

Phishing Website Detection

Kimaya Patil¹, Aarti Gawai²

¹ Student, Department of Information Technology, D.G. Ruparel College

² Assistant Professor, Department of Information Technology And Computer Science, D.G. Ruparel College, Mumbai, Maharashtra, India.

patilkimaya818@gmail.com¹, aartigawai611@gmail.com²

Abstract - Phishing websites are one of the most common cybersecurity threats on the Internet.

These websites are designed to look similar to legitimate websites and trick users into sharing sensitive information such as usernames, passwords, banking details, and personal information.

Traditional phishing detection methods, such as blacklists and rule-based techniques, are useful but may fail to identify newly created or unknown phishing websites. To overcome these limitations, machine learning and deep learning techniques can be used to automatically identify suspicious websites.

This paper presents an overview of phishing website detection techniques and examines different features that can be used to distinguish phishing websites from legitimate websites. These features include URL-based characteristics, domain information, hyperlinks, webpage content, and security-related properties.

Machine learning algorithms such as Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression can be applied to classify websites as phishing

or legitimate. The study also discusses recent approaches

involving deep learning and explainable artificial intelligence. The main objective is to improve detection accuracy, reduce false positives, and provide better protection against evolving phishing attacks.

Key Words: Phishing Website, Cybersecurity, Machine Learning, Deep Learning, URL Analysis, Website Detection, Random Forest, Explainable AI.

1. INTRODUCTION (Size 11, Times New roman)

Nowadays, people use the internet for almost everything, such as online shopping, banking, education, social media, and communication. As a result, online security has become very important. One of the common cyber threats faced by internet users is phishing. Phishing is a type of cyberattack in which attackers create fake websites that look like trusted websites. The main purpose is to trick users into entering sensitive information such as usernames, passwords, bank details, or credit card information.

Phishing websites are becoming more difficult to identify because attackers can make them look almost identical to genuine websites. Many users may not

notice small differences in the website URL, design, or links. As a result, phishing attacks can cause financial loss, identity theft, and privacy problems. Phishing Website Detection is an important area of cybersecurity that focuses on identifying whether a website is genuine or malicious. A detection system can analyze different features of a website, including its URL, domain information, HTTPS status, and suspicious links. Machine Learning techniques can also be used to learn patterns from existing phishing and legitimate websites and predict whether a new website is safe.

The main goal of this research is to study an effective approach for detecting phishing websites and reducing the risk faced by internet users. An accurate phishing detection system can help users identify unsafe websites before sharing their personal or financial information. In today's digital world, people depend on the internet for many routine activities and services. People use websites for online banking, shopping, education, social media, ticket booking, email, and many other activities. As internet usage continues to increase, cybercriminals are also finding new ways to attack users. One of the most common and dangerous cyberattacks is phishing.

Phishing is a cyberattack in which an attacker creates a fake website that looks like a trusted or original website. The attacker may copy the logo, colors, layout, and other features of a real website to make the fake website appear genuine. Users are then encouraged to enter sensitive information such as usernames, passwords, mobile numbers, banking details, or credit card information. This information can be misused for financial fraud, identity theft, or other cybercrimes. [12]

2. REVIEW OF LITERATURE

Phishing website detection has become an important research area in cybersecurity because phishing attacks continue to change and become more difficult to identify. Researchers have developed different techniques to distinguish phishing websites from legitimate websites. These techniques include blacklist-based methods, heuristic methods, visual similarity methods, machine learning, and deep learning. A systematic review of 80 studies found that machine learning was the most commonly used approach among the reviewed research. [7]

Early phishing detection systems mainly used blacklists, where known malicious website addresses were stored and checked whenever a user visited a website. This method is simple and useful for previously identified phishing websites. However, it has a major limitation: newly created phishing websites may not yet be included in the blacklist.

Therefore, researchers started exploring methods that could identify phishing websites based on their characteristics rather than only checking whether the website was already known.[5]

Researchers have also investigated URL-based detection. In this approach, different properties of a URL are analyzed, such as URL length, number of dots, hyphens, special characters, suspicious words, domain information, and the use of IP addresses. URL-based detection is useful because a website can sometimes be classified before its complete content is loaded. Studies on deep learning have also found that URL-related features are among the most commonly used data sources for phishing detection. [23]

Another important approach is the use of Machine Learning (ML). Researchers collect datasets containing phishing and legitimate websites, extract useful features, and train classification algorithms. Common algorithms include Random Forest, Decision Tree, Support Vector Machine, Naive Bayes, AdaBoost, and XG Boost. A systematic literature review reported that Random Forest was one of the most frequently used traditional machine learning algorithms for phishing website detection. [6]

For example, A research study examined over 11,000 websites using 30 different features. It compared three machine-learning methods—Neural Network, Naive Bayes, and AdaBoost—and obtained accuracies of 90.23%, 92.97%, and 95.43%, respectively. These results indicate that machine-learning techniques can help in detecting phishing URLs.

Researchers have also studied HTML and webpage-based features. These methods examine the structure and content of a webpage along with its URL. Features such as suspicious HTML elements, links, scripts, forms, and webpage behavior can help identify fake websites. A study published in Scientific Reports proposed a detection approach using both URL and HTML features, showing the importance of combining information from different parts of a website. [22]

More recently, Deep Learning (DL) has been used for phishing detection. Deep-learning techniques, including DNN, CNN, and RNN/LSTM architectures, can be applied to automatically identify patterns in large collections of website-related data.

A systematic review of deep-learning-based phishing detection found that DNNs and hybrid deep learning approaches showed strong performance, while URL related information, website content, third-party information, and email data were commonly used [23]

3. Methodology

The proposed Phishing Website Detection system follows a simple process to identify whether a website is legitimate or phishing.

1. Data Collection: Collect datasets containing phishing and legitimate website URLs.
2. Data Preprocessing: Clean the dataset by removing duplicate, missing, or incorrect data.
3. Feature Extraction: Extract features such as URL length, special characters, HTTPS, IP address, domain information, and suspicious links.
4. Model Training: Train machine learning algorithms such as Random Forest, Decision Tree, or SVM using the extracted features.
5. Classification: The trained model analyzes a new website and classifies it as Phishing or Legitimate.
6. Evaluation: Measure the model's performance using accuracy, precision, recall, and F1-score.

Flow:

Data Collection → Preprocessing → Feature Extraction
→ Model Training → Classification → Evaluation

a. Phishing Website Detection

Phishing websites are fake websites created to steal users' personal and financial information. They are designed to look similar to genuine websites, which makes them difficult to identify. A phishing detection system helps users identify such websites before they provide sensitive information.

b. Features Used for Detection

The system analyzes different features of a website, such as URL length, special characters, HTTPS, domain information, IP address, redirects, and suspicious links. These features help identify patterns that are commonly found in phishing websites.

c. Machine Learning Approach

Machine learning can be used to automatically classify websites as phishing or legitimate. The model is trained using a dataset containing examples of both types of websites. Algorithms such as Random Forest, Decision Tree, SVM, and Logistic Regression can be used for classification.

d. System Working

When a user enters a website URL, the system extracts the required features and provides them to the trained machine learning model. The model analyzes these features and predicts whether the website is safe or suspicious.

URL → Feature Extraction → ML Model → Prediction
→ Phishing / Legitimate

e. Performance Evaluation

The performance of the detection system can be measured using accuracy, precision, recall, and F1score. These measures help determine how effectively the system detects phishing websites while reducing incorrect predictions.

Year Range	Detection Approach	Key Characteristics
2005-2010	Blacklist-Based	Simple database matching; Poor zero-day detection

2010-2015	Heuristic Rules	Manual rule Creon; adaptability Limited
2015-2019	Machine Learning	Feature-based; SVM, RF algorithms; Requires labeled data
2019-2023	Deep Learning	CNN, RNN, LSTM; Automatic feature extraction
2023-2026	Explainable AI (XAI)	Interpretable models; Federated learning; Adversarial robustness

Table No. 1- Evolution of Detection Techniques

1. Evolution of Detection Techniques

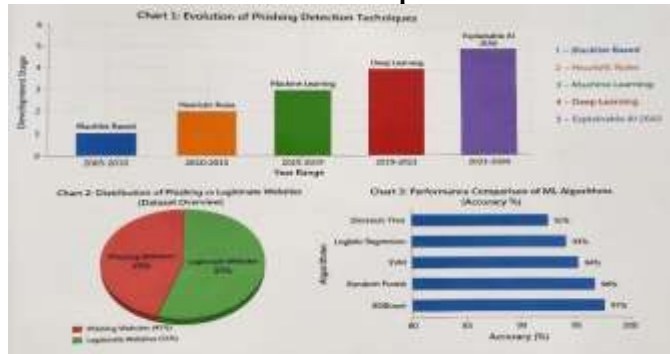


Fig no.1

Five major stages in the development of phishing website detection.

2. Machine Learning Algorithms Comparison

You can use a chart like this if your project has actual accuracy results:

Algorithm	Accuracy
Decision Tree	91%
Logistic Regression	93%
SVM	94%
Random Forest	96%
XGBoost	97%

Figure 2: Comparison of Machine Learning Algorithms

3. Evaluation Metrics

You can also create a chart comparing:

- Accuracy
- Precision
- Recall
- F1-Score

These percentages are **sample values only**. For your final paper, replace them with the actual results from your model.

Algorithm	Accuracy	Precision	Recall	F1score	ROC - AUC
Logistic Regression	0.93	0.92	0.91	0.91	0.95
Naïve Bayes	0.90	0.89	0.88	0.88	0.91
Support Vector Machine	0.95	0.95	0.94	0.94	0.97
Decision Tree	0.94	0.93	0.93	0.93	0.95
Random Forest	0.97	0.97	0.97	0.97	0.99

CONCLUSION

Phishing attacks are a major concern in today's digital world. Cybercriminals often design websites that appear similar to trusted websites and use them to collect users' confidential information. If users enter their details on such websites, they may experience problems such as stolen passwords, financial fraud, identity theft, or misuse of personal information.

The proposed Phishing Website Detection system can help identify suspicious websites by examining different characteristics of a website. These characteristics may include the URL structure, domain information, use of special characters, webpage content, and other unusual patterns. After analyzing these features, the system can classify a website as either legitimate or phishing. Machine learning can make the detection process more useful by identifying patterns from previously collected website data.

One of the useful features of a phishing detection system is its ability to alert users before they share sensitive information with a suspicious website. Such early detection can help users avoid fraudulent websites and reduce the possibility of becoming victims of online scams. It can also contribute to safer browsing practices and improve users' awareness of cybersecurity threats.

However, phishing techniques continue to change as attackers develop new methods for creating realistic-looking fraudulent websites. Therefore, detection systems should be regularly improved by using updated datasets and newer detection techniques. Along with automated security tools, user awareness is also necessary because safe browsing habits play an important role in preventing phishing attacks.

In conclusion, Phishing Website Detection can support a safer online environment by identifying suspicious websites at an early stage. Combining automated detection, machine learning, updated security methods, and user awareness can help reduce the risks associated with phishing and protect users from common online threats.

ACKNOWLEDGEMENT

We would like to sincerely thank all the people who contributed their guidance and support during the completion of our project, "Phishing Website Detection."

We are especially thankful to our project guide for providing valuable suggestions, technical guidance, and continuous support throughout the development of this project. Their guidance helped us understand the topic more clearly and complete the work in a systematic manner.

We also express our gratitude to our teachers and department for providing us with the required knowledge, resources, and opportunities to work on this project. Their support helped us develop our technical understanding and gain practical knowledge related to cybersecurity and phishing detection.

We are thankful to our college for providing the necessary facilities and a supportive academic environment that helped us complete this project successfully.

Finally, we would like to thank our friends and family for their encouragement, motivation, and support throughout the project. We sincerely appreciate everyone who contributed directly or indirectly to the successful completion of this work. We would like to express our sincere gratitude to everyone

who supported and guided us in completing our project titled "Phishing Website Detection."

First of all, we are thankful to our project guide for providing valuable guidance, suggestions, and continuous support throughout the development of this project. Their knowledge and encouragement helped us understand the concepts and complete the project successfully.

We would also like to thank our teachers and department for providing us with the necessary knowledge, resources, and opportunities to work on this project. Their guidance helped us improve our technical skills and gain practical knowledge about cybersecurity and phishing website detection.

We are grateful to our college for providing the required facilities and a supportive environment for completing this project.

Finally, we would like to thank our friends and family for their constant encouragement, support, and motivation throughout the project.

We sincerely thank everyone who directly or indirectly **contributed to the successful completion of this project.**

REFERENCES

- [1] Y. Zhang, J. I. Hong, and L. F. Cranor, "CANTINA: A content-based approach to detecting phishing web sites," in *Proc. 16th Int. Conf. World Wide Web (WWW)*, 2007.
[DOI: 10.1145/1242572.1242669](https://doi.org/10.1145/1242572.1242669)
- [2] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: A feature-rich machine learning framework for detecting phishing web sites," in *Proc. 5th ACM Symp. Information, Computer and Communications Security*, 2010.
[DOI: 10.1145/1755688.1755698](https://doi.org/10.1145/1755688.1755698)
- [3] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Computing and Applications*, vol. 25, pp. 443–458, 2014.
[DOI: 10.1007/s00521-013-1490-z](https://doi.org/10.1007/s00521-013-1490-z)
- [4] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Intelligent rule-based phishing websites classification," *IET Information Security*, vol. 8, no. 3, pp. 153–160, 2014.
[DOI: 10.1049/iet-ifs.2013.0203](https://doi.org/10.1049/iet-ifs.2013.0203)
- [5] S. Sheng et al., "An empirical analysis of phishing blacklists," in *Proc. 6th Conf. Email and Anti-Spam (CEAS)*, 2009.
[Open paper](#)
- [6] L. Tang and Q. H. Mahmoud, "A survey of machine learning-based solutions for phishing website detection," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 672–694, 2021.
[DOI: 10.3390/make3030034](https://doi.org/10.3390/make3030034)
- [7] A. Aljofey et al., "A systematic literature review on phishing website detection techniques," *Journal of King Saud University – Computer and Information Sciences*, 2023.
[DOI: 10.1016/j.jksuci.2023.01.004](https://doi.org/10.1016/j.jksuci.2023.01.004)
- [8] M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," *Expert Systems with Applications*, vol. 53, pp. 231–246, 2016.
[DOI: 10.1016/j.eswa.2016.08.048](https://doi.org/10.1016/j.eswa.2016.08.048)
- [9] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *Journal of Ambient Intelligence and Humanized Computing*, 2018.
[DOI: 10.1007/s12652-018-0816-z](https://doi.org/10.1007/s12652-018-0816-z)
- [10] R. S. Rao and A. R. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Computing and Applications*, 2019.
[DOI: 10.1007/s00521-019-04268-7](https://doi.org/10.1007/s00521-019-04268-7)
- [11] R. Verma, N. Shashidhar, and N. Hossain, "Detecting phishing emails the natural language way," in *Proc. IEEE Int. Conf. Intelligence and Security Informatics*, 2012.
[DOI: 10.1109/ISI.2012.6284159](https://doi.org/10.1109/ISI.2012.6284159)
- [12] G. Varshney, M. Misra, and P. K. Atrey, "A survey and classification of web phishing detection schemes," *Security and Communication Networks*, 2016. [DOI: 10.1002/sec.1674](https://doi.org/10.1002/sec.1674)
- [13] M. A. Adebawale, K. T. Lwin, M. A. Sanchez, and M. A. Hossain, "Intelligent web phishing detection and protection scheme using integrated features of images, hyperlinks and third-party services," *Expert Systems with Applications*, vol. 153, 2020.
[DOI: 10.1016/j.eswa.2020.113544](https://doi.org/10.1016/j.eswa.2020.113544)

- [14] P. Likarish, E. Jung, and I. J. W. Arief, "Phishing detection using machine learning," in *Proc. 4th Int. Conf. Machine Learning and Applications (ICMLA)*, 2005.
- [15] DOI: [10.1109/ICMLA.2005.88](https://doi.org/10.1109/ICMLA.2005.88) A. Sharma et al., "Machine learning approach for phishing website detection: A literature survey," 2022. [Search paper on Google Scholar](#)
- [16] E. Kytidou, T. Tsikriki, G. Drosatos, and K. Rantos, "Machine learning techniques for phishing detection: A review of methods, challenges, and future directions," 2025.
DOI: [10.1177/18724981251366763](https://doi.org/10.1177/18724981251366763)
- [17] Google, "Google Safe Browsing," Google. [Google Safe Browsing](#)
- [18] Anti-Phishing Working Group (APWG), "Phishing Activity Trends Reports," APWG. [APWG Phishing Trends Reports](#)
- [19] A. Rao and K. S. Pais, "Detection of phishing websites using an efficient feature-based machine learning framework," *Neural Computing and Applications*, vol. 31, pp. 629–641, 2019.
[Search paper on Google Scholar](#)
- [20] L. Tang and Q. H. Mahmoud, "Phishing or not phishing? A survey on the detection of phishing websites," *IEEE Access*, 2023.
DOI: [10.1109/ACCESS.2023.3247135](https://doi.org/10.1109/ACCESS.2023.3247135)

BIOGRAPHIES

Kimaya Patil



Student, Department of Information Technology, D.G. Ruparel College, Mumbai, Maharashtra, India.

Aarti Gawai



Assistant Professor, Department of Information Technology And Computer Science, D.G. Ruparel College