ISSN 2347 - 3983
Volume 9. No.10, October 2021
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter069102021.pdf
https://doi.org/10.30534/ijeter/2021/069102021

A Review of Ensemble Learning-Based Solutions for Phishing Website Detection

Aditya Soni [1], Jyoti Tiwari [2], Ritambhara Patidar [3]
[1] M.E. Computer Engineering SGSITS Indore, India, adityasoni195@gmail.com
[2] Assistant Professor Computer Science & Engineering Department, SGSITS Indore, India, jyotimona23@gmail.com
[3] Assistant Professor Computer Science & Engineering Department, SGSITS Indore, India, NA

Received Date : September 05, 2021 Accepted Date : September 25, 2021 Published Date : October 07, 2021

ABSTRACT

Phishing is the deception of a trustworthy person in an electronic connection in order to obtain confidential information from individuals or organisations usernames, passwords, and credit card numbers are just a few examples. Phishers imitate legitimate websites by creating websites that are visually and semantically identical. As technology advances, phishing techniques have become more sophisticated, necessitating the use of antiphishing measures to detect phishing attacks. To solve the phishing attacks problems. We got the data for the Phishing website from the Kaggle open source website, which is a Google Limited Liability Company-owned online community of data scientists and machine learning experts (LLC). We are using Ensemble learning to detecting website. We are also analize accurary. We compared the results of multiple machine learning methods for predicting phishing websites.

Keywords :Phishing, Phishing Websites, Detection, Ensemble Learning.

1. INTRODUCTION

People's lives have been made easier by the availability of financial services such as banking on the Internet. As a result, maintaining the security and safety of such services is critical. Phishing is one of the most serious dangers to web security. Phishing is a technique for obtaining user credentials by impersonating a legitimate website or service on the internet. There are several varieties of phishing assaults, including Spear phishing, which targets specific individuals or businesses, Clone phishing, which involves copying an original email with an attachment or link into a new email with a different (potentially malicious) attachment or link, Whaling, and so on.

Phishing can result in significant financial losses. According to the Microsoft Consumer Safer Index (MCSI) research for 2014, the yearly global effect of Phishing and other identity crimes is projected to reach almost USD 5 billion [16]. Similarly, the Internal Revenue Service (IRS) has issued a warning about an increase in phishing attacks, claiming a 400 percent increase in reported incidents. Several methods have been proposed to combat phishing, ranging from online user education to enhanced phishing detection systems.

The conventional technique of phishing detection has failed due to the complex and dynamic nature of phishing attacks. The Anti-Phishing Working Group claims that (APWG), 239,910 different phishing reports were reported in 2018 [15]. Over the previous high point in June 2016 was 211,032 [14], the number of reports submitted increased by 12%. Despite taking precautions to avoid phishing, this happened. Further investigation revealed that each phishing attempt was unique from the others.

As a result, finding a mechanism to adjust our phishing detection systems as new attack patterns are discovered becomes critical. Because they allow a system to find new patterns from data, machine learning algorithms are an excellent answer to the challenge of phishing detection. Although numerous publications have attempted to detect phishing attacks using 10 machine learning in recent years, we intend to go one step further and build a software solution that can be easily installed on end user computers to detect phishing attempts.

In order to complete our job, On a dataset of characteristics that describe traits typically associated with phishing pages, we will test three machine learning algorithms, choose the best model based on its performance. The following is the layout of the project report. The Previous Work section describes traditional approaches to phishing detection as well as some of the machine learning approaches that have been attempted in recent years.

2. BACKGROUND THEORY

Phishing is the deceit of a trustworthy individual over an electronic connection [13] in order to get sensitive data such as usernames, passwords, and credit card numbers for malicious objectives. Phishing assaults include email phishing, internet phishing, spear phishing, Whaling, Tab napping, Evil twin phishing, and other sorts of phishing. To avoid being a victim of a phishing scam,, a variety of anti-phishing techniques should be utilised. Blacklist, heuristics, visual similarity, machine learning, and other anti-phishing technologies are examples.

2.1 Blacklist method

This is the most common technique, in which a database of phishing URLs is maintained, and if a URL is found in the database, it is flagged as phishing and a warning is provided; otherwise, it is considered authentic. Because it checks whether the URL is recorded in the database, this technique is easy and quick to create. However, the list-based strategy can be bypassed with a tiny change in URL, and the list must be updated frequently to fight new attacks.

2.2 Heuristic based method

This is a blacklist plugin that can identify new attacks by utilising features gathered from phishing websites to detect phishing attacks. However, there is a problem in that it is impossible to identify all new attacks, and it is simple to bypass once an attacker understands the method or features employed. Furthermore, because the site may or may not contain common traits, this has a low detection rate.

2.3 Machine learning

With large datasets, this technique works well. This also overcomes the present approach's drawbacks and allows for the identification of zero-day threats. Machine Learning-based classifiers are highly accurate, with a 95% accuracy rate. Performance is affected by the amount of the training data, the feature set, and the type of classifier. This has the disadvantage of failing to detect when an attacker hosts their website on a compromised domain.

In the topic of phishing detection, numerous studies have been undertaken. The majority of research has centred on improving the accuracy of phishing website detection. Using a number of different classifiers. There are a number of different classifiers that are utilized KNN, SVM, Decision Tree, ANN, and Naive Bayes are some of the terms used in machine learning.

3. LITERATURE SURVEY

Amani Alswailem et al.[1] employed a kaggle dataset as well as URL and Document Object Model (DOM) objects to represent the website's features. The URL that was utilised to extract the features of the URL and page rank. While the DOM is used to extract content page features, it is a connection between scripts and website pages that contain logical structure of documents and allow programmers to access and manipulate the DOM.

Waleed Ali et al.[2] They suggested a method for evaluating features that used two methods: There are two types of assessments: wrapper-based and filter-based. Filter-based evaluation techniques choose significant qualities based on statistical measurements to evaluate and balance aspects without categorization information. Filter-based assessment methods employ filters in their evaluation operations. To be used later in a classification, the important features are chosen with a significant dependency on the target class and little inter-correlation.

Muhammet Baykara et al.[3] Intrusion detection systems are the recommended solution to the problem. They analysed the previous results of the Bayes classifier as well as detection capability, and this proposed method checks if a website is phishing or not. There is no automation system in place.

Mona Ghotaish et al.[4] proposed method for developing a web-based phishing percentage detection system. They haven't been compared to other approaches or classifiers. This proposed solution involves physically inspecting a website to see if it is phishing or not. There is no automation system in place.

Chunlin et al. [5] suggested a method that focuses on character frequency features primarily. They used a combination of statistical URL analysis and machine learning techniques to get a more accurate classification of harmful URLs. They also evaluated at six machine-learning algorithms to see if the suggested method, which has a precision of 99.7% and a false positive rate of less than 0.4 percent, was effective.

Three new elements were proposed by Ahmad et al.[6] to increase the detection accuracy of phishing websites In this study, the author used both well-known and unique features to classify phishing and non-phishing websites. Finally, the author concludes that integrating these novel characteristics with decision tree machine learning classifiers will improve this work.

Pradeepthi et al.[7] This study looked into numerous classification methods and found that tree-based classifiers are the most accurate and deliver the best results for phishing URL detection. Lexical, URL, network, and domain-based elements are also used by the author.

M. Amaad et al.[8] For phishing website classification, we created a hybrid methodology. This paper tested the suggested model in two stages. In phase 1, they apply classification algorithms independently and select the top three models based on accuracy and other performance metrics.They blended each individual model with the best three models in phase 2 to create a hybrid model that is more accurate than the individual models.

Hossein et al.[9] Fresh-Phish, an open-source framework, was created. This system may be used to create phishing website machine-learning data. They worked with a smaller set of features and wrote the query in Python. They generate a large labelled dataset and use it to test a variety of machine-learning classifiers. The accuracy of machine-learning analysis is really good.

Gupta et al. [10] suggested a novel anti-phishing strategy based solely on client-side features. The proposed method is quick and dependable because it does not rely on a third party and extracts features solely from URLs and source code.

Mohammad et al. [11] created a methodology for detecting phishing websites that automatically extracts significant information without the need for human interaction. In this study, the author concludes that extracting features with their programme is far more efficient and dependable than hand extraction.

S.Aarthi et al.[12] To evaluate website URLs, the suggested system employs the URL Mining method. The system is separated into three modules: classifier, feature extraction, and feature analyzer, as shown. The user accesses the webpage in the classifier module, and the system then analyses it. The suspicious URL properties including length, address, and time are extracted in the feature extraction module.

3.1 Table

Table 1 : Literature Review

S. No.	**Paper Name**	**DATASET**	**Method and Classifier used**	**Parameter Used**	**Results**
1.	Detecting Phishing Websites Using Machine Learning	URL	DOM use for extract content of website	DOM , page rank, url	98% accuracy
2.	Phishing Website Detection based on Supervised Machine Learning with Wrappers Features Selection	phishing and legitimate websites	Random Forest Classifier , wrapper features selection	Accuracy	87% accuracy
3.	Detection of phishing attacks	Phishing Attacks website	Intrusion detection	Accuracy	94% Accuracy
4.	Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code	URL	Webpage Source Code	Accuracy	90% Accuray
5.	Finding effective classifier for malicious URL detection	URL	Character Frequency	False rate, Precision	99% Accuracy
6.	Feature Extraction Process: A Phishing Detection Approach	Phishing Url	Random Forest	False Rate, Precision	99% Accuracy
7.	Performance Study of Classification Techniques for Phishing URL Detection	URL	Tree based classifer	Accuracy	98% Accuracy
8.	A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms	URL	Hybrid Model for Classification	High Accuracy, Low error rate	97% Accuracy
9.	A Framework for Auto-Det	Own build dataset	DNN, SVM	High level Accuracy	89% Accuracy

	ection of Phishing Websites				
10.	Towards detection of phishing websites on client-side using machine learning based approach	URL	Random forest	True positive rate	99% Accuracy
11.	An Assessment of Features Related to Phishing Websites using an Automated Technique	URL	KNN	Accuracy	97% Accuracy
12.	Classification of Phishing Website Based on URL Features	URL	URL Mining Algorithm	Accuracy	98.5% Accuracy

4. CONCLUSION

In this study, we conducted a comprehensive literature review on phishing website identification. As a result, we can conclude that ensemble learning is more appropriate than alternative approaches

REFERENCES

[1] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh,Aram Alsedrani ,"**Detecting Phishing Websites Using Machine Learning**" ,2nd nternational Conference on Computer Applications & Information Security 2019.

[2] Waleed Ali" **Phishing Website Detection based on Supervised Machine Learning with Wrappers Features Selection**", IJACSA (International Journal of Advanced Computer Science and Applications, Vol. 8 No. 9, Issue:2017

[3] Muhammet Baykara, Zahit Ziya Gürel "**Detection of Phishing Attacks**", 6th International Symposium on Digital Forensic and Security (ISDFS), 22-25 March, 2018

[4] Mona Ghotaish Alkhozae,Omar Abdullah Batarfi "**Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code**", International Journal of Information and Communication Technology Research, Volume 1 No. 6, October 2011

[5] Chunlin Liu, Bo Lang : **Finding effective classifier for malicious URL detection** : In ACM,2018

[6] Ahmad Abunadi, Anazida Zainal ,Oluwatobi Akanb: **Feature Extraction Process: A Phishing Detection Approach** :In IEEE,2013.

[7] Pradeepthi. K V and Kannan. A: **Performance Study of Classification Techniques for Phishing URL Detection**: In 2014 Sixth International Conference on Advanced Computing(ICoAC) IEEE,2014

[8] M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar, Saira Gillani : **A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms** :In International Conference on Computational Science and Computational Intelligence IEEE ,2016.

[9] Hossein Shirazi, Kyle Haefner, Indrakshi Ray: Fresh-Phish: **A Framework for Auto-Detection of Phishing Websites**: In (International Conference on Information Reuse and Integration (IRI)) IEEE,2017.

[10] Ankit Kumar Jain, B. B. Gupta : **Towards detection of phishing websites on client-side using machine learning based approach** :In Springer Science+Business Media, LLC, part of Springer Nature 2017

[11] Rami M. Mohammad, Fadi Thabtah, Lee McCluskey: **An Assessment of Features Related to Phishing Websites using an Automated Technique**:In The 7th International Conference for Internet Technology and Secured Transactions,IEEE,2012

[12] S.Aarthi , Narsepalli Vamsi Kishan , V.Surya Teja , N.V.Harsha Vardhan Gupta: **Classification of Phishing Website Based on URL Features** : International Journal of Emerging Technologies in Engineering Research (IJETER) May (2019)

[13] **Phishing definition**,
https://en.wikipedia.org/wiki/Phishing

[14] "**Anti-Phishing Working Group** (2016). Phishing Activity Trends Report (4 th Quarter 2016). Unifying the Global Response To Cybercrime. [online] APWG".

[15] " **Anti-Phishing Working Group** (2018). Phishing Activity Trends Report (4 th Quarter 2018). Unifying the Global Response To Cybercrime"

[16] Microsoft**, Microsoft Consumer safety report.** Available at https://news.microsoft.com/ensg/2014/02/11/microsoft-consumer-safety-indexreveals-impact-of-poor-online-safety-behaviours-insingapore/sm.001xdu50tlxsej410r11kqvksu4nz