Phishing Detection Using Machine Learning, Web Scraping, and Real-Time Alerts

Siya Patil[1], Shravani Pawar[2], Sujeet Pawar[3] , Abhishake Sawant[4] ,Prof. Amol T. Take[5]

[1,2,3,4,]Undergraduate Students Computer Engineering Department , Zeal College of Engineering and Research
[5]Professor, Computer Engineering Department , Zeal College of Engineering and Research

---

Abstract -Phishing is one of the most prevalent and persistent threats to cybersecurity, affecting millions of users and organizations globally. This survey explores recent advancements in phishing detection methods, focusing on machine learning, web scraping, and real-time alerts delivered through browser extensions. By integrating these techniques, the proposed system improves the detection of phishing websites and phishing emails in real-time, reducing the user's exposure to cyber threats. The paper also reviews the limitations of existing solutions and provides recommendations for future research directions.

Key Words: Phishing Detection, Machine Learning, Web Scraping,Real-Time Alerts, Browser Extension, Cybersecurity

1.INTRODUCTION

Phishing attacks are one of the most dangerous cybercrimes, affecting millions of users and organizations worldwide. Attackers create deceptive websites or emails to trick users into providing personal information, such as passwords or credit card numbers. Traditional detection mechanisms, such as blacklists, often fail to detect new phishing sites in real-time.
Machine learning and web scraping have emerged as effective techniques to enhance phishing detection. This survey discusses recent advancements and proposes an integrated approach that combines these technologies to provide real-time alerts to users via a browser extension. By leveraging machine learning models trained on phishing and legitimate website data, along with web scraping for real-time validation, the system aims to provide proactive protection

2.OBJECTIVE

The primary goal of this research is to develop a robust phishing detection system by integrating machine learning algorithms, web scraping techniques, and real-time alert mechanisms. The system will offer real-time protection through a browser extension and focus on detecting both phishing websites and phishing emails.

Objectives include:

1. Developing a machine learning model that analyzes website features such as URLs, SSL certificates, and HTML content to classify phishing websites.
2. Using web scraping to validate website legitimacy by extracting and analyzing real-time data.
3. Detecting phishing attempts in emails by analyzing suspicious patterns.
4. Implementing a browser extension to deliver real-time alerts, warning users about phishing websites and emails
5.

3.LITERATURE SURVEY

Table -1: Literature Survey table

Paper	Objective	Limitation
Castaño et al. (2023) [2]	Introduced PhiKitA, a dataset linking phishing kits with websites to improve phishing campaign detection.	The dataset had incomplete coverage due to cloaking techniques used by phishers, reducing data accuracy.
Zieni et al. (2023) [3]	Provided a comprehensive review of phishing website detection, focusing on list-based, similarity-based, and ML-based methods.	List-based methods were reactive, while similarity-based approaches had high computational and storage demands.
Mittal et al. (2023) [4]	Developed a model using BERT for feature extraction and logistic regression for phishing domain detection.	system required regular updates due to evolving phishing techniques. Features were also critical for the model's success.
Wei & Sekiya (2022) [1]	To evaluate the performance of various machine learning (ML) and deep learning (DL) algorithms for phishing detection. Ensemble ML methods were found to be more efficient in both detection	Deep learning models had high computational demands, limiting real-time use. The study did not focus on adversarial attacks

	accuracy and computational efficiency.	
Do et al. (2022) [5]	Proposed a taxonomy of deep learning algorithms for phishing detection and conducted a systematic review of 81 papers.	Challenges included long training times and the need for manual parameter tuning, impacting real-time detection.
Purwanto et al. (2022) [6]	Introduced PhishSim, a feature-free phishing detection method using Normalized Compression Distance (NCD) to compare HTML content.	Relied on HTML similarity, which limited its ability to detect zero-day phishing attacks.
Kabla et al. (2022) [7]	Developed Eth-PSD, a machine learning-based phishing scam detection system for Ethereum.	Focused solely on Ethereum scams, not addressing phishing on other blockchain platforms.
Dutta (2021) [8]	Proposed a machine learning-based URL detection technique using Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) models.	Performance decreased with larger datasets or when exposed to different types of phishing attacks.
Tang & Mahmoud (2021) [9]	Conducted a survey of ML techniques for phishing website detection, comparing approaches like decision trees, SVM, and neural networks.	The study lacked experimental results and highlighted the computational complexity of some methods.
Yang et al. (2019) [10]	Introduced a multidimensional feature-driven phishing detection method using CNN-LSTM models.	High detection time due to complex feature extraction, affecting real-time application.

4. PROPOSED SYSTEM ARCHITECTURE

The following diagram illustrates the architecture of the phishing detection system, which integrates machine learning, web scraping, and real-time threat analysis. The architecture begins with the browser extension that acts as the first line of defense, collecting data from websites and emails. The data is then sent to the primary backend, where initial security checks, such as SSL certificate validation, are performed. Based on the content type—whether it's a website or an email—the backend applies different detection techniques. For websites, all URLs and links are analyzed through the machine learning model to check for phishing indicators. For emails, the system uses a screen-capturing method to detect embedded links, which are then scanned by the same machine learning model. The unique aspect of this architecture lies in its dynamic approach, seamlessly adapting its methods to analyze both websites and emails, while providing real-time alerts to users. The entire process ensures that users receive immediate notifications about potential threats, ensuring a proactive defense against phishing attacks.

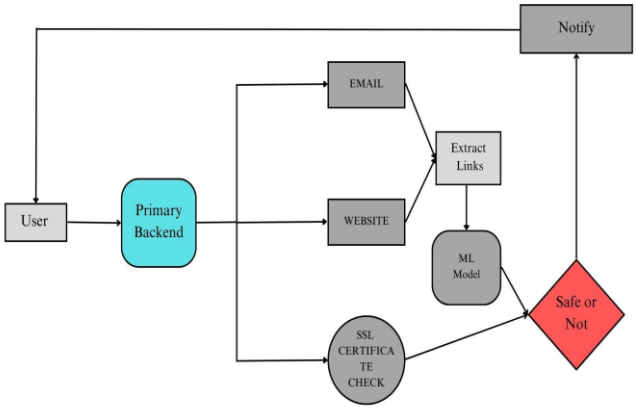

Fig -1: Proposed System Architecture.

5. METHODOLOGY

1. Browser Extension Initialization
When the user opens a website or an email, the browser extension immediately activates and sends the URL or email content to the backend server for analysis.

2. SSL Certificate Check
The primary backend scans the website's SSL certificate. If the SSL certificate is more than 2 years old, the system marks the site as potentially unsafe but does not classify it as phishing yet.

3. The backend then determines whether the content is a website or an email:

- If it is a website, the links and the main URL are extracted and sent for further analysis.
- If it is an email, a screen-capturing method (like Google Lens) is used to detect any embedded links without requiring the user's credentials.

4. Link and Content Analysis via Machine Learning Model
If the content is a website, the URL and all embedded links are sent to a machine learning (ML) model trained to detect phishing indicators (e.g., suspicious domains, URL length, SSL status). If the content is an email, only the links are analyzed by the ML model.

5. Phishing Classification
Based on the ML model's analysis, the system classifies the content (website or email) as:

- Phishing: The user is notified with a warning if the content is identified as phishing.
- Safe: The user is notified that the content is safe.

6. User Notification and Reporting
The system informs the user of the number of potential

phishing threats and provides appropriate warnings if any were detected.

6. COMPARISON WITH EXISTING SYSTEM

Table -2: Table below compares the proposed system with existing solutions, highlighting the integration of machine learning, web scraping, and real-time alerts.

Aspect	Existing Solutions	Proposed System
Phishing URL Detection	Google Safe Browsing, PhishTank	Machine learning models combined with real-time validation through web scraping [2]
Website Phishing Detection	Netcraft, heuristic-based detection	Web scraping integrated with machine learning for enhanced accuracy [5]
Email Phishing Detection	Gmail, Outlook (content-based filtering)	NLP-based email analysis combined with web scraping data [4]
Real-Time Alerts	PhishWHO, Google Safe Browsing	Real-time alerts provided through a browser extension [3]

7. EXPECTED OUTCOME

The proposed system is expected to achieve the following outcomes:

1. High-Accuracy Detection: The machine learning model will provide high accuracy in phishing website detection.
2. Real-Time Protection: Users will receive real-time alerts through a browser extension, reducing exposure to phishing threats.
3. Improved Website Validation: Web scraping will enhance website legitimacy validation, improving detection accuracy [5].

8. APPLICATION

The system can be applied in various sectors to enhance cybersecurity:

1. Email Clients: Integration with email clients like Gmail and Outlook to detect and block phishing emails [4].
2. Web Browsers: Real-time phishing protection for users browsing the web [2].
3. Network Security: Deployment at network gateways to prevent phishing websites from reaching users within an organization [7].
4. Banking: Protecting online banking customers from phishing attacks targeting login credentials and financial information [8].

9. CONCLUSION

This research paper presents a comprehensive system for phishing detection using machine learning, web scraping, and real-time alerts. The integration of these technologies enhances detection accuracy and provides proactive protection for users. Future research can focus on improving the system's scalability and real-time performance to handle large-scale, sophisticated phishing attacks.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Prof. Amol T. Take, Computer Engineering Department, Zeal College of Engineering and Research, for his continuous guidance, encouragement, and support throughout this research. His invaluable insights and expertise in the field of cybersecurity played a significant role in shaping the direction of this project.We are also grateful to the Computer Engineering Department of Zeal College of Engineering and Research for providing us with the resources and infrastructure necessary to conduct our research. Finally, we would like to extend our heartfelt thanks to our peers and mentors who contributed their valuable feedback and suggestions

REFERENCES

1. Y. Wei and Y. Sekiya, "Sufficiency of Ensemble Machine Learning Methods for Phishing Websites Detection," in IEEE Access, vol. 10,pp.124103-124113,2022,doi: 10.1109/ACCESS.2022.3224781.
2. F. Castaño, E. F. Fernañdez, R. Alaiz-Rodríguez and E. Alegre, "PhiKitA: Phishing Kit Attacks Dataset for Phishing Websites Identification," in *IEEE Access*, vol. 11, pp. 40779-40789, 2023, doi: 10.1109/ACCESS.2023.3268027.
3. R. Zieni, L. Massari and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites," in *IEEE Access*, vol. 11, pp. 18499-18519, 2023, doi: 10.1109/ACCESS.2023.3247135.
4. K. Mittal, K. S. Gill, R. Chauhan, M. Singh and D. Banerjee, "Detection of Phishing Domain Using Logistic Regression Technique and Feature Extraction Using BERT Classification Model," *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, Bangalore, India, 2023, pp. 1-5, doi: 10.1109/SMARTGENCON60755.2023.10442975.
5. N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions," in *IEEE Access*, vol. 10, pp. 36429-36463, 2022, doi: 10.1109/ACCESS.2022.3151903.
6. R. W. Purwanto, A. Pal, A. Blair and S. Jha, "PhishSim: Aiding Phishing Website Detection With a Feature-Free Tool," in *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1497-1512, 2022, doi: 10.1109/TIFS.2022.3164212.
7. A. H. H. Kabla, M. Anbar, S. Manickam and S. Karupayah, "Eth-PSD: A Machine Learning-Based Phishing Scam Detection Approach in Ethereum," in *IEEE Access*, vol. 10, pp. 118043-118057, 2022, doi: 10.1109/ACCESS.2022.3220780.

8.Dutta AK. Detecting phishing websites using machine learning techniques.*PLoS One*. 2021;16(10):e0258361. Published 2021Oct11.doi:10.1371/journal.pone.0258361

9.Tang L, Mahmoud QH. A Survey of Machine Learning-Based Solutions for Phishing Website Detection. *Machine Learning and Knowledge Extraction*. 2021; 3(3):672-694. https://doi.org/10.3390/make3030034

10.P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in *IEEE Access*,vol.7,pp.15196-15209,2019,doi:10.1109/ACCESS.2019.2892066.