustering approach was also done by [21]. In their work, they proposed three other different approaches, which were multilayer perceptron (MLP), J48 decision tree, and Naïve Bayes. The result showed that the prediction accuracy percentage for k-means clustering approach was higher compared to the other three approaches, which was 99%. The other characteristics, such as time taken of the production of the model, correctly classifying instances, and incorrectly classifying instances were also taken into consideration and compared. However, the production time of this approach took a longer time compared to MLP, but this approach showed the highest number of correctly classified instances and lowest number of incorrectly classified instances. Another work that use this approach was proposed by [17]. They used the relational K-mean clustering which deals with non-vector data. The study was done with different numbers of clusters: 5, 6, and 7 clusters. The result showed that the 5-cluster selection was better compared to the others, where both mean and standard deviation indicated an identical distribution of emails in different clusters. This showed that the accuracy level of classification of the email contents was higher when using the 5-cluster selection.

2.4. Naïve Bayes algorithm

The NB algorithm, also known as the Bayesian classifier, is a group of classification algorithm based on the Bayes' Theorem. It is based on a probabilistic classifier with strong independence assumptions between features. Naïve Bayes algorithm will share a common principle, where every characteristic being classified is independent of its value among any other characteristics. The general equation for Bayes' theorem can be expressed as follows [22]:

$$P(x|Y) = \frac{P(Y|x)P(x)}{P(y)} \quad (8)$$

where $P(x)$ is the independent probability of x prior probability, $P(Y)$ is the independent probability of Y, $P(X|x)$ is the conditional probability of Y given h: likelihood, and $P(Y|x)$ is the conditional probability of x given Y.

In a work performed by [23], they used Naïve Bayes classifier as a text classification method to filter the machine for spam emails of victims. Their method used tokens, which represented the words used in the spam and non-spam emails to calculate the probability of the emails; whether it was a spam or not. In her work used Naïve Bayes as a classifier to classify a web page. In her work, the proposed method gained the information of a web page from extracting the web page features based on the URL, source, and images [24]. Then, the ant colony optimization algorithm was applied to optimize the extraction process before Naïve Bayes was used to determine whether the web page was legit or fake. In his work proposed an improved method for spam email classification by determining whether the email contained a phishing URL or not [25]. In the proposed method, the intelligent water drop algorithm was utilized to construct the feature selection and then Naïve Bayes classifier was used to classify the email as legit or contained a phishing URL.

2.5. Random forest algorithm

The RF is based on an ensemble of learning methods created for classification and regression task by [26]. RF is a method that works by a set of decision trees, where the input of data will be added at the top of the tree and it will traverse up to down the tree to the smaller subsets. In the classification of phishing attack detection, the process works by the prediction of decision trees. For every input, the RF will randomly choose a subset of features that will be used in classification process, where RF makes the chosen process becomes unbiased through guesstimate. By doing this, the RF will improve the predictive accuracy and control over-fitting. The RF has variables which include forest size T, the depth of forest D, and the node (subset) i. The RF is normally given as follows:

$$h(v, \emptyset_i) \in \{true, false\} \quad (9)$$

where $v = (x_1, x_2, \ldots x_n) \in \mathbb{R}^n$ is the feature of vector. The $\emptyset_i$ is denoted as the optimal parameter of the i node. According to [27], RF combines multiple classification process, where each process contributes to a single assignation of the most frequent class to the input vector (x) to the class prediction denoted as $C_r^B{}_f$ and can be defined as follows:

$$C_{\mathrm{r}}^{B}{}_{f} = \text{ majority vote } \{C_b(\mathrm{x})\}_1^B \quad (10)$$

A work that used RF for the classification of phishing detection has been proposed by [28]. They investigated the performance of RF in classifying the phishing detection where their aim was to improve the prediction accuracy and use only fewer numbers of features. In their method, they only used 8 out of 15 features. The proposed method was tested on data that had 2,000 phishing and ham emails and the proposed method was able to achieve 99.7% accuracy. Another work that used RF was carried out in by [29]. They proposed an approach for spam filtering that had two stages, which were feature selection and email classification. In the first stage, they used an optimization method which was PSO based on the wrapper feature selection for the selection of features. The aim of the first stage was to reduce the number of features. In the second stage, they used RF to develop a filtering model by using the features selected in the first stage. They tested their proposed method on a dataset that had 9,346 ham and spam emails where every email had 79 features. Their work has proven the effectiveness in classifying ham and spam emails that contained phishing attacks. Another work that used RF was performed by [30]. They proposed a method in classifying phishing attacks based on URL. In their method, they used the URL information, which is the metadata of URL that included the number of slashes and keyword in the URL portion. Then, they applied the RF as a classifier to determine whether the URL was a phishing attack or a legitimate URL. They tested their method on two URL datasets that had 2,500 data with 31 features and 1353.

2.6. Support Vector Machine algorithm

The SVM algorithm, which is also known as SVM classifier, is a machine algorithm which is mostly used in classification problems. It is also a supervised learning technique, whereby it will classify the dataset that contains class labels and features. According to [31], SVM algorithm is a linear strong classifier which can identify two label classes in the dataset. This algorithm will produce a set of hyperplanes, in which the maximum marginal hyperplane will be considered at the end of the test. The SVM algorithm can be expressed in (11) and (12);

$$min\ \frac{1}{2}\ |w|^2 + c\sum_{i=1}^{n}\xi_i \quad (11)$$

$$y_i(wx_i - b) \gg 1 - \varepsilon_i \quad \varepsilon_i \geq 0 \quad (12)$$

where i is a range of 1, 2,…,n, n refers to the dimensionality of the feature, x is the input vector, w is the normal vector to the hyperplane, C refers to the capacity constant, and ε_i is the parameters for handling no separable data (inputs).

Since phishing websites are usually attached to spam emails, this algorithm will be suitable to help detect them. There are some attributes that are commonly used by SVM algorithm to detect phishing websites as listed in Table 1 [31].

Table 1. The attributes and their significance in phishing attack detection by SVM

No.	Attributes	Significance
1	Internet Protocol (IP) address	The website is phishing if the IP address is used in the domain name.
2	URL length	URL length that is more than 75 characters is considered as phishing websites.
3	Shortening service	Shortened link could confuse the user.
4	Having '@' symbol	Websites that contain the '@' symbols are usually phishing websites.
5	Double slash redirecting	The website can be categorized as a phishing website if there is '/1' at the end of its address.
6	Having sub domain	Websites having more than two levels and more than three dots (domain within a domain) could be phishing websites.
7	URL of Anchor	Phishing websites usually have different domains compared to legitimate websites, where the anchor tag is connected to the same domain as the source code.
8	Links in tags	This will lead to some infected websites.
9	Abnormal URL	Extracted from the database while the main identity of the legitimate websites is in the URL.
10	Age of domain	Websites that are more than six months of age can be classified as phishing websites.
11	Page rank	Phishing websites have low page rank.
12	Links pointing to page	Phishing websites usually have links pointing to zip files that contain malware, which will be downloaded automatically to the computer.

A work that used SVM was performed by [32]. They presented a novel approach that used lightweight phishing detection with URL-based as their feature and applied SVM as the classifier. Their method was able to achieve 95.80% of classification accuracy when it was tested on 2,000 datasets that consisted 1,000 legitimate and 1,000 phishing URLs. Another work that utilized the effectiveness of SVM was carried out by [33]. In their work, they developed a client phishing attack hybrid detection model. Their

method worked based on the web content feature and the IP address feature. They proposed a hybrid system that consisted of particle swarm optimization (PSO) and SVM. They used PSO to automatically optimize the selected parameters in their model and then SVM was used as a classifier to determine the phishing attack class. Used SVM in their work to classify phishing attacks [34]. They used kernel-based SVM method in their work by extracting features of webpages including textual properties, link structures, webpage contents, DNS information, and network traffic. Their method performed better in detecting phishing webpages where the proposed method was able to achieve the accuracy of around 95%.

3. COMPARATIVE ANALYSIS OF MACHINE LEARNING IN PHISHING ATTACK DETECTION

In choosing the suitable ML method for detecting phishing attacks, there are many factors that need to be considered. Normally, many factors will contribute to the performance and the accuracy of the method such as the processing speed, the accuracy of the classifier, the size and the complexity of data, the interpretability of the model produced by the ML method, and the easy implementation of the ML method to the specific problem. Table 2 lists out a comparative analysis of the machine learning methods.

Table 2. Comparative analysis of machine learning methods

Method	Advantages	Disadvantages
Artificial Neural Network	– The ANN allows for specifying the attribute and the type of learning performed in this approach [6]. – The ANN fault tolerant which is has an ability to work with incomplete knowledge and data that has noise [35]. – ANN able to develop an accurate model by using experimental data only [36]. – The ANN has a distributed memory which is suitable works in parallel processing [37-39].	– The order of the data attributes may affect the results obtained during the classification process [6]. – The ANN learning will be very slow if a very low learning rate is used. Besides, a high learning rate will cause oscillations in training and block the convergence of the learning process which become slow [37]. – Difficult to transform/model the problem to the network in ANN [40]. – The result produce by ANN is difficult to understand because ANN do not give a clue about the structure of their models and hard to predict the model [41, 42].
Decision Tree Algorithm	– Its simplicity to explain and interpret the feature relationships and interactions [43-45]. – The model produced by DT is easy to be interpreted and understood because it produces simple IF–THEN statements [46, 47]. – The DT is easy to be implemented compared to others [46]. – Requires less time in the classification process [43, 47].	– The DT does not support online learning and requires rebuilding the tree each time new samples exist; this will require a new process and rebuilding the tree requires more time [45]. – The classification result of DT is low compared to another ML methods [43]. – The DT becomes more complex as the number of features is increased [48]. – The DT is unable to deal with missing values [44].
k-means clustering	– The k-means clustering has an ability to prevent the drawback of linearly separable clusters in vector space [20]. – This approach is able to minimize clustering error in feature space [18].	– This approach is totally depending on the initial random assignments/search direction, which lead to poor result if the initialization is not proper define [49, 50]. – This approach is unable to classify if the website is phishing and considers it as 'may be'–uncertain and missing value [18].
Naïve Bayes Algorithm	– This method is easy to implement in the classification process and the process is fast [49, 50]. – Its simplicity and quick convergence. It is also an easy and straightforward method [45]. – Uses a small number of data to estimate the important features for classification process [47, 52, 53]. – Only requires less time in classification process compared to other methods [53, 54]. – Has the ability to handle missing values by assimilating the overall opportunities of the missing values [44].	– This approach requires higher computational resources and memory because this approach need to use all the input variable in classification process to get the higher accuracy [51]. – The NB cannot learn about the interactions and relationships between the features in each sample, where it leads to the low accuracy [45]. – Needs a large amount of data to get higher accuracy [53, 55, 56]. – The NB requires a large space to store data due to its instance-based nature where the NB stores all training samples in its process [56]. – The NB is not sensitive about the data because the NB is unable to show the relationship between the variables where the variables may totally depend on each other [57].
Random Forest Algorithm	– The speed and efficiency of RF is good when applied on large datasets and high dimensional problems with multi-class output [29, 58]. – Performance is better and robust compared to others where RF has higher accuracy especially on nonlinear problem [27, 35, 43, 59]. – A common problem in classification is overfitting, but RF is able to overcome this problem because if there are enough trees in the forest, RF will not overfit the model [35, 60].	– The main limitation of RF is that a large number of trees can make the algorithm slow and ineffective for real-time predictions. A more accurate prediction requires more trees, which results in a slower model [61]. – RF is a predictive modeling tool and not a descriptive tool, thus makes no description of the relationships data, making it hard to interpret the model [62]. – RF is sensitive where a small change of the parameter value can lead to a significant change of model and result [63]. – The result produce by RF is not consistent because RF use

Method	Advantages	Disadvantages
	– RF is easy to implement and the result produce by RF is easy to be interpreated [35].	random factor in bootstrapping, bagging and constructing tree, thus make difficult to prove the consistency of result produce by RF [64].
Support Vector Machine Algorithm	– The SVM is known for having higher accuracy in classification and its ability to classify data that is not linearly separable [45]. – Performs better compared to others when applied on high-dimensional data with minimum data [43, 65, 66]. – Good in handling large attributes and large amount of data [31]. – SVM is memory efficient and converge fast to find the optimum solution, because it uses a subset of training points in the decision process [67]. – It is a robust model to solve prediction problems since it maximizes margin [68].	– SVM demands a convex combination of kernels, thus making it time-consuming in the classification process [43, 45]. – Hard to interpret and difficult to understand the model produced by SVM [45, 69]. – Hard to implement and handle the numerical variables in the classification problem [46]. – The parameter in SVM is sensitive where it needs to be set correctly and will affect the classification accuracy if not set properly [44]. – SVM is a binary classifier and can be applied on binary classification problems. To apply on multi-class classification, SVM needs some modifications [51, 70].

3. CONCLUSION

his paper presents an overview of phishing detection. Phishing can be considered as an illegal activity by hackers to steal sensitive information of Internet users such as login credentials or bank account information by redirecting Internet users to the illegitimate websites. Usually, the hacker sends an email to Internet users that contains malicious software or URL to the fake website. Phishing detection is essential because it can prevent the hacker to steal Internet users' information. This paper discussed the four features that can be considered in the detection of phishing attacks, which included URL-based, domain-based, page-based, and content-based features. Phishing detection can be considered as a classification of security breach and one way to detect phishing attacks is by using the ML method. In the presented paper, six popular and widely used classification methods of ML have been presented and discussed in detail, where the discussion covered on mechanism of the methods, strengths, and weaknesses. The ML methods chosen are NB, SVM, DT, RF, k-means clustering, and ANN. Based on the discussion of the ML methods, it is hard to determine which method is the best one because each method has its own advantages and disadvantages. The selection of method depends on the problem and features selected because there is no single method that works best on every problem and can be applied on varities problem domain such as [71–91].

ACKNOWLEDGEMENTS

Special appreciation to Universiti Malaysia Pahang for the sponsorship of this study by approve the Ministry of Higher Education (MOHE) for Fundamental Research Grant Scheme (FRGS) with Vot No. RDU190113 and UMP Internal Grant with Vot No. RDU180307. Special thanks to the reviewers and editor who review this manuscript.

REFERENCES

[1] M. V. Pawar and J. Anuradha, "Network Security and Types of Attacks in Network," *Procedia Comput. Sci*., vol. 48, pp. 503–506, Jan. 2015.
[2] S. Kaui and A. Kaur, "Detection of phishing webpages using weights computed through genetic algorithm," in 2015 *IEEE* 3rd International Conference on MOOCs, Innovation and Technology in Education (MITE), pp. 331–336, 2015.
[3] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Tutorial and critical analysis of phishing websites methods," *Comput. Sci. Rev*., vol. 17, pp. 1–24, Aug. 2015.
[4] V. Suganya, "A Review on Phishing Attacks and Various Anti Phishing Techniques," *Int. J. Comput. Appl.,* vol. 139, no. 1, pp. 20–23, 2016.
[5] A. Aleroud and L. Zhou, "Phishing environments, techniques, and countermeasures: A survey," *Comput. Secur.,* vol. 68, pp. 160–196, Jul. 2017.
[6] R. P. Ferreira et al., "Artificial Neural Network for Websites Classification with Phishing Characteristics," *Social Networking.* 7(2) pp. 97–109, 2018.
[7] M. F. Darmawan, N. I. Jamahir, R. D. R. Saedudin, and S. Kasim, "Comparison between ANN and Multiple Linear Regression Models for Prediction of Warranty Cost," *Int. J. Integr. Eng.,* vol. 10, no. 6, 2018.
[8] R. M. A. Mohammad, T. L. Mccluskey, and F. Thabtah, "Predicting Phishing Websites using Neural Network trained with Back - Propagation," *In Proceedings of the 2013 World Congress in Computer Science, Computer Engineering, and Applied Computing*, Las Vegas, Nevada, pp. 1–6, 2013.
[9] M. Darmawan, H. Hasan, S. Sadimon, S. Yusuf, and H. Haron, "A Hybrid Artificial Intelligent System for Age Estimation Based on Length of Left Hand Bone," *Adv. Sci. Lett.,* vol. 24, no. 2, pp. 1047–1051, 2018.

[10] M. F. Darmawan, S. M. Yusuf, M. A. Rozi, and H. Haron, "Hybrid PSO-ANN for sex estimation based on length of left hand bone," in 2015 *IEEE Student Conference on Research and Development (SCOReD)*, pp. 478–483, 2015.
[11] A. A. Abdullah, L. E. George, and I. J. Mohammed, "Email Phishing Detection System Using Neural Network," *Research Journal of Information Technology*. 6(3):39-43 no. August, 2015.
[12] S. Gupta and A. Singhal, "Phishing URL detection by using artificial neural network with PSO," *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, Noida, 2017, pp. 1-6.
[13] J. Kozak and U. Boryczka, "Collective data mining in the ant colony decision tree approach," *Inf. Sci. (Ny).*, vol. 372, pp. 126–147, Dec. 2016.
[14] A. Priya and E. Meenakshi, "Detection of phishing websites using C4.5 data mining algorithm," in *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, Bangalore, 2017, pp. 1468-1472.
[15] R. Sankhyan, A. Shetty, L. Dhanopia, C. Kaspale, and G. Dantal, "PDS-Phishing Detection Systems," *Int. Res. J. Eng. Technol.*, vol. 5, no. 4, pp. 2429–2431, 2018.
[16] A. K. Shrivas and R. Suryawanshi, "Decision Tree Classifier for Classification of Phishing Website with Info Gain Feature Selection," *Int. J. Res. Appl. Sci. Eng. Technol.* , vol. 5, no. 5, pp. 780–783, 2017.
[17] I. Alsmadi and I. Alhami, "Clustering and classification of email contents," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 27, no. 1, pp. 46–57, Jan. 2015.
[18] M. N. Badadhe, M. S. More, and N. V Puri, "An Efficient Approach To Detecting Phishing A Web Using K-Means And Naïve-Bayes Algorithms With Results," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*. vol. 3, no. 5, pp. 1584–1589, 2014.
[19] A. Trevino, "Introduction to K-means Clustering," Oracle Data Science.com, 2016.
[20] K. Sahu and S. K. Shrivastava, "Kernel k-Means Clustering for Phishing Website and Malware Categorization," *Int. J. Comput. Appl.*, vol. 111, no. 9, pp. 20–25, Feb. 2015.
[21] M. Arab and M. K. Sohrabi, "Proposing a new clustering method to detect phishing websites," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 25, pp. 4757–4767, 2017.
[22] N. Kumar and P. Chaudhary, "Mobile Phishing Detection using Naive Bayesian Algorithm," *Int. J. Comput. Sci. Netw. Secur.*, vol. 17, no. 7, pp. 142–147, 2017.
[23] S. B. Rathod and T. M. Pattewar, "Content based spam detection in email using Bayesian classifier," in 2015 *International Conference on Communications and Signal Processing (ICCSP)*, pp. 1257–1261, 2015.
[24] R. Priya, "An Ideal Approach for Detection of Phishing Attacks using Naïve Bayes Classifier," *Int. J. Comput. Trends Technol.*, vol. 40, no. 2, p. 2016, 2016.
[25] M. Singh, "Classification of Spam Email Using Intelligent Water Drops Algorithm with Naïve Bayes Classifier," in *Progress in Advanced Computing and Intelligent Engineering*, pp. 133–138, 2019.
[26] L. Breiman, "Random Forests," Ma-chine Learn., vol. 45, no. 1, pp. 5–32, 2001.
[27] M. Bassiouni, M. Ali, and E. A. El-Dahshan, "Ham and Spam E-Mails Classification Using Machine Learning Techniques," *J. Appl. Secur. Res.*, vol. 13, no. 3, pp. 315–331, 2018.
[28] A. A. Akinyelu and A. O. Adewumi, "Classification of phishing email using random forest machine learning technique," *J. Appl. Math.*, 2014.
[29] H. Faris, I. Aljarah, and B. Al-Shboul, "A Hybrid Approach Based on Particle Swarm Optimization and Random Forests for E-Mail Spam Filtering," *in Computational Collective Intelligence*, pp. 498–508, 2016.
[30] S. Jagadeesan, A. Chaturvedi, and S. Kumar, "URL Phishing Analysis using Random Forest," *Int. J. Pure Appl. Math.*, vol. 118, no. 20, pp. 4159–4163, 2018.
[31] P. Patil, R. Rane, and M. Bhalekar, "Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm," in *2017 International Conference on Inventive Systems and Control (ICISC)*, pp. 1–4, 2017.
[32] M. Zouina and B. Outtaj, "A novel lightweight URL phishing detection system using SVM and similarity index," *Human-centric Comput. Inf. Sci.*, vol. 7, no. 17, 2017.
[33] Y. Li, S. Chu, and R. Xiao, "A pharming attack hybrid detection model based on IP addresses and web content," *Opt. - Int. J. Light Electron Opt.*, vol. 126, no. 2, pp. 234–239, Jan. 2015.
[34] R. Karnik and G. M. Bhandari, "Support Vector Machine Based Malware and Phishing Website Detection," *IJCAT-International J. Comput. Technol.*, vol. 3, no. 5, pp. 295–300, 2016.
[35] L. A. dos S. Gruginskie and G. L. R. Vaccaro, "Lawsuit lead time prediction: Comparison of data mining techniques based on categorical response variable," *PLoS One*, *vol.* 13, no. 6, pp. 1–26, 2018.
[36] M. Hemmat Esfe, M. Reiszadeh, S. Esfandeh, and M. Afrand, "Optimization of MWCNTs (10%)–Al2O3 (90%)/5W50 nanofluid viscosity using experimental data and artificial neural network," *Phys. A Stat. Mech. its Appl.*, vol. 512, pp. 731–744, Dec. 2018.
[37] S. Yang, Q. Feng, T. Liang, B. Liu, W. Zhang, and H. Xie, "Modeling grassland above-ground biomass based on artificial neural network and remote sensing in the Three-River Headwaters Region," *Remote Sens. Environ.*, vol. 204, pp. 448–455, Jan. 2018.
[38] I. Ali, F. Cawkwell, E. Dwyer, B. Barrett, and S. Green, "Satellite remote sensing of grasslands: from observation to management," *J. Plant Ecol.*, vol. 9, no. 6, pp. 649–671, 2016.
[39] R. C. Deo and M. Şahin, "Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland," *Renew. Sustain. Energy Rev.*, vol. 72, pp. 828–848, May 2017.

[40] F. Ghasemi, A. Mehridehnavi, A. Pérez-Garrido, and H. Pérez-Sánchez, "Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks," *Drug Discov. Today*, vol. 23, no. 10, pp. 1784–1790, Oct. 2018.
[41] R. Féraud and F. Clérot, "A methodology to explain neural network classification," *Neural Networks*, vol. 15, no. 2, pp. 237–246, Mar. 2002.
[42] L. Li, L. Sun, G. Ning, and S. Tan, "Automatic Pavement Crack Recognition Based on BP Neural Network," *Promet-Traffic&Transportation*, vol. 26, no. 1, 2014.
[43] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced Spectral Classifiers for Hyperspectral Images: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 8–32, 2017.
[44] Aized Amin Soofi and Arshad Awan, "Classification Techniques in Machine Learning: Applications and Issues," *J. Basic Appl. Sci.*, vol. 13, pp. 459–465, 2017.
[45] A. Yasin and A. Abuhasan, "An Intelligent Classification Model For Phishing Email Detection," *Int. J. Netw. Secur. Its Appl.*, vol. 8, no. 4, 2016.
[46] K. Kim, "A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree," *Pattern Recognit.*, vol. 60, pp. 157–163, Dec. 2016.
[47] Jasmina Novaković, Perica Strbac, and Dusan Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugosl. J. Oper. Res.* , vol. 21, no. 1, pp. 191–135, 2011.
[48] X. YANG, L. YAN, B. YANG, and Y. LI, "Phishing Website Detection Using C4.5 Decision Tree," in *International Conference on Information Technology and Management Engineering (ITME 2017)*, pp. 119–124, 2017.
[49] J. Qian, "A Survey on Sentiment Classification in Face Recognition," *J. Phys. Conf. Ser.*, vol. 960, no. 1, pp. 1–7, 2018.
[50] K. Singh, D. Malik, and N. Sharma, "Evolving limitations in K-means algorithm in data mining and their removal," *IJCEM (International Journal of Computational Engineering & Management)*. April 2011, vol 12, 105-109.
[51] D. Bzdok, M. Krzywinski, and N. Altman, "Machine learning: Supervised methods, SVM and kNN," *Nat. Methods*, pp. 1–6, 2018.
[52] Lopamudra Dey, Sanjay Chakraborty, Anuraag Biswas, Beepa Bose, and Sweta Tiwari, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier," *Int. J. Inf. Eng. Electron. Bus.*, vol. 8, no. 4, pp. 54–62, 2016.
[53] Meherwar Fatima and Maruf Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 9, pp. 1–16, 2017.
[54] B. Yang, Y. Lei and B. Yan, "Distributed Multi-Human Location Algorithm Using Naive Bayes Classifier for a Binary Pyroelectric Infrared Sensor Tracking System," in *IEEE Sensors Journal*, vol. 16, no. 1, pp. 216-223, Jan.1, 2016.
[55] E. Miranda, E. Irwansyah, A. Y. Amelga, M. M. Maribondang, and M. Salim, "Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier," *Heal. Inf. Res*, vol. 22, no. 3, pp. 196–205, Jul. 2016.
[56] S.Archana and K.Elangovan, "Survey of Classification Techniques in Data Mining," *Int. J. Comput. Sci. Mob. Appl.*, vol. 2, no. 2, pp. 65–71, 2014.
[57] Sau Loong Ang, Hong Choon Ong, and Heng Chin Low, "Classification Using the General Bayesian Network," *Pertanika J. Sci. Technol.*, vol. 24, no. 1, pp. 205–211, 2016.
[58] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Stat. Comput.*, vol. 27, no. 3, pp. 659–678, 2017.
[59] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *J. Ambient Intell. Humaniz. Comput.*, pp. 1–14, Apr. 2018.
[60] A. M. Youssef, H. R. Pourghasemi, Z. S. Pourtaghi, and M. M. Al-Katheeri, "Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia," *Landslides*, vol. 13, no. 5, pp. 839–856, 2016.
[61] Y. Tang, S. Krasser, Y. He, W. Yang, and D. Alperovitch, "Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis," in *IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference*, pp. 1–5, 2008.
[62] F. Cánovas-García, F. Alonso-Sarría, F. Gomariz-Castillo, and F. Oñate-Valdivieso, "Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery," *Comput. Geosci.*, vol. 103, pp. 1–11, Jun. 2017.
[63] L. Ma and S. Fan, "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests," *BMC Bioinformatics*, vol. 18, no. 1, p. 169, 2017.
[64] Y. Wang, S.-T. Xia, Q. Tang, and J. Wu, "A novel consistent random forest framework: Bernoulli random forests," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 8, pp. 3510–3523, 2018.
[65] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben, "Machine learning in manufacturing: advantages, challenges, and applications," *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, 2016.
[66] G. Guo, S. Z. Li, and K. L. Chan, "Support vector machines for face recognition," *Image Vis. Comput.*, vol. 19, no. 9–10, pp. 631–638, Aug. 2001.
[67] Y. Lin *et al.*, "Large-scale image classification: Fast feature extraction and SVM training," *CVPR 2011*, Colorado Springs, CO, USA, 2011, pp. 1689-1696.

[68] H. Byun and S.-W. Lee, "Applications of Support Vector Machines for Pattern Recognition: A Survey," in Pattern Recognition with Support Vector Machines, 2002, pp. 213–236.
[69] L. Auria and R. A. Moro, "Support vector machines (SVM) as a technique for solvency analysis," DIW Berlin Discuss. Pap., pp. 1–16, 2008.
[70] H. Källén, J. Molin, A. Heyden, C. Lundström and K. Åström, "Towards grading gleason score using generically trained deep convolutional neural networks," *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, Prague, 2016, pp. 1163-1167.
[71] A. Firdaus, N. B. Anuar, M. F. A. Razak, I. A. T. Hashem, S. Bachok, and A. K. Sangaiah, "Root Exploit Detection and Features Optimization: Mobile Device and Blockchain Based Medical Data Management," *J. Med. Syst.*, vol. 42, no. 6, 2018.
[72] M. F. A. Razak, N. B. Anuar, F. Othman, A. Firdaus, F. Afifi, and R. Salleh, "Bio-inspired for Features Optimization and Malware Detection," *Arab. J. Sci. Eng.*, 2017.
[73] M. F. A. Razak, N. B. Anuar, R. Salleh, A. Firdaus, M. Faiz, and H. S. Alamri, "'Less Give More': Evaluate and zoning Android applications," *Meas. J. Int. Meas. Confed.*, vol. 133, pp. 396–411, 2019.
[74] A. Firdaus, N. B. Anuar, A. Karim, and M. F. A. Razak, "Discovering optimal features using static analysis and genetic search based method for android malware detection," *Front. Inf. Technol. Electron. Eng.*, pp. 1–27, 2017.
[75] D. N. A/L Eh Phon, M. H. Abdul Rahman, N. I. Utama, M. B. Ali, N. D. Abd Halim, and S. Kasim, "The Effect of Augmented Reality on Spatial Visualization Ability of Elementary School Student," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 9, no. 2, pp. 624–629, 2019.
[76] D. N. E. Phon, M. B. Ali, and N. D. A. Halim, "Learning with augmented reality: Effects toward student with different spatial abilities," *Adv. Sci. Lett.*, vol. 21, no. 7, pp. 2200–2204, 2015.
[77] Rafei, N.S.I.M., Hassan, R., Saedudin, R.D.R., Raffei, A.F.M., Zakaria, Z., Kasim, S. "Comparison of feature selection techniques in classifying stroke document". *Indonesian Journal of Electrical Engineering and Computer Science*, 14 (3), pp.1244-1250, 2019.
[78] Ibrahim, A.O., Shamsuddin, S.M., Saleh, A.Y., Ahmed, A., Ismail, M.A. and Kasim, S. "Backpropagation Neural Network Based on Local Search Strategy and Enhanced Multi-objective Evolutionary Algorithm for Breast Cancer Diagnosis". *International Journal on Advanced Science, Engineering and Information Technology*, 9(2), pp.609-615, 2019.
[79] Omar, N.A., Kasim, S. and Fudzee, M. F. M. "A review of semantic similarity approach for multiple ontologies". *International Journal of Information and Decision Sciences*, 10(3), pp.212-221, 2018.
[80] Khaleel, M.K., Ismail, M.A., Yunan, U. and Kasim, S., 2018. "Review on Intrusion Detection System Based on The Goal of The Detection System". *International Journal of Integrated Engineering*, 10 (6).
[81] Zunaidi, W.H.A.W., Saedudin, R.R., Shah, Z.A., Kasim, S., Seah, C.S. and Abdurohman, M. "Performances Analysis of Heart Disease Dataset using Different Data Mining Classifications". *International Journal on Advanced Science, Engineering and Information Technology*, 8 (6), pp.2677-2682, 2018.
[82] Mohd, A.I., Mohamed Alhaj, A., Shahreen, K., Ashraf, O.I., Anik, H.A., Saima, A.L. and Ali, A. "An Enhancement of Multi Classifiers Voting Method for Mammogram Image based on Image Histogram Equalization". *International Journal of Integrated Engineering.* 2018 Vol. 10 No. 6. p. 209-213.
[83] Saedudin, R.R., Kasim, S., Mahdin, H. and Yanto, I.T.R., 2017. A Comparative Analysis of Rough Sets for Incomplete Information System in Student Dataset. *International Journal on Advanced Science, Engineering and Information Technology*, 7 (6), pp.2078-2084.
[84] Rahman, N.A.A., Hassan, R., Zakaria, Z. and Kasim, S., 2017. NIMSAD Framework to Evaluate the Role-based Goal Modelling. *International Journal on Advanced Science, Engineering and Information Technology*, 7 (5), pp.1728-1734.
[85] Hassan, R., Othman, R.M., Asmuni, H. and Kasim, S., 2017. Incorporating Multiple Biology based Knowledge to Amplify the Prophecy of Enzyme Sub-Functional Classes. *International Journal on Advanced Science, Engineering and Information Technology*, 7 (4), pp.1479-1485.
[86] Arieef, M.F., Kasim, S., Choon, Y.W., Mohamad, M.S., Deris, S. and Napis, S. "A Hybrid of Integer Differential Bees and Flux Balance Analysis for Improving Succinate and Lactate Production". *International Journal on Advanced Science, Engineering and Information Technology*. 2017, 7(4-2), pp.1615-1620.
[87] Kusairi, R.M., Moorthy, K., Haron, H., Mohamad, M.S., Napis, S. and Kasim, S. An Improved Parallelized mRMR for Gene Subset Selection in Cancer Classification. *International Journal on Advanced Science, Engineering and Information Technology*. 2017, 7(4-2), pp.1595-1600.
[88] Hassan, R., Kasim, S., Ning, N.K., Ramlan, R., Salamat, M.A. and Saedudin, R.R. "Analysis of Multi-Stakeholder Requirements Using Requirement Interaction Matrix". *International Journal on Advanced Science, Engineering and Information Technology*, 2017, 7 (4-2), pp.1498-1503.
[89] Muhamad, R., Samah, A.A., Majid, H.A., Sulong, G., Mohamad, M.S. and Kasim, S. "Review on Local Binary Patterns Variants as Texture Descriptors for Copy-Move Forgery Detection". *International Journal on Advanced Science, Engineering and Information Technology*, 2017, 7(5), pp.1678-1684.
[90] Chan, W.H., Mohamad, M.S., Deris, S., Zaki, N., Kasim, S., Omatu, S., Corchado, J.M. and Al Ashwal, H. "Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme. Computers in biology and medicine", 2016, 77, pp.102-115.
[91] S. Kasim, M. F. M. Fudzee, S. Deris and R. M. Othman, "Gene Function Prediction Using Improved Fuzzy c-Means Algorithm," *2014 International Conference on Information Science & Applications (ICISA)*, Seoul, 2014, pp. 1-4.