Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

HYPER PARAMETERS TUNING USING PARTIAL SWARM OPTIMIZATION ALGORITHM BASED ON RANDOM FOREST FOR URLs-BASED PHISHING DETECTION

DUMUA AL AHMARE
College of Computing and Informatics, Saudi Electronic University, Riyadh, KSA.
Email: g200005686@seu.edu.sa

SAIMA ANWAR LASHARI
College of Computing and Informatics, Saudi Electronic University, Riyadh, KSA.
Email: s.lashari@seu.edu.sa

ABDULLAH KHAN
Institute of Computer Sciences and Information Technology, the University of Agriculture Peshawar, Pakistan. Email: abdullah_khan@aup.edu.pk

SANA SALAH-UDDIN
Institute of Computer Sciences and Information Technology, the University of Agriculture Peshawar, Pakistan. Email: sanabatoor@gmail.com

Abstract

Creating phishing URLs is a common deception technique in phishing attack as appear to be legitimate website. Phishing URLs can cause serious dangers once loaded by the web browser such as drive-by download and crypto jacking attacks, therefore it is highly important to focus on identifying and preventing phishing URLs in early stages. The detection of phishing attack is a supervised classification process that make use of a labeled dataset to fit Machine Learning (ML) models and classify the data. Several security researchers came up with various ML techniques that able to detect and classify the website phishing However, phishing attack detection with high accuracy is still challenging. In this study, Sequential Forward Feature Selection (SFFS) technique is implemented to find the optimal set of features and Practical Swarm Optimization (PSO) hyper parameter tuning technique is developed to tune the hyper parameters of the Random Forest classifier to identify and detect phishing website by utilizing URLs -based features in tow phishing datasets. The result of the proposed technique showed the best and outperformed other ML techniques such as:(RF, LR, KNN, SVM and NBC) in terms of accuracy score as well as other classification performance measures.

Keywords: Machine Learning, RF, LR, KNN, SVM, Sequential Forward Feature and NBC.

1. INTRODUCTION

Nowadays, most of industries mainly rely on computers to perform various functionalities which greatly consist of valuable data. Data is targeted by various attacks while it is stored or in transit where these attacks can cause a huge damage to the industries' value as well as individuals when they compromise their confidentiality, integrity and availability [1]. Phishing attack is a popular cybercrime form, that pose significant threats to the privacy and security of users as it can performed using various channels, therefore, phishing attack can be defined as a network attack that marge both social engineering

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

and computer technology and the main objective of this attack is to steal the sensitive information of individuals and organization by tricking them to provide their sensitive information such as username, passwords, social security number and bank account number [2].

In addition, phishers using various tactics strategies including phishing e-mails, URLs, text massages, and phone calls as phishing, attacker designs contents that mostly similar the legitimate content in order to deceive the victim to provide sensitive information [1]. However, the main intent of phishers when carrying out the attacks is to perform more malicious actions such as sell the victim's personality to others, get financial profits and exploit the system's vulnerability [3]. As result of the rapid development of communication technologies and the global networks the number cyber-attacks including phishing are on raise. Therefore, phishing attacks considered a serious concern too many researchers. However, based on the report of phishing Activity Trends of the fourth Quarter of 2020, the observed number of phishing attacks are growing though 2020. Amazon, Prime Day phishing attack and Google Docs Invitation are examples of commonly phishing attacks [1]. Educating internet users and deploying technical defenses are in general parts of anti-phishing tactics, the technical defense strategies including list-based approach which involve the process of collecting, validating, and identifying phishing and legitimate websites then add them to the whitelist and blacklist to be shared with other users, this approach prevent the reuse of the identified phishing URLs. However, this approach cannot detect new phishing URLs. In the other hand, heuristic-based detection approach which works by extracting the webpage contents features and using third party services such as the rank of the page and domain age to the decide the phishing website. However, these services can be restricted and unable to detect Phishing Attack Mechanisms -Types Social Engineering based -Phishing Attack Malware-based Phishing Attack Figure 1.1:Phishing Attack Mechanisms 3 phishing websites [3]. Due to the existing weakness of these approaches many security researchers have involved the use of Machine learning approach to handle phishing attack in various shapes[4].

Phishing attack is serious security issue that pose a significant threat to the online users as it ranked as a top security threats. Phisher employs various techniques in order to lure the victim to provide personal information. Many investigations have been conducted against phishing that using characteristics that employed by attackers such as the websites URLs, content, source code, however, creating phishing URLs one of the techniques that used in phishing attack as attacker create a phishing URLs that mimic the legitimate one and jamming users [3].

Phishing URLs can cause serious dangers once loaded by the web browser such as drive-by download and crypto jacking attacks, therefore it is highly important to focus on identifying and preventing phishing URLs in early stages in order to prevent such attacks [5]. Due to the dynamic nature of the URLs, traditional detection approaches are not able to identify the new created URLs.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

Machine learning functioning by extracting features from the gathered data to determine relationship among them and using classification technique URLs are classified as either legitimate or phishing. However, achieving high accuracy in phishing detecting is still a challenge [4].

The development of machine learning approach has emerged various ML techniques for detecting phishing attacks in various channels such as spoofed pages, phishing emails and fake URLs. ML techniques that used in phishing which are supervised classification algorithms that use labeled dataset to fit the model and classify data [2].

The significant increase in the activities of websites phishing attacks and as websites phishing attacks pose a serious threat to the privacy and security to our information when using the internet are the main motivations behind this study. Machine learning approach has emerged various ML techniques to reduce the risk of phishing attacks, therefore several researchers came up with machine learning techniques that able to detect and classify the website phishing with high accuracy when the features are satisfying, commonly using supervised algorithms such as Naves neural network, Linear regression, Decision tree, Random Forest, Support vector machine and K-nearest neighbor, The main objective of this study to contract a Sequential Forward Feature Selection (SFFS) technique to find the optimal set of features and Practical Swarm Optimization (PSO) hyper parameter tuning technique to tune the hyper parameters of the Random Forest classifier to identify and detect phishing website.

This study focuses on detecting and preventing phishing URLs using hyper parameters tuning machine learning technique based on Random Forest algorithm that implemented using URLs based features of legitimate and phishing website. The main contribution of this paper as given below:

i. To implement Forward Feature Selection (SFFS) technique will be implemented to find the optimal set of features.

ii. To design and implement the proposed hyper parameters tuning techniques (Particle Swarm Optimization (PSO) based on Random Forest algorithm and compare it with the existing ML techniques, such as Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB).

iii. The performance of the proposed technique will compared with the other constructed ML techniques mention in (ii) in term of accuracy, precision, recall, and the confusion metrics (True positive (TP), True negative (TN), False positive (FP) and False negative (FN).

The publication still has four sections. The related work is explain in Section 2. While the training model and proposed methodology is explained in Section 3. The findings and analyses are presented in Section 4, and the conclusions and recommendations are presented in Section 5.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

2. RELATED WORK

Phishing strategies come in a variety of shapes and sizes. Website phishing is one of the most dangerous types of phishing because individuals are too preoccupied with finding a certain piece of information on these websites to notice some illicit phishing actions going on in the background. Based on the report of phishing Activity Trends of the fourth Quarter of 2020, the observed number of phishing attacks are growing though 2020. As phishing activates involve the use of phishing sites. Google Docs invitation is commonly known example of phishing attack where in May 2017 cybercriminals delivered forged invitation to the users of google to modify certain document, when recipients clicked the invitation, they were redirected to a third party-app to make the phishing process easier to get access to their sensitive information [1]. Phisher usually following constant process that consisting of planning phase, composed fake URLs, emails or text massages, attack phase, information gathering phase and finally fraud thus phishing process starts when the attacker planning for the attack by creating imitated website that similar to a legitimate one in order to deceive the victim and gather the required information, deliver that imitated website to the victim thought multiple channels such as emails and text messages then once victim is deceived can easily provide the required sensitive information to the attacker and attacker now is able to use that gathered information in the legitimate website and commits cybercrimes including financial fraud. phishing attacker can use various mechanisms to target online users and convince them to provide their sensitive information [1]. Phishing attack is a critical security issue that threaten intent users, where phisher attempts to take use of the user's vulnerability and that is too difficult to be mitigated, as result many of attempts that focusing on improving phishing detection systems. Machine Learning approach can handle the problem of phishing attacks by transforming the problem into a classification technique. Therefore, several works focused on improving the phishing detection using ML techniques with high accuracy performance. This section presents various ML techniques that used to help to improve the detection of web phishing attacks.

In [6] the author proposed an approach named meta-heuristic approach that helps to protect internet users from the web phishing attack, the data was collected from benchmark databases which are University of California Irvine and University of Hudders field with total number 13,756 records, this approach performed by first analyzing and ranking web features which are URLs, JavaScript code, HTLM, page images and text and the domain name of the webpage, second features extracting which involves extraction the most effective ranked features that important to improve the accuracy of detection, In the third step Random Forest classifies is applied using the selected subset data features. Finally, the accuracy of classifier was evaluated and achieved 96.33%.

Further in [7] the authors implemented an features selection as optimization techniques in order to improve the accuracy of the classifier, the experiment data was collected from UCI the repository of machine learning as it contains 1125 total number of records as 1185 considered as legitimate and 10030 as phishing , so, First, several features

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

selection filters algorithms including Gain Ratio, Info Gain Ratio, One R Attribute, Relief Attribute and Symmetric Uncertainty Attributes Evaluator were applied and their output are analyzed to find out the important features to be used in the classification phase, in the classification phase authors have applied many classification algorithms which are Random forest, K-NN, Decision Tree with and without the use of the proposed filters and result the classifier Random forest provided butter performance with high accuracy performance as it achieved 99% after implementing the features fileting.

Where else in [8] conducted a comprehensive analysis of different classification techniques which are, Decision Tree, Random forest, XGBOOST, Support Vector Machine and Multilayer perception, the experiment was performed on collected data that contains legitimate and phishing URLs as the address bar, Domain and webpage dependent are the main features that used, however the result of this experiment showed that the classifier XGBOOST performed very well when compared with others as the score of the classifier was 86% In the train set and 85% in the test set. In this paper [9], the authors have explored ability of ML techniques to identify phishing URLs. The dataset was collected and features extracted from 12 different sources which are Phish Tank for and Miller Smiles as 30 URLs features that used in this works categorized into Address bar features, abnormal features, JavaScript and HTML codes and domain features, so, they first have implemented the wrapper- based as feature selection technique as it contains the URLs metadata, in this project , several classical classification ML techniques are implemented which K-NN, RF, DT, SVC linear and one class SVM. As result Random Forest classifier overcome other classification techniques with about 96.87% accuracy as it classified as the most suitable classifier compared to others. In [10] this study the author propose the use of Extreme machine learning to handle phishing attack using URLs, the main objective of this experiment is to construct a real-time phishing detection, Extreme machine learning known as a feed- forward artificial neural network- ANN as it is a tool used with ML, ELM consists of layers of input, output and hidden layers and it is helpful to avoided the overfitting problem, furthermore, 30 URLs phishing features were used which are the Address bar features, abnormal features, JavaScript and HTML codes and domain features, so, the final result showed that ELM achieved accuracy 96.93% which is the best compared to other ML classification algorithms that implemented such as SVM.

In [4] this paper construct a machine learning model to handle Phishing URLs, so, the author implemented several ML algorithms such as Random forest, Decision Tree, Linear model, and Neural network on a dataset that collected from multiple sources which are Phish Tank and Miller Smiles as the it contains 2456 websites with 30 extracted features. Furthermore, all applied algorithms were compared based on the accuracy of the detection, the rate of true positive and true negative and the F measures, as result Random Forest classifier showed the best result in terms of accuracy when compared with Decision Tree, Linear model, and Neural network as it archived 95.70 %.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

In [11] this pepper the authors have discussed a framework that based on machine learning approach that using a hybrid URLs features in order to handle URLs phishing attacks, Moreover, the hybrid features were based on the length, letter counts, numbers and binary values as this concept helps to construct a robust classifiers, furthermore, many ML classification algorithms were implemented which are k-Nearest Neighbors, Support Vector Machine, Decision Tree, Logistic Regression, Stochastic Gradient Descent, Random Forest, Gradient Boosting classifier, Adaptive Boosting and Extreme Gradient Boosting. As the result showed that Adaptive Boosting performed the best with high accuracy as it achieved 99.7%.

Table 1: Comparison of ML techniques of recent studies in phishing detection.

Ref	Techniques	Features	Experiments Results
[4]	ML algorithms such as Random Forest, Decision Tree, Linear model, and Neural network.	2456 websites with 30 extracted features.	Neural network with accuracy score 95.70 %.
[6]	Meta-heuristic approach (Feature Selection Technique) +Novel ML techniques (DT, RF, SVM, KNN)	URLs, JavaScript code, HTLM, page images and text and the domain name of the webpage,	Random Forest classifies with accuracy score 96.33%.
[8]	Different classification techniques which are, Decision Tree, Random Forest, XGBOOST, Support Vector Machine and Multilayer perception	Address bar, Domain and webpage dependent Features.	XGBOOST with accuracy score classifies 86%
[10]	Extreme machine learning, artificial neural network-compared to other ML classification algorithms that implemented such as SVM.	features, abnormal features, JavaScript and HTML codes and domain features	ELM classifier with accuracy score 96.93%
[9]	Random Forest, K nearest neighbors, Decision Tree, Linear SVC classifier, One class SVM Classifier)	Address bar, abnormal features, JavaScript and HTML codes and domain features	Random Forest classifies with accuracy score 96.87%
[11]	ML techniques (k-Nearest Neighbors, Support Vector Machine, Decision Tree, Logistic Regression, Stochastic Gradient Descent, Random Forest, Gradient Boosting classifier, Adaptive Boosting and Extreme Gradient Boosting)	hybrid URLs features, based on length, letter counts, numbers and binary values	Adaptive Boosting classifies with accuracy score 99.7%.

However, phishing mechanisms can be classified as into two main categories which are first social engineering attack in this mechanism the attacker exploit the human errors and gain valuable information using URLs of imitated legitimate website that can be attached with phishing emails, advertisements and cracked licensed software. Second subterfuges using technical techniques which involve a malicious code can be delivered thought emails and websites as can be self-executable code and directly installed in the victim's PC in order to steal valuable information such as credentials, major technical techniques that used including Cross-site-Scripting in this technique the attacker mainly

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

using a malicious JavaScript code that embedded to generated websites that used by users to provide information, Malware based-phishing that involves a malicious code that installed in the victim's machine that store and deliver credentials to the phisher, Key logger that capturing and providing the phisher victim's actions including mouse motions and screenshots [12].

2.1 Phishing Detection Mechanisms

Cyber security is a field of information technology, and its main objective is to protect data, systems and other digital technologies from various attacks including unauthorized access to sensitive data. Web users are targeted by phishing attacks to obtain valuable information as phishing attacks continually expanding security risks and challenges are increasing as well. In general, phishing attacks can be avoided if fake websites are identified and consumers are carefully educated. Various anti-phishing mechanisms are developed to detect and prevent phishing attacks and each of these mechanisms has special advantages and disadvantages. 6 Based on the proposed taxonomy of anti-phishing detection mechanisms [12], anti-phishing detection mechanisms can be classified into two main categories which are:

I. Phishing Detection Based on Website Content

This mechanism improves the detection of phishing attacks by examine the content of the website using its features such as the website URLs, text including spilling, grammars and password, images, and the similarity of visuals. So, phishing detection based on website contents mechanism including the following various approaches that summarized in below table 2.

Table 2: Phishing Detection Approaches based on Website Contents

Phishing Detection Approaches based on website content	Description
Analyzing Website URLs	In this approach, the website's URL is examined by extracting features that help to determine if the link is legitimate or malicious. Examples of URLs features including existence of IP address rather than domain name, presence of special characters such as (@), dots numbers and hexadecimals, link's prefix and suffix, the length of the URL and the HTTPS in the domain. Moreover, this approach used by other approaches such as Rule-based approach and Machine learning approaches in order to detect phishing URLs.
Visual similarity-based approach- Analyzing Website Images	In this approach, the website's visuals including logs, CAPTCHAs and images are examined to determine if the website is legitimate or malicious.
Text similarity -based approach- Analyzing Website Text content	In this approach, the website's text content is mainly examined including page scripts, keywords and if the secure socket layer-SSL is enabled. Moreover, this approach used by Machine learning approach as phisher using similar keywords to the actual one.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

II. Phishing Detection Based on Non-Content Website

This mechanism detects phishing attacks by examine the website's features rather that its contents using various approaches that summarized in below table 3.

Table 3: Phishing Detection Approaches based on non-Content Website.

Phishing Detection Approaches based on non-content website	Description
List-based approach, Whitelist and Blacklists	In this approach, the new URL is checked, if the URL is existed in the whitelist, then considered as legitimate otherwise not and the same way using the blacklist if the URL existed in the blacklist considered as phishing URL otherwise not.
Domain Property -based approach	In this approach, the domain information is mainly used to decide whether the website is legitimate of not. Examples of domain information including the details of domain registration, the authority of the website certificate and the details of the website certificate. The process starts when the browser extension the clicked link to the server and extract domain information to be verified from google so, the user then will be alerted based on the result.
DNS- based approach	In this approach, the detection process is based on DNS information to decide the authenticity of domain name and the IP address.

2.2 Machine Learning in Phishing Detection

Phishing assaults are on the raise, making phishing detection a top need for developers and researcher, Traditional detection approaches such as the list-based approach including backlisting and whitelisting works by collecting both legitimate and phishing websites are considered in most antiviruses, intrusion detection systems and email spams filtering. However, the list-based detection approach is not quite efficient due to the continues changes in the behavior of website as result, determining whether a website is a legitimate or malicious become a major challenge [13]. Generally, accuracy and scalability are the overall limitations of the list-based detection approach [14]. In the other hand, Rule-based detection approach which works generating rules based on the process of analyzing the website including the URL's standard components such as the subdomain, used protocols, query parameters, path and port. For instance, if the used domain name just similar to another domain name. However, this approach requires third-party service request such as domain date and popularity in order to classify the website and phisher can easily learn about the rules when published as result new phishing URLs that not match these rules are generated in order to bypass the rule. Predication accuracy improvement is a major challenge that need to be solved in order to predict the new emerging phishing techniques, Therefore, machine learning approach has emerged to increase predications performance of detecting phishing websites [2]. Machine learning known as a multidisciplinary approach that was first used to form analytical models in supervised learning as it's important in a various applications including data mining and

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

image recognition. Mainly, it makes the use of various algorithms to build models and based on input data it makes predications without the use of explicitly programs, using part of the whole dataset the model trained as it called a training set then the model performance is 8 tested using the remaining part of the dataset. In terms of phishing detection, machine learning considered a helpful approach to determining whether a website is a legitimate or malicious by convert the problem into a classification task as classification in machine learning is supervised technique that helps to classify the new observation. The approach of Machine learning can be applied in different areas of phishing attack as phisher using various channels and techniques such as phishing emails and websites [9].

3. METHODS AND MATERIALS

In this study, will implement Sequential Forward Feature Selection (SFFS) technique to find the optimal set of features and Practical Swarm Optimization (PSO) hyper parameter tuning technique to tune the hyper parameters of the Random Forest classifier for the aim of URLs phishing detection and classification, as researcher believes that the use of hyper parameter tuning technique will improve the performance of the algorithm. Therefore, the process will go thought five major steps which are: 1. collecting data for the study, 2. Data preprocessing, 3. Features Selection, 4. ML Techniques Development, 5. Performance Evaluation. The figure below presents the overview of the methodology that used in this study.

3.1 Data Collection Phase

In this study, the proposed technique will be implemented in two different URLs based phishing datasets which intended to be used as for phishing detection systems that mainly using machine learning. The table 4 showing details of each dataset that are used in this study.

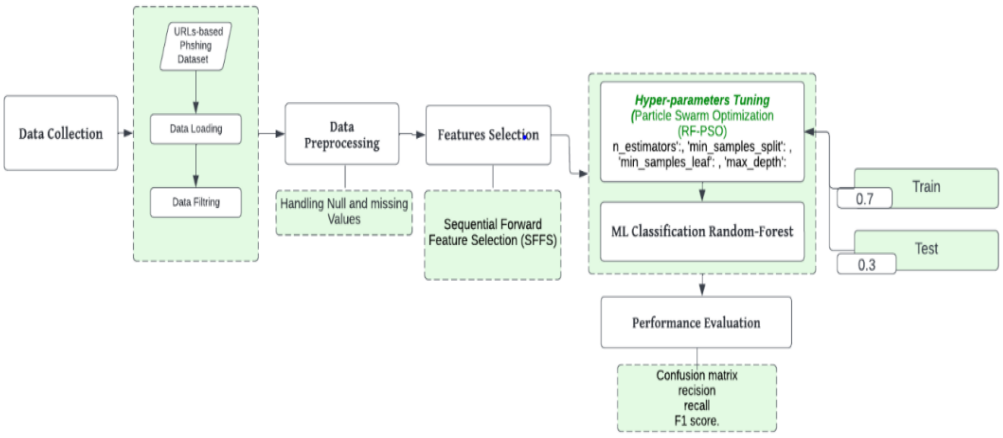


Figure 1: Hyper parameters tuning in Random Forest Algorithm for URLs-based Phishing Detection Methodology Process

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

Table 4: Details URLs based Phishing Datasets of the study

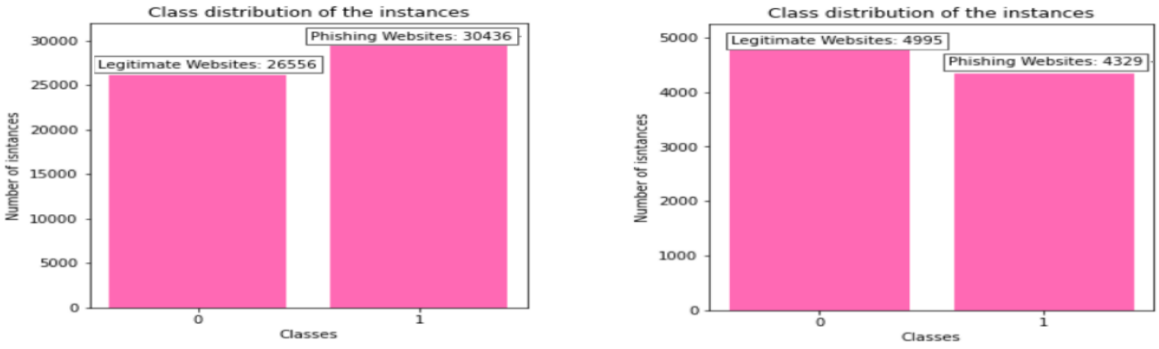
URLs phishing dataset	Dataset Volume	Features Categories
Dataset 1 [15]	58645 entries and 112 columns.	• URLs-based features that obtained by analyzing the structure and syntax of the URLs • Content based- features that extracted from the content of the URLs pages – External • Based features which extracted by querying external services.
Dataset 2 [16]	10000 entries and 50 columns.	• Attributes of URLs. • URLs parameters. • URL's domain name.

3.2 Data Preprocessing Phase

In this phase, all datasets will be tested for any missing values, and it will be handled and filled using the "median" value of that column if the data type of the column is numeric and for non-numeric columns the missing value will be filled using "mode", Null values and will be replaced by the "median" value of each attribute, duplicated values will be also removed. Furthermore, all columns will be examined for the categorical values using label encoding function that used to convert categorical value to number.

3.3 Features Selection Phase

Features selection process in ML has a great impact on the model's performance and important for classification to remove the irrelevant features. In this phase, Sequential Forward Feature Selection (SFFS) technique will be implemented to find the optimal set of features. SFFS is a technique of wrapper feature selection method. SFFS first works with an empty set of features and select the optimal features sequentially in every step so, once added SFFS is returned to the previous step to identify the worst features and remove them from the optimal set. Moreover, the accuracy score of each feature is calculated and the most accurate are only added to the optimal set [17]. Figure 3: show the Sequential Forward Feature Selection (SFFS) Workflow as below.



(a) Dataset 1 (b) Dataset 2

Figure 2: Class Distribution of Dataset 1 and Dataset 2

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

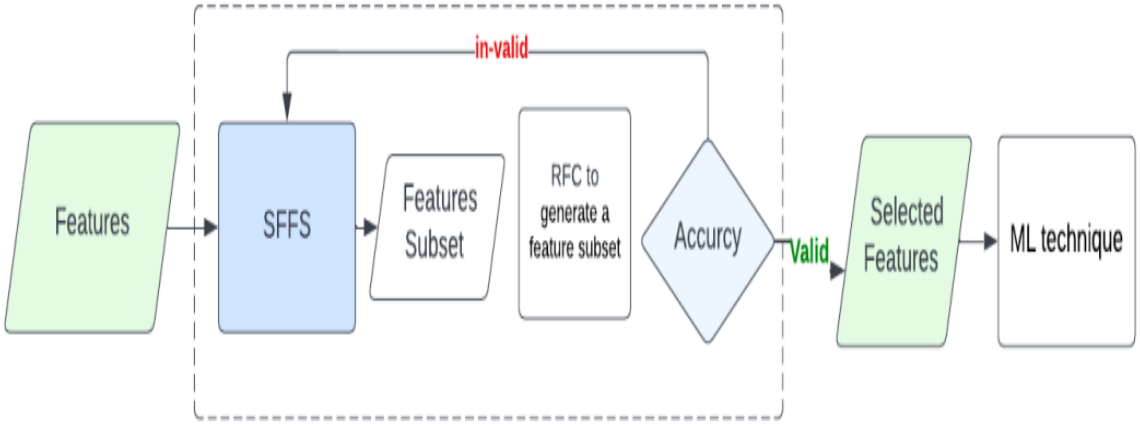


Figure 3: Sequential Forward Feature Selection (SFFS) Workflow

3.4 ML Techniques Development

Once the data is cleaned, missing values are filled and irrelevant features are removed, the data then will be divided into training dataset and testing dataset. In this study, the training dataset represents 70% of the dataset while the remaining 30% is used for testing purpose. After the data is being divided into training and testing. The hyper parameter tuning technique is used to find the best settings by attempting a variety of various combinations and evaluate each model's performance then the output of this stage will be fed to the Random Forest classification method to classify the URLs into phishing URL or legitimate URL.

3.4.1 Random Forest Algorithm

Random forest (RF) is an algorithm that commonly used to solve regression and classification problems, it mainly consists of decision tree from multiple samples and depending on majority vote for classification. Moreover, based on the decision's prediction RF algorithm determined the quality as it anticipates by averaging the output of several trees and as the number of trees grows the precision of the result improves.

RF algorithm has many characteristics, it can handle problems such as overfitting, sparse or missing data. In terms of classification problems, RF using an ensemble sparse or missing data. In terms of classification problems, RF using an ensemble research method to produce the desired result, it produces superior results as we want to know to which category the observation belongs to [9].

In this study, Python pandas 'scikit-learn package' is utilized to call the random forest classifier, the training dataset used to feed various decision trees for training purpose as the datasets including findings and attributes can be randomly chosen though node dissociation.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

The proposed algorithm as give as:

Input: A training set S=(x_i, y_i)…... (x_n, y_n) F feature, and number of trees in forest B

Output: D selected feature that have highest accuracy

1. *Start*
2. *Select M trees from the dataset*
3. *Construct a decision tree from the M trees*
4. *Repeat Step 1 and step 2 B time*
5. *At each node*
6. *Construct f as a tiny subset of F*
7. *Split on best feature in f*
8. *New recodes are given to the category that wins the most votes*
9. *End*

3.4.2 Hyper parameter Tuning Technique

Hyper parameter tuning technique is a process of finding the most optimal hyper parameter values for the learning algorithm as selecting the best hyper parameters is an important role for the performance of the trained model. In ML algorithms, the hyper parameter tuning technique is initialized before the model starts learning.

In Random Forest algorithm, the hyper parameter technique considers the number of decision trees in the forest or the number of nodes that each tree should have as well as the number of attributes that should be considered when splitting a node.

Moreover, hyper parameter tuning technique is based on experiment results rather than theoretical results, so, using various values of hyper parameter and then compare their results to find their best combination. In this study, Practical Swarm Optimization (PSO) hyper parameter tuning technique is implemented to tune the hyper parameters of the Random Forest classifier.

PSO use information sharing and collaboration among particles in group to find the best solution as Swarm in PSO is a collection of particles each is represented by a vector that holding its position, velocity and the best position then the performance score and the current position are calculated once the particle initialized.

As illustrated in figure 4 the dataset will be first divided into training and testing datasets then using the produced best hyper parameters the Random Forest classifier is tuned.

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

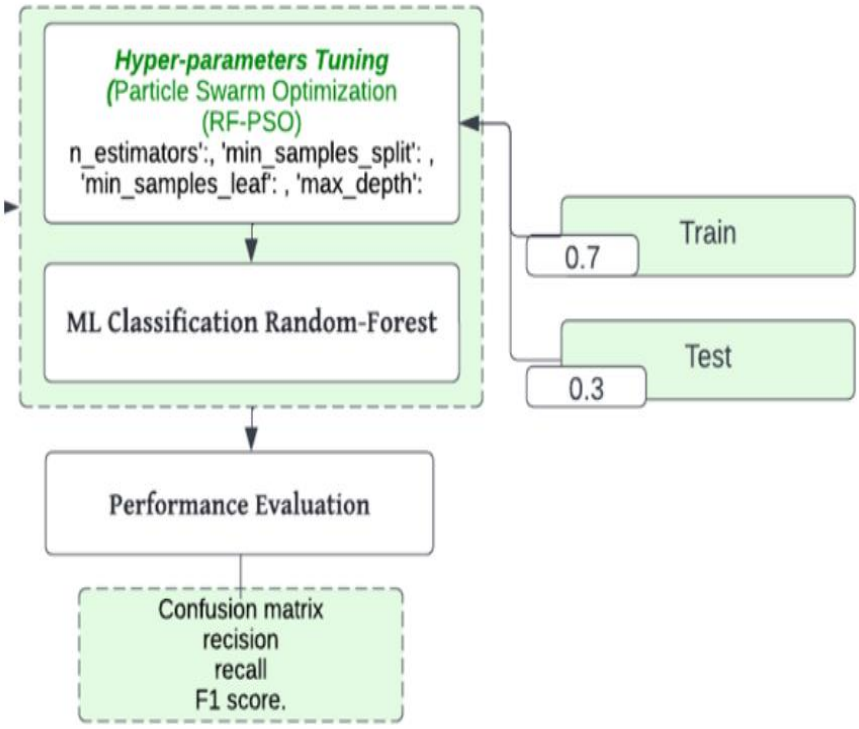


Figure 4: RF-PSO Hyper parameter Tuning Technique Process

3.5 Evaluation Parameters

To check the performance of used model various parameter are used in this paper. Efficiency of the proposed study are analysis and verifications of the performance of the suggested models are done using the following parameters.

i. Recall

Recall is the total number of positive values that were successfully detected is added together, and recall is computed by dividing that number by the total number of true positive and false negative values. "True Positive Rate" measurements refer to positive components that can be precisely detected. Cases with a high recall rate had the correct results.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (1)$$

ii. Precision

Precision is calculated by dividing the sum of true positives and false positives by the total number for correctly detected values.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

iii. F-Measure

To calculate F-Measure, recall & precision are utilized, as shown below:

$$\text{F-Measure} = \frac{2*Recall*Precision}{Recall+Precision} \quad (3)$$

iv. Accuracy

Accuracy reveals how confidently the model can distinguish between negative and positive classifications. It is calculable as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Here, the abbreviations TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively. In this study, mobile cost prediction dataset is used to evaluate the effectiveness of the proposed machine learning classifiers. The effectiveness of the suggested models is evaluated using accuracy, loss, precision, f-measure, and recall. The models employed both testing and training data. An 80:20 split was used to separate the data into testing and training groups.

4. RESULTS AND DISCUSSIONS

In this study, Sequential Forward Feature Selection (SFFS) technique is implemented to find the optimal set of features and Practical Swarm Optimization (PSO) hyper parameter tuning technique is developed to tune the hyper parameters of the Random Forest classifier in three URLs based phishing datasets. Therefore, in this section the result of this experiment is reported with respect of performance measures Accuracy (ACC), precision, recall, True Positive, False Positive, True negative and False negative. The proposed technique is first compared with other machine learning techniques that frequently used in classifying the phishing website second with other ML techniques that proposed by recent studies in URLs phishing detection.

4.1 Evaluation of Performance of ML Techniques with Default Hyper Parameters

The given table 5 provides information about two different datasets, their corresponding accuracy values, and the hyper parameters used for a Random Forest algorithm combined with Particle Swarm Optimization (RF-PSO). The table has three columns: "Dataset," "Accuracy," and "Hyper Parameters." "RF-PSO" refers to the combination of a Random Forest algorithm and Particle Swarm Optimization, which is a metaheuristic optimization technique. The table 5 provides the accuracy achieved by the RF-PSO algorithm on two different datasets. For Dataset 1, the accuracy achieved is 98.46%, and the hyper parameters used are 'n_estimators': 107, 'min_samples_split': 2, 'min_samples_leaf': 1, and 'max_depth': 112. For Dataset 2, the accuracy achieved is 99.57%, and the hyper parameters used are 'n_estimators': 126, 'min_samples_split': 2, 'min_samples_leaf': 1, and 'max_depth': 91. Accuracy represents the performance of the

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

RF-PSO algorithm in correctly predicting the target variable in the datasets. Higher accuracy values indicate better performance. Hyper parameters are the settings or configurations of the algorithm that can be adjusted to optimize its performance. In this case, the hyper parameters specify the number of estimators (decision trees) in the random forest ('n_estimators'), the minimum number of samples required to split an internal node ('min_samples_split'), the minimum number of samples required to be at a leaf node ('min_samples_leaf'), and the maximum depth of the trees ('max_depth'). Overall, the table 5 provides a summary of the accuracy achieved and the hyper parameter settings used for the RF-PSO algorithm on two different datasets.

Table 5: Accuracy Score of RF-PSO in the two Datasets

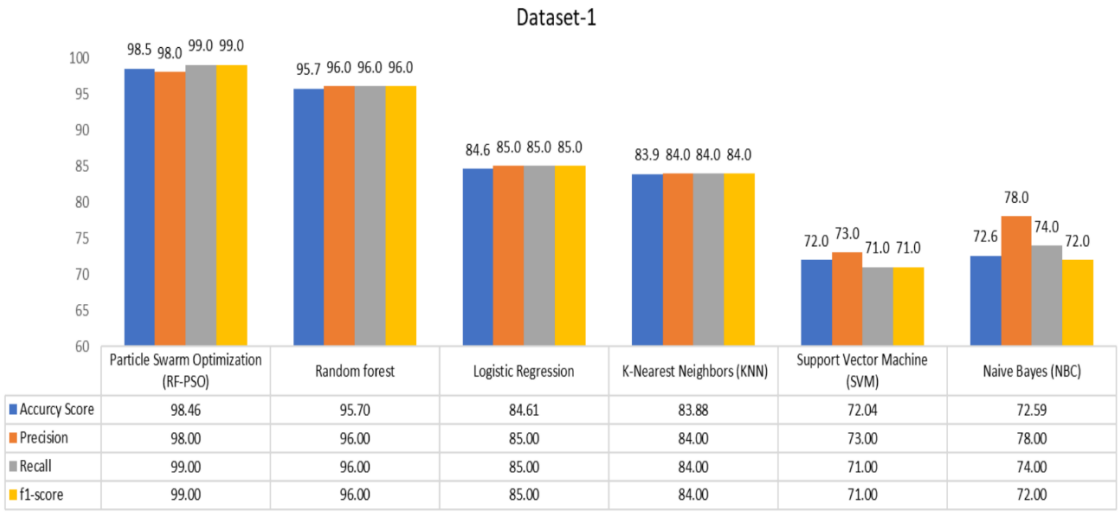
Dataset	Accuracy	Hyper Parameters
RF-PSO in Dataset 1	98.46	n_estimators': 107, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 112
RF-PSO in Dataset 2	99.57	n_estimators': 126, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 91

The given figure 5 represents the performance metrics of different algorithms on a dataset 1. The figure contains six columns: "Algorithms," "Accuracy," "Precision," "Recall," and "f1-score".The proposed RF-PSO algorithm achieved perform better as accuracy of 98.46%, precision: 99%, recall: 99%, and f1-score: 99%. Similarly RF achieved an accuracy up to 95%, precision: 96%, recall: 96%, f1-score: 96%. The LR has accuracy: 84.61%, Precision: 85%, Recall: 85%, f1-score: 85%. Where the KNN has the accuracy: 83.88%, precision: 84%, recall: 84%, f1-score: 84%. The SVM get the accuracy: 72.04%, precision: 73%, recall: 71%, f1-score: 71%. Furthermore the NB reached to achieve an accuracy: 72.99%, precision: 78%, recall: 74%, f1-score: 72%. Looking at the results, RF-PSO achieves the highest accuracy, precision, recall, and f1-score, indicating the best overall performance among the presented algorithms. RF follows closely behind RF-PSO in terms of performance metrics. LR and KNN show moderate performance. SVM and NB exhibit comparatively lower performance in terms of accuracy, precision, recall, and f1-score. In summary, the figure 5 provides a comparison of different algorithms' performance metrics on the dataset 1, indicating the strengths and weaknesses of each algorithm.

Similarly for the dataset 2 the figure 6 provides the performance metrics of different algorithms on a specific task. The table consists of five columns: "Algorithms," "Accuracy," "Precision," "Recall," and "f1-score." The RF-PSO achieved an accuracy: 97.57%, precision: 100%, recall: 100%, f1-score: 100%. Where RF has accuracy: 98.18%, precision: 98%, recall: 98%, f1-score: 98%. Although LR get the accuracy: 93.46%, precision: 93%, recall: 93%, and f1-score: 93%. Further the KNN obtained an accuracy: 93.92%, precision: 94%, recall: 94%, and f1-score: 94%. Where else the SVM has the accuracy: 94.67%, precision: 95%, recall: 95%, and f1-score: 95%. Finlay the NB get the accuracy: 82.06%, precision: 85%, recall: 85%, and f1-score: 85%. The figure 6 presents the evaluation metrics of various machine learning algorithms on dataset 2. Accuracy measures the proportion of correctly predicted instances out of the total instances. From

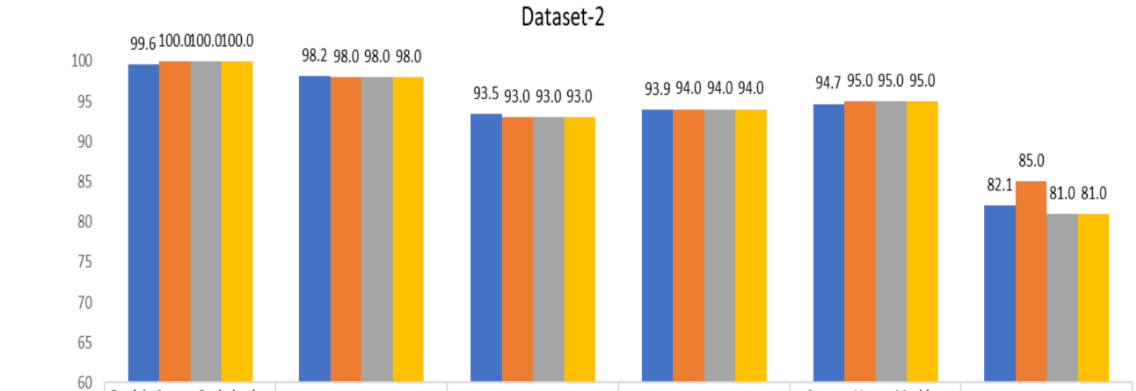
Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

the results, it show that the RF-PSO achieves a high accuracy, precision, recall, and f1-score, indicating strong performance across all metrics. RF also exhibits excellent performance in terms of accuracy, precision, recall, and f1-score. LR, KNN, and SVM demonstrate relatively good performance, although slightly lower than RF-PSO and RF. NB achieves a lower accuracy, precision, recall, and f1-score compared to the other algorithms, indicating comparatively weaker performance. Similarly table 6 and 7 show the confusion matrix detail of the both dataset 1 and 2.



	Particle Swarm Optimization (RF-PSO)	Random forest	Logistic Regression	K-Nearest Neighbors (KNN)	Support Vector Machine (SVM)	Naive Bayes (NBC)
Accurcy Score	98.46	95.70	84.61	83.88	72.04	72.59
Precision	98.00	96.00	85.00	84.00	73.00	78.00
Recall	99.00	96.00	85.00	84.00	71.00	74.00
f1-score	99.00	96.00	85.00	84.00	71.00	72.00

Figure 5: Comparative Analysis between Values of the Classification Report-Dataset 1



	Particle Swarm Optimization (RF-PSO)	Random forest	Logistic Regression	K-Nearest Neighbors (KNN)	Support Vector Machine (SVM)	Naive Bayes (NBC)
Accurcy Score	99.57	98.18	93.46	93.92	94.67	82.06
Precision	100.00	98.00	93.00	94.00	95.00	85.00
Recall	100.00	98.00	93.00	94.00	95.00	81.00
f1-score	100.00	98.00	93.00	94.00	95.00	81.00

Figure 6: Comparative Analysis between Values of the Classification Report-Dataset 2

Xi'an Shiyou Daxue Xuebao (Ziran Kexue Ban)/
Journal of Xi'an Shiyou University, Natural Sciences Edition
ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

Table 6: Comparative Analysis using Confusion Matrix in Dataset 1

ML Techniques		Confusion matrix Predicted P	N
RF-PSO	P	5179	71(FP)
	N	100	6049
Random Forest	P	5103	260
	N	213	5823
Logistic Regression	P	4616	747
	N	757	5279
K-Nearest Neighbors (KNN)	P	4872	491
	N	1260	4776
Support Vector Machine (SVM)	P	2993	2370
	N	880	5156
Naive Bayes (NBC)	P	5088	275
	N	2881	3154

Table 7: Comparative Analysis using Confusion Matrix in Dataset 2

ML Techniques		Confusion matrix Predicted P	N
RF-PSO	P	1027	1
	N	7	830
Random Forest	P	990	14
	N	24	837
Logistic Regression	P	944	60
	N	69	792
K-Nearest Neighbors (KNN)	P	977	27
	N	82	779
Support Vector Machine (SVM)	P	839	165
	N	102	759
Naive Bayes (NBC)	P	960	44
	N	299	562

5. CONCLUSIONS AND FUTURE WORK

Phishers are rapidly updating their technologies to trick internet users and reveal confidential information. Therefore, phishing attack detection and prevention is a serious challenge and considered a dynamic topic that includes many variables and requirements. Machine Learning approach based on various techniques have employed to handle the process of phishing attack detection. In this study, Sequential Forward Feature Selection (SFFS) technique were implemented to find the optimal set of features and Practical Swarm Optimization (PSO) hyper parameter tuning technique were developed to tune the hyper parameters of the Random Forest classifier and compared with other ML techniques using the default values of hyper parameters that specified by the Python scikit-learn library package. Moreover, the experiment was in two URLs based

ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

phishing datasets. As result the proposed RF-PSO technique outperformed other developed ML techniques (RF, LR, KNN, SVM and NBC) in terms of performance measures as it also helped to lower the number of the wrongly classified phishing websites (False Positive FP) and improve the number of correctly classified phishing websites (True negative TN). The computation time that needed to implement the proposed technique were too long (18 hours) as that considered the only disadvantage of this technique. This study will helpful in future studies to continue in this domain. Therefore, in the future work more hyper parameter tuning techniques such as Grid Search, Random Search, Bayesian Optimization and General algorithm would be explored and implemented for different ML techniques such as LR, KNN, SVM and NBC as all have a defined set of parameters that need to be tuned before initiating the process of learning in order to maximize the performance in classifying the phishing website.

Acknowledgment

The authors would like to thank the College of Computing and Informatics, Saudi Electronic University, Riyadh, KSA.

Data Availability

The dataset used in this research is taken from Datasets for phishing websites detection, and Phishing Dataset for Machine Learning: Feature Evaluation.

References

1) A. Shankar, R. Shetty, and B. J. I. J. o. A. E. R. Nath, "A review on phishing attacks," vol. 14, no. 9, pp. 2171-2175, 2019.

2) L. Tang, Q. H. J. M. L. Mahmoud, and K. Extraction, "A survey of machine learning-based solutions for phishing website detection," vol. 3, no. 3, pp. 672-694, 2021.

3) A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J.-P. J. E. Niyigena, "An effective phishing detection model based on character level convolutional neural network from URL," vol. 9, no. 9, p. 1514, 2020.

4) M. H. Alkawaz, S. J. Steven, A. I. Hajamydeen, and R. Ramli, "A comprehensive survey on identification and analysis of phishing website based on machine learning methods," in *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2021, pp. 82-87: IEEE.

5) H. Tupsamudre, S. Jain, and S. J. a. p. a. Lodha, "PhishMatch: A Layered Approach for Effective Detection of Phishing URLs," 2021.

6) M. A. J. M. J. o. E. E. R. El-Rashidy, "A smart model for web phishing detection based on new proposed feature selection technique," vol. 30, no. 1, pp. 97-104, 2021.

7) D. Mehanović and J. J. T. d. S. Kevrić, "Phishing Website Detection Using Machine Learning Classifiers Optimized by Feature Selection," vol. 37, no. 4, 2020.

8) A. Velamati, "Comparative study of machine learning algorithms for phishing website detection."

9) G. Harinahalli Lokesh and G. J. J. o. C. S. T. BoreGowda, "Phishing website detection based on effective machine learning approach," vol. 5, no. 1, pp. 1-14, 2021.

10) P. K. Kandi and P. J. I. J. Agarkar, "Detection of phishing websites using extreme learning machine based on URL," vol. 5, no. 6, 2020.

ISSN: 1673-064X
E-Publication: Online Open Access
Vol: 67 Issue 05 | 2024
DOI: 10.5281/zenodo.11273166

11) A. Ghimire, A. K. Jha, S. Thapa, S. Mishra, and A. M. Jha, "Machine learning approach based on hybrid features for detection of phishing URLs," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2021, pp. 954-959: IEEE.

12) R. Zaimi, M. Hafidi, and M. Lamia, "Survey paper: Taxonomy of website anti-phishing solutions," in *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2020, pp. 1-8: IEEE.

13) A. Bhagwat, K. Lodhi, S. Dalvi, and U. Kulkarni, "An implemention of a mechanism for malicious URLs detection," in *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2019, pp. 1008-1013: IEEE.

14) F. Song, Y. Lei, S. Chen, L. Fan, and Y. J. I. J. o. I. S. Liu, "Advanced evasion attacks and mitigations on practical ML-based phishing website classifiers," vol. 36, no. 9, pp. 5210-5240, 2021.

15) G. Vrbančič, I. Fister Jr, and V. J. D. i. B. Podgorelec, "Datasets for phishing websites detection," vol. 33, p. 106438, 2020.

16) C. L. J. R. O. Tan, "Phishing dataset for machine learning: Feature evaluation, mendeley data, v1," 2019.

17) P. Charoen-Ung and P. Mittrapiyanuruk, "Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning," in *Recent Advances in Information and Communication Technology 2018: Proceedings of the 14th International Conference on Computing and Information Technology (IC2IT 2018)*, 2019, pp. 33-42: Springer.