Grenze International Journal of Engineering and Technology, June Issue

Phishing Website Detection using Neural Networks

Sowmya K[1], Vijayalakshmi S Katti[2], Sushitha S[3] and Uday Kumar T[4]
[1,4]Nitte Meenakshi Institute of Technology, Department of MCA, Bengaluru, India
Email: sowmyaraik.somi@gmail.com, udaytgvt@gmail.com
[2]SJR College for Women, Department of Computer Science, Bengaluru, India
Email: viju.s.katti@gmail.com
[3]Nagarjuna Degree College, Department of MCA, Bengaluru, India
Email: bhatsushitha@gmail.com

***Abstract*— Phishing is a pervasive technique used by cybercriminals to deceive individuals into divulging personal information by using counterfeit websites designed to look like legitimate ones. These phishing websites are crafted to steal sensitive data such as usernames, passwords, and financial details by mimicking the appearance and language of authentic sites, making it challenging for users to discern the difference. As phishing techniques rapidly evolve alongside technological advancements, employing effective anti-phishing strategies .is crucial. Machine learning emerges as a powerful tool in counteracting phishing attacks, as attackers find it easier to trick victims into clicking seemingly genuine malicious links than to penetrate computer security systems directly. These links often feature the spoofed company's logos and authentic information, adding to their credibility. Our proposed method leverages machine learning, specifically the Gradient Boosting Classifier, to create an innovative approach for detecting phishing websites by analyzing URL features. This involves evaluating characteristics such as URL length, suspicious characters, and uncommon domain extensions, which help differentiate legitimate sites from phishing attempts. Our approach operates in real time, allowing for the swift identification and mitigation of threats. The results of our studies show that this method effectively distinguishes between legitimate and fraudulent websites, offering a reliable means of real-time protection against phishing attacks. This innovative solution not only enhances security but also provides a scalable approach to addressing the ever-evolving nature of phishing threats.**

***Index Terms*— Phishing Detection, Machine Learning, Neural Networks, Cyber security, Model Training.**

I. INTRODUCTION

Phishing is a type of cyber assault in which attackers impersonate genuine companies in order to acquire sensitive information such as usernames, passwords, and credit card numbers. These assaults are often carried out by constructing phony websites that have the appearance of legitimate sites. Detecting phishing websites is critical for preventing visitors from falling victim to these criminal actions.

Traditional phishing detection techniques are largely concerned with recognizing and mitigating phishing efforts using spam filters and blacklists, examines email content and behavioural patterns, Checks URLs for known malicious domains and suspicious redirection, as well as domain reputation.

Grenze ID: 01.GIJET.11.2.513_2

Uses email authentication protocols (SPF, DKIM, and DMARC) and multi-factor authentication (MFA) to authenticate origins and add security levels. Utilizes methods to assess and flag emails and URLs based on their reputation. Scans and sandboxes attachments to identify and isolate potential malware. Maintains lists of trusted or known harmful websites in order to control communication and reduce threats. These strategies try to detect and prevent phishing by assessing many signs and using multiple layers of defence.
In this project, we focus on using neural networks, a type of machine learning model inspired by the human brain, to detect phishing websites RNNs and LSTMs are neural networks that assess email content for suspicious language or patterns associated with phishing. CNNs and other models scan URL structures and domain names for anomalies or indicators of malicious intent. Neural networks can extract and interpret information from emails and webpages to distinguish between phishing and authentic material. Feedforward neural networks, CNNs, and RNNs are trained on phishing and legitimate datasets to accurately classify new data. Models can be updated with new phishing data to increase detection accuracy and respond to emerging attacks. Neural networks improve phishing detection by identifying complicated patterns that standard methods may miss.
An automatic method for identifying phishing websites that combines the use of Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Gradient Boosting Classifier. This system will analyse numerous aspects of URLs and use a holistic approach that makes use of the advantages of many algorithms. The integration of these techniques is intended to guarantee a high degree of accuracy and dependability in the differentiation of reputable websites from dangerous ones, hence augmenting the overall efficacy of phishing detection. A crawler was created to extract 7900 URLs from the Alexa Rank portal. The Phish tank dataset was also utilized to gauge the effectiveness of the suggested URL detector. The study's conclusion shows that, in comparison to the current deep learning techniques, the suggested strategy yields better outcomes. The suggested URL detector yielded a total of 7900 dangerous URLs.

II. Statement of the Problem

Phishing websites are a serious danger to online security because they use false identities to obtain private data. Because of the sheer volume of websites on the internet and their growing sophistication, identifying these fraudulent sites is a difficult task. Evolving phishing tactics frequently prove to be too much for traditional detection technologies to keep up with. Manual or rule-based detection is problematic due to the vast volume of webpages. Efficient processing and analysis of vast amounts of data require automated techniques. Phishing websites can closely resemble authentic websites, making it challenging to recognize distinguishing characteristics with conventional techniques.

III. Literature Survey

Phishing is a deceptive practice and a type of cyber-attack devised and carried out with the express intention of obtaining sensitive information by impersonating legitimate websites. Phishers deceive users by mimicking the original and legitimate contents in order to obtain personal information such as a security number, credit card number, password, and so on. Many anti-phishing strategies, such as blacklist- or whitelist-, heuristic-feature-, and visual-similarity-based methods, have been proposed as of today [1]. Phishing is the most common form of cybercrime, with the goal of persuading victims into providing accurate information such as account IDs, bank details, and passwords. This form of cyberattack is typically triggered by emails, instant chats, or phone calls. Existing anti-phishing techniques rely mostly on source code elements that necessitate scraping web page content, as well as third-party services that slow the classification of phishing URLs. Although machine learning techniques have recently been employed to identify phishing, they require manual feature engineering and are not experts at detecting emerging phishing offenses [2].
The frequency with which website security incidents have occurred in recent years has prompted us to research website security. The accuracy of phishing website detection has not been satisfactory, despite the availability of numerous phishing detection techniques. In this paper, we offer a unique neural network classification method-based phishing detection model. By using the design risk minimization concept, this detection model may achieve high accuracy and good generalization ability [3].
Due to the ongoing increase in phishing attempts and the number of phishing websites, people and businesses throughout the world are now more vulnerable to different types of cyberattacks. Consequently, enhanced cyber defence necessitates more potent phishing detection. Therefore, we provide a deep learning-based method in this research to enable high accuracy phishing site detection. The suggested method classifies real websites from phishing sites with excellent accuracy by using convolutional neural networks (CNN). We assess the models on

a dataset that we collected from 4,898 phishing and 6,157 legitimate websites [4]. The author of this paper suggested a machine learning-based URL detection method. The technique of recurrent neural networks is used to identify phishing URLs. Researchers used 5800 legitimate and 7900 malicious websites to test the suggested strategy. The results of the studies demonstrate that the suggested strategy outperforms more contemporary techniques for malicious URL detection [5].
In order to identify phishing websites, they have suggested a supervised learning strategy that makes use of deep learning algorithms. With the CNN (Conv2D) model, we obtained 93.6% accuracy, while with the regular neural network model, we obtained 94.8% accuracy. A dataset that we used was obtained from the machine learning repository at the University of California, Irvine [6].
Using the Scikit Learn package, eight popular classification methods were set up in Python and their classification accuracy was evaluated on all publicly accessible phishing datasets. Afterwards, three distinct ranking strategies were used to order classification algorithms according to accuracy on various datasets, and Welch's T-Test was used to determine whether the results indicated a statistically significant difference. This study presents the comparison results, which demonstrate the superior performance of neural networks and ensembles over other traditional techniques [7].

IV. PROPOSED SYSTEM

Advanced methods for enhancing the identification of phishing websites are provided by neural networks, especially those that are based on Python. Large volumes of data must be processed and analysed efficiently, which calls for automated techniques.
Complex properties from website data, such as HTML content, URL structure, and graphic aspects, can be automatically learned and extracted by neural networks. These features are essential for spotting phishing efforts. Enhance the identification of phishing websites by utilizing Python and neural networks, offering a stronger defence against a constantly changing cyber threat scenario.

V. METHODOLOGY

You can use Python and a combination of Gradient Boosting Classifier (GBC), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs) to detect phishing websites.

A. Gathering and pre-processing data

assemble a varied dataset that includes reputable and fraudulent websites. website URLs, the webpages' source code, screenshots, or HTML content that has been rendered.
Features to extract include the number of subdomains, domain length, and the existence of questionable keywords. Extract tags, attributes, and content structure from HTML material by parsing it. Utilize image processing techniques or convert screenshots to pixel values. Scale graphical elements to a predetermined range. Transform numerical forms into category data (like HTML tags). Tokenize and pad sequences for text data to guarantee consistent input size.

B. Model Building and Training

Spatial patterns and hierarchies can be captured by CNNs. CNN may be trained to distinguish between phishing and legal websites by using picture or structured content data. Network with Long Short-Term Memory (LSTM): Manage sequential data, including text from HTML pages or URLs. Long-term dependencies are captured in sequences by LSTMs. LSTMs can be useful in comprehending sequential patterns in URL data. GBC, or gradient boosting classifier: To increase overall accuracy, combine forecasts from multiple learners who are not as strong. GBC using a mix of characteristics taken from the LSTM and CNN models.

C. Fusion and Integration of Models

Integrate CNN, LSTM, and unprocessed feature sets. Concatenating the feature vectors that each model generates will do this. To improve detection accuracy, combine predictions from CNN and LSTM models using the Gradient Boosting Classifier.

D. Evaluation

Analyse models for phishing site detection efficacy utilizing metrics such as ROC-AUC, F1-score, accuracy, precision, and recall. To make sure that models generalize well across various data subsets, do k-fold cross-validation.

E. Python Libraries and Tools

CNN and LSTM construction and training may be done with Tensor Flow/Keras. PyTorch: For advanced approaches and the implementation of neural network models. Scikit-learn: For pre-processing and the implementation of gradient boosting classifiers. Pandas/Numpy: For the analysis and manipulation of data. OpenCV/Pillow: To handle visual features and process images.

F. Model Implementation

Place the trained model in a real-time setting so that users may use it to examine URLs as they come across them. This might be incorporated into email clients, security programs, or web browsers. Provide a straightforward API or user interface that enables users to submit URLs for phishing detection and get fast response.

G. Monitoring and Maintenance

Keep an eye on the deployed model's effectiveness to make sure it keeps up with the emergence of new phishing attack types. To keep the model effective against changing phishing techniques, periodically update it with new data and retrain it.

H. Reporting and Documentation

Record every step of the process, including data gathering, pre-processing, and modelling. Provide a detailed report of the model's performance, including insights into which features are most indicative of phishing websites and any challenges encountered during development.

VI. System Architecture

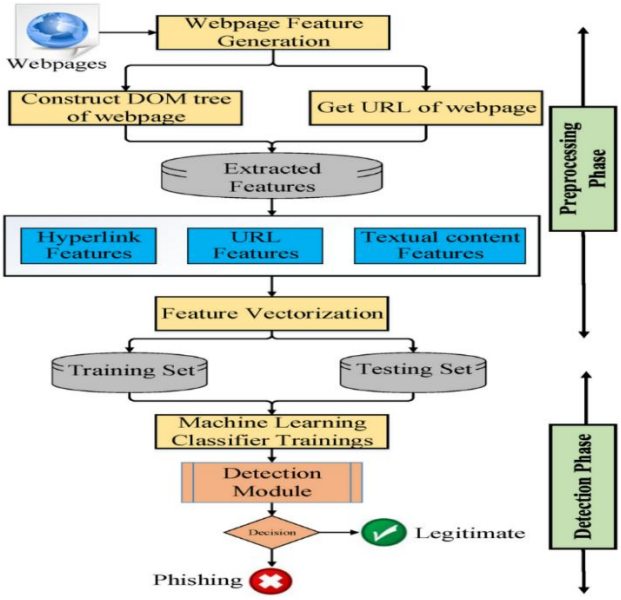

Figure 1

Fig 6.1: the system architecture outlines a robust pipeline for detecting phishing websites using neural networks. It integrates various layers to ensure data is collected, pre-processed, and analysed in an efficient manner, and employs a neural network to classify websites in real-time. The post-detection components allow for effective handling of detected phishing threats, while the feedback loop ensures continuous learning and adaptation to new phishing techniques.

VII. Implementation

The first step in the implementation is to compile a varied dataset from reliable sources, making sure it contains a variety of both authentic and phishing websites. Multiple features, including those based on domains, content, and URLs, can be extracted from this extensive dataset. These characteristics are essential for differentiating between authentic and phishing websites.

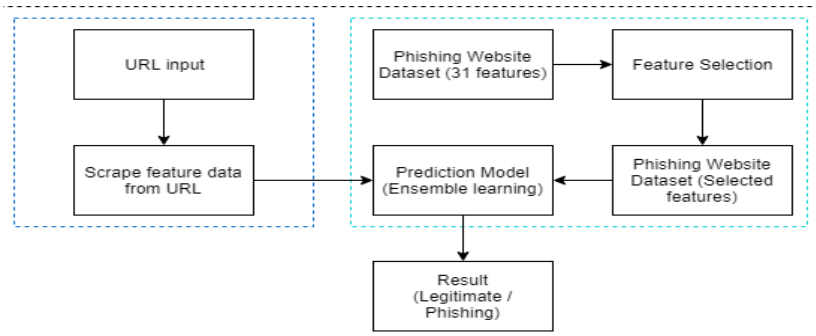

Figure 2

A. Users of the system

Registration and Authentication:
Users can register with their email address and password.
Users can log in and out safely.

Managing Roles
Role Assignment: Users' access permissions can be assigned by administrators to roles.
Accessibility to Features: Within the phishing detection system, varying roles grant access to particular features and functionalities.

Reports on Invoices
Reports on Invoice History: Users are able to create reports that include status updates and invoice history.

B. Managing Care of URLs

Through a protected portal, users will be able to submit URLs for phishing detection.
Machine learning models will be used by the system to analyse the submitted URLs.
Users will obtain comprehensive analysis results, encompassing possible hazards from phishing schemes.
Admins are able to assign responsibilities and rights for URL analysis, as well as manage user accounts.

C. Integration of Machine Learning

For URL analysis and phishing attempt detection, the system will make use of Gradient Boosting Classifier, Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs).
Users will have access to model performance indicators and forecast outcomes.
To increase accuracy, the system will enable ongoing model training and upgrades with fresh data.

D. Alerts and Reporting

Reports on detected phishing attempts, including URL details and detection confidence, will be generated by the system.
Email or in-app communications will be used to notify and inform users about phishing dangers. Reports can be exported in CSV and PDF formats for additional analysis.

VIII. RESULTS

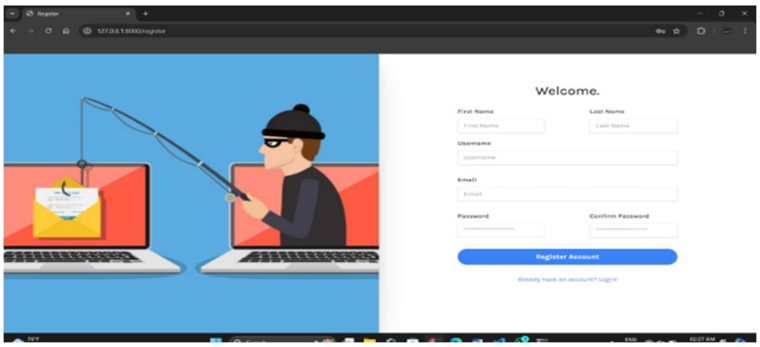

Figure 3 Registration form

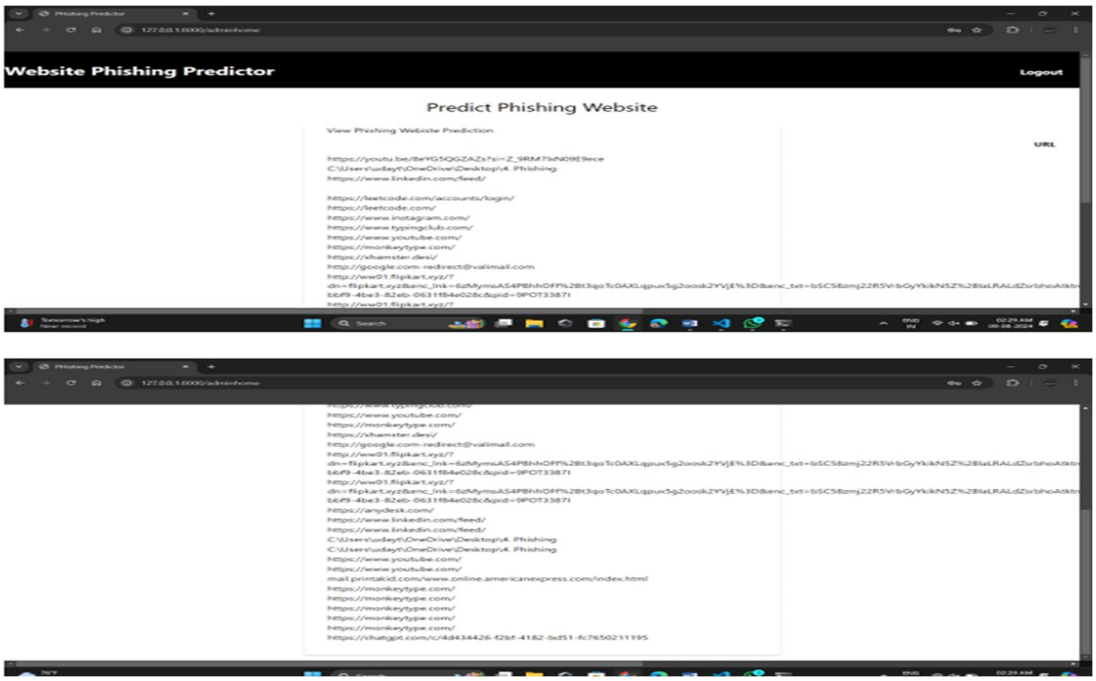

Fig 4: Admin page lists the phishing websites

IX. Conclusions

The phishing strategy was highlighted in the proposed study in the context of classification, where phishing websites are thought to automatically classify websites into a predefined range of class values based on a number of attributes and the class variable. The functionality of websites is used by ML-based phishing tactics to collect data that may be used to categorize websites and identify phishing sites. Although phishing cannot be completely eliminated, it can be lessened by addressing it in two ways: enhancing focused anti-phishing methods and strategies and educating the public about the ways in which phony phishing websites can be recognized and avoided. Machine learning techniques to counteract phishing assaults and strategies are crucial due to their constant evolution and complexity. The authors used the LSTM approach to distinguish between trustworthy and dangerous websites.

Future Enhancement

The goal of developing a safe, effective, and user-friendly payment processing solution was successfully met by the Payment Processing System (PPS) project. The group's dedication to security and innovation will not waver as they advance, guaranteeing that the PPS will continue to be a pioneer in payment processing technology. Crypto currencies: Including support for a range of crypto currencies will provide consumers more ways to make payments and keep up with new developments in the market. By utilizing machine learning algorithms to identify and stop fraudulent activity in real-time, organizations and individuals can lower their chance of suffering financial losses. By applying behavioural analytics, one can spot anomalous transaction patterns and alert authorities to possible fraud. Adapting the UI to mobile devices so that payments are made easily on all platforms. Adding tools that let users save their preferred methods of payment for a quicker checkout experience. Converting the system to an infrastructure based on the cloud in order to improve scalability and minimize downtime. Using load balancing strategies will guarantee that the system can accommodate higher transaction volumes without seeing a drop in performance. Investigating how block chain technology might improve transaction processing security and transparency. Creating tools to handle payments from Internet of Things (IoT) devices in order to support the expanding linked device market. Support System: To improve customer support and speed up issue resolution, AI-driven chatbots and support tickets are being added. Instructional Plans: Creating thorough training curricula for users and businesses to optimize PPS benefits. By concentrating on these upcoming improvements, the PPS will keep developing and adapting to the shifting demands of the financial sector, offering a safe, effective, and user-friendly payment processing solution.

REFERENCES

[1] Somesha, M., Pais, A. R., Rao, R. S., & Rathour, V. S. (2020). Efficient deep learning techniques for the detection of phishing websites. Sādhanā, 45, 1-18.

[2] Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J. P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. Electronics, 9(9), 1514.

[3] Feng, F., Zhou, Q., Shen, Z., Yang, X., Han, L., & Wang, J. (2024). The application of a novel neural network in the detection of phishing websites. Journal of Ambient Intelligence and Humanized Computing, 1-15.

[4] Yerima, S. Y., & Alzaylaee, M. K. (2020, March). High accuracy phishing detection based on convolutional neural networks. In 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS) (pp. 1-6). IEEE.

[5] Dutta, A. K. (2021). Detecting phishing websites using machine learning technique. PloS one, 16(10), e0258361.

[6] Siddiq, M. A. A., Arifuzzaman, M., & Islam, M. S. (2022, March). Phishing website detection using deep learning. In Proceedings of the 2nd International Conference on Computing Advancements (pp. 83-88).

[7] Vaitkevicius, P., & Marcinkevicius, V. (2020). Comparison of classification algorithms for detection of phishing websites. Informatica, 31(1), 143-160.