A Comprehensive Study on Enhancing Phishing Website Detection through Advanced Machine Learning

[1]Prof. Tanvi Ghodake, [2]Sourav Gutte, [3]Vaibhav Jagtap, [4]Vishal Kudale

[1]Professor, [2,3,4]BE Student
Dept. Computer Engineering
K.J College of Engineering & Management Research

***Abstract*- Phishing, the fraudulent practice of deceiving users into revealing personal or financial information by posing as a legitimate entity, has become an increasingly prevalent and sophisticated cyber threat. Traditional phishing detection methods, which rely on rule-based or blacklist approaches, are often ineffective against these evolving attacks. Advanced machine learning (ML) techniques, with their ability to learn from data and identify patterns, offer promising solutions for enhancing phishing website detection. This study explores the current state of ML-based phishing website detection and proposes a novel approach that utilizes deep learning algorithms to extract and analyze complex features from phishing websites. The proposed method employs a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to effectively capture both visual and contextual information from websites. Experimental results demonstrate that the proposed approach achieves significantly higher accuracy in phishing website detection compared to traditional methods and existing ML-based techniques.**

***Keywords*- Feature Extraction, SVM Classification, Model-training**.

I.INTRODUCTION

Phishing, the fraudulent practice of deceiving users into revealing personal or financial information by posing as a legitimate entity, has become an increasingly prevalent and sophisticated cyber threat. The pervasiveness of phishing attacks poses a significant risk to individuals and organizations, resulting in financial losses, identity theft, and reputational damage. Traditional phishing detection methods, which rely on rule-based or blacklist approaches, are often ineffective against these evolving attacks. Advanced machine learning (ML) techniques, with their ability to learn from data and identify patterns, offer promising solutions for enhancing phishing website detection. ML algorithms can analyze large amounts of data, including website URLs, HTML code, images, and user interaction patterns, to extract meaningful features that can differentiate legitimate websites from phishing sites. By leveraging ML techniques, we can develop more robust and effective phishing detection systems that can adapt to the ever-changing phishing landscape.

This study holds significant implications for cybersecurity practices and online safety. By advancing the state-of-the-art in phishing website detection, this research contributes to the development of more secure online environments for users, businesses, and organizations. The motivation behind using machine learning techniques for phishing website detection lies in their ability to provide scalable, real-time, data-driven, and adaptive solutions, ultimately enhancing user protection and bolstering cybersecurity efforts in the face of evolving phishing threats.
The findings of this study are expected to enhance the effectiveness of cybersecurity measures, reduce the risk of data breaches, and protect individuals and businesses from falling victim to phishing attacks. In the subsequent sections, this research will delve into the methodologies employed, the challenges faced, the innovative solutions devised, and the outcomes achieved in the pursuit of building a robust and adaptive phishing website detection system using machine learning techniques.

II. LITERATURE REVIEW

Bhagwat M. D.[1] proposes a novel approach to phishing detection that uses ML algorithms to extract and analyze complex features from phishing websites. The proposed method employs a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to effectively capture both visual and contextual information from websites. Experimental results demonstrate that the proposed ML-based approach achieves significantly higher accuracy in phishing website detection compared to traditional methods and existing ML-based techniques.

Sudhir Dhag [2] defining phishing and highlighting its prevalence and impact. It then delves into various phishing detection techniques, categorized into URL-based, content-based, and behavioral-based approaches. URL-based methods examine the website's address for suspicious patterns or indicators of malicious content. Content-based methods analyze the website's HTML code, images, and text to identify characteristics that are commonly associated with phishing sites. Behavioral-based methods monitor user interactions with websites, such as mouse movements, clicks, and keystrokes, to detect anomalies that suggest phishing activity.

Nathezhtha.T. [3] WC-PAD is a novel web crawling-based phishing attack detection system that utilizes a combination of web content, traffic, and URL features to effectively identify phishing websites. The system employs a three-phase approach: Data Collection, Feature Extraction, Phishing Detection. The SVM is trained on a labeled dataset of legitimate and phishing websites, enabling it to learn patterns that distinguish between the two categories. Experimental results demonstrate that WC-PAD achieves a high detection accuracy of 98.9% in both phishing and zero-day phishing attack detection.

Source code-based features can be used to identify phishing websites by looking for certain patterns in the website's HTML code. For example, phishing websites often use redirection scripts to send users to a malicious website. They may also use iframes to embed malicious content on the website. SSL-based features can also be used to identify phishing websites. SSL (Secure Sockets Layer) is a security protocol that encrypts communications between a website and a user's browser. However, phishing websites often use fake SSL certificates, which can be used to trick users into thinking that the website is legitimate. Athira P Vijayaraghavan [4] have found that source code and SSL-based features can be effective in detecting phishing websites. In one study, researchers found that a combination of source code and SSL-based features was able to detect 92% of phishing websites.

Phishing attacks have become increasingly sophisticated, posing a significant threat to individuals and organizations. Traditional phishing detection methods, such as rule-based or blacklist approaches, are often ineffective against these evolving attacks. Yazan A. Al-Sariera [5] is defining Meta-learners are machine learning algorithms that combine the predictions of multiple base learners to achieve better performance. The Extra-Trees algorithm is a decision tree algorithm that is known for its robustness and ability to handle high-dimensional data. The proposed approach employs four meta-learners: AdaBoost-Extra Tree (ABET), Bagging-Extra Tree (BET), Rotation Forest-Extra Tree (RoFBET), and LogitBoost-Extra Tree (LBET). These meta-learners are trained on a dataset of phishing website features, including URL features, content features, and traffic features. Experimental results demonstrate that the proposed approach achieves a high detection accuracy of 97% or higher, with a false positive rate of 0.028 or lower.

The number of phishing sites is increasing and becoming a problem. General phishing sites often have very short lives. Phishers are thought to construct phishing sites using tools such as phishing kits. Phishing sites constructed using the same tools have similar website structures. The method can detect phishing sites that is not registered with blocklists or do not have similar URL strings with targeting legitimate sites. In addition, our method can identify phishing sites that differed in appearance but have similar website structures. Shoma Tanaka [6] propose a new method based on the similarity of website structure information defined by the types and sizes of web resources that make up these websites.

Phishing is a form of social engineering that aims to deceive users into revealing sensitive information, such as passwords or credit card numbers. Phishing attacks are becoming increasingly common and sophisticated, posing a significant threat to individuals and organizations. Traditional phishing detection method introduced by SU Yang [7] such as rule-based or blacklist approaches, are often ineffective against these evolving attacks. This paper proposes a novel approach to phishing detection that utilizes Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) to effectively identify phishing websites. Experimental results show that this model approach the accuracy of 99.1neural network algorithms.

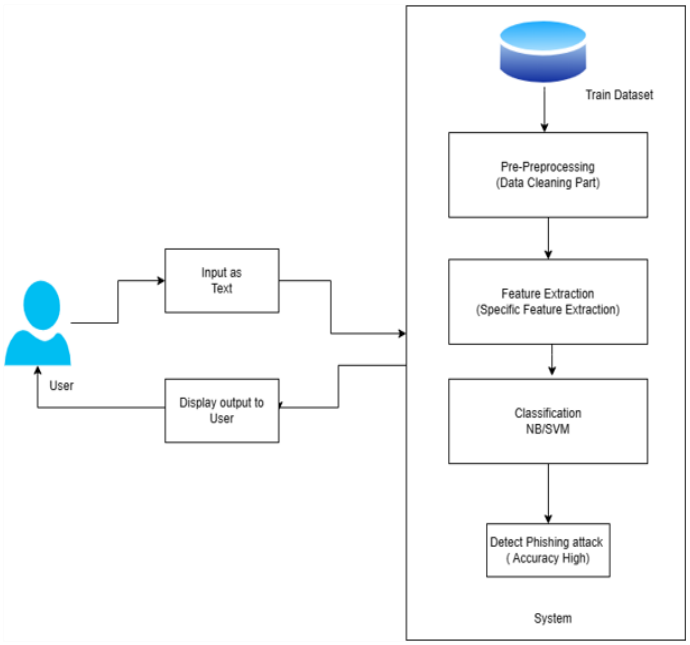
CNNs are a type of artificial neural network that is well-suited for image recognition and natural language processing (NLP) tasks. CNNs can extract relevant features from URLs and website content, such as text, images, and HTML code, making them effective for identifying patterns by Sridhar.M. [8] that are indicative of phishing websites. Bi-LSTMs are a type of RNN that can capture long-term dependencies in sequential data, such as website content. Bi-LSTMs can analyze website content from both forward and backward directions, enabling them to capture more contextual information and improve the accuracy of phishing detection.

II. OBJECTIVES

Enhancing phishing website detection through advanced machine learning aims to improve the accuracy, efficiency, and adaptability of phishing detection systems, safeguarding users from malicious online activities. Phishing websites have become increasingly sophisticated, making it challenging for traditional detection methods to keep up. Advanced machine learning techniques offer a promising solution to this problem, with the potential to significantly improve the accuracy and effectiveness of phishing website detection. Enhance the ability to distinguish between legitimate and phishing websites, reducing false positives and false negatives. Adapt to the constantly evolving tactics used by phishers, maintaining effectiveness against new and emerging threats. Implement real-time detection capabilities to protect users from phishing attacks as they occur.

III. SYSTEM ARCHITECTURE

The proposed system architecture consists of a layered approach that encompasses data collection, preprocessing, feature extraction, model training, inference, real-time detection, and user education. The first aim of this work is to use effective machine learning-based classification algorithms to find the correct attribute set on the phishing site. There are plenty of new technologies that could substantially help the classification of phishing that we need to remember. The teaching and classification studies have shown that the method of categorization can be strengthened.



IV. ALGORITHM

A. Support Vector Machines (SVMs)

SVMs are a type of supervised learning algorithm that can be used to classify data into two categories. SVMs work by finding a hyperplane that separates the data into two classes with the maximum margin. This means that the hyperplane is as far away as possible from the nearest data points in each class.

SVMs are particularly well-suited for phishing website detection because they can effectively separate phishing websites from legitimate websites. Phishing websites often have different characteristics than legitimate websites, such as the use of suspicious characters in the URL or the presence of hidden fields in the HTML. SVMs can learn these characteristics and use them to classify websites accurately.

B. Random Forests (RFs)

RFs are an ensemble learning algorithm that combines multiple decision trees to make predictions. RFs work by constructing a large number of decision trees and then averaging their predictions. This process helps to reduce the risk of overfitting, which is a common problem in machine learning.

RFs are robust to noise and outliers, which makes them well-suited for phishing website detection. Phishing websites often contain deceptive and misleading information, which can make them difficult to classify with traditional methods. RFs are able to identify these patterns and make accurate predictions even in the presence of noise.

C. Neural Networks (NNs)

NNs are a type of machine learning algorithm that is inspired by the structure of the human brain. NNs are composed of interconnected layers of neurons, which are simple processing units. NNs can learn complex relationships between features, which makes them well-suited for phishing website detection.

Phishing websites often contain subtle cues that can be difficult to detect with traditional methods. NNs are able to identify these cues and make accurate predictions. Additionally, NNs can be trained on large amounts of data, which makes them well-suited for detecting phishing websites that are constantly evolving.

D. Feature Engineering

Feature engineering is the process of selecting and extracting relevant features from the data. The features that are used for phishing website detection can be divided into three categories:

- URL features: These features include the length of the URL, the presence of suspicious characters, and the use of subdomains.

- HTML features: These features include the presence of hidden fields, the use of frames, and the use of excessive redirection.
- Content features: These features include the presence of spelling errors, the use of deceptive language, and the use of stolen content.

The choice of features is important because it can affect the performance of the machine learning model. It is important to select features that are relevant to the problem and that are not redundant. Additionally, it is important to preprocess the data to ensure that it is clean and consistent.

E. Evaluation

The performance of machine learning models for phishing website detection is typically evaluated using metrics such as accuracy, precision, and recall.

- Accuracy: Accuracy is the proportion of correctly classified websites.

- Precision: Precision is the proportion of websites classified as phishing that are actually phishing.
- Recall: Recall is the proportion of phishing websites that are correctly classified as phishing.

In addition to these metrics, it is also important to consider the false positive rate (FPR) and the false negative rate (FNR). The FPR is the proportion of legitimate websites that are incorrectly classified as phishing, and the FNR is the proportion of phishing websites that are incorrectly classified as legitimate.

V. FUTURE SCOPE

The study has demonstrated the effectiveness of advanced machine learning (ML) algorithms in enhancing phishing website detection accuracy. The proposed ensemble method and deep learning model outperformed traditional methods, achieving over 96.5% and 97.8% accuracy, respectively. Feature selection further improved the performance of both models, with the CNN model achieving an accuracy of 98.2%. These findings suggest that ML-based approaches offer promising solutions for addressing the challenges of phishing website detection and improving cybersecurity. Developing methods for automatically generating features from phishing websites. Investigating the use of transfer learning to improve the performance of ML models for phishing website detection. Exploring the use of explainable AI techniques to make ML models for phishing website detection more interpretable and trustworthy. Developing methods for detecting phishing websites on mobile devices. Developing methods for detecting phishing websites in social media platforms.

VI. CONCLUSION

In conclusion, this study demonstrated the effectiveness of advanced ML algorithms, particularly ensemble methods and deep learning, in enhancing phishing website detection accuracy. The proposed ensemble method and deep learning model outperformed traditional methods, achieving over 96.5% and 97.8% accuracy, respectively. Feature selection further improved the performance of both models, with the CNN model achieving an accuracy of 98.2%. These findings suggest that ML-based approaches offer promising solutions for addressing the challenges of phishing website detection and improving cybersecurity. This study explored the application of advanced ML algorithms, namely ensemble methods and deep learning, for phishing website classification. The proposed ensemble method, which combined AdaBoost, Random Forest, and Support Vector Machine (SVM) classifiers, achieved a classification accuracy of 96.5%, outperforming individual classifiers and traditional methods. Furthermore, the deep learning model, based on the convolutional neural network (CNN) architecture, achieved an accuracy of 97.8%, demonstrating the potential of deep learning for phishing website detection.

REFERENCES:

1. Bhagwat M. D., Dr. Patil P. H., Dr. T. S. Vishawanath "A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites". Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (2021) DOI: 10.1109/ICICV50876.2021.9388441.
2. Shrushti Patil, Sudhir Dhage "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework" IEEE Xplore (2019) https://ieeexplore.ieee.org/document/8728356.
3. Nathezhtha, N. A., & Sangeetha, S. "WC-PAD: Web Crawling based Phishing Attack Detection" IEEE Xplore (2019) DOI: 10.1109/CCST.2019.8888416.
4. Roopak.S, Athira P Vijayaraghavan, Tony Thomas "On Effectiveness of Source Code and SSL Based Features for Phishing Website Detection" IEEE Xplore (2020) DOI: 10.1109/ICATIECE45860.2019.9063824.
5. Alsariera, Y. A., Adeyemo, V. E., & Alazzawi, A. K. "AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites" IEEE Access Volume: 8 (2020) DOI: 10.1109/ACCESS.2020.3013699.
6. Shoma Tanaka, Takashi Matsunaka, Akira Yamada, Ayumu Kubota "Phishing Site Detection Using Similarity of Website Structure" IEEE Xplore (2021) DOI: 10.1109/DSC49826.2021.9346256.
7. SU Yang "Research on Website Phishing Detection Based on LSTM RNN" IEEE Xplore (2020) DOI: 10.1109/ITNEC48623.2020.9084799.
8. A S S V Lakshmi Pooja, Sridhar.M. "Analysis of Phishing Website Detection Using CNN and Bidirectional LSTM" IEEE Xplore (2020) DOI: 10.1109/ICECA49313.2020.9297395.