JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PHISHING WEBSITES DETECTION USING MACHINE LEARNING

[1]Arun Kumar M S, [2]Sumithra P N,[3]Vidya Y S,[4]Vaishnavi P N
[1]Assistant professor, Department of ISE, CIT, Gubbi, Tumakuru
[2] Student, Department of ISE, CIT, Gubbi, Tumakuru
[3] Student, Department of ISE, CIT, Gubbi, Tumakuru
[4] Student, Department of ISE, CIT, Gubbi, Tumakuru

Abstract: **Phishing, a widespread cybercrime, employs camouflaged websites to trick users into disclosing sensitive information or downloading malware. With the advancement of artificial intelligence, researchers increasingly rely on machine learning (ML) and deep learning (DL) algorithms for identifying phishing websites. This study conducts experiments to compare different ML and DL methods, with ensemble ML algorithms demonstrating superior accuracy and computational efficiency, even with reduced feature datasets. The paper also discusses why ensemble ML methods are well-suited for binary phishing classification in dynamic environments. Additionally, a proposed multilayered stacked ensemble learning technique achieves notable performance improvements across various datasets, with accuracies ranging from 96.79% to 98.90%. This method simplifies feature extraction and reduces processing overhead, resulting in high accuracy rates for detecting phishing websites. Despite ongoing advancements in anti-phishing techniques, challenges persist due to the evolving nature of phishing attacks, highlighting the ongoing need for research to effectively combat this cyber threat.**

Index Terms **– Phishing websites detection ,cybersecurity, machine learning, ensemble learning, deep learning.**

I. INTRODUCTION

In today's interconnected digital landscape, cybersecurity stands as a crucial defense against a plethora of cyber threats, safeguarding individuals, businesses, and nations. It encompasses a suite of technologies, procedures, and practices aimed at shielding computer systems, networks, data, and devices from unauthorized access, malicious attacks, and data breaches. This guide offers a comprehensive overview of cybersecurity, highlighting its significance, core principles, prevalent threats, and essential strategies for risk mitigation.

Significance of Cybersecurity: Cybersecurity is paramount for preserving the confidentiality, integrity, and availability of digital assets, while also fostering trust in digital interactions. Given the pervasive reliance on digital technologies across various sectors, effective cybersecurity measures are indispensable for protecting sensitive information, financial assets, and critical infrastructure from cyber threats posed by cybercriminals, hackers, state-sponsored actors, and insiders. Key Principles of Cybersecurity: Effective cybersecurity practices revolve around key principles such as risk management, defense in depth, least privilege, continuous monitoring, and incident response. These principles underscore the importance of understanding, assessing, and mitigating cyber risks, deploying layered defense mechanisms, limiting access privileges, continuously monitoring network activities, and having robust incident response plans in place. Common Cyber Threats: Cyber threats manifest in diverse forms, including malware, phishing, denial-of-service attacks, man-in-the-middle attacks, and data breaches. These threats aim to infiltrate, disrupt, or compromise computer systems, networks, and devices, leading to financial losses, reputational damage, and regulatory penalties. Essential Strategies for Cybersecurity: To effectively mitigate cyber risks and combat cyber threats, organizations and individuals should implement essential cybersecurity strategies such as secure configuration, user education and awareness, access control, patch management, network segmentation, encryption, and backup and recovery.

Comprehensive Overview of Phishing Attacks: Phishing is a prevalent cyberattack tactic wherein cybercriminals employ deceptive methods to trick individuals into divulging sensitive information. It encompasses various forms such as email phishing, spear phishing, whaling, vishing, and smishing, targeting individuals, employees, executives, and customers across diverse sectors including banking, healthcare, government, education, e-commerce, and technology. Effective methods include email filtering and authentication, anti-phishing software, employee training and awareness, two-factor authentication, URL filtering, continuous monitoring and incident response, security patches and updates, user reporting mechanisms, third-party security assessments, and collaboration and information sharing within the cybersecurity community.

A number of relevant research were explored recommended dataset and provides a quick overview of dataset preparation. The fourth chapter presents the outcomes of the experiment as well as a comparison of the classification job using six different machine learning approaches. There are talks and a 27% rise in the frequency of addressing research challenges in the fifth chapter. Furthermore, phishing is the second most expensive source of data breaches, according to IBM. In addition to

a breach caused by an attack that, on average, cost RM $4.65 million [3].

This study will create a dataset of phishing domain names using an existing intelligence database. This data collection can be used in future research as well. Different from the other data sets, the important list it generated was created by experts in security organizations and included data from national channels.
Using six different machine learning approaches, we classified the data according to eleven predefined criteria and tested the website URLs and domain names on the provided data set. We tried to determine which machine learning method would yield more accurate results using the available data. We compared the algorithm with the data content.

By implementing these strategies and staying vigilant, organizations can bolster their defenses against phishing attacks and mitigate the associated risks effectively.

II. RELATED WORK

Phishing is currently the largest issue with networks and the Internet. A number of researchers have attempted to provide tools that protect users from cyberattacks by blocking phishing URLs using black lists, white lists, deep learning, and machine learning. Machine learning-based and list-based phishing detection systems were the two types of systems that were previously presented and implemented in study. This part consists of two sections: the previous machine learning-based and list-based research.

LIST BASED PHISHING IDENTIFICATION SYSTEM

List-based phishing identification systems use white lists and blacklists, two distinct lists, to identify and classify phishing and legal URLs. Whitelist-based phishing identification systems provide safe and reliable websites in order to produce the required data. A suspicious website only needs to match the domains on the whitelist to be deemed harmful; if it doesn't, the user is acting threateningly and suspiciously. In [20], a whitelist-based system is developed, which compiles a whitelist by monitoring and recording the IP address of every website with a login form where users can submit their details. When the user utilizes this screen for login

On the Windows 2008 system, a notification regarding the incompatibility of registered information details occurs. Because of this, this system mechanism mistrusts respected websites when users visit them for the first time. Reference [21] developed a system that notifies users when a website is phishing and automatically updates and maintains the whitelist on a regular basis. The system performance is determined by the extraction of properties hidden in the link between the source code and the module corresponding to the IP address of the domain. According to the study's first results, the false negative score was 86.02 and the genuine positive rate was 1.48%. Based on the records of URLs referred to as phishing websites, blacklists were compiled. Record entries are gathered for list creation from a variety of sources, including user notifications, spam system detection, and third-party authorities.

The blacklist helps systems prevent hackers from obtaining their IP addresses and URLs. The next time, the attackers will have to use a different IP address or URL. because the blacklist-based approach detects their previous IP addresses or URLs.System security management can automatically update the blacklist on a regular basis to block new attackers by recognizing harmful URLs or IP addresses. As an alternative, these lists can be downloaded by people who want to upgrade their security system. Because blacklist-based systems cannot detect new or first-day assaults, they are mostly susceptible to zero-day attacks. Compared to machine learning-based systems, these intrusion detection systems have a lower false-positive rate. The blacklist makes it very accurate to identify intrusions or attacks on these systems.

Zhang et al. The sample data set of 3,000 websites was utilized by the model. In the study, four distinct approaches were used: the Naive Bayes classifier, Random Forests, Sequential Minimal Optimization, and Logistic Regression. The Sequential Minimum Optimization (SMO) algorithm technique has been demonstrated to be the most accurate, with an accuracy rate of 95.83%. It is unknown how this strategy will work if it is restricted to only Chinese or non-Chinese phishing e-business websites. Xiang et al. employed a pre-trained model named CANTINA in a machine learning framework to classify and apply an ID to identify phishing websites using URL, HTML DOM, and other data [24].

III. MATERIALS AND METHODS

We will first present the dataset that we have collected in this section. Next, the suggested approach is de-ned. Subsequently, the classification algorithms' trial outcomes and preference classification algorithms are showcased.

A. DATASET

It is common knowledge that accurate and well-organized datasets are essential for data-driven research. A review of the literature reveals that there are only a few studies examining a machine learning approach to phishing attacks. Fette et al. found 860 phishing emails and 6950 non-phishing emails [20], Zhang et al. found 3,000 phishing websites [13], Xiang et al. found 8,118 phishing emails and 4,883 legitimate total 13,001 web pages. Nevertheless, we found two significant issues with these datasets.

First, there is insufficient data for feature classification. The second problem is that there should be more real-world instances in these databases.

The lists that the research came to are out of date and treated more speculatively. Furthermore, the lists gathered for earlier surveys weren't done so by organizations having access to various sources. Specifically, the objective is to leverage diverse institution/organization data obtained from other target kits, ascertained by specialists through focused attacks, and utilized on a daily basis. This makes it clear that the aforementioned datasets could be helpful for certain inference tasks. More specifically, we wanted to solve the problem that attackers may manipulate the characteristics of phishing websites by creating a 124422 model that could assess many classification criteria on real-world samples. The current datasets were therefore insufficient for our investigation.

B. PREPARATION OF DATASET

We concentrate on identifying the malicious cybercriminal's source page—whether it be through inbound email, notifications, SMS, or another communication channel—during a phishing assault to steal corporate account information.

It is evident that URL and query-based statistics are frequently used while evaluating the state of the art research on phishing site detection. Furthermore, obfuscation and manipulation techniques are typically ineffective against the combination of URLs and query-based data. IP address, Sub-Domain, Pre x-Suf x, and URL length are the definition of URL analysis data. Examples of query-based statistics are the Statistics Report, Web Traffic, Google Search, Check Page Rank, and Who is Query .

IV.LITERATURE SURVEY

The University of Tokyo conducted experiments comparing various machine learning and deep learning algorithms for phishing website detection. Results indicate ensemble machine learning techniques excel in accuracy and resource efficiency, making them a strong choice for cybersecurity. Compared to other methods, ensemble algorithms demonstrate superior performance in identifying phishing websites. Machine learning and deep learning algorithms offer advanced capabilities in cybersecurity, enabling accurate threat detection. Researchers can enhance the efficiency of ensemble architectures by optimizing feature selection methods and refining model training processes.

The Multilayer Stacked Ensemble Learning Model (MLSELM) is designed to detect phishing websites, emphasizing the effectiveness of ensemble methods in handling diverse data. Previous research has explored various ensemble techniques, such as Random Forests and AdaBoost, achieving high accuracy rates. Data balancing methods like Random Under Sampling and Random Over Sampling further enhance detection effectiveness. MLSELM outperforms existing techniques in accuracy, precision, and F score across diverse datasets, bolstering cybersecurity against phishing attacks. Overall, ensemble learning, especially MLSELM, proves beneficial for identifying phishing websites and strengthening cybersecurity measures.

Names Features: Detecting Phishing Websites Using Machine Learning Methods" examines how machine learning can identify phishing websites by analyzing URL and domain name features. The study underscores the significance of categorizing URLs and domain names based on specific features to thwart cyber attacks. Prior research underscores the importance of host-based and lexical features in determining website validity. The authors emphasize the necessity of extensive datasets and expert input to create a model that accurately assesses classification features on real-world samples, thus bolstering cybersecurity measures against phishing attacks.

The survey classifies detection methods into list-based, similarity-based, and machine learning-based approaches, examining their techniques, data sets, strengths, weaknesses, and areas needing further study. It contrasts its method with previous surveys, pointing out their limitations and explaining how papers were chosen. By offering a wide-ranging view of relevant studies, it helps readers navigate the latest in phishing detection. Ultimately, it contributes by analyzing detection methods, data sets, strengths, weaknesses, and research gaps, making it a valuable tool for researchers, professionals, and anyone addressing phishing threats.

V. USE OF MACHINE LEARNING ALGORITHMS

This section goes into detail on the study's methodology and machine learning algorithm system. There are two pieces that make up the whole project. To begin, add non-phishing URLs to the master list. Accurate result analyses, near-far deviation estimations, and candlestick analysis were performed. Logistic Regression (LR), Linguistic Discriminant Analysis (LDA), Nearest Neighbor (KNN), Black Tree (DT), Support Vector Machines (SVM), and Random Forest (RF) techniques were employed to compare the acquired learning dataset in the fourth gauge. The term "accuracy" is utilized during the categorization process since the dataset separates phishing from legitimate sites. The Figure 4 underwent cross-validation. Analyzing and comparing algorithms. Volume 10.

The learning and research divisions are then created from the master list. After the learning list was created, separated into the control and training parts. Machine learning techniques are used to test the produced training list. By comparing the results acquired with the highest score—that is, by comparing the test results with the checklist results—the

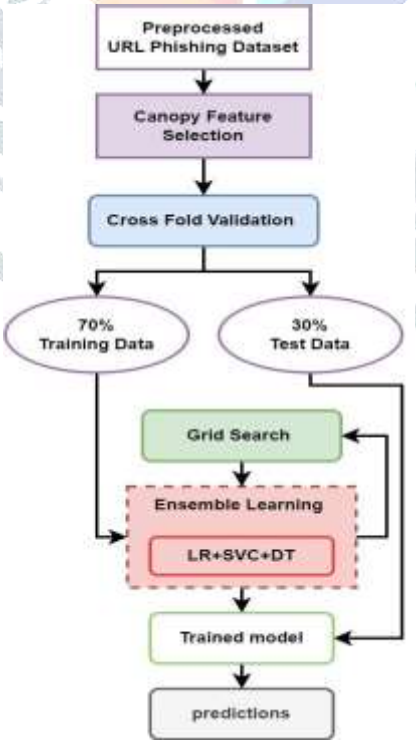
algorithm with the greatest score is determined. Figure 5 shows the graph of the distribution. There were two phases to the evaluation process for research and learning lists.

VI. METHODOLOGY

Understanding the Importance of Cybersecurity: To begin, let us discuss the critical role that cybersecurity plays in today's interconnected digital world. Call attention to how important it is to preserve the confidentiality, accessibility, and integrity of digital assets. Emphasize the value of cybersecurity in fostering trust in online transactions and protecting personal information across a range of sectors.

Defining Key Concepts: Provide an overview of the essential principles that support reasonable cybersecurity practices. Discuss concepts such as least privilege, incident response, continuous monitoring, risk management, and defense in detail. Stress the significance of these concepts for detecting and mitigating cyberthreats, establishing tiered defenses, and promptly reacting to security incidents.

Importance of Cybersecurity: To begin, let's discuss cybersecurity in greater detail, emphasizing how it safeguards digital assets and encourages confidence in online communications. Call attention to how pervasive digital technology is across a range of sectors.
Important Cybersecurity Principles: Discuss the essential concepts that form the basis of effective cybersecurity protocols, such as least privilege, continuous monitoring, risk assessment, and incident response. Give a succinct description of every principle.
Common Cyber Risks: Describe the typical cyberthreats that impact individuals, organizations, and nations. Malware, phishing, denial-of-service attacks, man-in-the-middle attacks, and data breaches are a few instances of these dangers. Provide examples and outline the potential repercussions.
Important Cybersecurity Strategies: Enumerate the most important techniques for lowering the risks and resisting cyberattacks. These could consist of patching and network security.



VI.CONCLUSION

Comparing Machine Learning (ML) and Deep Learning (DL) for phishing detection underscores DL's strengths in managing extensive data and automating feature generation, despite its intricate architecture and lengthy training periods. Ensemble ML methods like Random Forest (RF) enhance performance through model fusion and dynamic weight adaptation. While ML grapples with big data and feature extraction hurdles, automatic feature selection techniques offer potential in simplifying processes without compromising accuracy. Continuous feature updates remain critical due to the evolving nature of phishing threats. Future research aims to validate ensemble ML's efficacy across varied datasets and refine zero-day attack detection, ultimately deploying practical detection systems in real-world settings.

The introduction of a multi-layer stacked ensemble model for phishing detection illustrates high accuracy across diverse datasets. MLSELM displays superior performance with balanced data compared to imbalanced sets, surpassing baseline models significantly. Future research focuses on enhancing model effectiveness through feature selection algorithms and parameter

adjustment. Overall, the proposed model shows promise in enhancing phishing detection accuracy and emphasizes the need for further optimization in feature selection and parameter tuning.

A study utilizing web data to forecast phishing attempts employed a random forest model, achieving notable success rates in both trained and untrained datasets. Minimal data loss, appropriate ML techniques, and consistent dataset definitions contribute to the model's success. Results from the test dataset reveal high success rates in identifying both phishing and legitimate pages. The study underscores the importance of considering local and global attack factors in cybersecurity, providing a national resource for future studies to improve usability.

In conclusion, ongoing research endeavors are vital in effectively combating the ever-present and evolving threat of phishing. Advancements in ML-based detection methods, including ensemble models and automatic feature selection, offer promise in addressing the challenges posed by sophisticated phishing schemes. Additionally, focus on model interpretability, resilience against adversarial attacks, and individual education are crucial for advancing cybersecurity in the battle against phishing

REFERENCES

[1] A survey on the application of deep learning to cybersecurity was conducted by S. Mahdavi far and A. A. Ghorbani [1]. Published June 2019, Neurocomputing, vol. 347, pp. 149–176; doi: 10.1016/j.neucom.2019.02.056

[2] PhishSKaPe: A content-based strategy to evade phishing attacks, A. K. Jain, S. Parashar, P. Katare, and I. Sharma, Proc. Comput. Sci., vol. 171, pp. 1102–1109, Jan. 2020, doi: 10.1016/j.procs.2020.04.118

[3] In the 7th International Conference on Social Network Analysis, Management, and Security (SNAMS), December 2020, pages 1–8, R. Zaimi, M. Hafidi, and M. Lamia present a survey paper titled "Taxonomy of Website Anti-Phishing Solutions." doi: 10.1109/SNAMS52053.2020.9336559

[4] [PhishSKaPe: An approach focused on content to prevent phishing scams, Proc. Comput. Sci., vol. 171, pp. 1102–1109, Jan. 2020, doi: 10.1016/j.procs.2020.04.118 A. K. Jain, S. Parashar, P. Katare, and I. Sharma

[5] In the December 2020 edition of the 7th International Conference on Social Network Analysis, Management, and Security (SNAMS), R. Zaimi, M. Hafidi, and M. Lamia deliver a survey paper titled "Taxonomy of Website Anti-Phishing Solutions." doi: 10.1109/SNAMS52053.2020.9336559.

[6] E. Kirda, C. Kruegel, and E. Medvet. "Visual-similarity-based phishing detection." Published in [9]. Proc. 4th Int. Conf. Secur. Privacy Communication. Networks, September 2008, pp. 1–6. [53]

[7] A. P. E. Rosiello, E. Kirda, C. Kruegel, and F. Ferrandia. "A layout-similarity-based approach for detecting phishing pages." In Proc. 3rd Int. Conf. Secure. Privacy Communication. Net Workshops, 2007, pp. 454–463. [54]