International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

AI-Driven Phishing Website Detection with Blockchain-Powered Secure Threat Logging

Mrs. S. Vijayalakshmi[1], B. Amulya[2], Ch. Lahari[3], A. Anjali[4]

[1]Assisstant Professor, Department of Computer Science and Engineering, KKR AND KSR INSTITUTE OF TECHNOLOGY AND SCIENCES (AUTONOMOUS), GUNTUR
[2]Student, Department of Computer Science and Engineering, KKR AND KSR INSTITUTE OF TECHNOLOGY AND SCIENCES (AUTONOMOUS), GUNTUR
[3]Student, Department of Computer Science and Engineering, KKR AND KSR INSTITUTE OF TECHNOLOGY AND SCIENCES (AUTONOMOUS), GUNTUR
[4]Student, Department of Computer Science and Engineering, KKR AND KSR INSTITUTE OF TECHNOLOGY AND SCIENCES (AUTONOMOUS), GUNTUR

Abstract - Using counterfeit websites, phishing assaults have grown a major cybersecurity menace since they use trust of consumers to grab valuable data. Blacklists, which cannot identify new or developing phishing sites in real time, are frequently the foundation of conventional phishing detection approaches. Although machine learning-based techniques have shown potential, obstacles stay in guaranteeing accurate, live detection and secure record keeping of recognized hazards. This project introduces an Advanced Phishing Website Detection System that uses blockchain technology and machine learning to provide a secure and fast answer to these issues. From URLs, the system first extracts important characteristics, preprocesses them, and then uses machine learning models Random Forest, Multi-Layer Perceptron (MLP), XGBoost, and Support Vector Machine (SVM) to classify websites as phishing or genuine. Using a Flask-based web application, the trained models have high accuracy since MLP is used for real-time detection. The MLP model parses and examines the characteristics of a URL provided by a user; if it is found to be phishing, the URL is stored in a blockchain ledger to assure tamper-proof logging of phishing sites. If not the case, the user has website access. Users can also see every identified phishing site on a particular page, therefore raising knowledge and proactive protection. The incorporation of blockchain technology improves transparency, security, and trust in the detection system and offers a powerful, live, and scalable strategy for combating phishing threats.

Key Words: AI, Blockchain, Detection, Machine, Phishing, Real-Time, Security, Website.

1.INTRODUCTION

Phishing attacks, which take use of human weaknesses to trick users into disclosing private information including banking information, login passwords and personal information have become one of the most prevalent enduring cybersecurity risks [1]. Phishing has its roots in the early 1990s, when hackers started posing as reliable organizations through email-based deception tactics. Phishing techniques have changed over time, embracing increasingly complex strategies like smishing (SMS phishing), vishing (sounds phishing), and spear phishing [2]. Because of the increased hazards brought about by the extensive use of online platforms and digital services, phishing has become a popular tactic used by cybercriminals to attack people, companies, and even governmental organizations. Phishing is still a big problem today, costing billions of dollars every year as a result of successful attacks.

Over time, a number of detecting techniques have been developed to counteract phishing. Earlier methods used blacklist-based screening, which involved keeping databases of known fake URLs and comparing them to incoming web requests [3]. A significant drawback of our approach was that it was unable to identify recently emerging and unreported phishing sites. Eventually, heuristic-based detection developed, using pre-established rules and patterns to find dubious URLs based on email content, linguistic analysis, and domain attributes. Artificial intelligence (AI) and machine learning have been used more recently to enhance phishing detection, enabling systems to examine enormous volumes of data and identify phishing attempts that were previously undetected. Even with these developments, real-time detection and safe preservation of phishing site information remain challenges for many current methods [4]. Heuristic techniques, crowdsourced phishing databases, and machine learning classifiers are the mainstays of the contemporary phishing detection systems. Even while AI has greatly increased the accuracy of detection, several problems still exist, such as false positives, sluggish reaction times, and a lack of safe storage options. Furthermore, phishing websites that regularly alter their domain names or structure tend to go undetected. The lack of an invulnerable to decentralized way to effectively store and distribute phishing site records is another significant flaw in current systems [5]. By regularly changing website properties, attackers might evade detection in the absence of a secure recording mechanism, making it more difficult to monitor and eventually reduce phishing threats.

The growing complexity of phishing assaults and the urgent requirement for a more dependable prevention and detection system are the driving forces behind this effort. Conventional approaches frequently fall behind the constantly evolving phishing tactics, leaving customers open to online theft. Furthermore, attackers have repeatedly exploited the absence of a clear and safe system for tracking phishing websites. In order to create a thorough, scalable, and resilient phishing prevention system, our project will combine

blockchain technology for secure threat logging with deep learning for real-time detection [6]. The ultimate objective is to raise knowledge of cybersecurity, offer a strong defense against phishing attempts, and guarantee a safer online experience for users everywhere.

2. LITERATURE SUREVY

In the past, a number of studies have attempted to identify phishing websites using various techniques, from rule-based methods to machine learning to deep learning models. Blacklist-based methods, in which known phishing URLs were kept in databases and matched to recently visited websites, were a major component of early phishing detection systems. This method's inability to identify zero-day phishing attacks newly developed phony websites that haven't been detected yet was one of its major shortcomings [7]. Heuristic-based detection techniques were developed to get over this restriction. These techniques classified websites by examining characteristics like URL structure, domains age, HTTP/HTTPS usage, and linguistic patterns. Several studies looked into signature-based detection, which used predetermined attack signatures to identify malicious websites. However, these techniques had trouble with dynamic phishing websites, which often altered their characteristics. In subsequent advancements, researchers used text-based analysis and Natural Language Processing, or NLP, techniques to identify phishing efforts in emails and webpages by examining grammatical errors, content inconsistencies, and misleading language patterns [8]. Even though these methods enhanced phishing detection, they were not flexible enough to adjust to changing attack tactics and continued to have large false positive rates.

Deep learning and machine learning models have gained prominence in phishing detection due to developments in artificial intelligence. In order to classify URLs using extracted data, researchers have used neural networks, random forests, decision trees, and support vector machines (SVM), with greater detection rates than conventional techniques. In order to identify phishing websites using visual and sequential patterns, more recent research has investigated deep learning architectures including Convolutional Neural Networks, and Recurrent Neural Networks (RNNs) [9]. Nevertheless, in spite of these developments, the majority of current systems lack a safe, decentralized mechanism for storing phishing information and do not provide real-time detection capabilities. Some researchers have tried to log phishing websites using centralized repositories or cloud-based networks however they are susceptible to single points of failure, illegal changes, and tampering. Although blockchain integration has recently drawn attention for guaranteeing the openness and integrity of threat to cybersecurity records, its use in phishing detection is yet relatively unexplored [10]. By fusing blockchain-based security logging with real-time machine learning-powered phishing detection, this study expands on previous attempts and provides a more reliable, transparent, and impenetrable solution.

3. PROPOSED METHODOLOGY

To guarantee real-time discovery and safe storage of phishing website data, the suggested approach (as shown in Fig.1) for detecting phishing sites combines machine learning-based classification with blockchain technology. Among several others, the system first pulls key characteristics from a provided URL domain age, HTTPS presence, subdomain count, special character usage, and URL length. These extracted characteristics including feature selection, normalization, and handling of missing values are preprocessed to improve model performance. Next, the preprocessed data is used to train several machine learning models Random Forest, Multi-Layer Perceptron (MLP), XGBoost, and Support Vector Machine (SVM) that classify the URL as legitimate or phishing. MLP is chosen for real-time detection given its outstanding results among these. Developed as a Flask-based web app, users may input URLs which are then analyzed by the MLP model in real time. A URL marked as phishing is not only blocked but also recorded on a blockchain ledger, providing a unchangeable and tamper-proof record of spotted phishing sites. The customer can normally visit the website if the URL is secure.

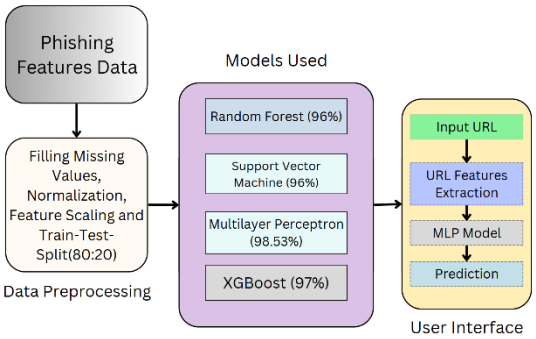


Fig.1 Proposed Methodology of System

The data movement within the system follows a defined pipeline to enable flawless execution. The user first enters the URL of a website, which is then parsed and examined in order to extract features. The trained MLP model then receives the extracted data and produces a classification result (legitimate or phishing). If the website is determined as phishing, the information it provides are hashed and saved on a blockchain ledger, prohibiting any unauthorized adjustments. The blockchain ledger keeps a distributed record of every phishing website that has been identified, giving users and cybersecurity professionals access to an open and verifiable list of dangers. To further raise awareness of cybersecurity, a special webpage is offered where users may browse all phishing websites that have been identified in the past.

A. Machine Learning Models

By allowing precise URL classification based on extracted data, machine learning is essential to the suggested phishing website detection system. Among the different models utilized, Random Forest is one of the most successful due to its capacity to manage complicated data structures and avoid overfitting. As an ensemble learning method, Random Forest builds several decision trees during learning and aggregates their predictions to improve accuracy [11]. The model is very tolerant to noisy data since it minimizes bias and variation by averaging the output of several trees. This resilience is especially helpful in phishing detection because hackers constantly alter website features to get around security. Features that are powerful predictors of phishing attempts, like URL length, HTTPS presence, number of subdomains, and unusual characters, are processed effectively by Random Forest. It is a good option for detecting phishing websites across many domains because of its ability to manage data that is both numerical and categorical.

The Multi-Layer Perceptron (MLP) [12], a kind of artificial brain network that is excellent at discovering intricate links within data, is another potent model utilized in this system. Because MLP is made up of several layers of linked neurons, it can identify non-linear patterns that more straightforward models would overlook. MLP is especially useful in phishing

detection because it can learn complex feature relationships, increasing the accuracy of classification. In contrast to conventional models, MLP uses hidden layers and backpropagation to improve its decision-making, enabling it to adjust to changing phishing tactics. MLP is used for continuous detection in the Flask-based website due to its robust performance. MLP analyzes the attributes that are retrieved when a user enters the URL of a website to evaluate whether it is a legitimate or phishing website.

Due to its excellent accuracy, scalability, and efficiency, XGBoost (Extreme Gradient Boosting) [13] is another machine learning algorithm used in this work. With regularization strategies that avoid overfitting, the gradient boosting algorithm XGBoost optimizes performance by building decision trees in an additive fashion. Its capacity to manage unbalanced datasets, which prevents the model from unfairly favoring trustworthy websites over phishing ones, is one of its main advantages in phishing detection. Because legal websites and phishing websites can have minor but misleading similarities, XGBoost's feature priority ranking aids in prioritizing the most important characteristics in classification. The model is a very dependable option for cybersecurity applications since it can identify concealed trends in phishing URLs.

Because of its outstanding performance in classification tasks especially when working with high-dimensional data the Support Vector Machine (SVM) model is also used [14]. Finding the appropriate hyperplane in an attribute space to distinguish between phishing and trustworthy websites is how SVM works. It is a useful complement to the collection of machine learning algorithms in this system since it is very good at handling non-linearly separable data. Because SVM optimizes the margin between classes, it is resilient to overfitting, which is one of its greatest advantages. SVM's capacity to establish a distinct division between categories improves the model's dependability in recognizing phishing attacks, where the distinction between authentic and fraudulent websites can be imperceptible.

Although each of these models created using machine learning makes a distinct contribution to phishing detection, the system is kept accurate, scalable, and impervious to fraudsters' evasion techniques. MLP is excellent at capturing intricate feature correlations, but Random Forest offers stability and interpretability. SVM guarantees ideal classification limits, while XGBoost improves detection by utilizing gradient boosting [15]. By integrating these models, the system successfully overcomes the drawbacks of conventional phishing detection techniques, greatly enhancing real-time response and classification accuracy. A more complete and flexible security system results from the usage of these several models, which also guarantee that even if one model has trouble with a certain phishing scam pattern, another can make up for it. In the end, combining these four potent machine learning models guarantees a multifaceted strategy for phishing detection that can tackle both established and new threats. The system offers a quick, efficient, and scalable way to detect and stop phishing attempts by utilizing feature engineering, preprocessing methods, and real-time inference. The detection system's integrity and credibility are further improved by the integration of blockchain technology for secure threat logging, which builds a strong cybersecurity foundation.

4. RESULTS

Several machine learning models, such as Random Forest, Multi-Layer Perceptron (MLP), XGBoost, and Support Vector Machine (SVM), were used to thoroughly assess the phishing website detection system. To ensure a comprehensive assessment of each model's efficacy, a dataset of extracted URL features was used for both training and testing. 97.5% for Random Forest, 98.53% for MLP, 97.6% for XGBoost, and 95.0% for SVM were the training accuracies, whereas 92.6% for Random Forest, 95.35% for MLP, 93.4% for XGBoost, and 97.8% for SVM were the testing accuracies (as shown in Fig.2 & 3). According to these findings, MLP and SVM fared better than the other models, exhibiting stronger generalization skills and resilience to phishing attempts. MLP was chosen for real-time analysis in the deployed system due to its excellent accuracy and flexibility with regard to unseen data.

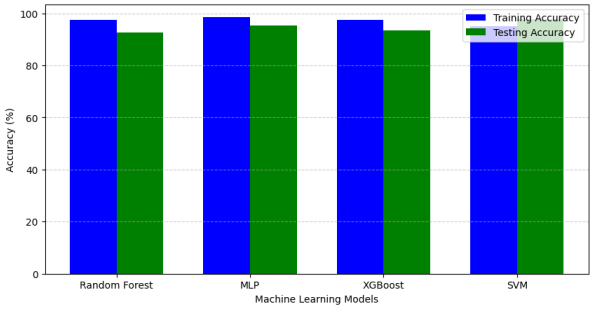


Fig.2 Model Training vs Testing Accuracy

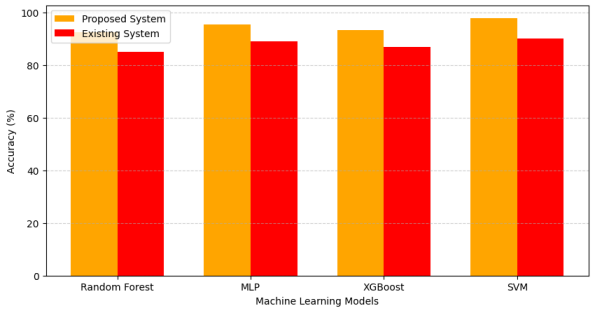


Fig.3 Comparison of Accuracy with Existing System

Flask, an efficient web framework that facilitates smooth communication between the user and the model of machine learning, was used to deploy the system during the real-time detection phase. Key data like domain age, HTTPS presence, number of subdivisions, and special character usage are instantly extracted by the system when a user inputs a URL. The trained MLP model receives these preprocessed features and determines if the webpage is phishing or authentic. The technology warns the user and blocks access if the website is identified as phishing, reducing the risk of cyberattacks. In order to minimize categorization and response delays, the real-time detection method is made to be quick and effective. The blockchain-based phishing website storage method is one of the system's main improvements. When a URL is flagged as phishing, the information is entered into a blockchain ledger, guaranteeing transparency, security, and immutability. This method keeps hackers from altering or erasing phishing records, which makes it a useful tool for both consumers and cybersecurity experts. Users can also obtain a continuously updated database of harmful URLs by visiting a dedicated page that shows all stored phishing websites. By keeping consumers updated on new phishing dangers, this tool enables them to make safer online choices.

Fig.4 Real Time URL Phishing Detections

Fig.5 Real Time URL Normal Detections

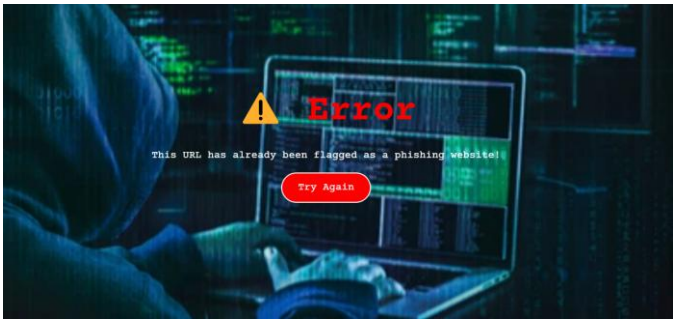
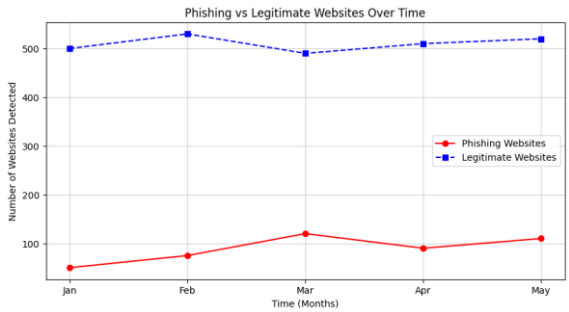


Fig. 6 Block Chain Ledger Data Store

A range of phishing and trustworthy websites, including unknown phishing domains, were used to test the system's real-time performance. Newly developed phishing websites that weren't in the training dataset were successfully identified and blocked by the algorithm. This demonstrates how flexible the MLP model is, since it can recognize intricate patterns in URL structures and identify phishing websites by their tiny differences (as shown in Fig.5 & 6). This machine learning-based method successfully detects zero-day phishing assaults by identifying novel harmful patterns, in contrast to conventional rule-driven detection systems that depend on blacklists. In order to assess the system's performance under high loads, stress testing was also performed. The Flask-based framework processed real-time queries with low latency when processing many URLs at once (as shown in Fig.6), guaranteeing that consumers received timely results. Phishing classification was found to have an average response time of less than one second, which makes it extremely useful in real-world applications. The system demonstrated the capacity for large-scale recognition of phishing applications by maintaining its speed and accuracy even when it was installed on a cloud server.



The system's intuitive interface, which is made for smooth interaction, is another important feature. With its straightforward yet user-friendly interface, the Flask application lets users enter URLs, browse previously identified phishing websites, and get real-time results. By guaranteeing that a phishing website cannot be changed or deleted once it has been registered, the blockchain integration improves accountability and trust. In addition to offering a proactive protection mechanism, this system teaches users how to identify typical attack patterns and comprehend phishing threats. The dynamic character of phishing attempts was a crucial finding from real-time analysis. To avoid discovery, attackers usually change SSL certificates, domain names, and website structures. Nonetheless, the learning-based methodology and feature extraction of the suggested system guarantee ongoing adaptation to such strategies. The system sustains its high detection accuracy over time by periodically retraining the MLP model with new data and continuously adding new phishing websites to the blockchain ledger.

CONCLUSION

The Advanced Phishing Site Detection System offers a reliable and effective defense against phishing assaults by effectively combining blockchain technology, real-time analysis, and machine learning. The solution guarantees high precision, real-time detection, and improved security for users by utilizing blockchain for immutable phishing data, Flask for deployment, and MLP for classification. The technology is more effective than conventional rule-based detection techniques since it employs a feature extraction-based methodology to detect zero-day phishing attacks. Incorporating a publicly available blockchain database not only stops phishing data from being manipulated, but it also gives consumers and cybersecurity experts a useful tool to keep up with new threats. Users are immediately protected from fraudulent websites thanks to the real-time analysis's quick reaction times, low latency, and flexibility in responding to changing phishing tactics. The system's success demonstrates its promise as a proactive, scalable cybersecurity approach that can be improved going forward. In subsequent research, we intend to enhance feature extraction methods, add deep learning models like LSTMs and transformers for more sophisticated phishing detection, and broaden the system to identify phishing messages in emails and social media networks.

REFERENCES

1. 1. Ghelani, Diptiben. "Cyber security, cyber threats, implications and future perspectives: A Review." *Authorea Preprints* (2022).
2. Im-erb, Kanika. "Understanding and Mitigating Phishing Attacks in the Digital Era."
3. Díaz-Verdejo, Jesús E., et al. "A critical review of the techniques used for anomaly detection of HTTP-based attacks: taxonomy, limitations and open challenges." *Computers & Security* 124 (2023): 102997.
4. Do, Nguyet Quang, et al. "Deep learning for phishing detection: Taxonomy, current challenges and future directions." *Ieee Access* 10 (2022): 36429-36463.
5. Alkhalil, Zainab, et al. "Phishing attacks: A recent comprehensive study and a new anatomy." *Frontiers in Computer Science* 3 (2021): 563060.
6. Lagos, Leonel, et al. *Secure Data Logging and Processing with Blockchain and Machine Learning*. No. DE-FE0031745. Florida International Univ.(FIU), Miami, FL (United States), 2023.
7. Chrysanthou, Anargyros, Yorgos Pantis, and Constantinos Patsakis. "The anatomy of deception: Technical and human perspectives on a large-scale phishing campaign." *arXiv preprint arXiv:2310.03498* (2023).
8. Damaiyanti, Susi. "Grammatical errors made by students in speaking English." *Journal of English Language Teaching and Learning (JETLE)* 2.2 (2021): 2-3.
9. Ghosh, Rajib, and Anupam Kumar. "A hybrid deep learning model by combining convolutional neural network and recurrent neural network to detect forest fire." *Multimedia Tools and Applications* 81.27 (2022): 38643-38660.
10. Boddapati, Mohan Sai Dinesh, et al. "Creating a protected virtual learning space: a comprehensive strategy for security and user experience in online education." *International Conference on Cognitive Computing and Cyber Physical Systems*. Cham: Springer Nature Switzerland, 2023.
11. Palimkar, Prajyot, Rabindra Nath Shaw, and Ankush Ghosh. "Machine learning technique to prognosis diabetes disease: Random forest classifier approach." *Advanced computing and intelligent technologies: proceedings of ICACIT 2021*. Singapore: Springer Singapore, 2021. 219-244.
12. Sathish, T., et al. "Characteristics estimation of natural fibre reinforced plastic composites using deep multi-layer perceptron (MLP) technique." *Chemosphere* 337 (2023): 139346.
13. Ali, Zeravan Arif, et al. "eXtreme gradient boosting algorithm with machine learning: A review." *Academic Journal of Nawroz University* 12.2 (2023): 320-334.
14. Kurani, Akshit, et al. "A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting." *Annals of Data Science* 10.1 (2023): 183-208.
15. Huber, Florian, et al. "Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches." *Computers and Electronics in Agriculture* 202 (2022): 107346.