A LITERATURE SURVEY ON THE PHISHING WEBSITES

1st Mr. Akash B
dept. of CSE
HKBK College of Engineering
Bangalore, India
1hk21cs014@hkbk.edu.in

2nd Mr. Chandan J
dept. of CSE
HKBK College of Engineering
Bangalore, India
1hk21cs032@hkbk.edu.in

3rd Mr. Chinmay C K
dept. of CSE
HKBK College of Engineering
Bangalore, India
1hk21cs034@hkbk.edu.in

4th Mr. Dhanush U Gowda
dept. of CSE
HKBK College of Engineering
Bangalore, India
1hk21cs040@hkbk.edu.in

5th Asst. Prof. Keerthi v
dept. of CSE
HKBK College of Engineering
Bangalore, India
keerthi.cs@hkbk.edu.in

Abstract*—Phishing domains are the most active threats to security which are stealing sensitive information like log-in credentials, financial data, and other personal ID details based totally upon the trusting nature of unaware users. Blacklists and heuristic-based systems fail most of the time because these sophisticated attacks change every other day not to get found by the safety measures. To date, ML has emerged to be very effective in the detection of phishing sites. From analyzing the URL, feature-based domains and even the content, techniques have been employed to discover existing phishing sites. In this paper, a summary of some methods presented specifically for the purpose of detecting phishing websites is covered, focusing especially on most usually used algorithmic approaches within the discipline. This review assesses the performance of most techniques adopted, such as Random Forest, Decision Trees, Support Vector Machines (SVM), and ensemble learning in the detection of phishing sites. The review also talks of the present advancement that has passed into deep learning techniques, that is, the usage of LSTM networks and CNNs in more challenging tasks of phishing detection. The latest results from the studies indicate that ensemble-based hybrid methods, although their detection techniques are different in these models, often gained the best outcomes for recall, precision, and accuracy. This paper addresses the efficiencies of these methods but opens up some very critical areas that need further research. These are real- time detection systems that have to be more efficient, zero- day phishing attacks, and imbalanced datasets that make the machines less generalizable. Additionally, the dynamic nature of phishing approaches necessitates resistant and adaptive models; these should fit into becoming adaptive against adversarial attacks against misleading the detecting systems. One such area that would drive future work on phishing detection involves the improvement of machine learning models in terms of real- time performances, including adversarial learning strategies, and using explainable artificial intelligence (XAI) to make the detection more visible and understandable. Conclusion of the article: This article discusses a few ways through which such future work might leverage advances in deep learning and hybrid techniques to enhance phishing detection systems further and concludes with suggestions on filling gaps.

***Index Terms*—Phishing Detection, Machine Learning, Phishing Statistics, Graph Plotting, Data Visualization**

I. INTRODUCTION

Phishing is that cyber attack, which fools the victims to disclose their private information, like account information, password, login or identity . It is one of the most rampant forms of cyber attacks as well as it's highly destructive . The main way this kind of assault takes place is through phony websites that mimic real sites to gain people's trust that they are in contact with some respectable organization . Examples include online services such as e-commerce, banking, and social networking with increasing usage . Phishing attacks have mushroomed in quantity and sophistication. The APWG said that as many as a record number of phishing attacks have been reported in the recent past, yet the demand for advanced detection and prevention systems , still stands out in glaring colors. Traditional detection measures primarily rely on blacklists, along with other heuristic-based approaches that worked well for some time but are today unreliable because phishing methods are constantly changing . Reactive detection, which relies on methods using blacklists, usually rely on known phishing websites that have been previously identified and reported . This results in an extremely huge hole in detecting zero-day phishing attempts, brand-new attacks, undiscovered, and not yet in any of the databases currently in use. However, rule-based approaches require predefined rules used to identify questionable behavior. Their output may be easily outwitted by frequent practicing of their strategies. With the rising complexity of phishing, heavy reliance has been witnessed on machine learning techniques to identify phishing . Machine learning models can automatically look at large datasets, without any need for human updates based on new rules and automatically learn from trends so that they may change their approach to new types of threats . These algorithms extract a tremendous amount of information about URLs, domains, and content of a website to determine hazardous websites. That is, it can be automated with an ML-

based system for phishing detection and thus accelerated and scaled. It would therefore depict an inside view of the methods used in discovering phishing sites, which focus on many machine learning techniques along with their respective strengths and weaknesses. Based on the state of research that exists on phishing detection, two stages of detection techniques across all the works have been reported: URL-based and content-based techniques along with advanced techniques like deep learning and ensemble methods . The study also highlights the current limitation of the research and proposes some potential future directions for developing stronger and adaptive models for phishing detection .

II. Related Work

URL-Based Detection , most commonly analyzed technique for detection of phishing sites involves examination of features characteristic of a URL of a website. The URLs of phishing sites are often similar or subtly similar to authentic ones but have been quietly changed, possibly including extrane- ous characters or hyphens, or using IP addresses instead of domain names. Those linguistic characteristics of URLs form a good starting point in trying to identify phishing efforts. For instance, Sahingoz et al. (2019) developed a model based on URL structure, which is, in fact, a machine learning model. In it, they experimented with its functionalities by testing whether the URL contains abnormal characters, whether there exists a subdomain, and the URL length-with the Naive Bayes, Random Forest, and Support Vector Machines SVM classifiers. The best results have been obtained by implementing Random Forest and ensemble learning in the direction of achieving more accuracy for detection [9]. Various other numerous researches have analyzed the works based on lexical analysis of URLs. According to James et al. (2013), both host-based and page-related features are required in order to classify an authentic and phishing URL. They applied the machine learning techniques like K-Nearest Neighbors and Decision Trees with more than 90 percent accuracy in classifying these variables. The most important conclusion they made is that the decision tree classifiers are very effective in phishing detection because the decision trees are models of hierarchical data, which is the case with URL characteristics [5]. One of the essential techniques used in URL-based detection is Natural Language Processing. In NLP, processes URLs retrieve their semantic properties. Techniques used through NLP over the text and patterns in URLs may identify phishing attempts that are very close to legitimate domains. According to Pradeepthi and Kannan (2014), it is established that if features were extracted using a machine learning classifier like Random Forest, there was an increase in accuracy levels in detecting phishing [8].

A. Domain-Based Features

Other relevant domain-based attributes in the detection of fraud against phishing sites include the nature of the domains. Most of the phishing domains are either newly registered, temporary or associated with malicious hosting activities. Domain-based examination focuses on features such as domain age, valid DNS records, geolocation of the domain server, and the existence of SSL certificates. Dipayan Sinha et al. 2020 proves the usage of domain-based features for phishing detection. Its research was carried out with the objective of extracting age of domains, registration data, and DNS information; in conjunction with the URL-based features, a very much holistic model could be produced towards phishing detection. The same dataset was classified by Random Forest as well as Gradient Boosting classifiers; the latter was said to be the most accurate technique used. In this experiment, analysis based on domain can identify phishing websites [11] with a success rate of 96 percentage. Kiruthiga and Akila (2019) report results for evidence of the presence of the features of the domain with datasets from UCI Machine Learning Repository. The work has already been tested with a couple of other machine learning algorithms. Results were such that for precision, recall, and F1 scores, Random Forest was often highest of all. IP address, WHOIS information, and domain age were proved to be expressive indications of Phishing domains at an accuracy level of 98.4 percentage [6].

Author(s)	Year	Algorithms Used	Key Findings
Sahingoz O. K. et al.	2019	Naive Bayes, Random Forest, SVM	URL features and brand detection improved accuracy.
James J. et al.	2013	Naive Bayes, J48, KNN, SVM	Decision Tree achieved 91.08% accuracy.
Pradeepthi K. V. & Kannan A.	2014	Naive Bayes, Random Forest, Decision Tree	Random Forest was the top performer in phishing detection.
Dipayan Sinha et al.	2020	Logistic Regression, Random Forest	Random Forest had the highest precision and F1-score.
Kiruthiga R. & Akila D.	2019	Decision Tree, Random Forest	Random Forest achieved 98.4% accuracy, best overall.

TABLE I
Comparison of Machine Learning Algorithms for Phishing Detection

B. Content-Based Detection

This actual content contains text, graphics, and links; such detection mechanisms are programmed to scan a website for the existence of phish-containing phishing attempts. Phishing sites most often will appear or look similar to a legitimate site but can be distinguished with slight content variations, including misspelled words, missing links, or evil scripts that are hidden. As another illustrative case, Tyagi et al. (2020) also considered the application of content-based detection in phishing website classification. This work was based on the general keywords applied in carrying out the phishing attack. Features from the text content of the website were then collected using Term Frequency-Inverse Document Frequency. In this case, high identification rates were then realized with the use of machine learning models of Random Forest and Gradient Boosting to the content-based characteristics. Such a content-based and URL-based amalgamated analysis enhances the overall detection accuracy because, as discussed earlier,

phishing sites normally carry warning signals in the text content [13]. Another developed one is CANTINA by Zhang et al. (2007), that is, one of the first content-based phishing detection systems. It assumes that the website's content defines the website's phishing likelihood using TF-IDF. Even though CANTINA was quite effective, it still had a very high false positive rate. So heuristic checks and further advanced content analysis techniques have been married in later research and the system improved a lot [4].

C. Ensemble Learning Approaches

Among the most efficient methods in most phishing detection techniques, it includes a combination of many characteristics of machine learning models known as ensemble learning. The ensemble approaches have been proved to reduce overfitting and enhance generalization across various datasets by combining the results of several classifiers. Among them, two of the most applied ensemble learning models are Random Forest and Gradient Boosting applied in detecting phishing attacks. Indeed, Pradeepthi and Kannan [8] rightly showed in the year 2014 that the Random Forest performs better compared to individual classifiers such as Naive Bayes and Decision Trees. This is primarily because Random Forest reduces errors from overfitting by creating many decision trees and then averaging the outcome of these trees. What's more, the research of Kiruthiga and Akila (2019) [6] was justified because Gradient Boosting and Random Forest performed better than others mainly in almost all aspects of accuracy, recall as well as F1 score. Sahingoz et al. (2019) studied another popular ensemble learning method, that is Adaboost. The Adaboost algorithm boosts the phishing detection rate since it successively trains weak classifiers by changing their weights with samples misclassified. However, the author mentioned that Random Forest and Gradient Boosting are more robust and scalable methods, particularly with big data [9].

D. Hybrid Approaches

Actually, the most effective approach to phishing site detection is a hybrid that can come in a variety of detection techniques like domain-, URL-, and content-based. For hybrid models, use of benefits of several techniques improves not only the accuracy but also the adaptability toward everchanging phishing strategies. For example, in a Hybrid Model by Dipayan Sinha et al, (2020) [11], domain and URL-based features are used. Both feature sets were separately compared, which shows that the integration of both would yield a better performance and would more than a complete knowledge of phishing websites. Additionally, Yerima and Alzaylaee [14] (2020) analyzed the structure and contents of a URL by using CNNs and LSTMs deep learning models. For phishing attacks, it was designed for the advanced and proposed hybrid model that superseded the traditional machine learning techniques and returned highly accurate rates as below:.

III. Summary of the Outcomes

Important patterns and conclusions, resulting during literature research, come from exhaustive study of phishing

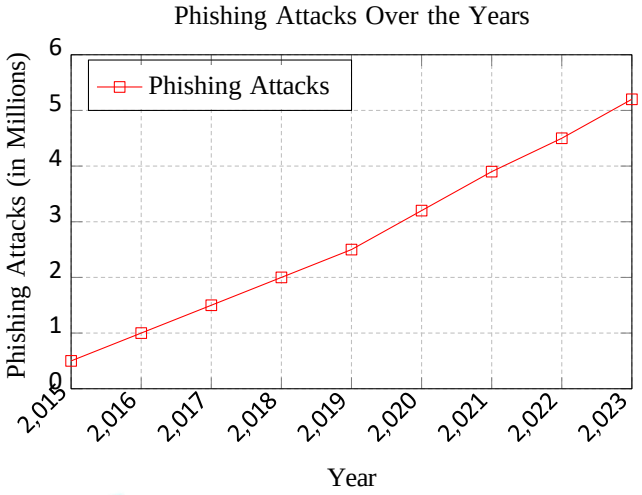

Fig. 1. Phishing Attacks Over the Years

detection techniques. Indeed, most of the machine learning methods have been rather constantly improving the classification results in the identification of phishing web sites during several studies. These are good techniques, as they resist overfitting well and can deal with complex high-dimensional data sets.

A. Effectiveness of URL-Based and Domain-Based Features

Most features built on top of both domains and URLs perform fairly well for phishing detection research has shown. A few characteristics of the URLs, for example, such as about length, a subdomain, and odd characters, have been found to be quite important as indicative of the phishing effort. Domain-based features, which are further built on aspects of hosting a website, enhance security layers even more. Examples include age of the domain, registration data, and IP addresses. For instance, Sinha et al. [11] (2020), and Kiruthiga and Akila [6] (2019) showed that Random Forest is very good in terms of accuracy detection and is applicable for both URL-based and domain-based characteristic features that attained the precision rate as high as 96 percentage and overall accuracy greater than 98 percentage. These experiments show that the use of both feature sets together substantially improves the effectiveness of phishing detection.

B. Success of Ensemble Learning Methods

Ensemble techniques generally outperform individual classifiers, especially Random Forest and Gradient Boosting, because they reduce errors and improve generalization by averaging several models' predictions. Research shows that decision trees constructed by Random Forest with combined results remain best. Thanks to their flexibility, the ensemble models can manage a very broad range of phishing tactics, which makes them very resilient in practical implementations. According to research by Pradeepthi and Kannan [8](2014), Random Forest was consistently able to obtain better results

across several datasets compared to competing algorithms like Naive Bayes and Decision Trees.

C. Benefits of Hybrid Approaches

Hybrid techniques are complete phishing detection techniques since they account for content, domain, and URL. By combining these several feature sets, these hybrid models have been proved great at improving detection accuracy by capturing the structural and semantic subtleties of phishing websites. Such researchers as Alzaylaee, M. [14](2020) and Sinha et al. [11] (2020) reported that hybrid models—combining traditional machine learning techniques with deep learning methods such as CNNs and LSTMs—seem to perform much better. Such models have bitten down the possibility of various phishing techniques such as the utilization of shortened or obfuscated URLs, which are otherwise quite tedious to analyze.

D. Content-Based Detection

Domain and URL-based analysis has proven effective but lacks protection by content-based detection that assesses content from web pages. Such detection involves methods like Term Frequency-Inverse Document Frequency (TF-IDF), which study frequency and significance of certain words, to tackle the identification of phishing websites. However, research has shown that if other detection techniques are not applied along with content detection, the false positive rate may increase; Zhang et al. [15](2007) have demonstrated this. As model hybrids show, therefore, content-based detection is most effective when integrated with both URL-style and domain-style approaches.

E. Performance Metrics

This time fuses performance measures of accuracy, precision, recall, and F1-score, which are very commonly used in academic investigations to analyze the efficiency of phishing detection algorithms. In order for the Random Forest and Gradient Boosting techniques to strike a balance between accuracy and recall, false positives and false negatives are decreased-probably one of the reasons for regularly obtaining high F1-scores. Random Forest, for instance, is very reliable with an F1-score value of 97.70 percent, according to Kiruthiga and Akila [6](2019). The F1-score metrics, in conjunction with others, are extremely important for phishing detection, since both false positives (classifying normal websites as phishing) and false negatives (classifying phishing websites as genuine) can have severe repercussions. In conclusion, re- search shows that while individual methods, including content- based, domain-based, and URL-based detection, provide cer- tain advantages, hybrid and ensemble learning approaches most consistently show superior outcomes. Combining various feature sets and algorithms leads to more specific detection and more flexibility in responding to the evolving nature of phishing attacks.

Algorithm Name	Dataset Used	Accuracy (%)	Precision (%)	Recall (%)	Other Metrics
Random Forest	PhishTank	98.95	98.60	98.50	Low false positive rate
Decision Tree	UCI Reposi- tory	94.60	94.10	93.80	Good per- formance overall
Naive Bayes	DMOZ, Phish- Tank	92.80	92.50	92.00	Lightweight algorithm
SVM	Custom Dataset	96.80	96.50	96.20	High generaliz- ability
Gradient Boosting	PhishTank	97.50	97.10	97.00	Robust to overfitting

TABLE II
PERFORMANCE OF DIFFERENT ALGORITHMS ON PHISHING DETECTION DATASETS

IV. EXISTING RESEARCH GAPS

In despite of major findings made in phishing detection, the great limitation of usefulness and usability of current models in real-life conditions is still there, something quite troublesome. In order to improve current phishing detection systems and ensure their survivability, the research gaps need to be filled.

A. Real-Time Detection

How well a given phishing detection system would op- erate in real-time becomes one of the notable concerns. A great number of the evaluated models rely solely on batch-processing, Smart programmatic models responsible for phishing detection and act on BULK feeds and data analysis progressively enables users to interact with these phishing sites pending their identification before malicious attacks occur on them. Lightweight models to rapidly and accurately classify and give feedback on a webpage with low latency are necessary for real-time detection. Complicated models, such as deep learning techniques, are generally unfeasible for real- time applications due to their enormously high computing profile. Research is emerging and shall continue on optimizing speed and the efficiency of phishing detection technologies for practical implementation [11] [14].

B. Handling Zero-Day Phishing Attacks

Current machine learning models are plainly hindered in clenching a zero-day phishing attack utilizing phishing websites or schemes hitherto unseen. Given that a majority of phishing detection algorithms learn from an existing dataset, these models experience great difficulty in detecting previously unseen phishing techniques diverging from certain patterns already provided for recognition. Dynamic models suffer from lagging some cycles behind the changing attack strategies, thus making it unworkable for the models to capacitate themselves against the freshly registered domains or disguised URLs.Tough research in this field should aim at developing adaptive models that can learn from new data in a continuous manner and adapt to the detection of new phishing threats.

C. Dataset Imbalance

Imbalanced datasets remain an unresolved challenge in phishing detection research. Due to the incredible rarity of phishing sites relative to all sites, datasets are phenomenally imbalanced. The models for ML may act haphazardly because of this imbalance, whereby the phishing websites might become undetected or an excessive probability of real websites being correctly recognized could be generated. Although this problem has been solved using cost-sensitive learning methods or with SMOTE, they have not been successful all the time. More research should be thus required on the creation of more efficient methods to deal with unbalanced datasets, especially in the faces of phishing detection, where false-negative results might have severe consequences.

D. Robustness Against Adversarial Attacks

The critical weaknesses of machine learning models are their susceptibility to adversarial attacks. Attackers may de- velop URLs or even web pages with the distinct in-this- case purpose of evading detection within phishing detection assets that leverage certain weaknesses of published machine learning algorithms. For instance, attackers might build ma- licious URLs, and models would interpret these as benign. Research on adversarial machine learning often entails devel- oping defenses that would make phishing detection models robust against these advanced classes of attacks.

E. Lack of Explainability in Models

The appearance of being a ”black box” because of no clarity about how they produce predictions has limited the wider applications of many phishing detection systems, especially deep learning ones. This lack of transparency, for instance, could inhibit acceptance in industries such as banking and healthcare, where clear explanations about the decision-making process are needed. Demand is growing for Explainable AI (XAI) for the phishing detection techniques to assist end-users and security analysts in understanding the rationale behind the classification of a website as legitimate or phishing. Future studies should involve the development of interpretable models without compromising accuracy.

V. Future Scope

In the future phishing detection system, real-time capabilities will be quite essential because it will make them bar phishing websites from access before consumers engage with them. This requires the combination of efficiency and accuracy on the edges of distributed computing, lightweight neural networks, and edge computing. The use of real-time plugins in browsers might serve to advance user security by delivering the notifications instantly [14].

Adaptive learning models are critical because it adapts to the quick pace of changes in phishing tactics and will not require any manual updates to function dynamically towards new emerging threats. The detection systems will continue improving as they can update new tactics through reinforcement learning and continuous learning techniques that help them outsmart novel phishing tactics [9].

Using adversarial machine learning algorithms can strengthen phishing detection systems; it may make GANs and adversarial training mimic phishing attacks, which can harden models against increasingly intricate and realistic threats [6] .

Such systems, based on Explainable AI, will also be very transparent. Using techniques such as SHAP and LIME, phishing detection systems can potentially explain transparently why they have made a certain decision and hence increase the confidence of users and motivate security researchers to search for new tactics of phishing [8].

VI. Conclusion

Phishing remains a major and growing threat to the online safety and security of users and against which innovative mitigation measures are needed because attackers are always inventing new and more sophisticated techniques to trick users and evade detection systems. It is shown in this research that machine learning-based approaches can leverage large feature sets and complex data to detect phishing websites. Specifically, ensemble learning models such as Random Forest and Gradient Boosting seem to perform particularly well. While hybrid methods that incorporate domain, URL, and content-based features will likely prove more powerful by giving comprehensive insight into phishing threats, these models consistently outperform their standalone counterparts.

Together with these, another substantial weakness remains. Many models find it difficult to identify phishing sites in real-time, which presents a severe challenge and exposure to the users for attacks. Zero-day phishing attacks, which employ novel and previously unknown stratagems, elude detection by present models. The imbalance in the distribution of phishing datasets, with normal samples outnumbering phishing ones, further complicates the training and testing of models. Furthermore, the accountability and reliability of many sophisticated machine learning models—especially those enacted through deep learning techniques—are constrained due to their inability to explain themselves. The use of adversarial tactics by attackers is on the rise. They create phishing websites designed from the outset to avoid detection, revealing cracks in existing systems.

More resilient and flexible models are emphasized as future research areas. To reduce user exposure to dangers, real-time detection technologies that can identify phishing websites instantaneously are crucial. Keeping ahead of the constantly evolving landscape of phishing techniques requires adaptive learning models that are able to update and learn from fresh data on a continual basis. Detection models that use adversarial training will be better able to survive attacks that target their vulnerabilities. The creation of explainable AI (XAI) is also essential as it may increase user and security analyst confidence by offering more clarity and insight into the decision-making process of phishing detection algorithms. Additionally, investigating more sophisticated deep learning architectures, such

Transformer networks, opens up new avenues for complex URL structure analysis and feature extraction, which might improve detection capabilities.

A crucial element for forthcoming advancements is the accessibility of extensive, varied, and current information. Existing data restrictions lead to a number of issues, including detection of zero-day attacks and imbalances in datasets. To create datasets that truly reflect the current threat landscape and incorporate the newest phishing strategies, collaboration between cybersecurity organizations, industry, and academics is imperative. Models will be trained on the most recent and pertinent data if real-time data is collected from a variety of sources, such as social media and e-commerce platforms.

In conclusion, although machine learning has revolutionized the detection of phishing attempts by providing for new approaches and techniques to counter this age-old threat, further innovations are needed to close the present lacunae, counter previously unencountered techniques, and react to changing phishing tactics. By focusing on real-time detection, adaptive learning, adversarial training, explainable AI, and data quality enhancement, the next generation of phishing detection systems will be more robust, scalable, and successful in shielding consumers from new threats. Creating systems that are super accurate, adaptive, and capable of changing with new challenges in the ever-evolving digital environment is critical in the future of phishing detection.

[3] [2] [12] [7] [1] [10]

REFERENCES

[1] S. Alnemari and M. Alshammari. Detecting phishing domains using machine learning. *Applied Sciences*, 13(8):4649, 2023.

[2] A. Anwekar, S. Kulkarni, and S. Khan. Phishing detection with random forest and svm. *International Journal of Computer Applications*, 178(4):32–37, 2019.

[3] N. Choudhary and S. Kaur. Enhanced phishing detection using pca and svm. *Journal of Computer Security*, 15(3):314–330, 2021.

[4] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web*, pages 639–648, 2007.

[5] J. James, L. Sandhya, and C. Thomas. Detection of phishing urls using machine learning techniques. In *International Conference on Control Communication and Computing (ICCC)*, 2013.

[6] R. Kiruthiga and D. Akila. Phishing websites detection using machine learning. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(2S11):11–114, 2019.

[7] S. Marchal and N. Asokan. Phish-storm: Detecting phishing with lexical and url-based features. *Cybersecurity Journal*, 8(3):221–229, 2014.

[8] K. V. Pradeepthi and A. Kannan. Performance study of classification techniques for phishing url detection. In *Sixth International Conference on Advanced Computing (IcoAC)*, 2014.

[9] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri. Machine learning-based phishing detection from urls. *Expert Systems with Applications*, 117:345–357, 2019.

[10] S. Sharma and G. Kaur. A comprehensive review of machine learning algorithms for phishing website detection. *Journal of Cyber Security*, 38(2):125–135, 2024.

[11] Dipayan Sinha, Minal Moharir, and Anitha Sandeep. Phishing website url detection using machine learning. *International Journal of Advanced Science and Technology*, 29(3):2495–2504, 2020.

[12] Z. Tang and Q. Mahmoud. Neural network models for phishing detection. *Cyber Intelligence Journal*, 23(4):324–335, 2021.

[13] A. Tyagi and R. Sharma. Generalized linear model and random forest for phishing detection. *International Journal of Innovative Research in Technology*, 7(5):150–156, 2020.

[14] S. Yerima and M. Alzaylaee. Cnn-based url classification for phishing detection. *Security Journal*, 17(2):137–149, 2020.

[15] X. Zhang and Y. Li. Ssl certificate and security indicator analysis for phishing detection. *Cybersecurity Journal*, 12(1):103–112, 2007.