WWW.IJRPR.COM

**International Journal of Research Publication and Reviews**

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Phishing Website Detection Using Machine Learning

***S Harika[a], Balasani Preetham[b], Rupani Mallesh[c], Voyepuram Sai Vyasa Teja[d] , N.Sridhar [e*]***

[a,b,c,d] Student, Department of AIML. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

[e] Professor, Department of AIML. Malla Reddy Engineering College, Maisammaguda, Hyderabad-500100

**ABSTRACT**

Phishing websites have proven to be a major security concern. Several cyberattacks risk the confidentiality, integrity, and availability of company and consumer data, and phishing is the beginning point for many of them. Many researchers have spent decades creating unique approaches to automatically detect phishing websites. While cutting-edge solutions can deliver better results, they need a lot of manual feature engineering and aren't good at identifying new phishing attacks. As a result, finding strategies that can automatically detect phishing websites and quickly manage zero-day phishing attempts is an open challenge in this field. While existing systems have been effective in detecting phishing websites with often exceeding 90% in well-balanced datasets. However, the accuracy of these systems can be impacted by issues such as over-fitting, an imbalanced data-set, and the evolving nature of phishing attacks.So In this proposed system adresses the approache that By combining multi-modal feature analysis the system provides a more accurate solution for phishing website detection.The Project focuses on the detection of phishing websites using machine learning algorithms, with the goal of identifying the algorithm that yields the highest accuracy for phishing detection. Random Forest with 97% and Decision Tree Classifier with 94% are the most fitted algorithms.

**Keywords-**Phishing, Machine Learning, cyberattacks, phishing websites ,Random-Forest, Decision Tree, maliciousness, Over-fitting

## 1. Introduction

The rapid evolution of the internet has led to a significant increase in cyber threats, with phishing websites being one of the most pervasive security challenges. Phishing attacks, where malicious websites deceive users into revealing sensitive information, pose a substantial risk to both businesses and individuals. These attacks undermine the confidentiality, integrity, and availability of critical data, often leading to financial loss and reputational damage. Over the years, numerous approaches have been proposed to detect phishing websites, ranging from rule-based methods to machine learning-driven techniques. While traditional methods have made significant strides, they often require extensive manual feature engineering and struggle to keep pace with the constantly evolving tactics employed by cybercriminals. Furthermore, many existing models struggle with challenges like overfitting, imbalanced datasets, and the detection of zero-day phishing attacks, where new threats are introduced without prior detection. This paper focuses on developing a machine learning-driven approach to improve phishing website detection by utilizing advanced algorithms that require minimal manual intervention and can adapt to new phishing techniques. By combining multi-modal feature analysis, which includes factors such as URL characteristics, website content, and hosting details, the system aims to enhance detection accuracy. The goal of the project is to identify the most effective machine learning algorithm for detecting phishing websites, with a particular focus on the Random Forest algorithm, which achieved an impressive accuracy of 97%, and the Decision Tree Classifier, which recorded a 94% accuracy rate. These results indicate that machine learning can play a pivotal role in combating phishing attacks, providing businesses and consumers with a reliable tool to safeguard their online presence and prevent data breaches.

## II. Literature Survey

The increasing prevalence of phishing attacks has necessitated the development of effective detection systems to protect users from fraudulent websites. Several studies have explored different machine learning techniques for phishing website detection, demonstrating the potential of automated systems to enhance online security. The Anti-Phishing Working Group (APWG) has been at the forefront of unifying global efforts against cybercrime, providing resources and data to improve phishing detection mechanisms [1]. Various types of phishing attacks, including spear phishing and pharming, have been identified, with administrators urged to watch out for these evolving threats to protect against data breaches [2].Lakshmanarao et al. (2021) proposed a novel machine learning fusion approach to detect phishing websites, demonstrating improved accuracy by combining different classifiers and algorithms [3]. Similarly, Chapla, Kotak, and Joiser (2019) explored URL-based phishing detection using fuzzy logic classifiers, offering insights into how specific features of URLs can be used to detect malicious sites [4].A comparative analysis conducted by Vaishnavi et al. (2021) focused on various machine learning algorithms, assessing their effectiveness in predicting malicious URLs, and highlighted the importance of selecting the right model for phishing detection tasks [5]. Microsoft's Consumer Safety Report further underlined the increasing impact of poor online safety practices, correlating the rise of phishing attacks with widespread user negligence in cybersecurity practices [6].The IRS email schemes (2016) and other similar phishing campaigns

have underscored the need for reliable detection mechanisms to mitigate the growing threat of phishing attacks targeting both individuals and businesses [7]. Studies like that of E.B. K.T. (2015) have proposed a machine learning and web mining-based approach for phishing URL detection, highlighting its practical applications for real-time threat detection [8].Furthermore, research by Wang et al. (2013) focused on static malicious JavaScript detection using support vector machines (SVM), another critical component in preventing phishing attacks that involve dynamic content on web pages [9]. Basnet et al. (2003) also examined the effectiveness of machine learning models in phishing attack detection, providing a foundation for the development of modern phishing detection systems [10].

**Table .1.** Literature Survey

| **Study** | **Key Contribution** | **Year** |
|---|---|---|
| APWG | Unified global efforts to combat cybercrime, providing resources and data to improve phishing detection systems. | 2012 |
| SysCloud | Highlighted various types of phishing attacks and the importance for IT administrators to monitor evolving phishing threats. | 2021 |
| Lakshmanarao et al. | Proposed a novel machine learning fusion approach for phishing website detection, improving accuracy by combining different classifiers. | 2021 |
| H. Chapla, R. Kotak and M. Joiser | Developed a machine learning approach for URL-based phishing detection using fuzzy logic, demonstrating its effectiveness in identifying phishing websites. | 2019 |
| Vaishnavi, D., Suwetha, S., Jinila, Y.B., Subhashini, R., Shyry, S.P. | Conducted a comparative analysis of machine learning algorithms for malicious URL prediction, highlighting key models for phishing detection. | 2021 |
| Microsoft | Released a Consumer Safety Report discussing the impact of poor online safety behaviors and its correlation with phishing attacks. | 2014 |
| IRS | Issued a warning on new phishing schemes targeting consumers during tax season, underlining the evolving nature of phishing attacks. | 2016 |
| E., B., K., T | Introduced a machine learning and web mining-based approach for phishing URL detection, demonstrating the application of web mining techniques in phishing detection. | 2015 |
| Wang Wei-Hong, L V Yin-Jun, CHEN Hui-Bing, FANG Zhao-Lin. | Explored static malicious JavaScript detection using support vector machines (SVM), contributing to the broader efforts in phishing and malicious web detection. | 2013 |
| Ram Basnet, Srinivas Mukkamala et al | Developed a machine learning-based approach to detecting phishing attacks, laying the groundwork for modern phishing detection techniques. | 2003 |

## III. Methodology

The methodology for this research focuses on leveraging machine learning techniques, statistical analysis, and data-driven approaches to analyze and predict phishing websites. The process involves several key steps, including data collection, preprocessing, feature engineering, model selection, training, and evaluation. Each of these steps contributes to developing an effective machine learning model to accurately classify phishing websites.

### *3.1. Data Collection*

This study utilizes a Kaggle job market dataset, which includes features such as URL components (domain name, URL length, presence of HTTPS), website metadata, WHOIS information, and sometimes page content or HTML structure. The quality and variety of this dataset are crucial for training a robust machine learning model.

| WebsiteFor | StatusBarCu | DisableRigh | UsingPopup | IframeRedir | AgeofDoma | DNSRecordi | WebsiteTra | PageRank | GoogleInde | LinksPointin | StatsReport | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | -1 |
| 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | 0 | -1 | -1 |
| 0 | 1 | 1 | 1 | 1 | -1 | -1 | 1 | -1 | 1 | -1 | 1 | -1 |
| 0 | -1 | 1 | -1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 | -1 | -1 |
| 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 0 | 1 | -1 |
| 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | -1 | 0 | -1 | 1 | 0 | 1 | -1 |
| 0 | 1 | 1 | 1 | 1 | -1 | 1 | 1 | 1 | 1 | -1 | -1 | 1 |
| 0 | 1 | 1 | 1 | 1 | -1 | -1 | -1 | -1 | 1 | 0 | -1 | -1 |
| 0 | -1 | 1 | -1 | 1 | 1 | -1 | -1 | -1 | 1 | 0 | 1 | -1 |
| 0 | 1 | 1 | 1 | 1 | -1 | -1 | 0 | -1 | 1 | 1 | 1 | -1 |
| 0 | 1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | 1 | -1 | 1 | 1 |

**Fig 1.** Sample Dataset

### *3.2. Data Preprocessing*

Data pre-processing is a cleaning operation that converts unstructured raw data into a neat, well-structured data set that may be used for further research. The data is split into 8000 training samples and 2000 testing samples, before the ML model is trained. The following supervised machine learning models were examined for this project's dataset training: Decision Tree , Random Forest, Logistic Regression , KNN and SVM.

### *3.3 Model Training*

To ensure accurate job market predictions, multiple machine learning models were trained and evaluated. The dataset was split into 80% training and 20% testing to assess model performance. The following models were implemented

#### *3.3.1 Logistic Regression*

Logistic Regression is a classification algorithm that predicts the probability of a given URL being phishing or legitimate. It uses a sigmoid function to model the probability of an event, which in this case, determines whether a URL is phishing or not.

$$\left(p\left(y=\frac{1}{x}\right)=\frac{1}{1+e^{-(wx+b)}}\right) \quad (1)$$

P(y=1/x) is the probability that the URL belongs to the phishing class,w is the coefficient vector for features, X is the feature set (URL components),b is the bias term.

#### *3.3.2 Support Vector Machine (SVM)*

SVM is a powerful classification algorithm that aims to find the optimal hyperplane that maximizes the margin between different classes (phishing and legitimate). It uses the concept of decision functions to separate data points. The decision function is given by:

$$f(X)=wTX+b \quad (2)$$

where w is the weight vector, X is the feature set, and b is the bias term.

#### *3.3.3 K Nearest Neighbors (KNN)*

KNN classifies jobs based on similarity to their nearest neighbors. The Euclidean distance formula determines the closest points:

$$d=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2} \quad (3)$$

Where d represents the distance between two points.

#### *3.3.4 Decision Tree Classifier*

Decision Trees are one of the simplest yet most powerful classification algorithms. They split the data into subsets based on feature values, creating a tree structure where each node represents a decision rule based on one feature.

f(X)=Decision Rule at Node(X) (4)

Where X represents the input features (URL components)

### *3.3.5 Random Forest*

Random Forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy.It reduces the variance of decision trees by averaging the predictions from several trees, each trained on different random subsets of the data. This results in a more robust and less overfitting-prone model.

$$f(X) = \frac{1}{N}\sum_{i=1}^{N} h_i(X) \qquad (5)$$

Where $h\_i\ (X)$ is the prediction from each tree and $1/N$ is the number of trees.

### *3.4 Model Evaluation*

The performance of the classification model has been evaluated using various evaluation metrics like accuracy, precision, recall, f1-measure.

**Table.2** The performance metrics used for classification and regression

| **Metric** | **Formula** |
|---|---|
| Precision (P) | $TP/(TP+FP)$ |
| Recall (R) | $TP/(TP+FN)$ |
| Accuracy | $(TP+TN)/(TP+TN+FP+FN)$ |
| F1-score | $2 * \frac{R * P}{R + P}$ |

## 4. Result Analysis and Discussion

The dataset consisted of 11053 numbers of values with 32 features, providing a diverse set of features for training the machine learning models. The data was split into an 80:20 training-to-testing ratio, ensuring a balanced evaluation of model performance. The Random Forest model outperformed all other algorithms, achieving an accuracy of 97%.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1 | 0.89 | 0.93 | 0.91 | 1399 |
| 1 | 0.95 | 0.92 | 0.93 | 1918 |
| accuracy | | | 0.92 | 3317 |
| macro avg | 0.92 | 0.93 | 0.92 | 3317 |
| weighted avg | 0.93 | 0.92 | 0.93 | 3317 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1 | 0.00 | 0.00 | 0.00 | 0 |
| 1 | 1.00 | 0.56 | 0.72 | 3317 |
| accuracy | | | 0.56 | 3317 |
| macro avg | 0.50 | 0.28 | 0.36 | 3317 |
| weighted avg | 1.00 | 0.56 | 0.72 | 3317 |

**Fig. 2.** Result of Logistic Regression and Support Vector Machine

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1 | 0.58 | 0.59 | 0.59 | 1458 |
| 1 | 0.68 | 0.67 | 0.67 | 1859 |
| accuracy | | | 0.63 | 3317 |
| macro avg | 0.63 | 0.63 | 0.63 | 3317 |
| weighted avg | 0.63 | 0.63 | 0.63 | 3317 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1 | 0.92 | 0.95 | 0.94 | 1415 |
| 1 | 0.96 | 0.94 | 0.95 | 1902 |
| accuracy | | | 0.95 | 3317 |
| macro avg | 0.94 | 0.95 | 0.95 | 3317 |
| weighted avg | 0.95 | 0.95 | 0.95 | 3317 |

**Fig. 3.** Result of K Nearest Neighbour and Decision Tree

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1 | 0.96 | 0.97 | 0.97 | 1438 |
| 1 | 0.98 | 0.97 | 0.97 | 1879 |
| accuracy | | | 0.97 | 3317 |
| macro avg | 0.97 | 0.97 | 0.97 | 3317 |
| weighted avg | 0.97 | 0.97 | 0.97 | 3317 |

**Fig. 4.** Result of Random Forest Classifier

we can observe that Random Forest outperforms all other models, achieving the highest accuracy, precision, and recall, making it the most reliable model for this dataset. Logistic Regression and Decision Tree show nearly identical strong results, indicating they are also solid choices. KNN and SVM has the lowest metrics but still performs decently, likely because its assumptions are less suited for this data distribution .

**Table 3.** Comparative Summary of Models

| Algorithm | **Accuracy (%)** | **Key Characteristics** |
|---|---|---|
| Logistic Regression | 92 | Works well with linearly separable data; interpretable but limited for complex relationships. |
| Support Vector Machine | 56 | Effective for small datasets; performs well with high-dimensional data. |
| K-Nearest Neighbor | 63 | Simple, non-parametric; sensitive to the choice of K and distance metric. |
| Decision Tree | 95 | Easy to interpret,Handles both categorical and continuous data. |
| Random Forest | 97 | Ensemble learning technique; reduces over-fitting; computationally expensive. |

## 5. Conclusion And Discussion

The findings of this study underscore the importance of machine learning in detecting phishing websites and improving cybersecurity measures. The analysis highlights the effectiveness of various machine learning algorithms in identifying phishing attempts, with Random Forest emerging as the most accurate model, achieving 97% accuracy. The results emphasize the critical role that URL features and machine learning-based classification methods play in accurately distinguishing between legitimate and phishing websites. Additionally, the study reveals that combining multiple features—such as URL structure, domain age, and web page content—enhances the detection capabilities, making the system more robust against evolving phishing tactics.Despite the promising results, challenges such as the handling of imbalanced datasets and the need for continuous updates to the model in response to new phishing strategies remain. The study also identifies the importance of real-time phishing detection systems that can quickly adapt to new attack patterns. Future work could explore the integration of advanced deep learning techniques, such as neural networks, for even higher accuracy and the development of adaptive systems that can dynamically learn from new phishing attempts.The project's findings offer valuable insights for both individuals and organizations. Users can leverage this information to better protect themselves against phishing threats, while businesses can integrate the predictive models into their security systems to safeguard sensitive data. Moreover, the research highlights the potential of machine learning-based phishing detection systems in real-world applications, such as browser extensions and web security services, to enhance online safety and protect against evolving cyber threats.

**References**

[1]. 'APWG | Unifying The Global Response To Cybercrime' (n.d.) available: https://apwg.org/

[2]. 14 Types of Phishing Attacks That IT Administrators Should Watch For [online] (2021)

[3]. Lakshmanarao, A., Rao, P.S.P., Krishna, M.M.B. (2021) 'Phishing website detection using novel machine learning fusion approach', in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Presented at the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 1164–116

[4]. H. Chapla, R. Kotak and M. Joiser, "A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier", 2019 International Conference on Communication and Electronics Systems (ICCES), pp. 383-388, 2019, July

[5]. Vaishnavi, D., Suwetha, S., Jinila, Y.B., Subhashini, R., Shyry, S.P. (2021) 'A Comparative Analysis of Machine Learning Algorithms on Malicious URL Prediction', in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Presented at the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 1398–1402

[6]. Microsoft, Microsoft Consumer safety report. https://news.microsoft.com/en-sg/2014/02/11/microsoft-consumersafety-index-reveals impact-of-poor-online-safety-behaviours-in-singapore/sm.001xdu50tlxsej410r11kqvks u4nz.

[7]. Internal Revenue Service, IRS E-mail Schemes. Available at https://www.irs.gov/uac/newsroom/consumers-warnedof-new-surge-in-irs-email-schem es-during-2016-tax-season-tax-industry-also-targeted.

[8]. E., B., K., T. (2015)., Phishing URL Detection: A Machine Learning and Web Mining-based Approach. International Journal of Computer Applications,123(13), 46-50. doi:10.5120/ijca2015905665.

[9]. Wang Wei-Hong, L V Yin-Jun, CHEN Hui-Bing, FANG Zhao-Lin., A Static Malicious Javascript Detection Using SVM, In Proceedings of the 2nd International Conference on Computer Science and Electrical Engineering (ICCSEE 2013).

[10]. Ram Basnet, Srinivas Mukkamala et al, Detection of Phishing Attacks: A Machine Learning Approach, In Proceedings of the International World Wide Web Conference (WWW), 2003.