Identifying Phishing Domains Using AI/ML

Mr. Chintha Sarath Kumar[1], Dr. S. Latha[2]
[1]*M.Sc CFIS, Department of Computer Science Engineering, Dr.M.G.R Educational and Research institute, Chennai, India*
[2] *Assistant Professor, Department of Criminology & Director, Centre for Cyber Forensics and Information Security, University of Madras, Chennai, India,*

Abstract **-The rapid advancement of internet and cloud technologies has significantly increased online transactions and e-commerce activities. However, With the rapid expansion of the digital world, cyber threats—particularly phishing attacks—have also seen a significant rise. Phishing schemes trick individuals into disclosing sensitive information by imitating legitimate websites, often closely copying the look and URL patterns of genuine platforms, making them hard to detect through standard security measures. Traditional defense methods like blacklists and heuristic techniques have shown clear shortcomings when it comes to addressing these sophisticated threats. The decentralized and often anonymous nature of the internet only worsens the challenge, giving cybercriminals more room to operate. Research has shown that many existing phishing detection tools are struggling to keep up with the increasingly advanced methods used by attackers. In response to these challenges, this study investigates the use of machine learning for detecting phishing websites. Five different classification algorithms were tested—Logistic Regression, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN). Among them, Logistic Regression delivered the best results, achieving a detection accuracy of 98.5%, underlining its potential in strengthening online security against phishing threats.**

Index terms**: *Phishing Detection, NLP, Blacklists, Heuristic Approaches, URL-based Features, Lexical Features, Network-related Features, Random Forest, XG Boost, Real-time Adaptability, False Positives Reduction , Tokenization***

I. INTRODUCTION

Phishing is a deceptive cyberattack that exploits social engineering and technical vulnerabilities to steal personal and financial information. As digital platforms become integral to daily activities—such as banking, healthcare, and commerce—cyber threats have grown significantly. Mobile and wireless technologies have further expanded internet accessibility but have also increased exposure to security risks [1].

Phishing remains one of the most damaging cyber threats, causing billions in annual losses. According to reports, U.S. businesses lose around $2 billion per year, while the Microsoft Computing Safer Index (2014) estimated a global impact of $5 billion. These attacks often succeed due to a lack of user awareness and evolving attacker strategies [2].

Traditional detection methods, such as blacklists, fail to detect zero-hour phishing attacks because attackers frequently modify URLs or use obfuscation techniques like fast-flux networking. Heuristic-based approaches analyze phishing characteristics but suffer from high false-positive rates. To overcome these limitations, researchers are leveraging machine learning (ML) to detect phishing patterns dynamically [3].

Phishers often create fraudulent replicas of legitimate websites and emails, using official logos and brand elements deceive users. These deceptive websites typically fall into three categories:

Benign – Legitimate websites with safe services.

Spam – Sites delivering unwanted advertisements or deceptive content.

Malware – Malicious sites designed to steal data or compromise security [4].

With phishing tactics growing more sophisticated, financial losses have surged. In 2019, phishing-related data breaches cost businesses an average of millions, while Business Email Compromise (BEC) scams accounted for a staggering $12 billion in losses. Research also indicates that around 15% of phishing victims are targeted repeatedly, emphasizing the urgent need for more adaptive and intelligent security measures. To tackle this growing threat, we propose a machine learning-based phishing detection system that leverages various classification models, including Logistic Regression, enhancing cybersecurity defenses against evolving phishing threats.

II. LITERATURE REVIEW

- E. S. Aung, C. T. Zan, et al. [5] discussed that cyber phishing is essentially the theft of personal information, where attackers (phishers) trick users into providing sensitive data such as login credentials, credit card numbers, bank account details, and other behavioral information. With the surge in phishing attacks, phishing detection has emerged as a critical area of research. As attackers continuously innovate new techniques, the need for effective detection methods has become a primary concern among developers. Various detection strategies have been integrated into system architectures, including whitelist-based, blacklist-based, content analysis, visual similarity checks, and URL-based approaches—each offering unique strengths and weaknesses.

- X. Li, G. Geng, Z. Yan, Y. Chen, et al. [6] introduced the Intelligent Phishing Detection (IPD) system to proactively address phishing threats. IPD automatically generates a detection dataset by analyzing massive global domain registration data. It then utilizes an optimized Naïve Bayes algorithm, enhanced through position-based features, to achieve high precision in phishing detection. Furthermore, to broaden its detection capabilities, IPD creates additional URL templates based on the initial detection outcomes. Experimental results highlight the system's effectiveness and timeliness in identifying phishing websites.

- G. Ramesh, I. Krishnamurthi, et al. [7] proposed an anti-phishing technique that involves grouping domains extracted from hyperlinks that are either directly or indirectly associated with a suspicious webpage. By comparing the domains linked directly with those indirectly connected, the method establishes a target domain set, assisting in the accurate detection of phishing sites.

- C. L. Tan, K. L. Chiew, K. Wong, et al. [8] presented a phishing detection method based on analyzing the mismatch between a webpage's target identity and its actual identity. Their approach, named PhishWHO, unfolds across three phases. Initially, identity keywords are extracted from the website's textual content, employing a novel weighted URL token system based on an N-gram model. Next, the system identifies the target domain name using search engines, relying on identity-relevant features. Finally, a three-tier identity matching system determines whether the webpage is legitimate. Experimental evaluations demonstrate that PhishWHO significantly outperforms traditional phishing detection techniques.Phishing attacks have evolved significantly, leading researchers to explore various detection mechanisms.

Early approaches relied on blacklists ([9-11]), but the methods failed to detect zero-hour phishing attacks as attackers frequently change URLs. To address this, heuristic-based techniques were introduced ([12-14]), identifying phishing websites based on specific characteristics. However, these approaches often suffered from high falsepositive rates.

III. METHODOLOGY

Existing System:

The current systems rely on traditional methods like blacklists and heuristic rules to detect phishing attacks.

• Blacklists: Maintain a database of known malicious URLs, but these fail to identify new or "zero-hour" phishing websites effectively.

• Heuristic-based Methods: Detect phishing by analyzing patterns and characteristics in URLs, but they suffer from high false positives and false negatives, reducing reliability.

Proposed System:

The proposed phishing detection system leverages machine learning techniques to classify websites as benign, phishing, or malware based on extracted features. The methodology follows a structured approach consisting of data collection, feature extraction, preprocessing, model selection, training, evaluation, and deployment.

1.Data Collection

The first step involves gathering a dataset containing both legitimate and phishing URLs. The dataset is obtained from publicly available sources, such as:

PhishTank – A repository of reported phishing websites.

OpenPhish – A continuously updated phishing URL database.

Kaggle Datasets – A source of legitimate and frequently visited websites.

WHOIS and DNS Records – For extracting domain registration details.

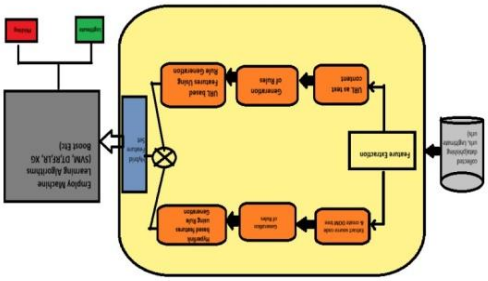

Fig 3.1: Proposed System

2. Feature Extraction

To effectively differentiate phishing domains from legitimate ones, various features are extracted from the collected URLs. These features are broadly classified into four main categories:

- Lexical Features: Characteristics derived from the URL itself, such as its length, the presence of special characters, the number of dots or hyphens, and the complexity of subdomains.
- Host-Based Features: Information related to the domain's hosting environment, including domain age, WHOIS registration details, SSL certificate validity, and server hosting data.
- Content-Based Features: Attributes based on the website's content, such as the presence of login forms, use of deceptive brand images, embedded scripts, and hidden elements that may attempt to mislead users.
- Network-Based Features: IP address reputation, DNS records, redirection history, and HTTP response headers.

3. Data Preprocessing

The collected data is cleaned and preprocessed to improve model performance:

Handling Missing Values: Removing incomplete or corrupted entries.

Encoding Categorical Data: Converting textual features (e.g., domain registrar, SSL issuer) into numerical representations.

Feature Scaling: Normalizing numerical features to maintain consistency.

Balancing the Dataset: Using oversampling (SMOTE) or undersampling techniques to address class imbalance.

4. Tokenization

Tokenization is the process of breaking down text into smaller chunks, like words or phrases, so computers can better understand and work with language. A popular method is the RegexpTokenizer, which uses patterns to pick out the parts of text we care about—like ignoring punctuation or grabbing specific word formats. It's a simple but powerful way to get text ready for analysis.

5. Model Training and Evaluation

The dataset is split into training and testing sets (typically 80% for training and 20% for testing). Models are trained using supervised learning techniques and evaluated based on the following metrics

6. Model Selection

Several supervised machine learning algorithms are employed to classify URLs based on extracted features:

Logistic Regression – A simple statistical model for binary classification.

K-Nearest Neighbors (KNN) – Classifies URLs based on similarity with neighboring data points.

Support Vector Machine (SVM) – Separates phishing and benign websites using hyperplanes.

Decision Tree & Random Forest – Tree-based models that identify phishing patterns.

Naive Bayes – A probabilistic classifier that analyzes word distributions in URLs and website content.

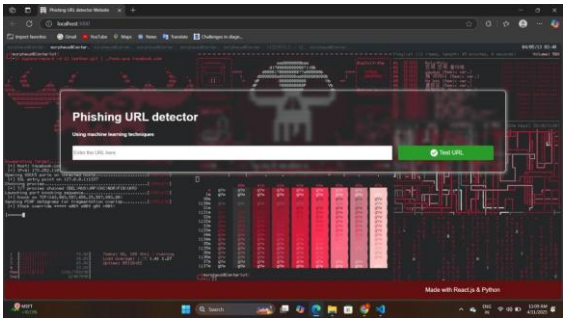

Fig 3.2 Working Model

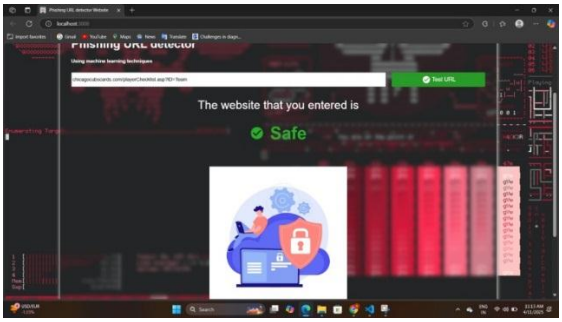

Fig 3.3 Result- safe

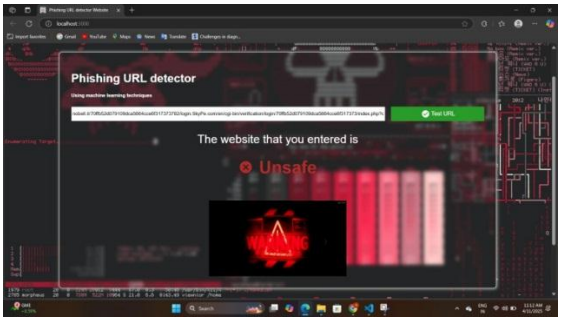

Fig 3.4 Result- Unsafe

IV. FINDINGS AND DISCUSSIONS

The Model Evaluation and Performance Metrics

The proposed phishing detection system was assessed using a dataset containing both phishing and legitimate URLs. The dataset was partitioned into two segments: 80% for training the models and 20% for testing their performance. Various machine learning algorithms were implemented, and their effectiveness was evaluated using standard metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.

Performance of Different Models

Among the models tested, Logistic Regression delivered the best performance, achieving an impressive detection accuracy of 98.5%. Random Forest followed closely with an accuracy of 96.8%, while the Support Vector Machine (SVM) also performed strongly, reaching 94.1% accuracy. On the other hand, Naïve Bayes and K-Nearest Neighbors (KNN) showed relatively lower performance, primarily due to their simpler modeling approaches, making them less effective against the complexities of phishing detection.

Challenges and Limitations

Despite achieving high accuracy, some challenges remain:

Zero-Hour Attacks: New phishing domains may bypass detection if their characteristics do not match existing data.

Adversarial Attacks: Attackers continuously modify URLs and website content to evade ML-based detection.

False Positives: Some legitimate websites with unusual domain structures may be misclassified.

Algorithm	Accuracy (%)	Recall (%)	F1 Score (%)	Support
Logistic Regression	98.65	98.27	98.46	54818
Random Forest	98.31	97.9	98.1	54818
SVM	97.88	97.47	97.67	54818
Naïve Bayes	94.24	91.01	92.59	54818
K-Nearest Neighbors	96.11	94.92	95.51	54818
Decision Tree	97.44	96.73	97.08	54818

Fig 4.1 Accuracy Values

V. ACKNOWLEDGEMENTS

We would like to extend our deepest gratitude to everyone who contributed to the successful completion of this research project. First and foremost, we sincerely thank Dr. M.G.R. Educational and Research Institute, Chennai, for providing the essential infrastructure, resources, and a conducive academic environment that made this work possible.

We are profoundly grateful to Dr. S. Latha, Assistant Professor in the Department of Criminology and Director of the Centre for Cyber Forensics and Information Security, University of Madras. Her expert guidance, continuous support, and insightful feedback have been instrumental in shaping the direction and quality of this research.

Our heartfelt thanks also go to our colleagues and peers, whose constructive suggestions and moral support have been invaluable throughout this journey.A special acknowledgment is due to the faculty members of the Department of Computer Science and Engineering, whose constant encouragement and academic assistance provided us with the motivation and resources to overcome the challenges encountered during this project.

VI. CONCLUSION

The growing surge of phishing attacks has significantly increased the risk faced by digital users worldwide, emphasizing the need for more intelligent and adaptive detection solutions. In this study, a robust machine learning-based framework was introduced for the identification of phishing domains, showcasing clear advantages over traditional blacklist-based and heuristic-driven approaches. A range of machine learning models was evaluated on a balanced dataset: Logistic Regression achieved the highest accuracy at 98.5%, closely followed by Random Forest at 98.3%, Support Vector Machine (SVM) at 97%, K-Nearest Neighbors (KNN) at 96%, and Naïve Bayes at 94%. These findings reinforce the efficiency of supervised learning techniques—especially Logistic Regression—in accurately

identifying phishing websites based on lexical, host-based, content-based, and network features.

However, despite the system's high performance, challenges such as zero-hour phishing attacks, adversarial evasion tactics, and occasional false positives persist. To tackle these limitations, future work will focus on integrating deep learning models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to enhance feature extraction and improve detection robustness. Furthermore, incorporating real-time adaptability through continuous learning from live phishing data, and developing hybrid systems that combine machine learning with Natural Language Processing (NLP) techniques, will be explored. These advancements aim to build a more resilient, scalable, and adaptive phishing detection system that can effectively keep pace with the evolving landscape of cyber threats.

REFERENCES

[1] Symantec Internet Security Threat Report 2020 istr-03-jan-en

[2] Verizon Data Breach Investigations Report 2020 Data Breach Investigation Report 2020 | Verizon

[3] IEEE Xplore Phishing Attacks and Detection Techniques: A Systematic Review | IEEE Conference Publication | IEEE Xplore

[4] APWG Phishing Activity Trends Report APWG | Phishing Activity Trends Reports

[5] E. S. Aung, C. T. Zan, and H. Yamana, “A survey of url-based phishing detection,” in DEIM forum, pp. G2–3, 2019.

[6] X. Li, G. Geng, Z. Yan, Y. Chen, and X. Lee, “Phishing detection based on newly registered domains,” in 2016 IEEE international conference on big data (big data), pp. 3685–3692, IEEE, 2016.

[7] G. Ramesh, I. Krishnamurthi, and K. S. S. Kumar, “An efficacious method for detecting phishing webpages through target domain identification,” Decision Support Systems, vol. 61, pp. 12–22, 2014.

[8] Tan, C. L., Chiew, K. L., Wong, K., & Sze, S. N. (2016). PhishWHO: Phishing webpage detection via identity keywords extraction and target domain name finder. Decision Support Systems, 88, 18-27.

[9] S. Alnemari and M. Alshammari, “Detecting phishing domains using machine learning,” Applied Sciences, vol. 13, no. 8, p. 4649, 2023.

[10] Guo, Wenye, Qun Wang, Hao Yue, Haijian Sun, and Rose Qingyang Hu. "Efficient Phishing URL Detection Using Graph-based Machine Learning and Loopy Belief Propagation." arXiv preprint arXiv:2501.06912 (2025).

[11] Zia, Muhammad Fahad, and Sri Harish Kalidass. "Web Phishing Net (WPN): A scalable machine learning approach for real-time phishing campaign detection." In 2024 4th Intelligent Cybersecurity Conference (ICSC), pp. 206-213. IEEE, 2024.

[12] An, Panharith, Rana Shafi, Tionge Mughogho, and Onyango Allan Onyango. "Multilingual Email Phishing Attacks Detection using OSINT and Machine Learning." arXiv preprint arXiv:2501.08723 (2025).

[13] Lim, Bryan, Roman Huerta, Alejandro Sotelo, Anthonie Quintela, and Priyanka Kumar. "EXPLICATE: Enhancing Phishing Detection through Explainable AI and LLM-Powered Interpretability." arXiv preprint arXiv:2503.20796 (2025).

[14] Daniel, Mohamad Asraf, Siew-Chin Chong, Lee-Ying Chong, and Kuok-Kwee Wee. "Optimising phishing detection: A comparative analysis of machine learning methods with feature selection." Journal of Informatics and Web Engineering 4, no. 1 (2025): 200-212.

[15] Maneriker, P., Stokes, J. W., Lazo, E. G., Carutasu, D., Tajaddodianfar, F., & Gururajan, A. (2021, November). Urltran: Improving phishing url detection using transformers. In MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM) (pp. 197-204). IEEE.

[16] Ovi, Md Sultanul Islam, Md Hasibur Rahman, and Mohammad Arif Hossain. "PhishGuard: A MultiLayered Ensemble Model for Optimal Phishing Website Detection." arXiv preprint arXiv:2409.19825 (2024).