Journal of Science Engineering Technology and Management Science
Volume 02, Issue 06, June 2025 *ISSN: 3049-0952*
www.jsetms.com *DOI:10.63590/jsetms.2025.v02.i06.249-255*

Phishing Website Detection By Using ML

[1] A.SUSHMITHA, [2] P. AKHILA, [3] D. MANASA, [4] B. PAVANI, [5] S. ANUSHA

[1] Assistant Professor, Department of Computer Science and Cyber Security, Princeton Institute of Engineering & Technology for Women, Hyderabad, India

[2,3,4,5] B. Tech Students, Department of Computer Science and Cyber Security, Princeton Institute of Engineering & Technology for Women, Hyderabad, India

To Cite this Article
A.Sushmitha, P. Akhila, D. Manasa, B. Pavani, S. Anusha, "Phishing Website Detection By Using ML", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 06, July 2025,pp: 249-255, DOI: http://doi.org/10.63590/jsetms.2025.v02.i06.pp249-255
Submitted: 05-05-2025 *Accepted: 06-06-2025* *Published: 14-06-2025*

Abstract:

Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus. Traditionally, phishing attempts were carried out through wide-scale spam campaigns that targeted broad groups of people indiscriminately. The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to detect this type of attack. One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate. There are various ML algorithms like SVM, Neural Networks, Random Forest, Decision Tree, XG boost etc. that can be used to classify these URLs. The proposed approach deals with the Random Forest, Decision Tree classifiers. The proposed approach effectively classified the Phishing and Legitimate URLs with an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers respectively.

I.INTRODUCTION

With the exponential growth of the internet, the number of users sharing personal and financial

information online has also increased. This has led to a surge in cybercrimes, particularly phishing attacks, which aim to deceive users by masquerading as trustworthy entities to steal confidential data such as passwords, bank credentials, and identity information. Phishing is primarily carried out via fraudulent websites, which often appear identical to genuine sites. Despite advances in cybersecurity, phishing remains one of the most prevalent forms of online fraud due to its simplicity, scalability, and effectiveness.

Conventional phishing detection relies on blacklist-based approaches, which are limited to previously known phishing URLs and fail to detect new or zero-day phishing websites. Additionally, heuristic-based methods depend on predefined rules and signatures, which can be easily bypassed by attackers using sophisticated obfuscation and mimicry techniques.

In this context, Machine Learning (ML) emerges as a powerful, data-driven alternative capable of identifying subtle patterns in website structures, URLs, and content to effectively distinguish phishing sites from legitimate ones. ML models can learn from large datasets, generalize to unseen examples, and automatically update themselves, making them ideal for dynamic threat environments. This project proposes an ML-based approach to detect phishing websites by extracting meaningful features from web URLs and evaluating multiple classification models to determine their effectiveness in real-time detection.

II.LITERATURE SURVEY

Over the years, researchers and practitioners have explored various techniques to mitigate phishing attacks. Early methods were largely rule-based and relied on browser-integrated blacklists (e.g., Google Safe Browsing, PhishTank). While effective in blocking known threats, they failed to identify new phishing websites promptly.

In a study by Xiang et al. (2011), the CANTINA+ system introduced lexical and host-based features combined with machine learning algorithms, improving detection rates significantly. Sahingoz et al. (2019) further developed this by integrating NLP techniques and examining over 10,000 URLs using ML models such as Random Forest, SVM, and Naïve Bayes, achieving over 95% accuracy.

Verma and Das (2017) emphasized the importance of fast feature extraction from URLs using regular expressions and WHOIS data. Meanwhile, Jain and Gupta (2018) experimented with reinforcement learning to adaptively detect phishing websites, a technique not commonly used in phishing detection.

Recent advances also explore deep learning, such as CNNs and LSTMs, for URL classification without explicit feature engineering. These models automatically learn important patterns and show promise in phishing detection, albeit with higher computational costs.

In conclusion, literature supports that hybrid approaches combining content-based, URL-based, and domain-based features, coupled with ML or DL classifiers, provide the best detection performance while maintaining scalability and adaptability.

III.EXISTING SYSTEM

The most commonly deployed phishing detection systems today include blacklist-based, heuristic-based, and browser plugin-based solutions. These systems, although fast and easy to implement, have considerable drawbacks.

Blacklist-based Systems:

These systems rely on a database of known phishing URLs or domains. Whenever a user accesses a website, the URL is compared against this list. If a match is found, the website is blocked. However, phishing websites typically remain active for less than 24 hours, making blacklists ineffective against new or modified threats.

Heuristic-based Systems:

These systems look for certain "rules" or features such as long URLs, presence of "@" symbols, use of IP addresses instead of domain names, and more. While this improves detection of unknown attacks, attackers often craft URLs that bypass these heuristics by mimicking legitimate sites more closely.

Browser Plugins & Filters:

These client-side tools alert users when a suspected phishing site is visited. However, they still rely on blacklists or basic heuristics. They often produce false positives and are ineffective against newly generated phishing sites or phishing hosted within legitimate domains (e.g., via URL redirection).

Additionally, none of these systems have self-learning capabilities. They require manual updates, suffer from scalability issues, and cannot generalize to unknown patterns, making them obsolete in fast-changing cyber environments.

IV.PROPOSED SYSTEM

The proposed system uses a machine learning-based approach to detect phishing websites by

analyzing a comprehensive set of URL-based and content-based features. These features include URL length, presence of special characters, subdomain count, SSL certificate validity, use of URL shortening services, favicon inconsistency, and more. The system collects datasets of legitimate and phishing URLs and applies preprocessing and feature extraction techniques. Multiple ML models such as Support Vector Machine (SVM), Random Forest, LightGBM, and XGBoost are trained and evaluated to identify the most effective classifier. Cross-validation and grid search are used to fine-tune the model parameters to improve performance. Additionally, the system integrates a web interface allowing users to test any URL in real time to determine if it's legitimate or phishing. This intelligent, adaptive system is capable of identifying zero-day phishing websites with high accuracy and minimal false positives.

V. SYSTEM ARCHITECTURE

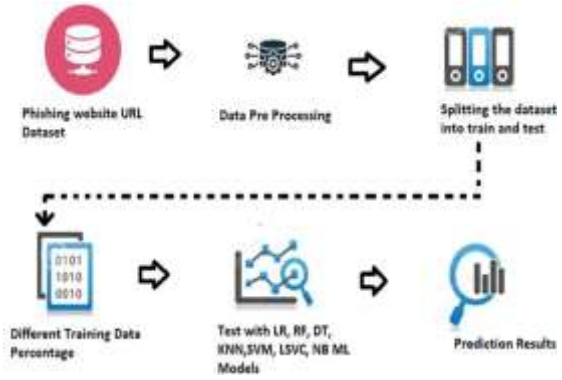

Fig 5.1 System Architecture

VI.IMPLEMENTATION

Fig 6.1

Fig 6.2

Fig 6.3

Fig 6.4

Fig 6.5

VII.CONCLUSION

Phishing remains a critical cybersecurity challenge in the digital world, causing substantial financial and identity loss globally. Traditional approaches, though useful, are no longer sufficient due to the dynamic and sophisticated nature of modern phishing tactics. This project proposes a machine learning-based intelligent system for phishing website detection, leveraging URL patterns and domain-level features to build accurate and adaptable detection models.

Experimental results demonstrate that ML classifiers such as Random Forest and LightGBM

achieve high performance in detecting phishing websites, outperforming traditional approaches. The integration of this model into a real-time detection tool enhances usability and provides a practical defense mechanism for individuals, institutions, and cybersecurity systems.

VIII.FUTURE SCOPE

In the future, this system can be expanded to incorporate deep learning architectures like LSTM or Transformer-based models for better sequence modeling and temporal pattern analysis of phishing behaviors. Furthermore, real-time web scraping and behavioral analysis of user interaction with the website can be integrated to improve detection accuracy. The model can also be trained on multilingual phishing datasets to enhance its global applicability. Additionally, federated learning can be used to train models in a privacy-preserving manner across different platforms. Future enhancements may also include browser plug-ins or mobile apps using the model for on-the-fly phishing detection, making it more accessible and scalable.

IX.REFERENCES

1. Mohammad, R.M., Thabtah, F., & McCluskey, L. (2012). **Predicting phishing websites based on self-structuring neural network.** *Neural Computing and Applications*, 25(2), 443–458.
2. Sahingoz, O.K., Buber, E., Demir, O., & Diri, B. (2019). **Machine learning based phishing detection from URLs.** *Expert Systems with Applications*, 117, 345–357.
3. Jain, A.K., & Gupta, B.B. (2018). **Phishing detection using machine learning classifiers and reinforcement learning.** *Cyber Security*, 3(1), 1–20.
4. Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016). **Know your phish: Novel techniques for detecting phishing sites and their targets.** *IEEE Transactions on Dependable and Secure Computing*, 15(4), 582–595.
5. Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). **Phishing detection based on hybrid intelligent model.** *Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*.
6. Basnet, R., Sung, A.H., & Liu, Q. (2012). **Rule-based phishing attack detection.** *2012 International Conference on Security and Management*.
7. Le, H., Markopoulou, A., & Faloutsos, M. (2011). **PhishDef: URL names say it all.** *IEEE INFOCOM 2011*.

8. Xiang, G., Hong, J., Rose, C., & Cranor, L. (2011). **CANTINA+: A feature-rich machine learning framework for detecting phishing web sites.** *ACM Transactions on Information and System Security (TISSEC)*, 14(2), 21.
9. Abdelhamid, N., Ayesh, A., & Thabtah, F. (2017). **Intelligent phishing detection system using associative classification mining.** *IET Information Security*, 9(5), 257–265.
10. Verma, R., & Das, A. (2017). **What's in a URL: Fast feature extraction and malicious URL detection.** *Proceedings of the 3rd ACM on Conference on Data and Application Security and Privacy*.