International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

DETECTING PHISHING WEBSITES WITH MACHINE LEARNING

[1]JALA SHILPA, [2] KAMMAMPATI RAKESH REDDY

[1]Assistant Professor &HOD, CSE, Tallapadmavathi College of Engineering, Somidi, Kazipet, Hanumakonda – 506003.Email-id: jala.shilpa2@gmail.com.

[2]Research Scholar, H.no: 24UC1D5803, CSE, Tallapadmavathi College of Engineering, Somidi, Kazipet, Hanumakonda -506003,Emai-id: kammampartirakeshreddy1405@gmail.com.

ABSTRACT

Phishing website detection using machine learning (ML) has become an effective approach to combating cybercrime and protecting users from malicious online threats. By leveraging ML algorithms, phishing websites can be accurately identified and classified, thereby ensuring the confidentiality, integrity, and availability of user and organizational data. Phishing is often the initial stage of cyberattacks, exploiting human vulnerability through social engineering to obtain sensitive information such as passwords and financial details.

This study proposes a model for phishing website detection based on feature extraction and ML algorithms. The datasets used for training and evaluation were obtained from PhishTank and the University of New Brunswick Dataset Bank, containing verified phishing, spam, malware, and defacement URLs. A combination of machine learning models and deep neural networks was implemented to classify URLs as legitimate or phishing. Experimental results demonstrate high classification accuracy, highlighting the effectiveness of the proposed approach. The developed model can assist individuals and organizations in verifying the authenticity of web links, thereby reducing the risk of phishing-related cyber threats.

Index Terms:-Phishing detection, machine learning, cybersecurity, URL classification, neural networks, feature extraction.

Received: 23-09-2025 Accepted: 27-10-2025 Published: 03-11-2025

1. INTRODUCTION

The Internet has become an integral part of modern life, serving as a vital medium for communication, commerce, and information exchange, particularly through social media platforms. However, as noted by Pamela (2021), while the Internet is a vast network of interconnected computers containing valuable data, the human element often represents the weakest link in cybersecurity. When users inadvertently disclose sensitive information or grant unauthorized access to their systems, even the most robust security mechanisms become vulnerable.

Social engineering attacks exploit this human vulnerability by manipulating users into compromising security. Imperva (2021) defines social engineering as one of the most prevalent forms of cyberattacks, wherein attackers deceive victims into interacting with malicious content such as fraudulent emails, instant messages, or text messages disguised as legitimate communications. Once a victim clicks on a malicious link, it may lead to the installation of malware, system hijacking via ransomware, or the unauthorized disclosure of sensitive information.

With rapid technological advancements, the frequency and sophistication of computer security threats have increased dramatically (Verkijika & De Wet, 2020). Users must therefore understand how phishing attacks operate and adopt effective defense strategies to protect themselves from being deceived. Phishing is not the only social engineering technique; attackers also employ several related

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

methods to extract personal information, including:

Vishing (Voice Phishing): This involves the attacker calling the victim to obtain confidential banking or personal information, often using spoofed caller IDs to appear legitimate (Symantec, 2020).

Smishing (SMS Phishing): This method utilizes text messages to lure victims into clicking links that redirect to phishing websites or download malware (KnowBe4, 2021).

Ransomware: In this type of attack, users are locked out of their data or systems until a ransom is paid. Ransomware is often distributed through phishing emails or compromised websites (Cisco, 2021).

Phishing has become a major concern for cybersecurity researchers because attackers can easily create fake websites that closely resemble legitimate ones. While security experts can often identify such sites, ordinary users frequently fall victim to these deceptive tactics, leading to the theft of sensitive information, particularly banking credentials (Gupta et al., 2018).

The success of phishing attacks largely stems from inadequate user awareness. Since phishing exploits psychological and behavioral weaknesses rather than purely technical flaws, complete prevention is difficult. Therefore, improving phishing detection techniques is essential for reducing cyber threats.

The primary objective of this research is to develop a system capable of detecting phishing websites created by attackers to steal user data. By identifying and alerting users about these threats in real time, the proposed model aims to promote safer browsing and protect users from unauthorized access or misuse of personal information.

2.LITERATURE SURVEY

Rao, Umarekar, and Pais [1] defined phishing as an attack that deceives online users through fake websites, emails, or other internet services to steal sensitive information such as passwords and credit card details. Similarly, Purbay and Kumar [2] described phishing as a fraudulent online activity aimed at stealing users' personal data such as usernames, passwords, and banking credentials. Their study proposed a detection approach that analyzes different components of URLs using both machine learning (ML) and deep learning techniques.

Alkawaz, Steven, and Hajamydeen [3] developed a phishing detection system focused on identifying blacklisted URLs—commonly associated with phishing websites—to alert users during browsing. This approach enhances web authentication and helps prevent users from being deceived by fraudulent sites.

Dey et al. [4] conducted a comparative analysis of various ML algorithms for detecting phishing emails and websites. Their experimental results demonstrated that the Multinomial Naïve Bayes (MultinomialNB) algorithm achieved the highest efficiency (98.06%) for phishing email detection, while the Decision Tree Classifier achieved 95.41% efficiency for phishing website detection.

Jain and Gupta [5] presented PHISH-SAFE, a machine learning–based anti-phishing system that relies on URL-based features to detect phishing websites. Their work demonstrated that URL-level analysis provides a robust foundation for automated phishing detection systems.

Nakano et al. [6] explored phishing mitigation through ML algorithms, emphasizing how fraudulent websites and emails can be automatically detected by analyzing their structural and behavioral attributes. Likewise, Jalil, Usman, and Fong [7] examined multiple phishing detection approaches—including artificial intelligence (AI)–based, third-party, heuristic, and content-based techniques—to address the growing persistence of phishing attacks.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Gupta et al. [8] tested several ML classifiers for phishing detection and achieved a high accuracy of 99.57% using the Random Forest algorithm. Their findings demonstrated the effectiveness of ensemble-based classifiers for identifying phishing URLs.

Sakhare, Bangare, and Purandare [9] extracted URL features such as length, dots, slashes, numbers, and special characters for model training. Their system incorporated real-time monitoring, enabling adaptive learning and improved resilience against evolving phishing tactics. Nanda and Goel [10] focused on phishing attacks that mimic official websites of government agencies, financial institutions, and e-commerce platforms to obtain users' personal information, such as credit card details and login credentials.

Mittal [11] highlighted that URL-based phishing attacks often exploit email, websites, and SMS channels. His study emphasized that feature engineering and statistical analysis of URLs are critical for identifying phishing links with high precision. Alshingiti et al. [12] discussed the broader challenge of cybersecurity, noting that phishing remains one of the most common methods of stealing personal information through imitation of legitimate websites.

John Smith [13] proposed a novel ML-based approach for phishing URL detection, utilizing advanced feature engineering and classifier evaluation techniques. His system achieved a high detection rate while maintaining a low false-positive rate, demonstrating the potential of ML in proactive phishing prevention.

3. EXISTING SYSTEM

The existing system of phishing detection techniques suffers from low detection accuracy and a high false alarm rate, especially when new and sophisticated phishing approaches are introduced. Among these, the most common method used is the blacklist-based approach, which is inefficient in responding to emerging phishing attacks. Since registering new domains has become increasingly easy, no blacklist can ensure a complete and up-to-date database for phishing detection.

In most cases, existing systems for phishing attack detection rely on a combination of manual inspection, blacklisting, and basic heuristic analysis. In the absence of advanced machine learning models, traditional methods depend heavily on human experts manually examining URLs and web content to identify potential phishing sites. Additionally, these systems may utilize blacklists containing known phishing URLs and domain names, which are regularly updated based on community reports and threat intelligence feeds.Heuristic analysis techniques are also used to detect common phishing indicators such as misspelled URLs, suspicious domain registrations, or abnormal web page structures. However, these conventional approaches often fail to keep pace with the rapidly evolving nature of phishing attacks. They lack the efficiency, scalability, and real-time detection capabilities that machine learning–based systems can provide in distinguishing between legitimate and phishing URLs with higher accuracy and reduced false positives.

DISADVANATGES

- Machine learning models require large amounts of labeled data to train effectively.
- Training and running machine learning models can be computationally expensive.

4. PROPOSED SYSTEM

The proposed system for phishing detection integrates rule-based analysis with machine learning techniques to accurately classify URLs as legitimate or phishing. It begins with data collection from multiple sources, creating datasets that capture features of both types of websites, focusing on URLs and domain-related

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

information. Key features such as IP address, URL length, presence of special characters, and domain identity are extracted for analysis, forming the basis of a rule-based inspection.

To enhance detection, a K-Means clustering algorithm is applied to identify patterns and characteristics typical of phishing URLs. A web application is developed as the user interface, built with HTML, CSS, and JavaScript, while the trained machine learning models serve as the back-end, analyzing user-input URLs and providing real-time feedback on their legitimacy. The system employs multiple machine learning models, including Decision Tree, Support Vector Machine, Random Forest, XGBoost, Gradient Boosting Classifier, and Multilayer Perceptron. These models were selected after comparative evaluations and are trained on website content–based features extracted from both phishing and legitimate datasets.

The implementation uses Python along with libraries such as pandas, scikit-learn, NumPy, seaborn, matplotlib, and python-whois, enabling a flexible and efficient phishing detection framework. This approach ensures higher accuracy, reduced false positives, and real-time protection for users against evolving phishing attacks.

ADVANTAGES:

- Machine learning models can adapt to new phishing techniques and patterns, improving effectiveness over time.
- They can be easily scaled to handle large volumes of data, making them suitable for real-time phishing detection.

5.SYSTEM ARCHITECTURE

Architectural design focuses on organizing a system and designing its overall structure, illustrating how various components work together to achieve the system's main objectives. It involves identifying the sub-systems that constitute the system and defining the framework for their control and communication.The proposed phishing detection system is depicted in Figure 1.which provides a graphical overview of its architecture. In this system, a user enters a URL link, which is then processed through multiple trained machine learning and deep neural network models. Among these models, the one achieving the highest accuracy is selected for deployment.The selected model is implemented as an API (Application Programming Interface) and integrated into a web application. This design allows users to interact seamlessly with the system through the web interface, which is accessible across various devices, including computers, tablets, and mobile devices.This architecture ensures that URL inputs are evaluated in real time, leveraging the most accurate model for detection. By combining multiple models with a dynamic selection mechanism and a user-friendly interface, the system provides robust and scalable phishing detection while offering accessibility across different platforms.

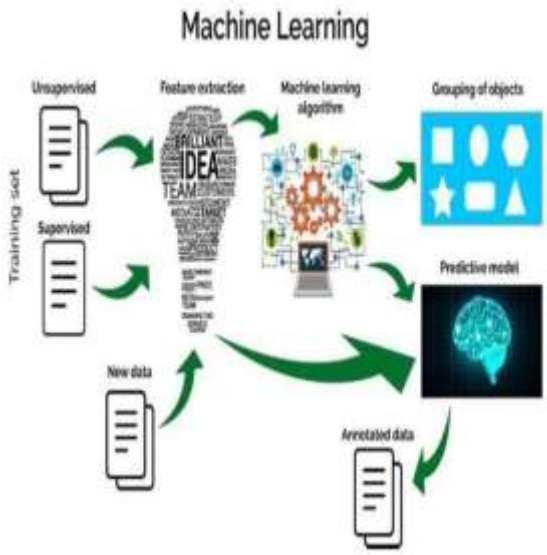

Fig. System Architecture

6.MODULES

- **Upload Dataset:** Import phishing and legitimate URL datasets from open-source sources.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

- **Data Preprocess:** Clean and prepare the dataset for analysis, handling missing or inconsistent data.
- **Feature Extraction:** Extract relevant features from URLs such as length, presence of special characters, SSL usage, etc.
- **Model Generation:** Train supervised ML and deep learning models (Decision Tree, Random Forest, SVM, XGBoost, MLP, Autoencoder) on the dataset.
- **Prediction:** Use the trained models to classify URLs as phishing or legitimate.

7.SCREENSHOTS

Home Page

Upload Url Data

Result

8.CONCLUSION

The most effective way to protect users from phishing attacks is through awareness and education. Internet users must follow security guidelines provided by experts and avoid blindly clicking on links that request sensitive information. Checking the URL before entering a website is essential.The proposed system provides a robust mechanism for detecting phishing websites using a combination of rule-based analysis and machine learning models, offering real-time classification and higher accuracy compared to traditional methods. Future improvements may include automatic detection of web pages, enhanced browser compatibility, and integration with mobile platforms. Additional features can be incorporated to better distinguish fake websites from legitimate ones.

The system's heuristics can be further enhanced to detect phishing attacks in the presence of embedded objects such as Flash content. Text and image extraction can be improved using Optical Character Recognition (OCR). Designing more effective inference rules for suspicious web pages and strategies for identifying phishing targets can further improve system performance. Combining dynamic and static features will also help achieve higher detection accuracy.

9.FUTURE SCOPE

By 2024 and beyond, phishing detection systems are becoming increasingly intelligent. Advanced machine learning algorithms can analyze large datasets to identify fake websites more accurately and quickly. These systems adapt to evolving phishing techniques, detecting threats before they can cause harm. Improvements in natural language processing (NLP) allow systems to evaluate website content for legitimacy.

Collaboration between cybersecurity experts, businesses, and government agencies is

enhancing the strength of these tools. Emerging technologies such as blockchain are being explored to verify website authenticity and prevent tampering. Future developments will focus on expanding these detection capabilities across multiple platforms, including web and mobile applications, to provide comprehensive protection against phishing attacks.

REFERENCES

1. Nanda, M., & Goel, S. (2024). URL based phishing attack detection using BiLSTM-gated highway attention block convolutional neural network. *Multimedia Tools and Applications*.
2. Jalil, S., Usman, M., & Fong, A. (2023). Highly accurate phishing URL detection based on machine learning. *Journal of Ambient Intelligence and Humanized Computing, 14*, 9233–9251.
3. Rao, R. S., Umarekar, A., & Pais, A. R. (2022). Application of word embedding and machine learning in detecting phishing websites. *Telecommunications*.
4. Nakano, H., et al. (2023). Canary in Twitter Mine: Collecting Phishing Reports from Experts and Nonexperts. *arXiv preprint arXiv:2303.15847*.
5. Mittal, S. (2023). Explaining URL Phishing Detection by Glass Box Models.
6. Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q.E.U., Saleem, K., & Faheem, M.H. (2023). A deep learning-based phishing detection system using CNN, LSTM, and LSTM-CNN. *Electronics, 12*(1), 232.
7. Purbay, M., & Kumar, D. (2021). Split Behavior of Supervised Machine Learning Algorithms for Phishing URL Detection. *Lecture Notes in Electrical Engineering, 683*.
8. Alkawaz, M. H., Steven, S. J., & Hajamydeen, A. I. (2020). Detecting Phishing Website Using Machine Learning. *16th IEEE International Colloquium on Signal Processing its Applications (CSPA)*.
9. Dey, N., Samhitha, S., Hariprasad, M., Anand, A., & Gadad, V. (2021). Analysis of machine learning algorithms by developing a phishing email and website detection model. *IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Bangalore.
10. Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In M. Bokhari, N. Agrawal, & D. Saini (Eds.), *Cyber Security. Advances in Intelligent Systems and Computing, 729*. Springer, Singapore.
11. Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. *Computers & Communications, 175*, 47–57.
12. Sakhare, N.N., Bangare, J.L., & Purandare, R.G. Phishing Website Detection Using Advanced Machine Learning Techniques.
13. Gandotra, E., & Gupta, D. (2021). An Efficient Approach for Phishing Detection using Machine Learning. *Algorithms for Intelligent Systems*. Springer, Singapore.
14. Hong, J., Kim, T., Liu, J., Park, N., & Kim, S.W. Phishing URL Detection with Lexical Features and Blacklisted Domains. *Autonomous Secure Cyber Systems*. Springer.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

15. Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J.P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics, 9*(9), 1514.
16. AlEroud, A., & Karabatis, G. (2020). Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks. *Proceedings of the Sixth International Workshop on Security and Privacy Analytics*.
17. Gupta, D., & Rani, R. (2020). Improving malware detection using big data and ensemble learning. *Computer Electronic Engineering, 86*, 106729.
18. Anirudha, J., & Tanuja, P. (2019). Phishing Attack Detection using Feature Selection Techniques. *Proceedings of ICCIP*, doi:10.2139/ssrn.3418542.
19. Sahingoz, O.K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications, 117*, 345–357.
20. Zamir, A., Khan, H.U., Iqbal, T., Yousaf, N., & Aslam, F. (2019). Phishing web site detection using diverse machine learning algorithms. *The Electronic Library, 38*(1), 65–80.
21. Almseidin, M., Zuraiq, A.A., Al-kasassbeh, M., & Alnidami, N. (2019). Phishing detection based on machine learning and feature selection methods. *International Journal of Interactive Mobile Technology, 13*(12), 171–183.
22. Tan, C.L., Chiew, K.L., & Wong, K. (2016). PhishWHO: phishing webpage detection via identity keywords extraction and target domain name finder. *Decision Support Systems, 88*, 18–27.
23. Gull, S., & Parah, S.A. (2019). Color image authentication using dual watermarks. In *Fifth International Conference on Image Information Processing (ICIIP)*, 240–245.
24. Giri, K.J., Bashir, R., & Bhat, J.I. (2019). A discrete wavelet based watermarking scheme for authentication of medical images. *International Journal of E-Health Medical Communication, 30*.