International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 14; Issue: 8; August- 2025

# DETECTION OF PHISHING WEBSITE USING MACHINE LEARNING

**Madhusha M**
PG, Student
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli,
Bengaluru- 560068
madhushammca2025@gmail.com

**Sujitha R**
Assistant Professor
Dept. of MCA
The Oxford College of Engineering,
Bommanahalli, Bengaluru- 560068
sujir5416@gmail.com

## ABSTRACT

In today's digital era, cloud computing has become the backbone of modern IT infrastructure, offering scalability, flexibility, and cost-efficiency. However, the widespread adoption of cloud services has introduced new challenges in securing user access, managing identities, and preventing unauthorized breaches. This project proposes a robust security framework that integrates **Multi-Factor Authentication (MFA)** with **Identity Governance** to ensure **secure cloud access management**. The core objective of this system is to strengthen access control by requiring users to verify their identity through multiple authentication factors such as passwords, biometrics, and one-time codes. Additionally, the identity governance layer ensures that access rights are granted based on predefined roles, policies, and user behavior analytics. The system continuously monitors user activity, enforces least privilege principles, and automates access provisioning and de-provisioning. By combining MFA with intelligent identity governance, the proposed solution mitigates the risk of data breaches, insider threats, and account compromises, ensuring that only authorized users gain access to cloud resources.

**KEYWORDS:** *Cloud Security, Access Management, Multi-Factor Authentication (MFA), Identity Governance*

## INTRODUCTION

With the rapid expansion of internet usage and digital services, phishing attacks have become one of the most prevalent cybersecurity threats. Phishing involves fraudulent attempts to acquire sensitive information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity through fake websites. Traditional rule-based approaches to detect such malicious websites are often ineffective due to the constantly evolving tactics employed by attackers.

To combat this issue, the designed system makes use of ML techniques to build an intelligent and automated Phishing Website Detection System. This system uses a publicly available dataset consisting of 11,054 website URLs, each labeled as either phishing or legitimate. Each URL is represented by 30 numerical features that capture various properties such as URL length, domain-related attributes, HTTPS presence, and embedded script behaviors.

## LITERATURE SURVEY

In the field of cybersecurity, phishing websites represent one of the most common threats, where attackers impersonate legitimate websites to steal personal information such as login credentials, credit card details, and other sensitive data. Various methods have been explored to detect phishing websites, including heuristic-based approaches, ML algorithms, and hybrid models that combine multiple techniques. Below are some notable works in this area:

1. Koh et al. (2010) - "Phishing Detection: A Literature Survey." This paper presents a comprehensive review of phishing detection techniques, discussing both traditional rule-based methods and ML approaches. The authors highlight the use of feature extraction techniques, such as URL analysis, content-based analysis, and domain name analysis. The study also emphasizes the role of ML algorithms in improving detection accuracy.
2. Xie et al. (2011) - "Phishing Webpage Detection using Multiple Classifiers." This research explored the use of multiple classifiers, including decision trees and support vector machines (SVM), to identify phishing websites. The authors used a feature set derived from the structure of the website's URL, the presence of domain name anomalies, and other content-based characteristics. The hybrid classifier model demonstrated outperforming single classifiers.
3. Moore et al. (2009) - "Improved Detection of Phishing Websites Using ML" Moore and colleagues proposed a ML-based method for phishing detection that utilized both structural and behavioral features of websites. The paper compared several classification techniques, including logistic regression, SVM, and decision trees, and identified Random Forest as one of the most accurate models for phishing detection due to its ability to handle large feature sets and provide robust classification results.

## EXISTING WORK

Traditional phishing detection systems largely rely on blacklist-based approaches, where known malicious URLs are stored in a database and incoming web requests are compared against this list. This can be seen in Safe Browsing API used by Google and Microsoft SmartScreen. To a certain degree effective, these systems on the other hand do not have the capability to detect newly launched phishing websites., since they depend on existing reporting and updates. Other rule-based solutions scan URL patterns

and web pages manually specified by cybersecurity professionals, however, such approaches are somehow lacking in flexibility and scalability to the changing phishing tactics. In addition to that, some web browsers have implemented phishing protection internally, although they rely strongly on centralized databases resulting in a lack of real-time detection and flexibility.

## PROPOSED SYSTEM

The proposed approach takes advantage of a combination of such features in order to precisely and effectively recognize phishing websites. In contrast to conventional blacklists / rule-based engines, the model is trained on a large, multi-dimensional dataset of >11,000 web entries that have 30+ features. Such characteristics include: URL features, domain-based features and technical features which are used to detect phishing activities.

The inner system includes the training of the classifiers like SVM and MLP by using the labeled data to differentiate the legitimate URLs with the phishing one. Having compared the models, the classifier that yields the best results (which in this case is MLP) is stored and included in a Flask web app that a user can use to enter any URL. The system then processes such a URL and gives real-time classification as to whether or not it is safe or not.

The method can detect phishing attempts more quickly, at scale, and can be more dynamic to capture URLs that had only minor changes or newly created ones that other methods would miss.

## METHODOLOGY

The process of identifying phishing websites will commence by having a reliable source of data comprising of both genuine and para-sitical websites. Information is usually retrieved using reputable sources like PhishTank, OpenPhish and Alexa top sites to achieve diversification and authenticity. After obtaining them, the next step will be the reduction of the extracted data on the basis of which human-readable features can be selected. Such features may be obtained based on the URL structure of the site (e.g., its length, or the presence of special characters, or suspicious keywords), information involving the domain (e.g., age, DNS records, and reputation), and page content (such as the presence of SSL certificates, form requesting sensitive data, or use of hidden elements).

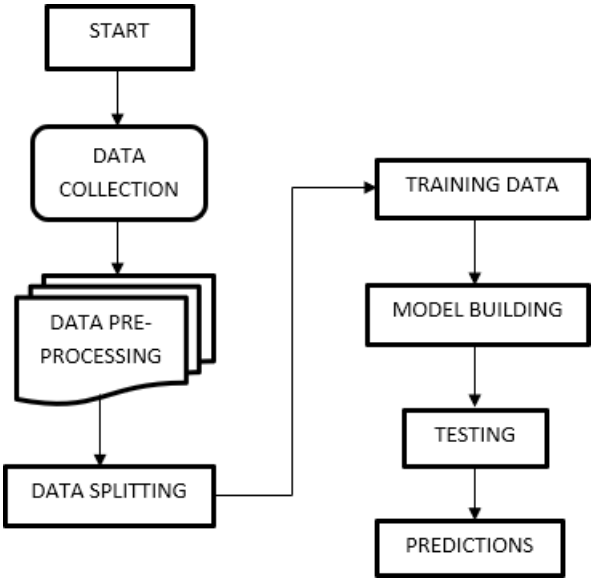


Fig 1. Block diagram of Phishing website

The presence of indicators such as frequent redirects or mismatched links are also solid indicators that the behaviour is part of a phishing attempt. The result of the feature extraction goes through a preprocessing phase, where inconsistencies may have to be removed, values need to be normalized, etc. Machine learning and deep learning algorithms like Random Forest, SVM, or neural networks then are used to classify the websites as phishing/clean. Evaluation of the models involves performance measurement through metrics such as accuracy, precision, recall and F1-score to make the models reliable. Having been successfully verified, the detection system can be implemented in the form of a browser extension, proxy filter, or cloud-based API used to provide protection on the fly. System effectiveness must be ensured by its continued updating with new patterns in phishing and its repeated re-training, so that it can be adjusted to the changing nature of cyber threats.

## EXPERIMENTAL RESULTS

The designed phishing website detection system was evaluated on some benchmark dataset; the dataset included both the phishing URL and legitimate URL. Post the pre-processing and feature extraction the dataset was split between the training and testing sets in the ratio of 70:30. Several machine learning algorithms such as Decision Tree, Random Forest, and Support Vector Machine were applied to test the effectiveness of the system. Among them, Random Forest had the highest accuracy rate, and it performs well in identifying phishing websites as opposed to benign websites.

The results were quantified on standard evaluation terms involving accuracy, precision, recall and F1-score. Its findings were that the model achieved a high detection accuracy, with balanced accuracy and recall, that is low false positive rate, indicating that the model was able to correctly identify phishing attempts, and with low incorrectly identified images. The area under the ROC-AUC score also supported the robustness of the model in classifications.

Besides the numerical assessment the system was experimented in a simulated environment that analyzed real time URLs. Automatic analysis of URL structure, domain reputation, and checking of the legitimacy of SSL certificates proved to be effective to identified.


Fig 2. Homepage

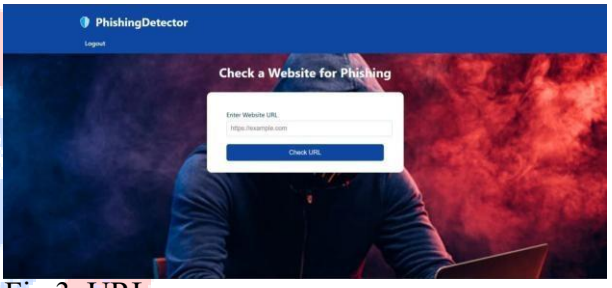


Fig 3. URL page

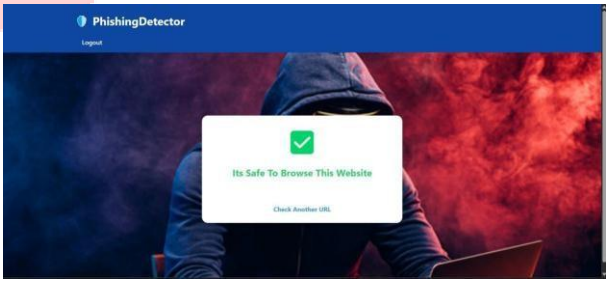


Fig 4. Result Page

## CONCLUSION

The ML is another security device in that the phishing Website Detection System is not only a security method, but also, a teaching tool since it keeps people informed on existing security threats and trends. Compared to the normal rule based method of detection which are always fixed and sometimes ineffective thus failing to thwart the new emerging tricks in the trade of phishing, the ML system brings a sense of flexibility and swift intelligence in detection. It is also free of serious attacks since the system learns on the basis of previous findings and constantly improves on its understanding base.

Its versatility and how such a system is scalable are other merits of such a system. It can be implemented as stand alone software, as an addition to browser, as well as into the security systems of large enterprises. This flexibility means that many individuals can use the solution- both the single internet extract that needs protection against phishing scams, and the large organizations that have several employees who need to be secured against phishing.

In addition to that, the system is quite usability and accessibility-oriented. The interface is simple to use and understand which could allow non-technical people to clear the safety of sites and does not need any knowledge of cybersecurity.

## REFERENCES

1. Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., & Nunamaker Jr, J. F. (2010). Detecting Fake Websites: The Contribution of Statistical Learning Theory. *MIS Quarterly*, 34(3), 435–461.
2. Jain, A. K., & Gupta, B. B. (2018). Phishing Detection: Analysis of Visual Similarity Based Approaches. *Security and Privacy*, 1(2), e9.
3. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting Phishing Websites Based on Self-Structuring Neural Network. *Neural Computing and Applications*, 25, 443–458.
4. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1245–1254.