International journal of application or innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/924-930

ISSN: 2319-4847

A Machine Learning Approach to Detecting Phishing on Websites

[1] Sahiba Sulthana, [2] SK. Sajitha

[1]Assistant Professor, Megha Institute of Engineering & Technology for Women, Ghatkesar.
[2] MCA Student, Megha Institute of Engineering & Technology for Women, Ghatkesar.

Article Info

Received: 30-04-2025 Revised: 06 -06-2025 Accepted: 17-06-2025 Published:28/06/2025

Abstract—

The Internet has become a vital tool in our contemporary lives, impacting both our personal and professional lives. A growing number of customers are opting to make their purchases online due to this. Users of the internet may be susceptible to several online dangers due to this reality. Financial loss, credit card theft, data breaches, brand harm, and consumer distrust in e-commerce and online banking are all possible outcomes of these dangers. One kind of cyber hazard is phishing, which is when a criminal creates a fake website in order to trick users into giving up important information (such as login credentials, passwords, and credit card numbers). Methods for identifying phishing attempts are the subject of this study. To identify a phishing attempt, this research used a machine learning strategy. Consequently, this study achieved a 94% success rate in detecting phishing attempts.

Keywords—phishing attack; phishing; website detection; malware; machine learning

Introduction

One of the most common forms of cybercrime, phishing attempts use every available communication channel to mislead victims into giving up personal information. In order to steal information that might do them or their companies harm, attackers use deception and make victims fall into their traps. The choice of communication channel is determined by the attacker's purpose and the data type. [1] Ransom demands and account termination threats are also part of it. Email spoofing is another misleading tactic that scammers use to trick victims into giving over personal information like passwords and credit card details. Critical information, like login passwords to online banking and credit card numbers, are the primary targets of phishing attacks. Online firms' reputations take a hit as a result of these fraudulent operations, which weaken confidence in online transactions. Attacks on computer systems may still occur even with data encryption. [2] in Preventing phishing attacks requires awareness and constant observation. It is possible to avoid injury by making it a habit to carefully browse the web and check the reliability of links. Users can be warned about malicious websites that try to steal their credentials through browser extensions and other technologies. Attacks including phishing schemes targeting cryptocurrency entities, including bitcoin exchanges and wallet providers, increased to 6.5%, necessitating the implementation of network technologies that restrict access to only specified websites. From the third to the fourth quarter, there was a 36% increase in the number of firms found to have been victims of ransomware. For corporate users, the most common types of phishing emails were those attempting to steal credentials (51.8%), response-based attacks (38.1%), and attempts to deliver malware (9.6%). [5] The APWG also recorded 316,747 assaults in December 2021. This is the most comprehensive monthly report in the APWG's history. Beginning in the year 2020, phishing scams have become more commonplace. Phishing attempts in the fourth quarter of 2018 were 23.2% more likely to target the financial sector than any other industry. There has been a persistently high volume of cyberattacks

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/924-930

ISSN: 2319-4847

targeting webmail and software as a service providers. assaults targeting bitcoin exchanges and wallet providers, among others, increased to 6.5% of all assaults. From Q3 to Q4, there was a 36% increase in the number of businesses found to have fallen victim to ransomware. 51.8 percent of the emails reported by business users were credential theft phishing attempts, 38.5 percent were response- based assaults (including BEC, 419, and gift card scams), and 9.6 percent were something else entirely. safety, even if it means sacrificing user ease. [1]To detect phishing attempts, this study used machine learning.

The methods used for detecting phishing websites are based on heuristics and gather information from websites in order to determine their legitimacy. Heuristics, in contrast to blocklists, can identify phishing sites while they are being built in real time. Heuristics that work for classifying websites depend on discriminating criteria. The heuristic method detects phishing websites by analyzing their HTML or URL signatures. Research on this method's efficacy is ongoing. the third Logistic Regression (LR), Bayesian Additive Regression Trees (BART), Classification and Regression Trees (CART), Random Forests (RF), and Neural Networks (NN) are some of the machine learning and data mining methods that are evaluated for phishing site prediction. To train and test classifiers, experiments were conducted using a dataset consisting of 1,172 phishing emails and 1,718 legitimate emails, using 43 different functions. Findings reveal that RF had the best accuracy rate at 7.72%, followed by CART at 8.13%, LR at 8.58%, BART at 9.69%, Support Vector Machines (SVM) at 9.90%, and NN at 10.33%. Nevertheless, the results show that no particular classifier stands out as being the most effective at identifying phishing websites. Bagging, AdaBoost, SVM, CART, NN, RF, LR, NB, and BART are some of the machine learning-based detection methods (MLBDMs) that are evaluated and compared in this review [4]. One thousand five hundred phishing and one thousand five hundred legitimate websites make up the dataset. The eight criteria that make up CANTINA's evaluation factors are as follows. [4]

Phishing Website attack trends

In December 2021, the Anti-Phishing Working Group (APWG) recorded 316,747 attacks, which is the highest monthly number in the group's reporting history. Since 2020 began, phishing schemes have grown in frequency. At 23.2% of all assaults in the fourth quarter of 2018, the financial sector was the most often targeted by phishers. The number of cyberattacks against SaaS and webmail providers has been quite high. The resource's percentage [19]. To facilitate benchmarking and model building, platforms such as Kaggle house community-contributed datasets of phishing websites[20]. Researchers may use GitHub's publicly accessible phishing datasets to study phishing patterns and improve detection methods [21]. Researchers may investigate phishing trends, create efficient detection algorithms, and test the efficacy of their approaches with the use of these databases. There are four stages to the empirical evaluation process in the dynamic analysis method: The dynamic analytic technique for identifying phishing websites is presented in Fig. 2. Gathering datasets of phishing websites is the first stage in identifying them. A larger number of datasets utilized in the experiment yields more accurate results, according to recent research. References [22], [23] Phishing datasets like Kaggle[24] and Git[25] are often used by researchers.

Website Vulnerabilities

Here we'll go over the most common security holes in websites that allow phishing attacks to happen: Exploitation of Insecure Sites (XSS) Attackers exploit cross-site scripting (XSS) vulnerabilities when they insert scripts that might execute unauthorized code in the victim's browser onto web pages that people see. Because of this, malicious actors may be able to steal login credentials or lead users to malicious phishing websites. "The Web Application Hacker's Handbook"[6] by D. Stuttard and M. Pinto is an all-inclusive resource for learning about and protecting against cross-site scripting vulnerabilities.

Next, attackers may take advantage of Cross-Site Request Forgery (CSRF) vulnerabilities to sneakily execute operations on a targeted website without the user's knowledge or permission. Things like completing purchases or filling out forms on phishing websites fall under this category. In order to protect oneself against CSRF attacks, one should read the article"RobustDefensesforCross-SiteRequestForgery"

International Journal of application or innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/924-930

ISSN: 2319-4847

written by A. Barth et al. [7] In addition, SQL injection vulnerabilities occur when an attacker is able to enter malicious SQL queries into a website's database by manipulating user-supplied input. By taking advantage of these security holes, malicious actors may get access to sensitive user data and utilize it in phishing scams. [8] Then, session hijacking vulnerabilities happen when malicious actors get the session ID of a user without their knowledge or consent. This gives them the ability to impersonate the user and carry out harmful activities, such as diverting them to phishing websites. The methods of session hijacking and successful countermeasures are covered in the article "Session Hijacking and Its Countermeasures" written by M. Naveed et al. [9].

.

Methodology

This study is using this research strategy because it can be easily utilized to incorporate new research breakthroughs by reverting to earlier stages with little loss. And if issues crop up at this level, the procedure allows for tweaks to any phase to address them. Last but not least, researchers may easily adapt this research technique to meet the needs of the study subject. [10]

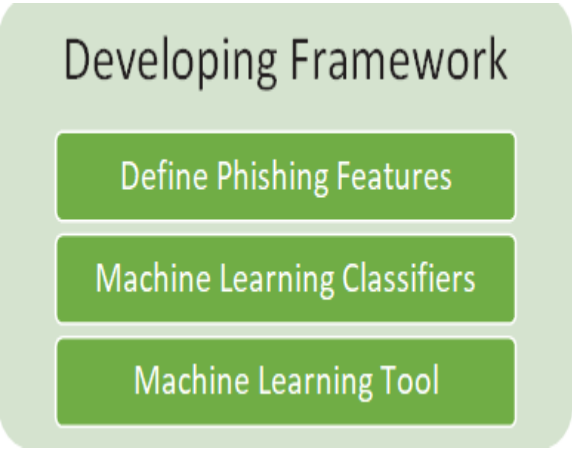

Fig.1.DevelopmentofPhishingWebsiteDetectionFramework

Define Phishing Features

Characteristics determined by URLs will be the main focus. The URL is the first thing to check when trying to figure out whether a website is phishable. Phishing domain URLs may display several distinctive qualities. We can learn more about the points' related traits by analyzing the URL. This project will investigate the following URL-based features:

i. Address Bar: How It Works
ii.Base Characteristics That Are Not Normal
iii. Web Development using JavaScript and HTML
iv. section four: functions that are exclusive to a certain domain

Description of the dataset

One of the most important things to do when trying to create reliable phishing detection is to collect relevant datasets. Both phishing and legal operations may be better understood with its help. A total of fifty-eight attributes culled from five thousand legitimate and five thousand fraudulent websites make up the dataset. Feature extraction using a browser automation framework is more reliable than regular expression-based parsing. "Legitimate" is now represented as"1" and “Suspicious" as"0." These numerical values indicate the transformation of the categorical values "Legitimate" and "Suspicious." The significance of each attribute is determined by analyzing the dataset using the Correlation Attribute Evaluation approach. This method assigns a numerical value to each attribute and ranks them accordingly. Several features have risen to the top spot because they are used so often in detection. the eleventh [12]

Classifiers and Tools for Machine Learning

The term "machine learning" refers to a subfield of artificial intelligence that can learn new things and improve existing ones without any human intervention. [31] The number 32. In a process called learning, it looks for patterns in datasets in order to generate predictions. Intrusion detection systems

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/924-930

ISSN: 2319-4847

often use several classifier types, which influence the learning process and the outcomes of predictions. There are two main methods for machine learning: supervised and unsupervised. Because there is labeled data (both phishing and regular), supervised machine learning is used to minimize mistakes in this study. Random Forest (RF), J48, Naïve Bayes, Logistics, and K-Nearest Neighbors (KNN) are the five classifiers that are used for comparison. An efficient collective learning technique, Random Forest trains several decision trees for classification or regression. in [13]

Due of its flexibility and cloud capabilities, Google Colab is used for training data sets. It works well with Python-based machine learning. In order to optimize speed, the memory-hogging machine learning technique relies on spreading GPU assets from Google servers to otherwise limited hardware on the programmer's end. The data set is stored on Google Storage's cloud drive architecture. The Colab online notebook is used for loading and training purposes. The trained model was then loaded into the Pi and verified using the data that was acquired. [14] [15] [16]

DYNAMIC ANALYSIS OF PHISHING WEBSITE AND DETECTION TECHNIQUE

The design model consists of four parts: data collection, factor identification, model testing, and result comparison. In the following section, we will take a quick look at each part.

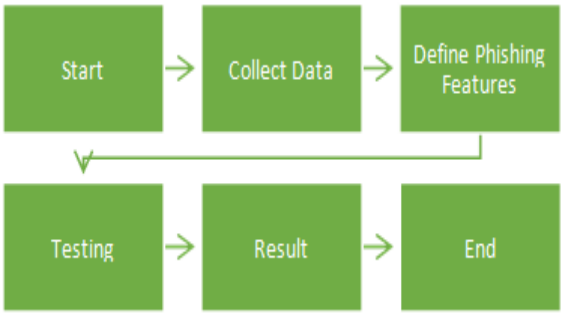

Fig.2.DynamicAnalysisProcessforDetectingPhishing Websites

At this stage, the proposed solution will be put into action with the help of the design model. Installing the study's software into a desktop, laptop, or mobile device—like Google Colab—is the first order of business at this stage.

There are four stages to the empirical evaluation process in the dynamic analysis method: The dynamic analytic technique for identifying phishing websites is presented in Fig. 2. Gathering datasets of phishing websites is the first stage in identifying them. In their studies and trials, researchers often use datasets from numerous phishing websites. The PhishTank dataset is one example of a popular tool for reporting and validating phishing URLs[17]. An extensive database of phishing URLs is made accessible to the public by OpenPhish[18]. Researchers have access to a significant resource thanks to the Anti-Phishing Working Group's (APWG) library of phishing URLs reported by people and organizations.

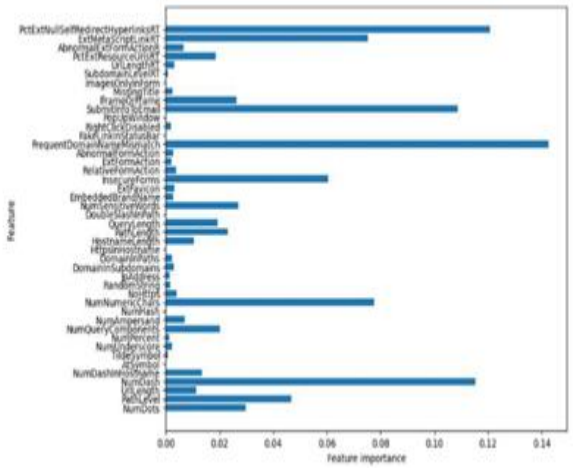

Fig.3.FeaturesRanking

After that, we classified the phishing website's components according to their salient features. This study employed the feature selection method to identify the most important attributes for accurate phishing website detection. Websites that are legitimate and those that are phishing use different techniques to distinguish between the two. Figure 3 displays a rundown of the characteristics of the phishing websites that were examined in the study.

International Journal of application or Innovation in engineering and management

Ijaiem.com/June 2025/ Volume 14/Issue 1/Article No-1/924-930

ISSN: 2319-4847

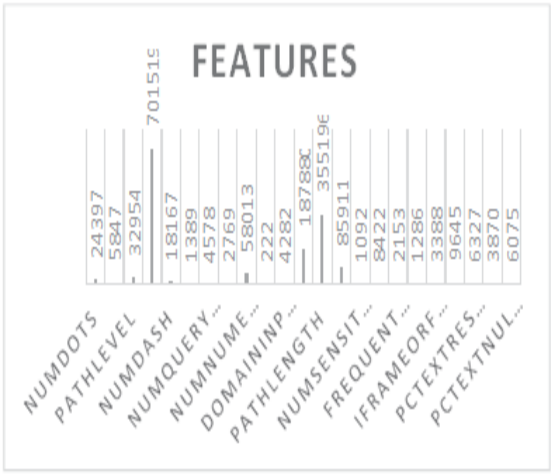

Fig.4.ChosenFeatures

The next stage is to define phishing characteristics by getting the most utilized data from the dataset in Figure 3 using correlation attribute assessment, which takes appraising a trait's worth by looking at how closely it ties to the category. Additionally, it displays the rank number for each attribute and provides a rating of the qualities from best to worst [12]. Since they are used often throughout the detection process, some features have the highest ranking, as shown in Figure 4.

The third stage is to test the dataset that will be used to evaluate the experiment after all the components have been included. We do testing and assessment to solve the issue statement and see whether we avoid the limitations of existing publications. The main goal of this inquiry is to prove the legitimacy of the results and claims by showing the most effective detection model that has been proposed. Also, by evaluating and testing, the research experiment might find limitations and weaknesses, which allows for more tweaks to get the right result. Using machine learning methods like random forest, J48, Naïve Bayes, KNN, and logistic regression, the last step is to examine the outcomes. From the data, these approaches derive insights and create predictions. For more precise findings, random forest builds a network of decision trees. J48 is a decision tree classifier that finds important traits and gives rules that anybody can understand. For big datasets, the probabilistic approach known as Naïve Bayes works well. KNN uses the closeness to existing examples to classify new occurrences. Models of logistic regression that account for inter-variable correlations in order to facilitate binary or multi-class categorization. In order to make well-informed decisions, these methods unearth patterns and give useful insights.

DISCUSSION

The findings are shown using three distinct machine learning classifiers: logistic, random forest, J48, Naïve Bayes. The accuracy, precision, and recall metrics were also used in this Python-based examination of the different measurements. Table I displays the results derived from the testing set's 25 phishing website characteristics using five selected classifiers.

TableI. The Related Study Classifiers Analysis

Classifiers	Accuracy	Precision	Recall	FPR	TPR
Random Forest	94.10%	0.978	0.904	0.021	0.904
J48	92.10%	0.917	0.926	0.084	0.926
Naive bayes	83.00%	0.921	0.771	0.071	0.771
KNN	92.21%	0.923	0.921	0.079	0.921
Logistic	89.50%	0.895	0.895	0.105	0.895

CONCLUSION

Finally, by comparing and contrasting previous methods with the present Machine Learning method for phishing website identification, this research is an important part of the whole inquiry. This chapter focuses on the techniques used to create the suggested malware detection approach and talks about how

current approaches need more refinement. To help internet users successfully recognize phishing websites, Chapter 3 presents a potential solution. In addition, the equipment and instruments that were used during the investigation are thoroughly described in this chapter. To make sure the approach for detecting phishing websites works, the next chapter will go over the steps of installing, testing, and evaluating it.

Chapter 4's results also show that the Random Forest algorithm beat the competition with a stunning accuracy rate of over 94%, as well as with precision, True Positive Rate (TPR), and Receiver Operating Characteristic (ROC) values. In several experiments, Naïve Bayes and Logistics had significantly lower performance, but J48 and KNN algorithms consistently performed over 90%. These findings point to the Random Forest algorithm as the best option for phishing attack detection. Within the context of the internet's revolutionary influence on human existence, the research highlights the need of tackling security concerns like phishing. The study optimizes feature datasets, uses machine learning classifiers, and achieves high accuracy using the random forest classifier, all with a focus on machine learning-based phishing website identification. Feature selection, lowering the false alarm rate, and investigating dynamic analysis methods are some of the improvement topics highlighted by the research. While thinking about dynamic analysis methodologies, future work should concentrate on improving the detection mechanism and prioritizing relevant feature selection.

REFERENCES

[1]. S. Hossain, D. Sarma, and R. J. Chakma, "Machine Learning-Based Phishing Attack Detection," 2020. [Online]. Available:www.ijacsa.thesai.org

[2]. M.N.Alam,D.Sarma,F.F.Lima,I.Saha,R.E.Ulfath,andS.Hossain,"Phishingattacksdetectionusingmachinelearningapproach,"inProceedingsofthe3rdInternationalConferenceonSmart Systems and Inventive Technology, ICSSIT 2020, Institute ofElectrical and Electronics Engineers Inc., Aug. 2020, pp. 1173–1179. doi: 10.1109/ICSSIT48917.2020.9214225.

[3]. N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, "Client-sidedefense against web-based identity theft." [Online]. Available:www.ebaymode.com

[4]. D. Miyamoto, H. Hazeyama, and Y. Kadobayashi, "AnEvaluationofMachine Learning-based Methods for Detection of Phishing Sites."

[5]. "Phishing E-mail Reports and Phishing Site Trends 4 Brand-DomainPairs Measurement 5 Brands & Legitimate Entities Hijacked by E-mail Phishing Attacks 6 Use of Domain Names for Phishing 7-9Phishing and Identity Theft in Brazil 10-11 Most Targeted IndustrySectors 12 APWG Phishing Trends Report Contributors thPHISHINGACTIVITYTRENDSREPORT ,"2022.[Online]. Available:http://www.apwg.org,

[6]. Dafydd Stuttard;Marcus Pinto, "Dafydd Stuttard, Marcus Pinto -Theweb application hacker's handbook_ finding and exploiting securityflaws- Wiley (2011)".

[7]. P.F. Syverson,Somesh.Jha, Xiaolan.Zhang, and A. Association forComputing Machinery. Special Interest Group on Security, CCS'08:proceedings of the 15th ACM Conference on Computer andCommunicationsSecurity :Alexandria,Virginia,USA,October27-31,2008.Associationfor ComputingMachinery,2008.

[8]. W. G. J. Halfond, J. Viegas, and A. Orso, "A Classification ofSQL Injection Attacks and Countermeasures," 2006.

[9]. L. and M. A. Vishnoi, "International Journal of ComputerScience & Information Security," International Journal ofComputerScience&InformationSecurity,vol.15,pp.1–425,2013,[Online].Available:https://sites.google.com/site/ijcsis/"Research MethodologyMethodsandTechniques(PDFDrivChoonLinTan,"PhishingDatasetforMachineLearning:Feature Evaluation," Mendeley Data, Mar. 24, 2018.

[10]. B. Espinoza, J. Simba, W. Fuertes, E. Benavides, R. Andrade, and T.Toulkeridis, "Phishing attack detection: A solution based on thetypical machine learning modeling cycle," in Proceedings - 6thAnnual Conference on Computational Science and ComputationalIntelligence, CSCI 2019, Institute of Electrical and ElectronicsEngineers Inc., Dec. 2019, pp. 202–207. doi:10.1109/CSCI49370.2019.00041.

[11]. M. Koppen, "Detecting malicious URLs using machinelearning techniques," in 2016 IEEE Symposium Series onComputational Intelligence, SSCI 2016, Institute of Electricaland Electronics

Engineers Inc., Feb. 2017. doi:10.1109/SSCI.2016.7850079.

[12]. [14] Institute of Electrical and Electronics Engineers, IEEECommunications Society;, Denshi Jōhō Tsūshin Gakkai(Japan). Tsūshin Sosaieti, and Han'guk T'ongsin Hakhoe,ICUFN 2019 : the 11th International Conference onUbiquitous and Future Networks : July 2 (Tue.)-July 5 (Fri.)2019, Zagreb, Croatia.

a. M. Kuroki, "Using Python and Google Colab to teachundergraduatemicroeconomictheory,"InternationalReviewofEconomics Education, vol. 38, p. 100225, Nov. 2021, doi:10.1016/J.IREE.2021.100225.

[13]. PrabanjanRaja,"WhatisGoogleColab?," ScalerTopics,Feb.11,2022.

[14]. PhishTank,"PhishingWebsitesDatabase,"May2023.

[15]. OpenPhish,"OpenPhishDataset,"May 2023.

[16]. "Anti-PhishingWorkingGroup(APWG),"May2023.

[17]. Kaggle,"PhishingWebsitesDataset, "May2023.

[18]. GitHub,"PhishingDatasets,"May2023.

[19]. N. Mallios, E. Papageorgiou, M. Samarinas, and K. Skriapas,"Comparison of machine learning techniques using theWEKA environment for prostate cancer therapy plan," inProceedings of the 2011 20th IEEE International Workshopson Enabling Technologies: Infrastructure for CollaborativeEnterprises, WETICE 2011, 2011, pp. 151–155. doi:10.1109/WETICE.2011.28.

[20]. I.RishandI.Rish,"AnEmpiricalStudyoftheNaïve Bayes ClassifierPredicting conversion to psychosis inclinicalhighriskpatientsusingresting-statefunctionalMRIfeaturesViewprojectClinical MachineLearningbasedonCardiorespiratory modelsandsimulationViewprojectAnempirical studyofthenaiveBayes classifier." [Online]. Available:https://www.researchgate.net/publication/228845263

[21]. N.S.Altman,"AnIntroductiontoKernelandNearest-NeighborNonparametric Regression,"1992.

[22]. S.L. R.X.S.David W.Hosmer Jr., "Applied_Logistic_Regression,"2013, doi: 10.1002/9781118548387.

[23]. A.Zabidi,M.I.M.Jaya,W.I.S.W.Din, H.A.Hassan,andI.

[24]. M. Yassin, "Non-Linear Autoregressive with Exogenous Input(Narx) Chiller Plant Prediction Model," Proc. - 2021 Int. Conf.Softw. Eng. Comput. Syst. 4th Int. Conf. Comput. Sci. Inf. Manag.ICSECS-ICOCSIM 2021, vol. 1, pp. 388–393, 2021.

[25]. N.S.Zaini,D.Stiawan,A.F.MohdFaizalAbRazak, S. K. Wan IsniSofiah Wan Din, and T. Sutikno, "Phishing detection system usingmachine learning classifiers," Indones. J. Electr. Eng. Comput. Sci.,vol. 17, no. 3, pp. 1165–1171, 2020.