An Ensemble Approach for Detecting Phishing Attacks

Himanshi Agrawal^{1*}, Rajni Ranjan Singh²

^{1,2}Department of Computer Science & Engineering, MITS, Gwalior, India

*Corresponding Author: agrahimanshi231@gmail.com

DOI: https://doi.org/10.26438/ijcse/v9i7.5359 | Available online at: www.ijcseonline.org

Received: 16/Jul/2021, Accepted: 20/Jul/2021, Published: 31/Jul/2021

Abstract— In cyberspace, phishing is one of several cybercrimes that often target internet users all over the world. Phishing performs by trying to trick the victim into accessing a web page which looks original, then instructing them to send important data. For prevention, it is essential to build a phishing detection system (PDS). Recent phishing detection system based on data mining and machine learning techniques. Development of an effective detection system while minimizing false positives and negatives is still a challenge. Instead of using single classification approach it would be better to use ensemble approach. In this work an ensemble approach is utilized to build a phishing website classification system. Bagging also known as Bootstrap Aggregating is a meta algorithm established to enhance the machine learning algorithms performance. To detect phishing website various classification models have been developed and implemented. It is observed that combination of Bagging, AdaBoost and j48 gives best results that is 97.2% accuracy.

Keywords— Meta-algorithm, classification, web phishing, website, internet, cyber security.

I. INTRODUCTION

Cybercriminals use fear, anxiety, curiosity, or trust to persuade user to update malware or risk losing personal information through email or phone in even a phishing scam. Cybercriminals frequently impersonate a faithful an official government agency, friend. even a famous business. Hackers appreciate taking advantage of our isolation and confusion yet they can't intimidate everybody abruptly. Hackers are sent out fraud job termination meeting alerts via Zoom across one latest phishing scam. When users obtain a suspicious email or meeting reminder, contact certain trustworthy people, such as co-workers and supervisors, since check contents of the suspicious email. Consequently, IT departments need to ensures that remote employees are protected by same centralized safeguards in an office environment. Multiple levels of monitoring should be included to assist workers in defending toward phishing attacks and several other types of cyber attacks. It's a wickedly brilliant tactic. Hackers are well aware that human error is responsible for over 90% of data breaches. With several persons working at home, cut off from regular contact with IT security, and usually under edge from stress or anxiety now it is the appropriate moment for hackers since put personal awareness towards the tests.

However this period, hackers have impersonated trustworthy tech platforms to respond to the telecommuting and remote work realities. Visitors of Google Meet, Zoom, Skype, are now the victims of misleading cyber crime. The impact of these latest phishing attacks are enormous. In mid-April, Google's Threat Analysis Study revealed blocking 18 million COVID-19-themed malware and phishing emails daily. Even when stay-at-home guidelines

identified by 4% of them becoming suspicious or potentially malicious. Hackers are faking Zoom meeting notifications and sending out manipulated COVID-19 email warming to use such fraud domains. Those who react to such warnings are more likely to download malware have their data security compromised in any way. Although spam blockers help to reduce the effect of scammers, no technology can completely ensure protection from the manipulation used in phishing attacks. It's how hackers use social engineering, a kind of psychological manipulation, to encourage and manipulate individual users. Educating workers on the signals of phishing scams and supporting us in improving their individual cyber security safety is the only cheap and effective strategy to fight them. Individual people at any and all levels of authority must be highly careful when opening emails or warnings whose tend to come from government agencies, health experts or companies in order to protect of there privacy. Customers should also be careful when reacting to videoconferencing meeting invites as well known. Finally, never download suspicious files! While this might

were first implemented, users have shown a 50 percent rise

in the number of ID Experts users reporting becoming goals by scams and phishing attacks. Over the last three

weeks immensely 1,700 Zoom-related domains have been

seem people would be shocked however most of the people unconsciously download malware files simply so reason original email appears to be legitimate on a first glance. Strange download requests should always be tested thoroughly. Whenever a service users have utilized with many years abruptly requests whether they download a new app or upgrade via a special connection it's basically a scam. Phishing is a big threat which can cost both money and peace of mind to individuals and companies. Hackers are constantly modifying their tactics in order to take advantage of our most vulnerable points. We must be vigilant to stay ahead of these criminals, particularly during the pandemic.

The purpose of the contribution statement is the statement should:

- 1. The addition of a meta algorithm performs for detecting website phishing attacks.
- 2. Many combination of various classification models have been tested for find best model.
- 3. Founding higher accuracy results of our model as comparison to base paper.

The Rest of the paper is organized as follows, Section I contains the introduction of phishing attacks came in websites, Section II contain the related work of web phishing detection system (WPDS), section III explain the detection of phishing attacks in methodology with flow chart, Section IV describes results and discussion about how to detect web phishing attacks, Section V concludes research work with future directions.

II. RELATED WORK

Humans' Internet requirements are gradually becoming a fundamental requirement which is having a significant effect on addiction. F. Furedi 2015), the presence of online culture has a visible effect on different operations who need to be taken out [1]. As a result, presently offered different informations must be uploaded ahead and bought available to everyone at any time at any location.

H. Hasan, S. Hayikader, M. Chewae, and J. Ibrahim 2015), the website is among the technology which supports and facilitates internet activities [2]. Today's website use has undergone a number of developments. P. Patil, R. Rane, and M. Bhalekar and M. Ganesan and P. Mayilvahanan 2017), websites are no longer only used to relay information; they are often use as the mechanism for communication and social networking i.e. social media as well as transaction media such as I-banking and ecommerce those are transactions of banking [3], [4]. R. Pompon, M. Levin, S. Boddy, and D. Walkowski 2019) who use internet world and facilities will benefit greatly from the development of website technology which will provide considerable safety and security. There are a slew of threats hidden under the ease of access of using the website as a contact and transaction medium if the user is not prudent. Web phishing (WP) is among the many harmful effects that users face while visiting a website [5].

Pritesh Saklecha, Jagdish, Raikwar 2018) To Identifying some acceptable and/or adoptable detection and prevention approaches whereby a system automatically detects a phishing web URL using data mining techniques, as well as several phishing techniques and their effects on our life today [6]. P.Priyadevi, V.Lalithadevi, M.sughashini,2018) A new approach for detecting phishing WebPages in genuine as they are accessed by a browser to solve those limitations. It is based on the assumption that phishers have intrinsic boundaries as a result of the restrictions cybercriminals meet whenever creating a webpage [7].

WP is a threat used by social engineering organisations. M. Karabatak and T. Mustafa 2018) in general web phishing works by directing victims to a web page which looks identical to real web page [8]. A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery 2017, 2018), web phishing may also be through of e-mail and pop-up messages with the objective of gaining personal details such as usernames and passwords, credit card numbers, and other valuable data [9]. In 2018, Indonesia ninth ranked in the world for WP attacks with a rate of about 71%. R. Pompon, D. Walkowski, S. Boddy, and M. Levin 2019) this rate is expected to rise by around 5% during the holiday season. Phishing attacks attacking e-commerce, financial institutions, and shipping websites (shipping) are most well-known [5]. To reduce the incidence of WP attacks and to prevent security breaches as a consequence of web phishing is important to build a framework which can detect them.

Researchers used various data mining techniques in their research to create a phishing site detection method. A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery 2017, 2018) as a result the C4.5 algorithm's accuracy in classifying web phishing is 94.1% with SSL and HTTPS on the websites gained to being the most influential attribute [9]. R. M. Mohammad, F. Thabtah, and L. McCluskey 2014) correlation-based attribute selection is used in another study which also utilizes data mining techniques [10]. The result was 94.31% classification performance as the C4.5 algorithm was used, and the 94.50% as the K-Nearest Neighbour (KNN) algorithm was used. The being of ensemble-algorithm germane to the data class on the other hand is relevant. L. Rahman, N. A. Setiawan, and A. E. Permanasari 2017) the use of a ensemble-algorithm in education of the data classification process, on the other hand results in a crucial improvement in the classification performance [11]. A. F. Nugraha, L. Rahman 2019) the result was 95.5% classification performance as the decision tree was used, 97.1% classification performance as bagging algorithm used in WEKA [12].

As a consequence, ensemble-algorithms will be used in this research to integrate the web phishing detection system (WPDS) classification performance. The use of bagging methods are among the meta-algorithms that will be evaluated in this study. The accuracy of the results will be used to test the study's classification performance. The accuracy value will also be contrasted across processing which mere use of classification techniques and the research methods that the integration of a ensemblealgorithm to determine which method has the best output and can be used as a benchmark to build phishing web detection system. The Phishing Websites Data Set collects

from UCI Machine Learning Repository [13] will be used in order to testing in this report, and it is free to download.

III. METHODOLOGY

As per Figure 1, the researchers tested an output which began with the 10 cross fold validation, performed through the modelling and analysis process, and ended with evaluation.



Figure 1. Research and methodology

The stage of preparation is associated with preparation of data for research purposes. Web Phishing Datasets are used in this study they can be downloaded by the UCI Machine Learning website [13]. This is a benchmark dataset which categorises website pages into 2 categories: legitimate websites and phishing websites [14]. As exhibited in Table 1, the phishing websites detection system contains 11055 data split into four groups, each group having 30 features.

Table 1. Phishing Web Features

Section	Feature Name	Description
Address Bar Based Features	Using Internet Protocol Address	If website address part uses a hexadecimal code or an IP address is called phishing.
	Use a long URL for mask the URL misleading form.	If the URL address is longer than 54 characters, the website will be included in phishing note.
	Using URL Shortening	The use of redirected addresses is based on the concept of the "Tiny URL."
	The presence of the "@" symbol in a URL	If there is a "@" symbol in the URL use for detection of phishing.
	To redirect a URL, use the "//" symbol.	If the use HTTP on the 6th position or the "//" sign is on the 7th position in the URL which is listed as phishing website.
	Add the dash symbol (-) as a prefix or suffix in that domain.	If the domain adding a dash (-) symbol for the phishing website categorization.
	Domain and sub domain numbers	When in the domain and sub domain the number of dots (.) is excessive, it's a warning that the site becomes phishing.
	Using HTTPS	If a website uses https and has not expired certificates so web is considered legitimate; otherwise considered as a phishing website.
	Domain registration	When a website is known to have an expired domain with a limited period of time is remaining time will be known as phishing. It will be considered a legitimate website if the timer has also not elapsed yet.
	Favicon usage	When the website's favicon is collected from external domain, website would identified become

VOI.9(7), JUI 2021, L-15514. 2547-2095					
	a phishing website.				
ing nonstandard rt	Whether there's any open ports besides the HTTP (80) port, those will also be added to the phishing group.				
e being of an	HTTPS as domain name not as a				

Vol 0(7) Jul 2021 E ISSN: 2347 2603

	Using nonstandard	Whether there's any open ports
	port	besides the HTTP (80) port,
		those will also be added to the
	The heine of on	UTTPS as domain nome not as a
	HTTPS token into	protocol
	the URL's domain	protocoli
	part.	
Abnormal	Percentage of	Both media and objects on a
Based	Requesting URL	legitimate website are encoded in
Features		similar domain name address.
	Number URL of	The anchor is determined by the
	Anchor	number of <a> tags which reflect
		the number of links for other
		$\langle a \rangle$ tags the more destination of
		phishing URLs are identified.
	The <meta/> ,	The meta> tag is used for
	<link/> , and	provide website metadata
	<script></script>	

	the rest have a page rank of over than 0.2.
Google Index	Google pages indexes all legitimate websites.
Amount of the Links pointing to the page	The amount of the links leading to a web page means that it is trustworthy.
Statistical reports- based feature	Phishing websites will also be identified on the websites like Phisthtank and StopBadware, which formulate statistical reports on phishing
	attacks on a periodic basis.

All of the features at table 1 have a 1 value instead of legitimate websites, 0 value instead of suspicious phishing websites, and -1 value instead of phishing websites. Data labels have the same value, with 1 value instead of legitimate websites and -1 value instead of phishing websites. The training data will be use of 10-fold cross-validation among the next stage of modelling and analysis. Modelling and analysis phase is a next stage which entails using a begging ensemble-algorithm until performing the classification process with the decision tree algorithm as the classifier. This study's begging meta-algorithm is discussed below.

A. BAGGING

Bagging also known as Bootstrap Aggregating is a meta algorithm established to enhance the machine learning algorithms performance in particular classification [15]. To minimise the rate of error in evaluating predicted results, bagging utilizes two stages of processing. The Bootstrap is first stage that operates on the sampling principle as many times as the return in the entire population of information to generate a training data set. The training data system results are certainly analyzed into the second stage which involves aggregating or determining the most votes by calculate the predicted value produced.

Produce the following equation for each n=1 to N bootstrap sample:

$$\{(z_1^1, z_2^1, \dots, z_s^1)\}, \{(z_{1,}^2, z_2^2, \dots, z_s^2)\}, \dots, \{(z_1^n, z_2^n, \dots, z_s^n)\}_{[1]}$$

 $z_1^1, ..., z_s^1$ denotes that the sample in order to each bootstrap is N. Each bootstrap sample size is the same as training data use prior to the bootstrap process is processed. After that, every bootstrap will be trained use of the learning model to determine importance of weak-learners (w).

$$w_1, w_2, \dots, w_n \tag{2}$$

The established weak-learner values are then analyzed use of Aggregating model by produce the strong model (s), which employs statistical value equation below.

$$s_n = \arg\max\{n\binom{n}{w_n}\}\tag{3}$$

B. BOOSTING

Boosting training dataset instances is evaluated the overall model accuracy update value or not for correctly classified instances. Successive models are trained and added until a minimal level of accuracy is reached or no more performance can be improved. Each model's skill is evaluated and these values are utilized when aggregating all of the models' predictions on new data [16].

AdaBoost ensemble model works with the decision tree. AdaBoost use an iterative approach on the collection of algorithms for converting weak to strong learners. Developing of the boosting ensemble algorithms lies between sequential weak learner training process to get small to small classification error [17]. For binary class classification [18]:

$$F(x) = sign\left(\sum_{m=1}^{M} \alpha_m f_m(x)\right)$$
(4)

The value of m-weak learner is $F_M(X)$, while the appropriate weight for M weak learners is α_m . The formulation of AdaBoost model for each m=1 to M :

1. The error function minimize based on equation (5)

$$W_e = \sum_{y_i \neq k_m(x_i)} w_i^{(m)} exp(a_m)$$
⁽⁵⁾

2. To update the weight in equation (6)

$$\alpha_m = \frac{1}{2} ln \left(\frac{1 - e_m}{e_m} \right) \tag{6}$$

With equation (7)

$$(e_m) = \frac{W_e}{w} \tag{7}$$

3. If the result is a misclassification for updating use the equation (8)

$$w_{i}^{(m+1)} = w_{i}^{(m)} exp(\alpha_{m}) = \sqrt[w_{i}^{(m)}]{\frac{1 - e_{m}}{e_{m}}}$$
(8)

If there is no misclassification update use the equation (9)

$$w_i^{(m+1)} = w_i^{(m)} exp(-\alpha_m) = \sqrt[w_i^{(m)}]{\frac{e_m}{1 - e_m}}$$
(9)

C. J48 CLASSIFICATION APPROACH

Decision tree j48 is the carried out from ID3 algorithm (Iterative Dichotomiser 3) processed by WEKA project team. J48 operates a decision node using required estimation of the class [19]. J48 containing several nodes such as root node, intermediate node and leaf node. J48 each node having a decision usually leads with our result. Decision tree split the input place of a detection system into manually unique domains having level, value to depict

or elaborate its data points. Separating criterion is utilized to calculate those features is the best to separated training dataset tree portion for reach to a particular node [20].

Table 2. Confusion Matrix					
Actual	Prediction Classification				
Classification	Positive	Negative			
Positive	True Positive (TP)	False	Negative		
		(FN)	_		
Negative	False Positive (FP)	True	Negative		
-		(TN)	_		

For binary class classification confusion matrix Table 2 includes data which contrasts the system prediction results founded onto model with the actual results, which operate effectively as the phishing web detection system ground truth. Confusion matrices are used in four different ways to determine classification performance:

- True Positive (TP) is a metric that calculates the sum of data classified positively through system to true positive data.
- The number of data detected negatively by system compared to real positive data is called False Negative (FN).
- Number of data detected negatively from machine compared to real data that is positive is called False Negative (FN).

• True Negative (TN) calculates sum of data observed negatively to real data that is negative value.

The high percentage produced in 4 categories of the confusion matrix, whereas TP, FP, FN and TN can be used to calculate classification performance measures founded on accuracy, precision, recall, and F1-Measure metrics. Whenever, in that analysis the metric was only used by determine accuracy value which described where well the resulting model classified phishing web features.

Use equation (10) by measure the accuracy value focused onto 4 classification results in this confusion matrix, that represents the percentage of the total which is correctly classified relative from each classifications generated through the system.

$$Akorasi = \frac{TP}{TP + TN + FP + FN}$$
(10)

IV. RESULTS AND DISCUSSION

The UCI Machine Learning Web Phishing Detections system is converted into a detection system which can be analyzed with using a bagging ensemble algorithms in the classification phase.

Algorithm	Accuracy%	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Bagging + AdaBoost + Decision Stump	92.4	0.925	0.075	0.925	0.925	0.925	0.950	0.990	0.981
Bagging + AdaBoost + Hoeffeding Tree	94.9	0.949	0.051	0.949	0.949	0.949	0.899	0.992	0.992
Bagging + AdaBoost + REP Tree	96.9	0.970	0.030	0.970	0.970	0.970	0.940	0.996	0.996
Bagging + AdaBoost + Random Tree	97.1	0.972	0.028	0.972	0.972	0.972	0.943	0.996	0.996
Bagging + AdaBoost + j48	97.2	0.973	0.027	0.973	0.973	0.973	0.945	0.997	0.997

All of the results mentioned in Table 3 this analysis works on training dataset of web phishing utilizing 10-fold crossvalidation is described inside the methodology section. Table 3 shows accuracy, TP Rate, FP Rate, Precision, Recall, F-measure, MCC, ROC area, PRC area of all algorithms. Decision Stump, Hoeffeding tree, REP tree, Random tree, j48 are applied with the combination of AdaBoost in the Bagging classifier on UCI Machine Learning Repository phishing web detections.



Figure 2. Accuracy %



Figure 3. Precision



Figure 4. Recall

The classification performance using machine learning algorithm in order to the meta-algorithm implementation classification process performed in WEKA Figure 2 show the accuracy value, Figure 3 shows Precision, Figure 4 shows Recall, value of phishing detection system are mentioned in Table 3.. The accuracy value provided from every model at this study is comparative analysis for determine the best model which can be utilized in creation of different WPDS [5]. That is illustrates a performance assessment graph based on the study's proposed scenario. The accuracy value of experimental results to get 97.2% classification performance is the best performance obtained by using bagging ensemble algorithm for phishing web detection system.

V. CONCLUSION AND FUTURE SCOPE

The objective of this work is to see how well a classification based on the addition of a meta algorithm performs in detecting the presence of website phishing. Many combination of various classification models have been tested and it is observed that ensemble based approach gives better result in comparison with single classifier. In future, work may be extended by adding suitable pre-processing approaches to improve the datasets as well as features selection approach to improve the classification accuracy.

© 2021, IJCSE All Rights Reserved

Vol.9(7), Jul 2021, E-ISSN: 2347-2693

REFERENCES

- [1] F. Furedi, "How the Internet and social media are changing culture," **2015.** [Accessed: 22-Apr-2019].
- [2] M. Chewae, S. Hayikader, H. Hasan, and J. Ibrahim, "How Much Privacy We Still Have on Social Network?," Int. J. Sci. Res. Publ., vol. 5, no. 1, pp. 1– 5, 2015.
- [3] P. Patil, R. Rane, and M. Bhalekar, "Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm," Proc. Int. Conf. Inven. Syst. Control. ICISC 2017, pp. 1–4, 2017.
- [4] M. Ganesan and P. Mayilvahanan, "Cyber Crime Analysis in Social Media Using Data Mining Technique," Int. J. Pure Appl. Math., vol. 116, no. 22, pp. 413–424, 2017.
- [5] R. Pompon, D. Walkowski, S. Boddy, and M. Levin, "2018 Phishing and Fraud Report: Attacks Peak During the Holidays" 2018. [Accessed: 20- Apr-2019].
- [6] Pritesh Saklecha, Jagdish Raikwar, "Prevention of Phishing Attack using Hybrid Blacklist Recommendation Algorithm", International Journal of Computer Sciences and Engineering, Vol.6, Issue.6, pp.188-191, 2018.
- [7] P.Priyadevi, V.Lalithadevi, M.Sughashini, "An Efficient and Usable Client-Side Phishing Detection Application", International Journal of Computer Sciences and Engineering, Vol.06, Special Issue.02, pp.398-401, 2018.
- [8] M. Karabatak and T. Mustafa, "Performance comparison of classifiers on reduced phishing website dataset," in International Symposium on Digital Forensic and Security (ISDFS), 2018, pp. 1–5.
- [9] A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery, "Intelligent phishing website detection using random forest classifier," 2017 Int. Conf. Electr. Comput. Technol. Appl. ICECTA 2017, vol. 2018-January, pp. 1–5, 2018.
- [10] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Intelligent rule-based phishing websites classification," IET Inf. Secur., vol. 8, no. 3, pp. 153–160, 2014.
- [11] L. Rahman, N. A. Setiawan, and A. E. Permanasari, "Feature Selection Methods in Improving Accuracyof Classifying Students' Academic Performance," 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. (ICITISEE)., no. 1, pp. 267– 271, 2017.
- [12] A. F. Nugraha, & L. Rahman, "Meta-Algorithms for Improving Classification Performance in the Web-phishing Detection Process". In 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 271-275). IEEE.
- [13] UCI Machine Learning, "UCI Machine Learning Repository: Phising Websites Data Set," 2019. [Accessed: 20-Apr-2019].
- [14] R. M. Mohammad, F. Thabtah, and L. Mccluskey, "Phishing Websites Features," Ieee. pp. 1–7, 2013.
- [15] L. Breiman, "Bagging predictors," Dep. Stat. Univ. Calif., no. 2, p. 19, 1994.
- [16] L. Chen, "Basic Ensemble Learning (Random Forest, AdaBoost, Gradient Boosting)- Step by Step Explained," 2016. [Accessed: 20-Apr-2019].
- [17] Y. Freund and R. E. Schapire, "A Short Introduction to Boosting," J. Japanese Soc. Artif. Intell., vol. 14, no. 5, pp. 771–780, 1999.
- [18] V. Estivill-Castro, M. Lombardi, and A. Marani, "Improving binary classification of web pages using an ensemble of feature selection algorithms," ACM Int. Conf. Proceeding Ser., 2018
- [19] Md. Nurul Amin, Md. Ahsan Habib "Comparison of Different Classification Techniques Using WEKA for Hematological Data" American Journal of Engineering Research (AJER).
- [20] N. Saravanan, V. Gayathari "Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)" In 2018 International Journal of Computational Intelligence and Informatics.

Vol.9(7), Jul 2021, E-ISSN: 2347-2693

AUTHORS PROFILE

Ms. Himanshi Agrawal Research scholar, Department of computer science & Engineering, Madhav Institute of Technology & Science, Gwalior.



Mr. Rajni Ranjan Singh, Assistant Professor, Department of Computer Science & Engineering, Madhav Institute of Technology & Science, Gwalior, M.P., India. His research interest includes AlgorithDesign, Network Security, Network Forensics and Computer Networks.

