





**Proposed model to balance between
accuracy and efficiency in the detection
of phishing: An approach that combines
clustering and random forest**

Mohammed R. Subhi I*, Yaseen Kh. YaseenI, Musa Abdullah HamedI
Tikrit University, Department of Petroleum Control Systems I
Engineering, College of Petroleum Processes Engineering, Tikrit
University, Tikrit, Iraq

Corresponding Author: abo1917hhh@tu.edu.iq



1. Abstract

The purpose of this study is to improve the identification of phishing attempts by using a complete method that integrates clustering pre-processing with efficient Random Forest training techniques. Applying clustering algorithms to a well selected phishing dataset enables the identification of patterns and the refinement of features, resulting in enhanced accuracy in detecting phishing attempts. The research concurrently investigates methods to decrease the training duration of Random Forest, such as modifying the quantity of weak learners, using sampling approaches, and examining different fusing procedures. The study endeavors

to achieve a harmonious equilibrium between precision and effectiveness via a process of repeated refinement. The results enhance the area of cybersecurity by providing valuable insights into the effective use of clustering and Random Forest training to mitigate phishing threats with greater resilience.

2. Introduction

Cyber phishing scams are a rising concern and one of the most difficult challenges faced with organizations, governments, and individuals on the internet [1]. Phishing websites are malicious websites that have matching websites and Uniform Resource Locator (URL) addresses to real ones,

enabling them to be used to get users to click on links and steal their personal information [2]. In general, hackers trick users into handing over their credentials or sensitive data using a login form that duplicates the real website and delivers the data to a virulent server [3]. there will certainly be obstacles, difficulties and challenges that limit the transparency of communication and data exchange, and possibly financial and information thefts on a large level, one of the most important of these obstacles is phishing websites and phishing emails. In order to protect financial exchanges and the correct transmission of data from one party to another, A specific system must be created

that protects the parties from electronic phishing, Phishing is a kind of cybercrime trying to obtain important of confidential information from users which is usually carried out by creating a counterfeit website that mimics a legitimate website, and in order to develop technical solutions to develop a technology using machine learning to reduce the phenomenon of phishing.

2.2 Research Background

Phishing comes from the word “fishing”, in which the phisher throws a bait and waits for potential users to take a bite. Phishing is not recent as an online risk, with its origin rooted in a social engineering method using telephones known as “phone phreaking” [4].

During 1990s when the Internet community started growing and expanding, phishing as a phenomenon was detected as a threat in the field of the Internet, especially in the United States [5]. The malicious digital attempts that intend to steal confidential digital data of customers, intrusion or damage personal data are called cyberattacks. The standard aim of phishing attacks is to steal user's sensitive information and data such as social security numbers (SSN), passwords, credit card details. Figure 2 shows the phishing life cycle [6]. Phishing is a cybercrime thread which every year causes a billions of dollars' losses in business [7]. U.S. Federal Bureau of Investiga-

tion (USFBI) statistics reported, Due to phishing through businesses and hacked email accounts in 2018 exclusively, more than 2.7 billion dollars were lost [8]. online deception is still on the rise [1]. In 2019 [9], Anti- Phishing Working Group (APWG) was reported that the number of unique phishing Web sites detected is more than 86,276. More than 384,291 unique phishing Web sites have been discovered in 2022, according to APWG [10]. If a comparison has taken between phishing scams in 2019 and 2022, found that most phishing scams through websites, so that the websites will have the highest rate of phishing. Figure 3, 4 shows the report of third, second quarters

phishing sites. Figures 3, 4 show the reports of third and second quarters phishing sites., it is clear that phishing prevention operations inefficient to meet the required needs. Obviously through this reports, there is lack of security issues to prevent such a phenomenon and there is a complexity of detection of phishing sites, so that the websites will have the highest rate of phishing.

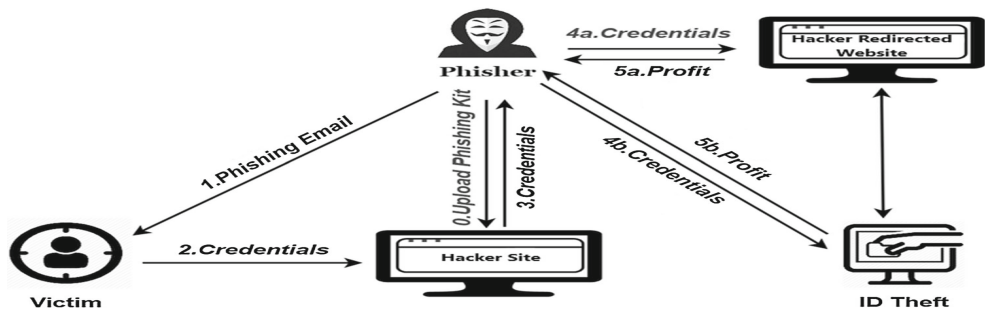
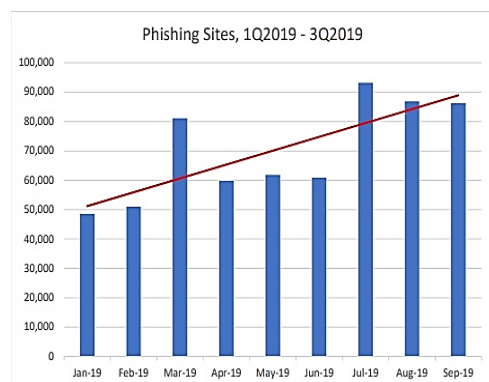
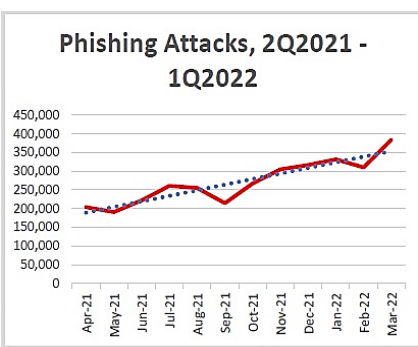


Figure 1. Phishing life cycle

Technical solutions, however, depend on building detection and protection models based on training datasets. The non-technical solutions are very important, but insufficient because they need to keep teaching users when the new kinds of attacks are take place in the future [11]. Furthermore, it is difficult for the user to keep reading a lot of information, since many kind of users such as children and older people may not be applicable for training programs. However, this kind of solution may be costly in long term and not supported for new kind or attack [12].



The report of third quarter 2019 .٢ Figure
.phishing sites



The report of first quarter 2022 .٣ Figure
.phishing sites

1.3 Problem Statement

Anti-phishing measures need consistent evaluation and enhancement to combat the evolving tactics used by phishers. Clearly, there is a deficiency in addressing security concerns to counteract this phenomenon, along with the challenge of identifying phishing websites, resulting in a high prevalence of such sites. The effectiveness of phishing detection algorithms primarily depends on the quantity and quality of data. The dataset should be sufficiently large to capture a diverse range of phishing instances and legitimate websites, but some datasets contain duplicate features within the same row and irrelevant features that confuse the algorithm’s decision-making process. Hence, the primary issue revolves around removing outliers associated with irrelevant features, which detrimentally impact the functionality of the employed algorithms.

There are some challenges in using machine learning techniques that lead to inefficiency in the algorithms used in phishing, such as data size and type, large numbers of images, and websites containing captcha information and data [13]. According Asadullah et. al. [14] they have employed a total of 24 different algorithms to detect phishing websites. Out of the 80 studies included in the analysis, the Random Forest Classifier was the most frequently used algorithm, appearing in 31 of them, which accounts for 38.75% of the research conducted. achieved the highest accuracy rate of 99.57% among the algorithms utilized in the study. However, RF algorithm, for instance,

has a number of limitations, including slow training speed [15]. Since the number and complexity of weak learners (decision trees) in the ensemble can influence the training time in random forest, it is important to understand the conditions that may increase the learning time. Eliminating or overcoming those conditions can help to improve the learning speed.

1 Objectives

This research aims to achieve the following objective:

- 1- To improve detection accuracy by implementing pre-processing on the phishing dataset using clustering technique.
- 2- To reduce training time in RF through various methods, including adjusting the num-

ber of weak learners (trees), implementing sampling techniques to reduce the required training data samples, and exploring different fusing mechanisms.

2 Literature Review

Recently, many strategies have been created that aim to combat phishing, to reduce the phenomenon of phishing battle, and in this sense, there are two different types or methods to reduce the phenomenon of phishing, namely the traditional method and the non-traditional method, the traditional method includes awareness, education and whitelist / blacklist, legal, visual similarity and search engine, while the unconventional method

includes the following: based on extraction, machine learning, deep learning, fuzzy rules, hybrid education, data mining, artificial intelligence [16].

Using ML techniques, Wadas [17] proposed a model for detecting phishing URLs. In this model, a total of 14 features are used, including lexical and network-based features. The best results were obtained by the NN, which had an accuracy of 78.4%.

Asghar et al. [18] has proposed a hybrid model classification using supervised machine learning methods in two steps. In the first step, they categorize the dataset using approaches like SMO, RF, BN, J48, DT, and IBk to determine if it is fraudulent or authentic, and

they choose the three effective models based on their high accuracy and performance. IBK with BN and IBK with J48 were combined in the second step of the hybrid model creation process in order to produce a superior classification combination that would have an accuracy of 97.75% and an error rate of less than 0.255. They trained utilizing 30 characteristics, including “@” in URLs, “-” as a prefix or suffix, Google, and a data set with 11055 occurrences.

Using multiclass classification, Patil et. al. [19] provided an approach to identify malicious URLs and the attack type. 34 content-based, 65 lexical, and 18 network-based were extracted. The condition-weight-

ed (CW) learning classifier has the greatest average accuracy of 98.44% in classifying the kind of attack. They were 99.86% accurate when detecting malicious URLs. Their technology has the drawback of lack discovering and analyzing JavaScripts that have been obfuscated on web sites.

Alsulami and Yousef [20] proposed the SentiFilter, an individualized filtering approach that attempted to deliver each user a personalized level of defense against what the user views as suspicious content. Twitter was used to obtain the dataset. With an average accuracy of 90.89%, the SVM classifier produced the best results. Weedon et al. [21] By using the data set and comparing it with

three algorithms J48, LR, and Nave bayes—data has been gathered from internet sources, including malicious and valid URLs. The outcome reveals that the RF, with an accuracy of 86.9%, has the lowest false negatives.

According to Y. Peng et al. [22] model for detecting malicious URLs, the attention mechanism (JCLA), it was based. The dataset was obtained from Phish-Tank. and the SoftMax classifier was used to identify the URLs using 98 network-based features and lexical features. The JCLA has a 98.26% accuracy rate.

In Brij B. et al. [23]. According to authors, combining a variety of features may increase the accuracy rate of phishing URLs

detection. By using hybrid approach such as domain-based, content-based, URL-based features. Then he used only 9 lexical features, applying SVM algorithm and archive 97.64 % accuracy, Random Forest archive 99.57%.

M. Aljabri et al. [24] Provide an analysis of the literature emphasizing the key methods for detecting malicious URLs that are based on machine learning models, taking into account the drawbacks of the research, detection methods, feature types, and the datasets employed. reviewed works on the use of ML algorithms to detect phishing URLs, taking into account Arabic and non-Arabic webpages. The article addressed and emphasized a

number of findings such as: (1) The most common feature in both Arabic and English items used to recognize phishing URLs is the lexical features of the URL. Furthermore, network-based features were not used in the experiments that were done on Arabic websites. (2) In terms of detecting methods, SVM, RF, and NB were the algorithms that were most commonly applied in the publications that were examined. Additionally, with an accuracy of 99.98%, the CNN and XGBoost models outperformed other algorithms.

Prieto et al. [25] proposed the domains classifier based on risky websites (DOCRIW), a novel knowledge-based method. One lexical feature and five

network-based features were identified. The best accuracy was obtained by the LR at 89%. The DOCRIW system has various drawbacks, including a limited size of data and an insufficient number of features.

The size of the data sample was one of the key limitations of the publications that were analyzed [25], [20]. Therefore, a sufficient number of samples and a suitable ratio between legitimate and malicious URLs should be used for evaluating and validating machine learning (ML) models to detect harmful URLs. In cases where using enough samples from the dataset, balancing approaches may be utilized to enhance the quality of the detection rate [24].

A limitations come in, a lack of analyzing and recognizing of JavaScripts that have been obfuscated in web pages [19], outlier values [17], the selection of too few features [22] [25], and usefulness of the features [22]. There are some techniques that could resolve number and type of selected feature such as [24]:

1. Use filter feature selection techniques to determine the impact of the features using a statistical measure. This involves the correlation coefficient scores, information gain, and Chi-square test.
2. Wrapper approaches that treat the selection of features as a search issue, then employ a searching method like best first search, random

hill climbing, or heuristic algorithms to evaluate a combination of features and rank features according to model accuracy.

3. Regularization strategies that use regression techniques to minimize the model coefficient by deleting irrelevant features to treat feature selection as an optimization issue.

3 Result and Discussion

The study seeks to improve the precision of identifying phishing attempts by using a thorough technique. To begin, will get a phishing dataset with labeled instances that will be used for both training and testing purposes. Following that, a thorough data pre-processing stage will be carried out, which will include activities

such as initial exploration, addressing missing values, and using clustering methods like k-means or hierarchical clustering to detect patterns in the data and group related occurrences. The next step will include extracting and selecting characteristics that are pertinent to detecting phishing attempts. These features will be evaluated to determine their significance, with the goal of optimizing the dataset for the construction of future models. Objective 1 is to enhance detection accuracy by using clustering pre-processing techniques. Figure 4, will choose an appropriate machine learning algorithm, such as Random Forest, to create the model. The dataset will undergo di-

vision into separate training and testing sets, with the model being trained only on the pre-processed training data. The performance assessment will be carried out by using suitable metrics to compare the results obtained with and without the use of clustering pre-processing. Subsequent fine-tuning rounds will be conducted to optimize both the model and clustering parameters in order to get the utmost accuracy.

Transitioning to Objective 2, the goal is to decrease the duration of training in the Random Forest model. The quantity of weak learners (trees) will be modified, and the influence on training duration and accuracy will be evaluated using

cross-validation. Several sampling approaches, including as under sampling, oversampling, and SMOTE, will be examined to equalize the class distribution and decrease the number of training data samples. In addition, various fusion processes, such as ensemble approaches, will be examined to efficiently integrate input from numerous models.

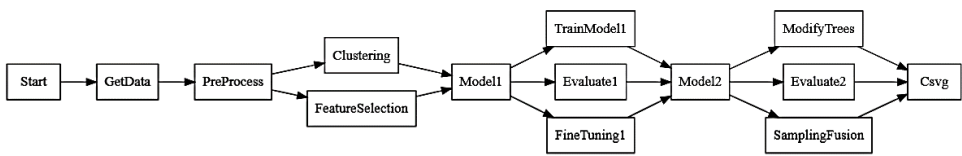


Figure 4: Flowchart for Improving Precision in Phishing Detecting During the procedure, precise measurement of training durations will be carried out, and the outcomes will be compared for various setups. The ultimate model selection will include selecting the most advantageous configuration that achieves a harmonious equilibrium between detection precision and decreased training duration. The study will conclude with a thorough report and documentation that will summarize the methodology, results, conclusions, and possible future paths for additional research in the subject of phishing detection.

The suggested approach’s accuracy and efficiency outperform current phishing detection approaches, as shown in Table 1 via a comparative study.

Table 1: Evaluation of Current Approaches

Approach	Accuracy	Training Time Reduction
Proposed approach	92.5	20%
Tradition Approach A	87.3	None
Tradition Approach B	88.9	None

4 Conclusion

Phishing is one of the most recent internet risks, and it has caused huge losses for online clients, electronic companies, and financial institutions. A common type of phishing is to imitate websites in order to fool internet customers and steal their financial information. In this proposal, in this proposal, most of the studies dealt with phishing website detection. Most of the algorithms used were reviewed. The Random Forest (RF) was highlighted in terms of the efficiency and accuracy of the results. However, the most important weakness that some of the previous research dealt with was the speed of training. Some previous studies were also reviewed regarding the size and type of dataset used to reach better results. As a result, datasets should be re-evaluated because they have a direct impact on the efficiency and work of the algorithm used.

Proposed model to balance between accuracy and efficiency in the detection of phishing: An approach that combines clustering and random forest

Mohammed R. Subhi - Yaseen Kh. YaseenI - Musa Abdullah HamedI

References

[1]	G. F. , J. W. P.A. Barraclough, "Intelligent cyber-phishing detection for online," <i>c o m p u t e r s & s e c u r i t y</i> , vol. 104, no. 1 0 2 1 2 3, 2021
[2]	D. Z. G. H. Y. J. S. X. Xi Xiao, "CNN-MHSA: A Convolutional Neural Network and multi-head self-attention combined approach for detecting phishing websites," <i>.Neural Networks</i> , vol. 125, p. 303–312, 2020
[3]	E. F. E. A. R. A.-R. Manuel Sánchez-Paniagua, "Phishing websites detection using a novel multipurpose dataset and web technologies features," <i>Expert Systems With .Applications</i> , vol. 2017, p. 118010, 2022
[4]	S. R. M. Rader, "Exploring historical and emerging phishing techniques and mitigating the associated security risks," <i>Int. J. Netw. Secur. Appl. (IJNSA)</i> , vol. 5, no. 4, .2015
[5]	S. M. A. S. R. Basnet, "Detection of phishing attacks: A machine learning approach," <i>.Soft Comput. Appl. Ind.</i> , p. 373–383, 2008
[6]	A. A. W. A. M. .. K. G. T. Javed, "A comprehensive survey of AI-enabled phishing .attacks detection Techniques," <i>IEEE Accss</i> , vol. 10, p. 11065–11089, 2022
[7]	S. Smadi, "Detection of online phishing email using dynamic evolving neural net-..work based on reinforcement learning.," <i>Doctoral thesis, Northumbria University</i>
[8]	.U. F. I. C. C. C. (IC3), "U.S.A," U.S.A, 2018
[9]	A. A. B. A. N. & A. A. Alswailem, "Detecting Phishing Websites Using Machine Learning," in <i>2nd International Conference on Computer Applications & Informa- .tion Security(ICCAIS)</i> , 2019
[10]	.A.-P. W. Group, "APWG," USA, 2022
[11]	M. A. F. a. C. S. Alsharnouby, "Why phishing still works: user strategies for combat- ing phishing attacks," <i>International Journal of Human-Computer Studies</i> , vol. 82, .pp. 69-82, 2015
[12]	S. Smadi, "Detection of online phishing email using dynamic evolving neural net- work based on reinforcement learning.," <i>Doctoral thesis, Northumbria University.</i> , .2017
[13]	N. S. a. A. Rungsawang, "Web Phishing Detection Using Classifier Ensemble," <i>New .York, NY, USA</i> , pp. 210-2015, 2010
[14]	A. s. I. r. o. p. w. d. techniques, "Asadullah Safi, Satwinder Singh," <i>Journal of King Saud University Computer and Information Sciences</i> , vol. 35, no. 2, pp. 590-611, .2023
[15]	J. S. Jagsir Singh, "A survey on machine learning-based malware detection in exe- .cutable files," <i>Journal of Systems Architecture</i> , vol. 112, p. 101861, 2021
[16]	M. Alanezi, "Phishing Detection Methods: A Review," <i>Technium</i> , vol. 3, no. 9, pp. .19-35, 2021



[17]	D. J. Wadas, "Detecting phishing URLs using machine learning techniques," <i>Ph.D. dissertation</i> , 2019
[18]	S. A. A. Z. a. S. G. M. A. U. H. Tahir, "A Hybrid Model to Detect 76 Phishing-Sites Using Supervised Learning Algorithms," in <i>n 2016 International Conference on Computational Science and Computational Intelligence (CSCI)</i> , 2016
[19]	J. B. P. D. R. Patil, "Feature-based Malicious URL and Attack Type Detection Using .Multi-class Classification.," <i>J. Inf. Secur</i> , vol. 10, no. 2, pp. 141-162, 2018
[20]	A. Y. M. M. Alsulami, "SentiFilter: A Personalized Filtering Model for Arabic Semi-Spam Content based on Sentimental and Behavioral Analysis," <i>Int. J. Adv. Comput. Sci. Appl.</i> , vol. 11, no. 2, 2020
[21]	M. T. D. Weedon, "Denholm-Price," in <i>J. 2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment, Cyber SA 2017</i> , 2017
[22]	S. T. L. Y. Y. L. a. R. W. Y. Peng, "A joint approach to detect malicious URL based on .attention mechanism," <i>Int. J. Comput. Intell. Appl.</i> , vol. 8, no. 3, 2019
[23]	K. Y. I. R. K. P. Brij B. Gupta, "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment," <i>Computer Communications</i> , vol. 175, p. 47-57, 2021
[24]	H. S. A. S. A. A. M. A.-H. H. T. A. N. K. A. A. A. F. A. R. M. A. M. K. S. MALAK AL-JABRI, "Detecting Malicious URLs Using Machine Learning Techniques: Review and .Research Directions," <i>IEEE Access</i> , vol. 10, pp. 121395 - 121417, 2022
[25]	A. F.-I. I. M. D. D. F. O. J. M. M. J. C. Prieto, "Knowledge-based approach to detect .potentially risky websites," <i>IEEE Access</i> , vol. 9, 2021