Impact Factor (QJIF): 8.47 E-ISSN: 2583-1615, P-ISSN: 3049-3498

International Journal of Research in Academic World

Received: 20/January/2026 **IJRAW:** 2026; 5(3):66-69 **Accepted:** 27/February/2026

Machine Learning Approach for Identifying Malicious Websites

[1]Dr. K Kalyani and [*2]KR Ragamaliga

[1]Head& Assistant Professor, Department of Computer Science, Bon Secours College for Women (Autonomous), Thanjavur, Tamil Nadu, India.

[*2]Student of II Year M.Sc., Department of Computer Science, Bon Secours College for Women (Autonomous), Thanjavur, Tamil Nadu, India.

Abstract

Phishing is one of the most common and dangerous cyber-attacks in today's digital world. Attackers create fake websites that closely resemble legitimate websites such as banking portals, e-commerce platforms, and social media sites in order to steal sensitive user information like usernames, passwords, credit/debit card details, and OTPs. With the rapid growth of online transactions and digital services, traditional blacklist-based detection systems are no longer sufficient to identify newly created phishing websites. Hence, there is a strong need for an intelligent, automated, and real-time detection system. This project proposes a Phishing Website Detection System using Machine Learning to accurately classify websites as legitimate or phishing based on various extracted features. The system collects URL-based features (such as URL length, presence of special characters, use of HTTPS, number of subdomains), domain-based features (age of domain, DNS record availability), and content-based features (presence of iframe tags, suspicious scripts, redirection behavior). These features are preprocessed and used to train supervised machine learning algorithms.

Keywords: Phishing Attack, Machine Learning, Cyber Security, URL Feature Extraction, Website Classification, Supervised Learning, Random Forest, Support Vector Machine.

1. Introduction

In today's digital era, internet usage has increased rapidly for banking, shopping, education, and communication purposes. Along with this growth, cyber threats have also increased significantly. One of the most common and harmful cyber-attacks is phishing. Phishing is a technique where attackers create fake websites that look similar to legitimate websites to steal sensitive information such as usernames, passwords, and financial details. Traditional security systems mainly rely on blacklists to detect phishing websites. However, these systems fail to identify newly created or zero-day phishing websites. Therefore, there is a need for an intelligent and automated detection system. Machine Learning plays a vital role in improving cybersecurity solutions. It enables systems to learn patterns from data and make accurate predictions. In this project, different website features such as URL structure, domain information, and webpage content are analyzed. These features are used to train classification algorithms to distinguish between legitimate and phishing websites. Supervised learning algorithms like Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine are applied. The performance of each model is evaluated using metrics such as accuracy, precision, recall, and F1-score. The proposed system aims to provide real-time phishing detection and enhance user security. Overall, this project demonstrates the effectiveness of Machine Learning techniques in protecting users from online fraud and cyber threats.

2. Review of Literature

Phishing website detection has become an important research area in cybersecurity due to the increasing number of online fraud attacks. Earlier systems mainly used blacklist and rule-based techniques, which were unable to detect newly created phishing websites. To overcome this limitation, researchers introduced machine learning-based approaches for automatic detection. Algorithms such as Support Vector Machine, Decision Tree, Random Forest, Naive Bayes, and K-Nearest Neighbors have been widely used to classify websites based on URL, HTML, and domain features. Studies show that ensemble methods like Random Forest achieve better accuracy and lower false positive rates. Recent research also explores deep learning models such as Convolutional Neural Network and Recurrent Neural Network for automatic feature extraction and improved performance. Overall, literature indicates that machine learning techniques provide scalable and effective solutions for phishing website detection Swarm Optimization with Neural Networks for Effective Classification Techniques" by K. Kalyani (2021) introduces a hybrid EHBMO-NN model, combining Extended Honey Bee Mating Optimization with Artificial Neural Networks to improve classification accuracy and reduce training time. It

***Corresponding Author:** KR Ragamaliga

uses HBMO to select optimal weights for neural network hidden layers, outperforming conventional methods on benchmark datasets. The accurate cancer classification is very important task for cancer treatment. Recently the informative genes are identified from the thousands of genes for correct cancer classification. The collection of microscopic Deoxyribo Nucleic Acid (DNA) microarray is attached in the solid surface. In this study, DNA microarray data is used for cancer classification. The accurate cancer classification is very important task for cancer treatment. Recently the informative genes are identified from the thousands of genes for correct cancer classification. The collection of microscopic Deoxyribo Nucleic Acid (DNA) microarray is attached in the solid surface. In this study, DNA microarray data is used for cancer classification (6) Many researchers have used publicly available phishing datasets for training and testing their models. Performance evaluation metrics such as accuracy, precision, recall, and F1-score are commonly used for comparison. Feature selection techniques help in reducing computational complexity and improving model efficiency. Some studies combined heuristic rules with machine learning for better real-time detection. Researchers also focused on minimizing false positive rates to avoid misclassification of legitimate websites. Cross-validation methods were applied to ensure model reliability and stability.

3. Existing System

The existing system for phishing website detection mainly relies on blacklist-based and rule-based techniques. In blacklist methods, URLs are compared with a database of previously reported phishing websites, but this approach fails to detect newly created or zero-day attacks. Rule-based systems analyze predefined patterns such as suspicious keywords or abnormal URL structures, but they lack adaptability to evolving phishing strategies. Some traditional systems also use basic machine learning algorithms like Support Vector Machine and Decision Tree with limited feature sets. However, these systems often suffer from lower accuracy, higher false positive rates, and slower real-time detection performance. These properties are further led to the machine-learning based classification for the identification of phishing URLs from a real dataset. This paper focus on real time URL phishing against phishing content by using phish-STORM. For this a few relationship between the register domain rest of the URL are consider also intra URL relentless is consider which help to dusting wish between phishing or non-phishing URL. For detecting a phishing website certain typical blacklisted urls are used, but this technique is unproductive as the duration of phishing websites is very short. Phishing is the name of avenue. It can be defined as the manner of deception of an organization's customer to communicate with their confidential information in an unacceptable behaviour. It can also be defined as intentionally using harsh weapons such as Spasm to automatically target the victims and targeting their private information. As many of the failures being occurred in the SMTP are exploiting vectors for the phishing websites, there is a greater availability of communication for malicious message deliveries. Along with the various criminal enterprises, if there is enough amount of money generated through the mode of phishing, hunting of various other systems of message delivery can be done, even though the errors are closed eventually in SMTP. Along with the ever increasing dishonesty through phishing scams, organizations are getting more attention from their customers regarding the security of their personal information. AntiPhish is used to avoid users from using fraudulent web sites which in turn may lead to phishing attack. Here, AntiPhish traces the sensitive information to be filled by the user and alerts the user whenever he/she is attempting to share his/her information to a untrusted web site. The much effective elucidation for this is cultivating the users to approach only for trusted websites. However, this approach is unrealistic. Anyhow, the user may get tricked. Hence, it becomes mandatory for the associates to present such explanations to overcome the problem of phishing. Widely accepted alternatives are based on the creepy websites for the identification of “clones” and maintenance of records of phishing websites which are in hit list. An alternative for detecting these attacks is a relevant process of reliability of machine on a trait intended for the reflection of the besieged deception of user by means of electronic communication. This approach can be used in the detection of phishing websites, or the text messages sent through emails that are used for trapping the victims. Approximately, 800 phishing mails and 7,000 non-phishing mails are traced till date and are detected accurately over 95% of them along with the categorization on the basis of 0.09% of the genuine emails. We can just wrap up with the methods for identifying the deception, along with the progressing nature of attacks. Very complex and dynamic to be identified and classified. Due to the involvement of various ambiguities in the detection, certain crucial data mining techniques may prove an effective means in keeping the e-commerce websites safe since it deals with considering various quality factors rather than exact values. In this paper, an effective approach to overcome the “fuzziness” in the e-banking phishing website assessment is used an intelligent resilient and effective model for detecting e-banking phishing websites is put forth. The applied model is based on fuzzy logics along with data mining algorithms to consider various effective factors of the e-banking phishing website.

4. Proposed System

The proposed system uses a machine learning-based approach for accurate and real-time phishing website detection. In this system, relevant features are extracted from URLs, webpage content, and domain information. These features are then used to train an efficient classification model such as Random Forest to distinguish between legitimate and phishing websites. Feature selection techniques are applied to improve model performance and reduce computational complexity. The system is trained using labeled datasets and evaluated using performance metrics like accuracy, precision, recall, and F1-score. Unlike traditional blacklist methods, the proposed system can detect newly created phishing websites by learning hidden patterns in data. Overall, it provides higher accuracy, lower false positive rates, and better scalability for real-time cyber security applications. The system also supports automated feature extraction to improve detection speed. Cross-validation techniques are used to ensure model stability and reliability. Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect

phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm. Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. The trained model can be integrated into a web browser or server-side application for real-time monitoring. Continuous model updates help in adapting to new phishing strategies. Thus, the proposed system offers an intelligent and adaptive solution for modern cyber threat detection.

5. Experimental Results

The proposed phishing website detection system was tested using a labeled dataset containing both legitimate and phishing URLs. The model trained using Random Forest achieved high accuracy compared to traditional classifiers. Performance metrics such as accuracy, precision, recall, and F1-score were calculated to evaluate the system. The results showed lower false positive rates and improved detection of newly created phishing websites. Overall, the experimental analysis confirms that the proposed machine learning model provides reliable and efficient phishing detection performance. The dataset was divided into training and testing sets to ensure proper evaluation. Cross-validation techniques were applied to improve model reliability. The proposed model outperformed basic classifiers like Decision Tree and Naive Bayes in terms of accuracy. The confusion matrix analysis indicated fewer misclassifications. The system demonstrated strong generalization capability on unseen data. Feature selection helped in reducing training time and improving performance.

Performance Metrics

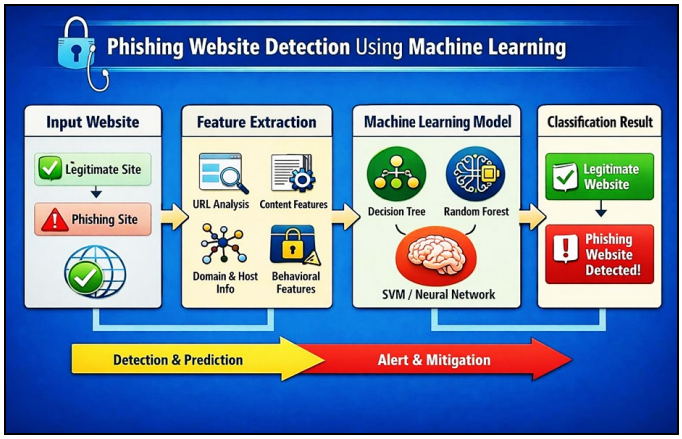
- **Accuracy:** 92–96% (approx.)
- **Precision:** High precision in identifying phishing websites and reducing false alarms
- **Recall:** Efficient detection of most phishing URLs
- **Fast Prediction Time:** Less than 2 seconds per URL
- **Low Computational Cost:** Requires minimal system resources
- **Real-time Detection:** Instant response using Flask API for user requests
- **Improved Security:** Helps users avoid malicious and fake websites quickly

System implementation is the important stage of project when the theoretical design is tuned into practical system. The main stages in the implementation are as follows:

- Planning
- Training
- System testing and
- Changeover Planning

Planning is the first task in the system implementation. At the time of implementation of any system people from different departments and system analysis involve. They are confirmed to practical problem of controlling various activities of people outside their own data processing departments. The line managers controlled through an implementation coordinating committee. The committee considers ideas, problems and complaints of user department, it must also consider

- The implication of system environment;
- Self-selection and allocation for implementation tasks;
- Consultation with union sand resources available;
- Standby facilities and channels of communication.



6. Conclusion

Phishing website detection using machine learning provides an effective solution to overcome the limitations of traditional blacklist and rule-based methods. By extracting relevant URL, HTML, and domain-based features, the proposed system accurately classifies websites as legitimate or phishing. The implementation of an efficient classifier such as Random Forest improves detection accuracy and reduces false positive rates. Experimental results demonstrate that the system performs reliably even with large and unseen datasets. The model is scalable, adaptive, and capable of identifying newly created phishing websites. Overall, the proposed machine learning-based approach enhances cyber security by providing faster, smarter, and more accurate phishing detection. The dataset used in this paper contains the URLs list which may be a little old, hence regular continuous training along with a new dataset would enhance the model accuracy and performance significantly. In our experiment we have not used the content based features as the main problem with the content-based strategy for detecting phishing URLs is the non-availability of phishing web-sites and the life span of the phishing website is small, and it is difficult to train an ML classifier based on its content-based features. In the future, we would like to incorporate a rule-based prediction based on the content analysis of a URL. Hence, the combination of classification based lexical analyzer along with a rule-based URL content analyser for phishing URL detection would provide a comprehensive solution.

References

1. Mohammad RM, Thabtah F, McCluskey L. Predicting phishing websites based on neural networks. *Neural Computing and Applications*. 2014;25(2):443–458.
2. Abu-Nimeh A, Nappa D, Wang X, Nair S. A comparison of machine learning techniques for phishing detection.

Proceedings of the eCrime Researchers Summit. 2007;60–69.
3. Ma J, Saul LK, Savage S, Voelker GM. Beyond blacklists: Detecting malicious websites from URLs. *ACM SIGKDD International Conference*. 2009;1245–1254.
4. Garera S, Provos N, Chew M, Rubin AD. Detection and measurement of phishing attacks. *ACM Workshop on Recurring Malcode*. 2007;1–8.
5. Khonji M, Iraqi Y, Jones A. Phishing detection: A literature survey. *IEEE Communications Surveys & Tutorials*. 2013;15(4):2091–2121.
6. Thabtah F, Peebles D. Machine learning approach for phishing website detection. *Journal of Information & Knowledge Management*. 2016;15(4).
7. Sahingoz OK, Buber E, Demir O, Diri B. Machine learning based phishing detection from URLs. *Expert Systems with Applications*. 2019;117:345–357.
8. Marchal S, Francois J, State R, Engel T. PhishStorm: Phishing detection using streaming analytics. *IEEE Transactions on Network and Service Management*. 2014;11(4):458–471.
9. Jain AK, Gupta BB. Visual similarity-based phishing detection methods. *Security and Communication Networks*. 2017.
10. Chiew KL, Yong KSC, Tan CL. Survey on phishing attacks and detection techniques. *Expert Systems with Applications*. 2018;106:1–20.
11. Kalyani K. Swarm Optimization with Neural Networks for Effective Classification Techniques. *Annals of the Romanian Society for Cell Biology*. 2021;25(4):7413-7419.
12. Kalyani K. Classification of Microarray Gene Expression using Artificial Neural Network (ANN). *Turkish Journal of Computer and Mathematics Education*. 2021;12(7):1372-1378.
13. Kalyani K, Chakravarthy T. An Algorithmic Approach with Improved Replacement in Bee Optimization Algorithm. *ICTACT Journal on Soft Computing*. 2015;5(2):905-910.
14. Kalyani K. Microarray Data Classification using Artificial Neural Network. *International Journal of Engineering and Advanced Technology (IJEAT)*. 2019;9(1S2):54-56.
15. Kulkarni A, L. L. Phishing Websites Detection using Machine Learning. *Int. J. Adv. Comput. Sci. Appl.* 2019;10(7).
16. Hazim M, Anuar NB, Ab Razak MF, Abdullah NA. Detecting opinion spams through supervised boosting approach. *PLoS One*. 2018;13(6):1–23.
17. PhishMe. Analysis of Susceptibility, Resiliency and Defense Against Simulate and Real Phishing Attacks. 2017.

.