IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 5, March 2025

Machine Learning Algorithm-Based Feature Selection Optimisation for Phishing Website Detection

Sonali C. Bhabad, Krutika V. Karad, Nikita D. Kashmire
Guru Gobind Singh Polytechnic, Nashik, Maharashtra, India

Abstract: *Phishing attacks continue to be one of the most common cybersecurity risks, and attackers are always improving their strategies to trick users into divulging private and sensitive financial information. Machine learning (ML) has shown itself to be a useful technique in tackling the crucial problem of phishing website detection. However, the features used for training have a significant impact on how well ML model's function, and feature selection is essential to increasing model accuracy and efficiency. The main goal of this study is to employ machine learning techniques to optimise feature selection for phishing website identification.*

Keywords: Phishing

I. INTRODUCTION

Phishing attacks continue to be one of the most common cybersecurity risks, and attackers are always improving their strategies to trick users into divulging private and sensitive financial information. Machine learning (ML) has shown itself to be a useful technique in tackling the crucial problem of phishing website detection. However, the features used for training have a significant impact on how well ML model's function, and feature selection is essential to increasing model accuracy and efficiency. The main goal of this study is to employ machine learning techniques to optimise feature selection for phishing website identification.

We offer a thorough framework that makes use of both conventional and cutting-edge feature selection methods, such as Genetic Algorithms, Mutual Information, and Recursive Feature Elimination (RFE), in order to extract the most pertinent features from a wide range of unprocessed data, including URL structure, domain registration details, and website content. We seek to improve the performance of several ML classifiers, such as Support Vector Machines (SVM), Decision Trees, and Random Forests, by lowering the dimensionality of the dataset while preserving the discriminative ability of the features.

The experimental findings show that feature selection optimisation enhances the accuracy, precision, and recall of phishing detection models in addition to speeding up the training process. This work also emphasises the trade-off between feature reduction and classification performance, offering important information for creating phishing detection systems that are more effective and scalable. The results provide a new method for feature engineering in the field of phishing website detection, which makes it more applicable to practical uses where high detection rates and computational efficiency are essential.

Phishing, a pervasive form of cybercrime, continues to pose a significant threat to individuals and organizations worldwide. It involves deceptive attempts to acquire sensitive information such as usernames, passwords, credit card details, and other personal data. The methods employed by phishers are constantly evolving, making it crucial to understand the techniques used, the effectiveness of current detection methods, and the best strategies for prevention.

This analysis will synthesize findings from various research papers to provide a comprehensive overview of the current state of phishing attacks and the ongoing efforts to combat them.

Types and Techniques of Phishing Attacks

Phishing attacks manifest in various forms, each leveraging different techniques to deceive victims. A common approach is email phishing, where attackers send fraudulent emails designed to mimic legitimate communications from

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 5, March 2025

trusted sources. These emails often contain malicious links or attachments leading to fake websites that collect sensitive information. Spear phishing, a more targeted approach, involves customizing emails to specific individuals or organizations, increasing the likelihood of success. Another variant, smishing, utilizes SMS messages to deliver phishing attempts (Putra, 2024). The sophistication of these attacks has increased significantly, with phishers employing advanced techniques to bypass security measures (Nirmal, 2023). For example, the use of wildcard SSL certificates can evade Certificate Transparency (CT) checks (Nirmal, 2023). Furthermore, the persistence of recurring scams and the variations within phishing campaigns present challenges for rule-based filters. The attackers' ability to leverage legitimate services indirectly to deliver phishing attacks adds another layer of complexity.

The success of phishing attacks hinges on the attacker's ability to manipulate the victim's psychology and emotions. Attackers often exploit psychological vulnerabilities, such as trust and urgency, to influence victims' decisions. They might employ persuasive techniques based on principles of authority, scarcity, or social proof to increase the credibility of their messages. Understanding these psychological factors is crucial for developing effective countermeasures. The impersonation of legitimate websites, often indistinguishable to the average user, is a common tactic. The creation of replica web pages that closely resemble authentic websites is a significant loophole in the cyber world, allowing phishers to operate successfully. This highlights the need for improved user education and awareness training to enhance the ability of individuals to identify phishing attempts.

Detection Methods: A Comparative Analysis

Several approaches have been developed to detect phishing attacks, ranging from simple blacklist and whitelist techniques to sophisticated machine learning algorithms. List-based approaches rely on maintaining databases of known phishing websites and emails. However, these methods are limited by their inability to detect new or evolving phishing attempts.

Similarity-based methods compare the characteristics of suspected phishing websites or emails to known examples. However, the dynamic nature of phishing makes this approach challenging. Machine learning techniques, such as Random Forest, XGBoost, and Logistic Regression, have shown promising results in detecting phishing websites and URLs. Studies using these techniques have reported high accuracy rates, often exceeding 95%. For example, Vajratiya et al. achieved a remarkable 99.97% accuracy using mutual information-based logistic regression). Other studies have employed deep learning models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs), demonstrating high accuracy in identifying phishing websites. However, challenges remain, including the need for continuous model updates to adapt to evolving phishing techniques and the potential for attackers to develop methods to evade detection.

The selection of appropriate features for machine learning models significantly impacts their performance. Research indicates that certain features, such as URL characteristics, HTML elements, and domain-based features, are particularly effective in distinguishing between legitimate and phishing websites. Feature selection methods, such as the Info Gain Attribute Eval, can improve the efficiency and accuracy of phishing detection systems. Furthermore, hybrid approaches combining multiple machine learning techniques or incorporating human expertise can enhance detection capabilities. The development of a hybrid ensemble feature selection method, for example, has demonstrated improved performance compared to previous studies. These studies highlight the importance of a multi-layered approach to phishing detection that combines technological solutions with human vigilance.

Prevention Strategies: A Multifaceted Approach

Preventing phishing attacks requires a multifaceted approach that combines technological solutions with user education and organizational. Technological defences include email filters, two-factor authentication (2FA), and encryption. These measures can significantly reduce the effectiveness of phishing attacks, but they are not fool proof. Email filters can block suspicious emails, but advanced phishing techniques can often bypass these filters. 2FA adds an extra layer of security, making it more difficult for attackers to access accounts even if they obtain login credentials (Putra, 2024). Encryption protects sensitive information transmitted over the internet, reducing the risk of interception by attackers (Putra, 2024).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 5, March 2025

User education and awareness are crucial components of any effective phishing prevention strategy. Studies have shown that individuals with higher levels of cyber security awareness are less susceptible to phishing attacks. Training programs that educate users about common phishing techniques, such as identifying suspicious emails and URLs, can significantly reduce the risk of victimization. These programs should focus on fostering critical thinking skills and promoting scepticism towards unsolicited communications. Furthermore, personalized training programs tailored to the specific needs and knowledge gaps of individual users may be more effective than generic approaches. The development of explainable systems and platforms that facilitate real-world studies can contribute to improved phishing education and awareness.

Organizational policies also play a vital role in preventing phishing attacks. Strict policies and procedures regarding email and internet usage can help to minimize the risk of employees falling victim to phishing attempts. These policies should include guidelines for handling suspicious emails, reporting phishing attempts, and using secure passwords. Regular security audits and vulnerability assessments can identify weaknesses in organizational security infrastructure and help prevent phishing attacks. The implementation of robust incident response plans is also essential for mitigating the impact of successful phishing attacks. These plans should outline procedures for containing the attack, investigating the incident, and recovering from the damage. Post-attack analysis is critical for learning from past incidents and improving organizational resilience against future attacks.

The human element remains a critical vulnerability in phishing defences. Organizations often rely on a single-layer defence against information security threats, including phishing. This approach is inadequate against sophisticated modern-day phishing attacks. A holistic approach that considers human factors, organizational aspects, and technological controls is necessary to effectively combat phishing threats. However, weaknesses can arise in each of these elements due to human involvement, creating gaps that successful phishing attacks can exploit. Bridging these gaps requires a better understanding of how these linkages can be managed more effectively. The application of relevant theories and best practices can enhance the human element of phishing prevention. This includes a more comprehensive understanding of user behaviour and the development of tailored training programs that address individual knowledge gaps and cognitive biases. Furthermore, fostering open communication channels between employees and IT departments can encourage reporting of suspicious activity and facilitate prompt responses to potential threats.

II. RESEARCH GAPS AND FUTURE DIRECTIONS

Despite significant advancements in phishing detection and prevention, several research gaps remain. The dynamic nature of phishing techniques necessitates continuous innovation in detection methods. Future research should focus on developing more robust and adaptable machine learning models that can effectively identify new and evolving phishing attacks.

Improving the interpretability of machine learning models is also crucial for understanding their decision-making processes and enhancing their reliability. Furthermore, research should investigate the effectiveness of different user education and awareness programs and explore ways to tailor these programs to the specific needs of different user groups. The development of innovative tools and platforms that facilitate real-world studies on phishing susceptibility and detection can help to inform the design of more effective prevention strategies. Finally, exploring the use of behavioural biometrics and contextual AI in phishing detection could lead to more effective and personalized protection. The integration of user education with advanced technical defences is critical for building a robust and adaptable defence against phishing attacks. A dynamic, multi-faceted approach combining ongoing user education, cutting-edge technical solutions, and proactive policy measures is needed to combat the ever- evolving threat of phishing.

The impact of phishing on various sectors, such as the business sector in KSA, warrants further investigation. Understanding the specific vulnerabilities and challenges faced by different industries can inform the development of targeted prevention strategies. The development of more sophisticated detection methods that can effectively identify phishing attacks in blockchain networks is also a critical area of future research. The use of explainable AI techniques can provide valuable insights into the factors that influence

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 5, March 2025

phishing susceptibility, enabling the development of more effective countermeasures. Further research is needed to understand the interplay between individual contextual factors and phishing susceptibility, particularly regarding the influence of technology competencies and routine internet activities. Replication of studies in diverse countries and contexts, along with the application of innovative tools, is needed to enhance the generalizability and practical applicability of findings. Moreover, research into the effectiveness of different persuasion principles and time pressure on phishing detection is needed to better understand user vulnerabilities and develop targeted interventions. Finally, research needs to address the limitations of existing methods in detecting zero-day attacks and develop more robust and adaptable solutions. This includes investigating the potential of hybrid frameworks that combine multiple detection models to enhance robustness and effectiveness.

III. CONCLUSION

Phishing remains a significant and evolving threat to cybersecurity. The techniques used by attackers are constantly adapting, making it crucial to develop robust and adaptable detection and prevention strategies. This analysis highlights the importance of a multifaceted approach that combines technological solutions, user education, and organizational policies. While significant progress has been made in developing machine learning-based detection methods, challenges remain, particularly in detecting new and evolving phishing techniques. Future research should focus on addressing these challenges, improving the interpretability of machine learning models, and developing more effective user education and awareness programs. Only through a collaborative effort involving researchers, organizations, and individuals can we hope to effectively combat the persistent threat of phishing.

REFERENCES

[1]. Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. A. (2021). Phishing attacks: a recent comprehensive study and a new anatomy. None. https://doi.org/10.3389/fcomp.2021.563060

[2]. P.M, D., M, M., B, N., R.S, S., M.E, P., & A, M. (NaN). Identification of phishing attacks using machine learning algorithm. E3S Web of Conferences. https://doi.org/10.1051/e3sconf/202339904010

[3]. Zieni, R., Massari, L., & Calzarossa, M. (NaN). Phishing or not phishing? a survey on the detection of phishing websites. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3247135

[4]. Tafa, I., Koi, E., Aliaj, R., & Muzhika, S. (NaN). Analysis of email phishing in session hijacking. None. https://doi.org/None

[5]. Mahmood, A., Pandey, V., Raj, R., & Mishra, G. S. (2024). Detection of phishing sites using machine learning techniques. None. https://doi.org/10.47392/irjaeh.2024.0034

[6]. Seth, P. & Damle, M. (2022). A comprehensive study of classification of phishing attacks with its ai/i detection. None. https://doi.org/10.1109/IIHC55949.2022.10060305

[7]. Saka, T., Jain, R., Vaniea, K., & Kkciyan, N. (NaN). Edinburgh research explorer phishing codebook a structured framework for the characterization of phishing emails. None. https://doi.org/None

[8]. Putra, F. P. E., Ubaidi, U., Zulfikri, A., Arifin, G., & Ilhamsyah, R. M. (2024). Analysis of phishing attack trends, impacts and prevention methods: literature study. Brilliance Research of Artificial Intelligence. https://doi.org/10.47709/brilliance.v4i1.4357

[9]. Nirmal, K., Janet, B., & Kumar, R. (2023). Effectiveness of certificate transparency (ct) check and other datapoints in countering phishing attacks. International Conference on Computing for Sustainable Global Development. https://doi.org/None

[10]. Jari, M. (2022). A comprehensive survey of phishing attacks and defences: human factors, training and the role of emotions. None. https://doi.org/10.5121/ijnsa.2022.14502

[11]. Black, J. & Sarno, D. M. (2023). The influence of time pressure and persuasion principles on phishing detection. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. https://doi.org/10.1177/21695067231192442

[12]. Sarker, O., Jayatilaka, A., Haggag, S., Liu, C., & Babar, M. A. (2023). A multi-vocal literature review on challenges and critical success factors of phishing education, training and awareness. Elsevier BV. https://doi.org/10.1016/j.jss.2023.111899

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 5, March 2025

[13]. Vajrobol, V., Gupta, B. B., & Gaurav, A. (2024). Mutual information based logistic regression for phishing url detection. Elsevier BV. https://doi.org/10.1016/j.csa.2024.100044

[14]. Alnemari, S. & Alshammari, M. (2023). Detecting phishing domains using machine learning. Applied Sciences. https://doi.org/10.3390/app13084649

[15]. Mittal, S., Agarwal, R., Saini, M. L., & Kumar, A. (2023). A logistic regression approach for detecting phishing websites. None. https://doi.org/10.1109/ICAICCIT60255.2023.10466221

[16]. Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S., Ramana, S., & Joga, K. (NaN). Phishing detection system through hybrid machine learning based on url. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3252366

[17]. Alshingiti, Z., Alaqel, R., Al-Muhtadi, J., Haq, Q. M. U., Saleem, K., & Faheem, M. H. (2023). A deep learning-based phishing detection system using cnn, lstm, and lstm-cnn. Electronics. https://doi.org/10.3390/electronics12010232

[18]. S, J. & Eliyas, S. (2023). Detecting phishing attacks using convolutional neural network and lstm. None. https://doi.org/10.1109/ICACITE57410.2023.10183234

[19]. Do, N., Selamat, A., Krejcar, O., Herrera-Viedma, E., & Fujita, H. (NaN). Deep learning for phishing detection: taxonomy, current challenges and future directions. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3151903

[20]. Bhoyar, V., Dharak, K., & Gawali, D. (2024). Detection of phishing website using machine learning. Shivkrupa Publication"s. https://doi.org/10.48175/ijarsct-15004

[21]. Kadhim, H. Y., Al-saedi, K., & Al-Hassani, M. D. (2019). Mobile phishing websites detection and prevention using data mining techniques. International Journal of Interactive Mobile Technologies. https://doi.org/10.3991/ijim.v13i10.10797

[22]. Alazaidah, R., Al-Shaikh, A., AL-Mousa, M. R., Khafajah, H., Samara, G., Alzyoud, M., Al-shanableh, N., & Almatarneh, S. (2024). Website phishing detection using machine learning techniques. None. https://doi.org/10.18576/jsap/130108

[23]. Jayaraj, R., Pushpalatha, A., Sangeetha, K., Kamaleshwar, T., Shree, S. U., & Damodaran, D. (2023). Intrusion detection based on phishing detection with machine learning. Elsevier BV. https://doi.org/10.1016/j.measen.2023.101003

[24]. Frauenstein, E. D. & Solms, R. V. (NaN). Using theories and best practices to bridge the phishing gap. European Information Security Multi-Conference. https://doi.org/None

[25]. Tally, A., Abbott, J., Bochner, A. M., Das, S., & Nippert-Eng, C. (2023). What mid- career professionals think, know, and feel about phishing: opportunities for university it departments to better empower employees in their anti-phishing decisions. None. https://doi.org/10.1145/3579547

[26]. Ribeiro, L., Guedes, I., & Cardoso, C. (2023). Which factors predict susceptibility to phishing? an empirical study. Elsevier BV. https://doi.org/10.1016/j.cose.2023.103558

[27]. Jimmy, F. (2024). Phishing attackers: prevention and response strategies. None. https://doi.org/10.60087/jaigs.v2i1.249

[28]. S, M., Haniah, S., Koti, S. M., Sahani, S., & N, S. S. (2024). A review on phishing threats and data securityin online trading systems using artificial intelligence techniques. International Conference on Advanced Infocomm Technology. https://doi.org/10.1109/ICAIT61638.2024.10690690

[29]. C., O. O. & E, A. C. (NaN). Awareness of phishing attacks in institutions of higher learning: a review of types and technical approaches. International journal of research and innovation in applied science. https://doi.org/10.51584/ijrias.2024.910031

[30]. (NaN). The impact of phishing on the business sector in ksa : analytical study. International Journal of Advanced Trends in Computer Science and Engineering. https://doi.org/10.30534/ijatcse/2021/471022021

[31]. Joshi, K., Bhatt, C. M., Shah, K. A., Parmar, D., Corchado, J., Bruno, A., & Mazzeo, P. (2023). Machine-learning techniques for predicting phishing attacks in blockchain networks: a comparative study. Algorithms. https://doi.org/10.3390/a16080366

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 5, March 2025

[32]. Fan, Z., Li, W., Laskey, K. B., & Chang, K. (2024). Investigation of phishing susceptibility with explainable artificial intelligence. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/fi16010031

[33]. Arivukarasi, M., Manju, A., Kaladevi, R., Hariharan, S., Mahasree, M., & Prasad, A. B. (2023). Efficient phishing detection and prevention using support vector machine (svm) algorithm. International Conference on Communication Systems and Network Technologies. https://doi.org/10.1109/CSNT57126.2023.10134735

[34]. Geest, R. J. V. D., Cascavilla, G., Hulstijn, J., & Zannone, N. (2024). The applicability of a hybrid framework for automated phishing detection. Elsevier BV. https://doi.org/10.1016/j.cose.2024.103736