# DETECTION OF PHISHING WEBSITE USING MACHINE LEARNING

*Ms.J.Maheswari[1], Ms.J.Dhivya[2], Ms.S.M.Amsaveni[3], Ms.S.Gayathri[4]*

*[1]Assistant Professor, Department of Computer Science & Engineering, Dhirajlal Gandhi College of Technology, Salem, Tamilnadu , India*

*[2,3,4]UG Scholar, Department of Computer Science & Engineering, Dhirajlal Gandhi College of Technology, Salem, Tamilnadu , India*

---

**ABSTRACT -** Phishing is a fraudulent practice in which scammers impersonate legitimate individuals or organizations to obtain sensitive information from unsuspecting victims. This can occur via various communication channels such as email, text messages, or social media. Phishing is a popular tactic among cybercriminals because it is easier to persuade someone to click a malicious link that appears to be authentic than it is to bypass computer security measures. One common approach used in phishing attacks is to create fake websites and emails that closely resemble legitimate ones. This is achieved by using logos, slogans, and other elements that are typically associated with the targeted organization or individual. When users click on the links provided in these emails, they are directed to fake websites where they are asked to provide sensitive information such as login credentials, bank account details, or other personal information. To detect phishing websites, machine learning models such as Random Forest, Decision Tree, and Multilayer Perceptron are utilized and compared for their accuracy and efficiency. These models analyze various features of the website URL to determine if it is a phishing website or not. Features such as the length of the URL, the presence of certain keywords, and the type of top-level domain are taken into account when analyzing the website. The existing machine learning models used to detect phishing websites have some limitations. For example, they have low latency and lack a specific user interface. Furthermore, there is a need for an effective comparison of different algorithms used for detecting phishing websites. Phishing is a dangerous and fraudulent tactic used by cybercriminals to steal sensitive information. Machine learning models are used to detect phishing websites, but there is a need for improvement in their accuracy, efficiency, and user interface. It is essential to stay vigilant and double-check the authenticity of emails, websites, and communication channels to avoid falling victim to phishing attacks.

***Key Words:*** *Phishing, cybercriminals, fake websites, malicious link, machine learning models, Random Forest, Decision Tree, Multilayer Perceptron, phishing website ,falling victim, phishing attacks.*

## 1. INTRODUCTION

In today's fast-paced world technology has become an essential part of everyone's life. Technology has been greatly escalating and thereby making our experiences comfortable. Nowadays our presence and business have been dependent on the internet and various online platforms. People perform various activities in their day-to-day life that includes accessing online shopping websites, banking websites, educational websites, and social media. Nonetheless, all these websites ask for our data and some of them consist of sensitive information that may be bank details or card details. And as a result of all this, hackers have found an easy way of attacking other personal information and tracking their behavior. There are many types of attacks including Man-in-the-middle Attacks, Dos Attacks, SQL injection, Phishing Attacks, and many more. Out of all these websites, phishing has been considered a great threat to a user's vital information. The social engineering trick is used to manipulate the users and thereby duping them with the legitimate-looking URL which is a fake URL. It is difficult for a naive user to spot whether the URL is legitimate or fake. Research has shown that there has been a great boost in phishing attacks. Some researchers have a heuristic-based approach while some use a Machine learning-based approach. Machine learning has two different approaches i.e., Supervised Learning and Unsupervised Learning. In this paper, we have focused on

supervised learning where we have provided a 2000 dataset of phishing URLs and a 2000 dataset of legitimate URLs from phishtank.com. Also, we have made use of the Random Forest Algorithm due to its high accuracy, robustness, and good performance. And based on characteristic classification the system will differentiate the provided URL and will conclude whether the given URL is legitimate or a phishing URL. By which the user will be able to figure out that he might endanger his information if he visits that particular URL. And hence this system helps in guarding and thereby providing a possible solution towards the issue.

## 2. SYSTEM IMPLEMENTATION

**2.1 EXISTING SYSTEM:** In recent times, machine learning has played a pivotal role in the detection of phishing websites. Advanced machine learning algorithms have been developed and applied to analyze various features and patterns inherent in website content and behavior. These algorithms can effectively learn from vast amounts of data, enabling them to accurately identify and classify potential phishing sites. By analyzing factors such as URL structure, content analysis, website reputation, SSL certificate validity, and user behavior, machine learning models can identify subtle indicators and similarities with known phishing patterns. This approach allows for real-time detection and proactive mitigation of phishing attempts. The continuous evolution and refinement of machine learning models contribute to their increasing effectiveness in identifying sophisticated and previously unseen phishing techniques. By leveraging the power of machine learning, organizations and individuals can enhance their defenses against phishing attacks and safeguard sensitive information from falling into the wrong hands.

**2.2 PROPOSED SYSTEM:** A proposed system for enhancing online security aims to detect phishing websites by combining several techniques. The system employs URL analysis to identify suspicious patterns and misspellings, while the content of web pages is scrutinized for forgery or inconsistencies. The main objective of this project is to create a domain authentication system that can differentiate between legitimate and fake websites created for fraudulent purposes. Multiple machine learning models are tested to solve this problem, with the Multilayer Perceptron (MLP) providing the highest accuracy of 99%, with suitably balanced precision and recall. The system also monitors website behavior for suspicious activities such as early data requests or unexpected redirects, and utilizes blacklists and whitelists of known phishing and legitimate websites. User feedback mechanisms are incorporated, and machine learning models are trained on historical data. Browser extensions are also developed to provide real-time warnings to users. While no system can offer absolute accuracy, employing a combination of these techniques and regular updates can significantly improve the detection of phishing websites.

## 3. MODULE DESCRIPTION

### 3.1 DATA COLLECTION

Data collection is one of the most important steps in any machine learning workflow as the efficiency of our production model is directly proportional to the quality of our training dataset. In order to train a production ready classification model which can optimally predict malicious phishing URLs, we need to train the model on similar datasets.

### 3.2 FEATURE EXTRACTION

Feature extraction in machine learning is a fundamental process that involves transforming raw data into meaningful features suitable for training machine learning models. The process begins by considering the raw data in its original form, which could be numerical values, text documents, images, audio signals, or sensor readings.

### 3.3 DATA PREPROCESSING

Data preprocessing is a crucial step in machine learning (ML) that involves transforming raw data to make it suitable for training ML models and improving their performance.

### 3.4 MODEL TRAINING & EVALUATION

After the data has been brought to a suitable format after processing and feature engineering, we can start with training machine learning models. Depending on the type of data given, the accuracy and performance of various machine

learning algorithms can also vary, therefore we must train multiple algorithms and gauge their performance to choose the one model which gives the best performance.

### 3.5 USER INTERFACE

A user web interface (UI) is the visual and interactive part of a website or web application that allows users to engage with the system. It serves as the gateway for users to access the functionalities and content provided by the web application.

## 4. CONCLUSION

The use of machine learning techniques for detecting phishing websites can significantly enhance online security. The proposed system combines several methods, including URL analysis, content analysis, monitoring of website behavior, and user feedback mechanisms. Multiple machine learning models are tested, with the Multilayer Perceptron providing the highest accuracy. By utilizing a combination of these techniques and regular updates, the system can effectively differentiate between legitimate and fraudulent websites and provide real-time warnings to users. However, it is important to note that no system can offer absolute accuracy, and vigilance is still required to avoid falling victim to phishing attacks.

## 5. FUTURE ENHANCEMENT

In future enhancements for detecting phishing websites using machine learning, additional machine learning algorithms and techniques can be explored to improve accuracy and efficiency. For instance, deep learning models and ensemble techniques can be employed to increase the precision of detection. Furthermore, the system can be integrated with user behavior analysis and anomaly detection methods to identify suspicious user actions that may indicate a phishing attack. Additionally, natural language processing (NLP) techniques can be used to analyze the text in web pages and identify subtle variations in language that may indicate fraudulent activity. Overall, the future enhancements for the detection of phishing websites using machine learning will focus on improving accuracy, speed, and user-friendliness, to provide better protection against phishing attacks.

## 6. REFERENCES

1. "Phishing website detection using machine learning techniques" by C. Vijayakumar and S. Arumugam. International Journal of Computer Applications, Vol. 140, No. 1, April 2016.

2. "Detection of Phishing Websites using Machine Learning Techniques" by R. H. Magar and S. S. Pawar. International Journal of Computer Applications, Vol. 171, No. 5, July 2017.

3. "Phishing website detection using machine learning classifiers" by M. Saranya and K. Saravanan. International Journal of Computer Science and Mobile Computing, Vol. 7, No. 2, February 2018.

4. "Phishing website detection using machine learning techniques based on URL features" by A. H. Al-Waisy and H. S. Al-Khafaji. International Journal of Advanced Computer Science and Applications, Vol. 9, No. 7, July 2018.

5. "Detecting Phishing Websites Using Machine Learning" by S. N. Alharthi and H. M. Al-Hassan. Journal of King Saud University - Computer and Information Sciences, Vol. 32, No. 3, May 2020.

6. T. Hasan, F. Faruque, M. Islam, M. F. Hossain, and R. Hasan, "A Deep Learning Based Detection and Prevention System for Phishing Websites," in Proceedings of the 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pp. 185-192, 2019.

7. A. R. M. Luthfi, S. A. Fauzi, and N. Hidayati, "Phishing website detection using machine learning," in Proceedings of the 6th International Conference on Computer Science and Computational Intelligence (ICCSCI), pp. 1-6, 2020.