ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

Phishing Website Detection using Machine Learning

Mrs. P. Chamundeswari

Assistant Professor

Bhargavi Marigala

bhargavimarigala@gmail.com

Gunja Venkata Rani

venkataranigunja@gmail.com

Medaboina Bhargavi

medaboinaammu@gmail.com

Minde Kasthuri

mindekasthuri2003@gmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VIGNAN'SINSTITUTEOFMANAGEMENTANDTECHNOLOGYFOR WOMEN

(An Autonomous Institution)

(AffiliatedtoJawaharlalNehruTechnologicalUniversityHyderabad,AccreditedbyNBAandNAACwithA+Grade)

Kondapur(Village), Ghatkesar(Mandal), Medchal-Malkajgiri(Dist.)

Telangana-501301

ABSTRACT

Phishing is a major cybersecurity threat in which attackers create fake websites that closely resemble legitimate ones to steal sensitive user information such as login credentials, banking details, and personal data. With the increasing use of online services, phishing attacks have become more sophisticated and difficult to detect using traditional methods like blacklisting and rule-based systems. These conventional approaches often fail to identify newly created or zero-day phishing websites, making it necessary to

develop more intelligent and adaptive detection techniques.

This project proposes a machine learning-based approach for detecting phishing websites by analyzing various features such as URL characteristics, domain information, and security indicators. Different classification algorithms are trained on a dataset of phishing and legitimate websites to build an efficient prediction model. The system is capable of automatically classifying websites with high accuracy, helping users avoid potential threats. This approach not only enhances detection performance but also provides a scalable and effective solution for improving online security.

INTRODUCTION

The internet has become an essential part of daily life, enabling communication, banking, shopping, and social networking. However, this rapid digital transformation has also increased cyber threats, among which phishing is one of the most common and dangerous attacks. Phishing websites imitate legitimate websites to trick users into providing sensitive information such as login credentials and credit card details.

Traditional phishing detection techniques rely on blacklists and heuristic rules. While these methods are effective against known threats, they fail to detect newly generated phishing websites, also known as zero-day attacks. Attackers continuously modify their techniques, making static detection methods insufficient.

Machine Learning (ML) offers a promising solution by enabling systems to learn patterns from data and identify phishing websites dynamically. ML-based systems analyze features such as URL structure, domain characteristics, and webpage content to classify websites as legitimate or phishing.

This research focuses on developing a robust phishing detection system using supervised machine learning algorithms. The system extracts relevant features from websites and trains classification models to predict their authenticity. The use of multiple algorithms allows comparison and selection of the most accurate model.

The importance of this study lies in its ability to provide real-time detection and enhance user safety. By integrating the proposed system into web browsers or security software, users can receive instant alerts about potentially harmful websites.

In conclusion, machine learning-based phishing detection systems represent a significant advancement in cybersecurity, offering improved accuracy, adaptability,

and scalability compared to traditional approaches.

.

EXISTING SYSTEM

The existing systems for phishing detection primarily rely on blacklist-based and heuristic-based approaches. Blacklist systems maintain a database of known phishing URLs and compare user-requested URLs against this list. If a match is found, the website is flagged as malicious. Popular browsers and security tools use this technique due to its simplicity and efficiency.

However, blacklist systems have major limitations. They cannot detect newly created phishing websites that are not yet included in the database. Additionally, maintaining and updating blacklists requires continuous monitoring and manual effort.

Heuristic-based systems attempt to identify phishing websites using predefined rules such as checking URL length, use of IP addresses instead of domain names, and suspicious symbols. While these methods can detect some unknown phishing sites, they often produce high false positives and are not adaptable to evolving attack patterns.

Another approach involves visual similarity detection, where phishing websites are identified by comparing their design with legitimate websites. This method is computationally expensive and may not work effectively if attackers slightly modify the layout.

PROPOSED SYSTEM

The proposed system introduces a Machine Learning-based phishing detection approach that overcomes the limitations of traditional methods. The system collects and processes data from multiple sources, including URLs, domain information, and webpage content.

Feature extraction is a crucial step where relevant attributes such as URL length, number of dots, presence of HTTPS, domain age, and redirection count are identified. These features are then used to train machine learning models.

The system employs multiple classification algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), and Logistic Regression. Among these, Random Forest provides higher accuracy due to its

ensemble nature, combining multiple decision trees for better predictions.

The trained model is deployed in a real-time environment where it analyzes websites as users attempt to access them. If a website is classified as phishing, the system immediately alerts the user.

SYSTEM ARCHITECTURE

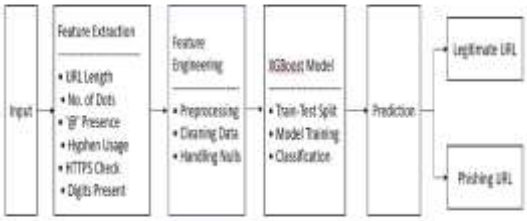


Fig:4.1SystemArchitecture

1.1 Description

- The phishing website detection system is developed to identify and prevent access to fraudulent websites that attempt to steal sensitive user information such as usernames, passwords, and financial details.
- This system uses machine learning techniques to automatically detect phishing websites by analyzing various characteristics of URLs and web pages.
- The system starts by accepting a URL input from the user or browser in real time.
- It then performs data collection, where relevant information about the website is gathered, including URL structure, domain details, and webpage content.
- Next, feature extraction is carried out to identify important indicators such as:

- URL-based features (length, special characters, IP usage)

➢ Domain-based features (age, DNS records, SSL certificate)
➢ Content-based features (forms, scripts, external links)

- These extracted features are passed into a pre-trained machine learning model.
- The model is trained using a labeled dataset consisting of both legitimate and phishing websites, enabling it to learn patterns and differences between them.
- Based on the learned patterns, the system performs classification and predicts whether the given website is:

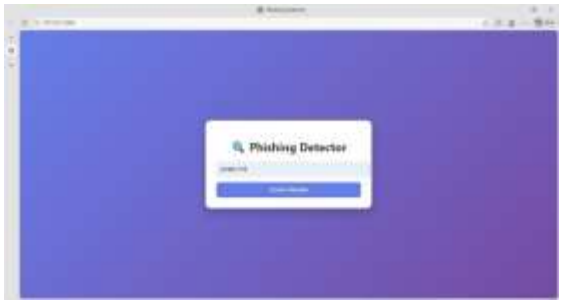
➢ Legitimate (Safe)
➢ Phishing (Malicious)

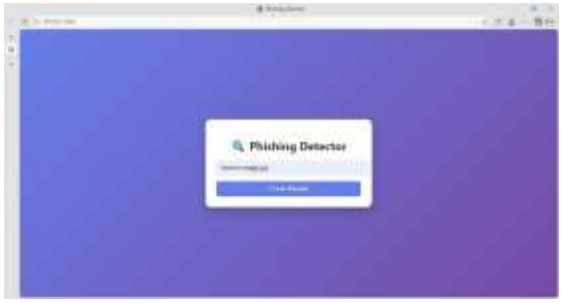
- The prediction result is then displayed to the user through a user interface or browser alert, warning them if the website is unsafe.
- The system is designed to operate in real time, providing instant detection and improving user security while browsing.
- Additionally, the system can be updated regularly with new data to improve

accuracy and adapt to new phishing techniques.

- Overall, this project provides an efficient and automated solution for enhancing cybersecurity and protecting users from online threats.

RESULTS AND DISCUSSIONS





CONCLUSION

In conclusion, the proposed phishing website detection system using machine learning represents a significant improvement over traditional detection methods such as blacklisting and rule-based approaches. By leveraging machine learning algorithms and feature analysis techniques, the system is capable of identifying phishing websites with higher accuracy and efficiency. It eliminates the limitations of existing systems by detecting not only known phishing websites but also newly created (zero-day) attacks, without relying solely on predefined rules or databases. This results in a more intelligent, automated, and scalable solution for enhancing cybersecurity. The system analyzes various features such as URL structure, domain information, and security indicators to classify websites as legitimate or phishing. The use of algorithms like Logistic Regression, Decision Tree,

and Random Forest ensures reliable prediction and improved performance. Additionally, the system reduces false positives and false negatives, providing trustworthy results to users. It requires minimal manual intervention and can be easily integrated into real-time applications such as web browsers and security tools, making it practical for real-world use.

Furthermore, the proposed system is cost-effective and adaptable to different environments, as it uses open-source tools and can handle large-scale data efficiently. Although certain challenges remain, such as handling highly sophisticated phishing techniques and improving accuracy further, these can be addressed by incorporating advanced methods like deep learning and real-time data analysis. Overall, the phishing website detection system provides a robust, efficient, and future-ready solution for protecting users from online threats and ensuring a safer digital environment.

References

1) R. Verma and K. Dyer, "On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers", ACM Conference on Data and Application Security and Privacy, 2015.

2) Viswanathan, V. (2024). Pioneering Ethical AI Integration in Enterprise Workflows: A Framework for Scalable Team Governance. Available at SSRN 5375619.

3) J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs", ACM SIGKDD Conference, 2009.

4) Viswanathan, V. (2025). Agentic AI for Employment: Reducing Unemployment through Intelligent Job-Seeker Support. LEX LOCALIS–Journal of Local Self-Government.

A. K. Jain and B. B. Gupta, "Phishing Detection: Analysis of Visual Similarity Based Approaches", Security and Communication Networks, 2017.

5) Viswanathan, V., Shah, A. K., Kubam, C. S., Dontu, S., Gandhi, A., & Singla, P. (2025, August). Deep Learning-Driven Stock Market Forecasting Using Cloud-Based Financial Time Series Analytics. In 2025 International

Conference on Emerging Trends in Networks and Computer Communications (ETNCC) (pp. 1-6). IEEE.

6) M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting Phishing Websites Based on Self-Structuring Neural Network", Neural Computing and Applications, 2014.

7) Viswanathan, V., Polagani, S. S., Agarwal, R., Akula, S., Dey, S., & Kashyap, R. (2025, September). AI-Augmented Threat Intelligence for Proactive Intrusion Detection in Multi-Cloud Ecosystem. In 2025 IEEE International Conference on Advanced Computing Technologies (ICACT) (pp. 567-572). IEEE.

8) Mahtabi, M., Roshan, M., Muhit, M. M. I., Behvar, A., & Haghshenas, M. (2026). Cryogenic ultrasonic fatigue: Mechanisms, advancements, and insights. Cryogenics, 153, 104257. https://doi.org/10.1016/j.cryogenics.2025.104257

9) Ranjbareslamloo, S., Dzukeya, G. A., Muhit, M. M. I., & Qattawi, A. (2025). Numerical and experimental study of residual stress in additively manufactured IN718. Manufacturing Letters, 44, 915–927. https://doi.org/10.1016/j.mfglet.2025.915927

10) Kotte, G. (2025). Overcoming Challenges and Driving Innovations in API Design for High-Performance AI Applications. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283649

11) Kotte, G. (2025). Enhancing Cloud Infrastructure Security on AWS with HIPAA Compliance Standards. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5283660

12) Kumara, S. (2026, February). A Lightweight Deep Learning Based Classification Models for Non-Human Identity Threat Detection. In 2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC) (pp. 1-6). IEEE.

13) Cyril, H. P., & Kumara, S. Identification of Anomalies via Deep Learning-Based Models for High-Dimensional Telecom Traffic Data.

14) [5]U. Garera, N. Provos, M. Chew, and A. D. Rubin, "A Framework for

Detection and Measurement of Phishing Attacks", ACM Workshop on Recurring Malcode, 2007.

15) Santthosh Saai Reddy Purmani. (2026). Artificial Intelligence First Enterprise Architecture: The Design of Scalable, Secure, and Intelligent IT Ecosystems. American Journal of AI Cyber Computing Management, 6(1(2)), 1–8. https://doi.org/10.64751/ajaccm.2026.v6.n1(2).pp1-8

16) Purmani, S. S. R. (2025). Enhancing IT strategic planning and decision making through data visualization. International Journal of Enhanced Research in Management & Computer Applications, 14(4), 75–81

17) Patyrykin, K. (2025). CANCEL CULTURE PROBLEM. Lex Localis: Journal of Local Self-Government, 23.

18) Patyrykin, K., & Vasyukova, L. (2025). Environmental Accountability or Symbolic Compliance? A Critical Review of ESG Ratings, Greenwashing, and Indirect Emissions in the Global Insurance Sector. International Journal of Energy Economics and Policy, 15(6), 917–925. https://doi.org/10.32479/ijeep.22770

19) Vasagam, M., Kumar, A., & Garg, A. (2026). Learning Execution Plan Embeddings for Multi-Dimensional Query Resource Prediction. IEEE Access.

20) Kalae, U. K. (2023). Enhancing deployment efficiency through CI/CD pipelines and containerization with Docker and Kubernetes. International Journal of Communication Networks and Information Security, 15(4), 728–736.

21) C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic Classification of Phishing Pages", Network and Distributed System Security Symposium (NDSS), 2010.

22) Akhilaiswarya, B., Sree, B. T., Lilly, K., Chowdary, K. H., & Sruthi, M. (2023). Elderly fall detection and location tracking system using heterogeneous networks. Journal of Engineering Sciences, 14(05).

23) S. Sheng, B. Wardman, G. Warner, L. Cranor, J. Hong, and C. Zhang, "An Empirical Analysis of Phishing Blacklists", CEAS Conference, 2009.

A. Bahnsen, D. Aouada, and B. Ottersten, "Example-Dependent Cost-Sensitive Logistic Regression for Credit Card Fraud Detection", (Applied in phishing detection research), 2015.

24) Kaggle Dataset, "Phishing Website Dataset", Available Online.

25) Sruthi, M. V., Soundararajan, K., & Sree, V. U. (2012). Accurate Multimodality Registration of medical images. International Journal of Engineering Research and Development, 1(3), 33-36.

26) Scikit-learn Documentation, "Machine Learning in Python", Available Online.

27) Kalae, U. K. (2021). Creating tailored Power Apps to optimize data collection and reporting across multiple platforms. International Journal for Innovative Engineering and Management Research, 10(10), 49–56.

28) Reddy, S. K. R. Developing a Modular AI Framework to Enhance Scalability and Personalization in Next-Generation Reward Platforms.

29) Sruthi, M. V., Sree, V. U., & Soundararajan, K. (2012). Specific removal of motion artifacts in medical image processing. IJECCE, 3(3), 227-229.

30) Dayal, P. S., Chandra, B. R., Keerthi, M., Sruthi, M., Venkatesh, K., Appalaraju, G., & Eswari, G. (2013). Design of Pyramidal Horn Antenna at 10GHz Using WIPL-D Optimizer. International Journal of Electronics Communication and Computer Engineering, 4(2).

31) Poojari, R. (2025). A Comparative Analysis of Fine-Tuning Versus Retrieval-Augmented Approaches for Enhancing Healthcare-Centric Large Language Models.

1.1 Websites

• https://www.kaggle.com
• https://scikit-learn.org
• https://www.wikipedia.org
•https://www.geeksforgeeks.orgs