

A comprehensive review of interpretable machine learning techniques for phishing attack detection

Pankaj Chandre, Pallavi Bhujbal, Ashvini Jadhav, Bhagyashree Dinesh Shendkar, Aditi Wangikar, Rajneeshkaur Sachdeo

Department of Computer Science and Engineering, MIT School of Computing, MIT Art Design and Technology University, Pune, India

Article Info

Article history:

Received Apr 25, 2024

Revised Jun 13, 2025

Accepted Jul 10, 2025

Keywords:

Cybersecurity

Decision-making processes

Detection methodologies

Interpretable machine learning

Phishing attacks

ABSTRACT

Phishing attacks remain a significant and evolving threat in the digital landscape, demanding continual advancements in detection methodologies. This paper emphasizes the importance of interpretable machine learning models to enhance transparency and trustworthiness in phishing detection systems. It begins with an overview of phishing attacks, their increasing sophistication, and the challenges faced by conventional detection techniques. A range of interpretable machine learning approaches, including rule-based models, decision trees, and additive models like Shapley additive explanations (SHAP), are surveyed. Their applicability in phishing detection is analyzed based on computational efficiency, prediction accuracy, and interpretability. The study also explores ways to integrate these methods into existing detection systems to enhance functionality and user experience. By providing insights into the decision-making processes of detection models, interpretable machine learning facilitates human supervision and intervention, strengthening overall system reliability. The paper concludes by outlining future research directions, such as improving the scalability, accuracy, and adaptability of interpretable models to detect emerging phishing techniques. Integrating these models with real-time threat intelligence and deep learning approaches could boost accuracy while preserving transparency. Additionally, user-centric explanations and human-in-the-loop systems may further enhance trust, usability, and resilience in phishing detection frameworks.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Pankaj Chandre

Department of Computer Science and Engineering, MIT School of Computing

MIT Art Design and Technology University

Loni Kalbhor, Pune, India

Email: pankajchandre30@gmail.com

1. INTRODUCTION

Phishing attacks pose a significant threat to cybersecurity, targeting individuals, organizations, and critical infrastructures worldwide. These attacks use deceptive techniques to fool users into disclosing private information, including bank account information and login credentials [1]. Although traditional machine learning techniques have been used to detect phishing attempts, their interpretability and transparency issues frequently restrict their efficacy [2]. The increasing complexity of phishing techniques is driving the demand for sophisticated detection mechanisms that can identify subtle patterns of attack [3]. As a result, approaches for interpretable machine learning have surfaced as viable remedies, providing transparent models that provide light on the decision-making process [4], [5]. This study provides a thorough assessment and analysis of current methods in order to investigate the function of interpretable machine learning in phishing attack

detection [6]. The introduction lays forth the goals and framework of this research, which prepares the reader for a thorough analysis of interpretable machine learning techniques for thwarting phishing attacks. Phishing attacks continue to be a major cybersecurity concern, with millions of incidents reported globally each year. According to industry reports, phishing attacks accounted for over 36% of data breaches in recent years, causing billions of dollars in financial losses for individuals, businesses, and organizations. The growing sophistication of phishing techniques, such as spear phishing and advanced social engineering tactics, has made traditional detection approaches less effective, necessitating the development of more robust and interpretable machine learning models.

Phishing attacks are dishonest tactics employed by bad actors to fool people into divulging private information, like passwords, bank account information, or personal information [7]. Because these assaults prey on human flaws rather than technical ones, they represent serious challenges to cybersecurity [8], [9]. Developing strong defences against phishing attacks requires an understanding of their nature and techniques. Interpretable machine learning techniques play a vital role in enhancing the transparency and explainability of phishing detection systems [10]. It becomes more difficult to comprehend how sophisticated machine learning models make decisions as phishing assaults continue to advance in sophistication [11]. By using interpretable methodologies, security analysts can better identify and mitigate phishing attacks by being able to trust and interpret the predictions provided by these models. The primary objective of this paper is to provide a comprehensive review and analysis of interpretable machine learning techniques for phishing attack detection. It seeks to examine the state of the field, pinpoint important approaches, and assess how well they work to solve the problems caused by phishing scams. The paper is organised so that an overview of phishing assaults is given first, and then the significance of interpretable machine learning approaches is discussed. To assist readers in understanding the following sections, it concludes by outlining the precise goals and parameters of the work.

2. BACKGROUND AND RELATED WORK

2.1. Explanation of phishing attacks, their types, and common characteristics

The section 2.1 delves into the intricacies of phishing attacks, encompassing their various types and common characteristics. Cybercriminals use phishing attacks as a hostile tactic to trick people into disclosing private information like bank account information, login passwords, or personal information. This section explains the many types of phishing assaults, such as spear phishing, email phishing, and pharming, which are designed to take advantage of security system flaws or human weaknesses. Additionally, the section describes the characteristics that set phishing assaults apart, emphasising their manipulative strategies and deceitful nature. To avoid discovery, these attacks frequently use social engineering techniques to entice gullible victims with captivating stories or pressing requests while imitating trustworthy communication channels. Furthermore, red flags such as dubious uniform resource locators (URLs), phoney websites, or fabricated sender identities are often present in phishing attempts and are crucial markers for both detection and mitigation procedures. Through an extensive explanation of the subtleties of phishing assaults, their typologies, and distinguishing features, this part provides a solid basis for comprehending the always changing realm of cyber dangers. Table 1 provides a structured overview of different types of phishing attacks, their descriptions, and common characteristics, which can aid in understanding the diverse methods employed by attackers to deceive unsuspecting victims.

Table 1. Summary of phishing attacks, their types, and common characteristics

Phishing attack type	Description	Common characteristics
Email phishing	Involves sending deceptive emails to users, typically impersonating legitimate entities such as banks or companies, to trick them into divulging sensitive information or performing harmful actions.	Spoofed sender addresses, urgent or alarming messages, requests for personal information, links to fake login pages
Spear phishing	A targeted form of phishing where attackers customize their messages for specific individuals or organizations, often using information gathered from social media or other sources to increase credibility and effectiveness.	Personalized content, contextually relevant information, impersonation of trusted contacts, enhanced social engineering tactics
Whaling	A subtype of spear phishing that targets high-profile individuals within organizations, such as executives or senior management, with the goal of obtaining sensitive corporate data or financial assets.	Impersonation of senior executives, requests for confidential business information Use of executive titles or authority
Vishing	Utilizes voice communication channels, such as phone calls or VoIP services, to deceive individuals into disclosing sensitive information or performing actions under false pretenses.	Automated voice messages or recordings, caller id spoofing, urgent or threatening tone, requests for verification codes or personal details

2.2. Review of existing literature on phishing attack detection methods

Molay [12] propose a novel technique for effortlessly identifying phishing websites on the client side through a redesigned browser architecture called the embedded phishing detection browser (EPDB). Using merely the URL, we extract 30 distinct features of a website using a rule-of-extraction framework. These attributes are then used by a random forest classification machine learning model to determine the validity of the website. The goal of this client-side strategy is to improve upon the weaknesses seen in current anti-phishing methods. With the addition of a specific section for in-the-moment phishing detection activities, the EPDB improves security without compromising the current user experience. By using prototypes, we can identify phishing websites with an astounding 99.36% accuracy rate, giving users of the internet the highest level of protection.

Mohith *et al.* [13] presents a novel anti-phishing strategy that leverages hybrid features extracted from URL and hyperlink information to detect phishing websites without relying on third-party systems. Conventional anti-phishing techniques, including whitelisting or blacklisting, have trouble efficiently identifying new domains or zero-hour phishing attempts. By concentrating solely on client-side features, the suggested method overcomes these difficulties and allows real-time detection without the need for intricate dependencies. Utilising the extreme gradient boosting (XGBoost) methodology, experimental findings show that the suggested method is effective, reaching a high detection accuracy of 99.17%. To aid in trials, a new dataset is also created, demonstrating the usefulness of the suggested strategy in strengthening cybersecurity defences against phishing attacks.

Guptta *et al.* [14] addresses the persistent challenge of phishing email detection by applying knowledge discovery principles and machine learning techniques. It assesses six machine learning techniques using features that have been carefully chosen, and it adds two new features to the body of current literature. The study obtains exceptionally low false positive and negative rates by thorough analysis, with naïve Bayes demonstrating the lowest true positive rate. Notably, with a high accuracy of 99.4% for phishing detection, neural networks appear as the most promising method. All things considered, the study emphasises how machine learning may improve the identification of phishing emails and points out areas that need more research and cybersecurity measure improvement.

Research by Paliath *et al.* [15], in response to the escalating threat of phishing attacks targeting internet-connected devices, researchers have turned to machine learning as a potential solution. Nevertheless, prior methods sometimes depended on numerous characteristics, making them unfeasible for devices with limited resources. A novel method for detecting phishing has been devised to tackle this difficulty, necessitating just nine linguistic characteristics for successful identification. Utilising the ISCXURL-2016 dataset, which has more than 11,000 examples of authentic and fraudulent URLs, the methodology was examined using multiple machine learning classifiers. The random forest algorithm produced the best accuracy of 99.57%, which is impressive and shows how successful this simplified method is in thwarting phishing attacks.

While interpretable machine learning models offer significant advantages in transparency and trustworthiness, their implementation in phishing attack detection presents several challenges. One of the key trade-offs in cybersecurity applications is between model interpretability and accuracy. Traditional deep learning models, such as neural networks, often achieve high detection rates but lack explainability, making it difficult for security analysts to trust their decisions. On the other hand, interpretable models like decision trees and rule-based systems provide clearer explanations but may sacrifice predictive performance, particularly when dealing with complex and evolving phishing tactics. Balancing accuracy and interpretability remain a critical challenge in designing effective phishing detection systems.

Another major challenge is adapting interpretable models to the dynamic nature of phishing attacks. Cybercriminals continuously refine their attack strategies, leveraging advanced obfuscation techniques and social engineering methods to bypass detection. As a result, interpretable models must be frequently updated to maintain their accuracy while preserving transparency. Unlike deep learning models that can automatically adapt through retraining on large datasets, interpretable models often require manual feature engineering and rule adjustments, which can be time-consuming and resource-intensive.

2.3. Discussion on the limitations of traditional machine learning approaches in this context

Table 2 summarizes key limitations of traditional machine learning approaches in phishing detection. Rule-based methods struggle to detect sophisticated and evolving phishing techniques as they rely on predefined patterns. Supervised learning algorithms depend heavily on labeled datasets, making them ineffective against zero-day attacks due to their reliance on historical data. Manual feature engineering often misses subtle indicators and fails to capture complex relationships in high-dimensional data. Additionally, the lack of explainability in many models hinders trust and validation, while poor generalization limits their accuracy and reliability in real-world scenarios where phishing techniques constantly evolve.

Table 2. Summary of the limitations of traditional machine learning approaches for phishing attack detection

Traditional machine learning approach	Limitations
Rule-based methods	Limited ability to handle complex and evolving phishing techniques. These methods rely heavily on predefined rules and patterns, making them less effective against sophisticated attacks that may not conform to predefined rules.
Supervised learning algorithms	Dependency on labeled datasets, which can be scarce and expensive to obtain. Phishing attacks are diverse and constantly evolving, making it challenging to construct comprehensive labeled datasets that capture the full spectrum of attack variations. Additionally, supervised algorithms may struggle with detecting previously unseen or zero-day phishing attacks due to their reliance on historical data.
Feature engineering	Manual feature selection and extraction require domain expertise and may overlook subtle but crucial indicators of phishing. Moreover, traditional feature engineering techniques may not adequately capture the complex relationships between features in high-dimensional data, limiting the performance of machine learning models.
Lack of explainability	Many traditional machine learning algorithms lack transparency and interpretability, making it difficult to understand the reasoning behind their predictions. This lack of explainability hinders trust and makes it challenging for cybersecurity experts to validate and interpret the model's outputs, especially in critical decision-making scenarios.
Generalization	Traditional machine learning models may overfit to the training data or fail to generalize well to unseen data, leading to reduced detection accuracy and reliability in real-world settings. This limitation is particularly problematic in the context of phishing attack detection, where the diversity and dynamics of attack patterns require models to adapt and generalize effectively across different environments and scenarios.

3. INTERPRETABLE MACHINE LEARNING TECHNIQUES

When it comes to phishing attack detection, interpretable machine learning approaches are ones that not only generate correct predictions but also make the process of making those predictions transparent and easy to understand [16]. By improving our understanding of the fundamental elements that go into classifying phishing assaults, these strategies hope to make it simpler for security analysts to decipher and rely on the model's conclusions. Key aspects of interpretable machine learning techniques include:

- Importance of qualities: these methods identify the characteristics or qualities of data that have the greatest bearing on whether an incident qualifies as a phishing assault. Analysts can obtain insights into the nature of phishing attempts by pinpointing crucial features.
- Model explainability: because of their transparent decision-making process, interpretable models like decision trees, rule-based systems, and linear models are favoured. They offer comprehensible justifications for the choices chosen, which promotes confidence and helps to spot possible weak points.
- Local explanations: interpretable techniques offer explanations at the instance level in addition to concentrating exclusively on the global behaviour of the model. This enables analysts to comprehend the rationale behind a given instance's classification as a genuine or phishing attack, enabling focused actions.
- Visualisation: complex machine learning processes are made simpler by using visual representations of model decisions, feature importance, and decision limits. Analysts can make more informed decisions by using graphic displays to help them intuitively identify trends and abnormalities.
- Performance vs. interpretability trade-off: to achieve transparency, interpretable machine learning models frequently give up some predicted accuracy. In real-world applications, striking the correct balance between interpretability and performance of the model is essential.

3.1. Introduction to interpretable machine learning and its relevance in cybersecurity

The introduction of "Interpretable machine learning techniques for phishing attack detection: a comprehensive review and analysis" serves as the foundation for understanding the significance of interpretable machine learning in the context of cybersecurity, particularly in combating phishing attacks [17]. It starts out by explaining how cyber dangers are changing and how phishing is becoming a more common and significant attack vector. Due to the intricacy and sophistication of phishing attempts, interpretable machine learning is being investigated as a potential remedy. The section explores the idea of interpretable machine learning, emphasising how it differs from conventional black-box models. Transparency and explainability are given priority in interpretable machine learning techniques, making it easier for practitioners and security analysts to understand how models make judgements. In cybersecurity, this transparency is critical since effective threat mitigation depends on the user's ability to understand and trust model results. The introduction also emphasises how crucial interpretable machine learning is to fostering cooperation between automated systems and human analysts. Interpretable machine learning enables analysts to improve overall cyber resilience by validating and fine-tuning detection strategies by offering insights into model predictions and decision-making processes [18]. It also discusses compliance

concerns and the regulatory environment, highlighting the necessity of transparent and accountable artificial intelligence (AI) systems in cybersecurity applications.

3.2. Overview of various interpretable machine learning models suitable for phishing detection

"Interpretable machine learning techniques for phishing attack detection: a comprehensive review and analysis" aims to explore and evaluate various interpretable machine learning models that are suitable for detecting phishing attacks. For cybersecurity applications, interpretable machine learning models are essential because they provide light on the model's decision-making process and make it simpler for security analysts to comprehend and believe the predictions made by the model. With an emphasis on their suitability for phishing detection, we will present an overview of many interpretable machine learning approaches in this study, including decision trees, rule-based models, linear models, and ensemble methods. By displaying decision rules in a hierarchical framework, decision trees provide transparency and make it possible for analysts to comprehend the reasoning behind each choice [19]. Conversely, rule-based models offer clear rules that domain specialists may understand with ease. We will also talk about linear models, like logistic regression, which provide clear and straightforward feature importance representation. By combining predictions from several base models, ensemble techniques such as random forests and gradient boosting offer interpretability and accuracy. We will examine the benefits and drawbacks of each method for phishing detection, taking into account aspects like scalability, model complexity, and interpretability of features [20].

3.3. Detailed explanation of each technique, including decision trees, rule-based models, LIME, and SHAP

3.3.1. Decision trees

The most discriminative qualities are used to divide the feature space in decision trees, which are simple, straightforward models. Decision trees were used to divide the information into subsets for the purpose of phishing attack identification [21]. These subsets were identified by criteria such as URL attributes, domain age, or the inclusion of suspicious phrases. A choice is taken based on a feature value at every node in the tree, which causes more splits until the ultimate decision is reached at the leaf nodes. Analysts can follow the decision-making process and comprehend the reasoning for categorising cases as authentic or phishing thanks to this hierarchical structure.

3.3.2. Rule-based models

Rule-based models are very interpretable since they define detection rules as if-then statements. These regulations clearly lay forth criteria based on characteristics that point to phishing activity, like strange patterns in URLs or differences in domain names [22]. A rule might say, for instance, "classify the URL as phishing if it contains an internet protocol address (IP) address and lacks hypertext transfer protocol secure (HTTPS)". These models offer transparent decision-making procedures by following pre-established guidelines, which enable analysts to confirm and assess the logic underlying each classification.

3.3.3. Local interpretable model-agnostic explanations

A post-hoc interpretability method called local interpretable model-agnostic explanations (LIME) was created to clarify specific predictions made by intricate black-box models. It functions by producing locally relevant, interpretable explanations for model predictions, concentrating on a particular case of interest [23]. LIME approximates the behaviour of the model locally by varying the input features surrounding the instance and tracking the resulting changes in the model's output. LIME helps analysts comprehend model decisions by providing insights into why a specific instance was labelled as phishing or not, by highlighting the most relevant features leading to the prediction.

3.3.4. Shapley additive explanations

Another post-hoc interpretability technique called like Shapley additive explanations (SHAP) allocates the contribution of each feature to the output prediction of the model [24]. It estimates feature importance values by utilising ideas from cooperative game theory, notably shapley values. To provide a thorough knowledge of feature impact, SHAP computes the marginal contribution of each feature to the prediction outcome across all potential permutations. SHAP clarifies the fundamental principles behind phishing attack detection models by quantifying the impact of specific variables on model predictions. This helps analysts to properly test and trust the models' functioning.

These interpretable machine learning algorithms, in essence, offer varying degrees of transparency and insight into the phishing attack detection models' decision-making process. While LIME and SHAP offer post-hoc explanations for sophisticated black-box models, decision trees and rule-based models provide clear decision rules, improving the interpretability and reliability of phishing detection systems. Table 3 shows that the summary of explainable AI techniques.

Table 3. Summary of explainable AI techniques

Technique	Description	Advantages	Limitations
Decision trees	Decision trees are hierarchical tree structures where internal nodes represent features or attributes, branches represent decisions or rules, and leaf nodes represent outcomes. They're interpretable and can easily show how a decision is made based on feature values.	Easy to interpret and understand, can handle both numerical and categorical data, automatically handles feature selection and interaction, can be visualized for better understanding.	Prone to overfitting, especially with complex datasets, can create biased trees if the dataset is imbalanced, may not capture complex relationships in the data effectively.
Rule-based models	Rule-based models use a set of rules to classify instances. These rules are usually in the form of "if-then" statements, making them highly interpretable. They're easy to understand and can directly map feature values to class labels, aiding in explaining the model's decision-making process.	Highly interpretable and transparent, easy to implement and deploy, can handle both numerical and categorical data, provides explicit rules for decision making.	May suffer from overfitting if the rule set becomes too complex, limited expressiveness compared to other models like neural networks or ensemble methods requires domain expertise to design effective rules.
LIME	LIME is a model-agnostic technique that explains individual predictions of black-box machine learning models by approximating them with interpretable surrogate models locally. It generates explanations in the form of simple, interpretable rules or explanations that can help users understand why a model made a particular prediction.	Provides local interpretable explanations for complex models, allows users to understand model predictions at the instance level, can be applied to any black-box model without access to internal model parameters.	May not always accurately represent the global behavior of the model, computationally expensive for large datasets or complex models, requires selecting a representative subset of instances for explanation generation, which may introduce bias.
SHAP	SHAP values provide a way to explain the output of any machine learning model by attributing the prediction outcome to each input feature. They represent the average contribution of a feature value to the prediction across all possible permutations of features. SHAP values offer global interpretability by showing the impact of each feature on the model's output.	Offers global interpretability by quantifying the contribution of each feature to model predictions, accounts for interactions between features, provides a consistent explanation method across different models and datasets.	Computationally expensive for large datasets or models with many features, interpretability might be challenging when dealing with highly correlated features, interpretation might not always be intuitive for non-technical users, requires careful normalization of input features to ensure meaningful SHAP values.

4. PROPOSED METHODOLOGY

The Figure 1 illustrates the architecture of a phishing attack detection system using explainable AI. Let us break down how this system works: i) external data sources: phishing email repositories are among the external data sources that the system consumes. The main dataset used to train and evaluate the phishing attack detection model consists of these emails; ii) email data: the dataset of phishing emails is represented by this component. It includes a variety of characteristics and elements that were taken from these emails, including metadata, embedded URLs, email text, and sender information; iii) feature extraction: the system carries out feature extraction after gathering the email data. This procedure entails formatting the unprocessed email data so that it can be entered into the machine learning model. Email features that can be extracted include sender reputation scoring, URL analysis, and textual content analysis; iv) explainable AI model: an explainable AI model is then fed the feature-extracted data. Machine learning models called "explainable AI" are intended to provide explanations for their decisions that are comprehensible to humans in addition to producing precise forecasts. This model learns patterns and traits typical of phishing attacks and explains why specific emails are flagged as legitimate or phishing in the context of phishing attack detection; v) model interpretation: to comprehend the logic underlying the predictions made by the explainable AI model, its output is interpreted. Gaining confidence in the model's judgements and being aware of its advantages and disadvantages need completing this step. Techniques for interpreting models may involve the visualisation of decision boundaries, feature importance analysis, and SHAP values; and vi) decision: ultimately, a determination is reached concerning the categorization of an email as a phishing attempt or not, considering the interpretations supplied by the model. This choice could result in several different things happening, like reporting the email as suspicious, preventing users from accessing embedded URLs, or notifying system administrators.

In summary, this architecture builds a phishing assault detection system that not only recognises possible threats but also provides an explanation for their flagging. It accomplishes this by combining data gathering, feature extraction, machine learning modelling, interpretation, and decision-making. Transparency plays a key role in improving system confidence as well as enabling human monitoring and action when needed.

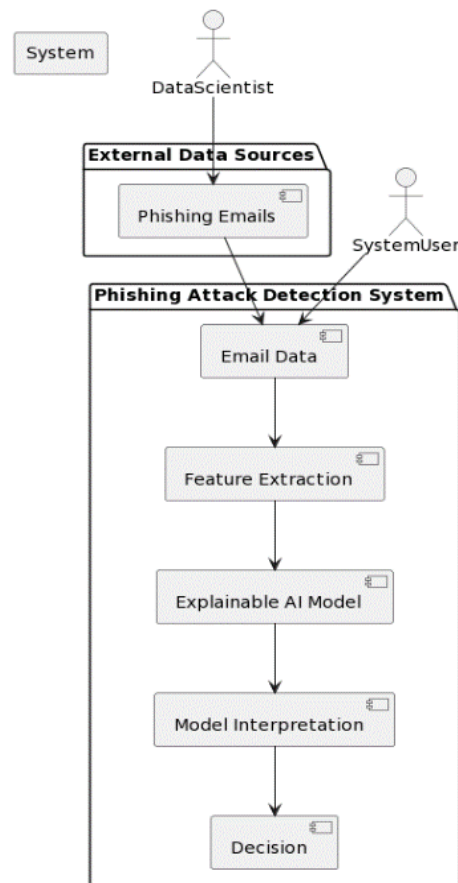


Figure 1. Architecture for interpretable machine learning techniques for phishing attack detection

5. CHALLENGES AND FUTURE DIRECTIONS

5.1. Identification of challenges and limitations associated with interpretable machine learning techniques for phishing detection

A comprehensive review and analysis provides a thorough examination of the challenges and limitations pertaining to interpretable machine learning methods when applied to the task of phishing attack detection. While interpretable machine learning models offer transparency and insight into decision-making processes, they also face several challenges and limitations in the context of phishing detection. Phishing attacks often involve complex features such as URL structure, HTML content, and linguistic patterns [25]. For instance, detecting subtle variations in domain names (e.g., "paypal.com" instead of "paypal.com") requires models to discern nuanced differences, which can be challenging for interpretable ML algorithms.

Suppose, for example, that an interpretable machine learning model marks an email as suspicious because it contains specific keywords or URL patterns [26], [27]. It might, however, find it difficult to explain precisely why the classification judgement was made based on these features, which would undermine the detection system's credibility [28]. While an interpretable machine learning model based on decision trees may be able to recognise simple phishing attempts with reasonable accuracy, it may not be able to identify more complex attacks that call for more intricate feature interactions. On the other hand, intricate ensemble models such as gradient boosting could provide better accuracy but are not as comprehensible. Interpretable machine learning techniques, including rule-based classifiers, may not be able to keep up with the increasing complexity and diversity of phishing attacks in real-time, which could cause delays in detection and response.

Phishing emails that are explicitly created to trick interpretable machine learning models by tampering with features that are crucial to categorization can be created by adversaries [29], [30]. They might, for example, disguise harmful URLs to look like safe ones to avoid being picked up by rule-based classifiers. When applied in a different organisational context with distinct phishing tactics and patterns, an interpretable machine learning model that was trained on a particular dataset containing phishing examples from a particular industry may find it difficult to generalise, which will reduce detection accuracy [31], [32].

Although an interpretable machine learning model may offer justifications for its choices, non-professional users may still find it difficult to understand and rely on these justifications, particularly in the case of intricate feature interactions or when the model's logic deviates from human intuition [33], [34]. For example, specific expertise in machine learning and cybersecurity may be needed to understand the meaning of specific linguistic cues or HTML elements in phishing emails.

5.2. Exploration of potential future research directions to address these challenges and improve detection accuracy and interpretability

5.2.1. Hybrid models

By integrating machine learning with conventional rule-based techniques or investigating the integration of several machine learning models, it is possible to capitalise on the advantages of various techniques and improve both interpretability and accuracy. Research on explainable AI techniques, such SHAP values and LIME, can shed light on how machine learning models decide, which will increase confidence and comprehension. It is crucial to create strong models resistant to adversarial assaults that are particular to phishing detection. Enhancing model robustness against adversarial manipulations of phishing emails or websites could be the focus of future research. Researching active learning strategies to choose informative samples for model training in a clever way may result in better model performance and a more effective use of labelled data, particularly in situations where there is a shortage of labelled data. To ensure strong performance in a variety of real-world settings, research endeavours ought to focus on improving the models' capacity to generalise across various phishing attack kinds, domains, and languages.

Investigating techniques that incorporate human knowledge into the machine learning pipeline can take advantage of the advantages of both automated algorithms and human intuition, enhancing the overall interpretability and accuracy of detection. It is critical to create real-time detection systems that can promptly recognise and stop phishing assaults as soon as they happen. The integration of automatic response mechanisms and speed optimisation of model inference could be the main areas of research. It is critical to look at privacy-preserving machine learning methods to safeguard private user data during the training and inference stages of models, particularly when it comes to situations involving surfing or personal email data.

In reviewing interpretable machine learning techniques for phishing attack detection, several limitations are evident. Firstly, many techniques still struggle with scalability and efficiency when applied to large datasets. Additionally, the interpretability of some models may be compromised in complex scenarios, limiting their practical utility. Future research should focus on improving the scalability of interpretative methods and developing techniques that balance interpretability with high performance. Further exploration into hybrid models that combine interpretability with advanced detection capabilities could enhance effectiveness. These improvements will have significant implications for creating more reliable and user-friendly phishing detection systems, ultimately strengthening cybersecurity defenses.

6. CONCLUSION

This comprehensive review highlights the critical role of interpretable machine learning techniques in phishing attack detection, emphasizing the need for transparency and trustworthiness in cybersecurity systems. By analyzing various models such as rule-based approaches, decision trees, and additive models like SHAP, the study demonstrates how interpretability enhances detection accuracy while enabling human oversight. The findings underscore the importance of explainable AI in improving phishing detection capabilities, making security systems more transparent and trustworthy. As phishing tactics become increasingly sophisticated, leveraging interpretable models ensures that detection decisions are understandable, facilitating better decision-making by security analysts and end-users alike. Despite these advancements, several challenges remain, warranting further research. Future efforts should focus on developing adaptive phishing detection models capable of automatically learning and responding to new attack patterns while maintaining interpretability. Integrating explainable AI with real-time detection systems, especially in dynamic environments like social media and corporate networks, is another crucial research avenue. Additionally, enhancing user trust in phishing detection frameworks through intuitive model explanations and interactive visualizations could improve adoption and effectiveness. Addressing these challenges will contribute to the development of more robust and transparent phishing detection systems, strengthening cybersecurity defenses against evolving threats.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to MIT Art, Design and Technology University, Pune, India, for providing the necessary resources, guidance, and support throughout the course

of this research work. Their encouragement and academic environment have been instrumental in completing this study successfully.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Pankaj Chandre	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Pallavi Bhujbal	✓	✓					✓		✓					
Ashvini Jadhav	✓		✓				✓		✓				✓	
Bhagyashree Dinesh Shendkar	✓		✓						✓		✓	✓	✓	
Aditi Wangikar	✓			✓	✓				✓		✓	✓		✓
Rajneeshkaur Sachdeo	✓			✓	✓				✓	✓				✓

- C : Conceptualization
M : Methodology
So : Software
Va : Validation
Fo : Formal analysis
- I : Investigation
R : Resources
D : Data Curation
O : Writing - Original Draft
E : Writing - Review & Editing
- Vi : Visualization
Su : Supervision
P : Project administration
Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

There is no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] V. K. Raghu *et al.*, “Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models,” *Thorax*, vol. 74, no. 7, pp. 643–649, Jul. 2019, doi: 10.1136/thoraxjnl-2018-212638.

[2] T. Yadav and R. Sachdeo, “Enhanced face age progression and regression model using hyper-parameter tuning-large scale GAN by hybrid heuristic improvement,” *The Imaging Science Journal*, vol. 72, no. 8, Sep. 2024, doi: 10.1080/13682199.2023.2254134.

[3] G. J. W. Kathrine, P. M. Praise, A. A. Rose, and E. C. Kalaivani, “Variants of phishing attacks and their detection techniques,” in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, Apr. 2019, pp. 255–259, doi: 10.1109/ICOEI.2019.8862697.

[4] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, “A systematic literature review on phishing email detection using natural language processing techniques,” *IEEE Access*, vol. 10, pp. 65703–65727, 2022, doi: 10.1109/ACCESS.2022.3183083.

[5] R. J. V. Geest, G. Cascavilla, J. Hulstijn, and N. Zannone, “The applicability of a hybrid framework for automated phishing detection,” *Computers & Security*, vol. 139, Apr. 2024, doi: 10.1016/j.cose.2024.103736.

[6] S. Makubhai, G. R. Pathak, and P. R. Chandre, “Prevention in healthcare: an explainable AI approach,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 5, pp. 92–100, May 2023, doi: 10.17762/ijritcc.v11i5.6582.

[7] E. J. Williams and A. N. Joinson, “Developing a measure of information seeking about phishing,” *Journal of Cybersecurity*, vol. 6, no. 1, Jan. 2020, doi: 10.1093/cybsec/tyaa001.

[8] M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature survey,” *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2091–2121, 2013, doi: 10.1109/SURV.2013.032213.00009.

[9] K. Demertzis and L. Iliadis, “Cognitive web application firewall to critical infrastructures protection from phishing attacks,” *Journal of Computations & Modelling*, vol. 9, no. 2, pp. 1–26, 2019.




[10] P. R. Chandre, P. N. Mahalle, and G. R. Shinde, “Machine learning based novel approach for intrusion detection and prevention system: a tool based verification,” in *2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN)*, IEEE, Nov. 2018, pp. 135–140, doi: 10.1109/GCWCN.2018.8668618.

[11] J. Gikandi, J. Kamau, D. Njuguna, and L. Sawe, “Sentence level analysis model for phishing detection using KNN,” *Journal of Cyber Security*, vol. 6, no. 1, pp. 25–39, 2024, doi: 10.32604/jcs.2023.045859.




- [12] K. Molay, "Best practices for webinars: creating effective web events with Adobe® Connect™," *White Paper*, pp. 1-14, , 2009. [Online]. Available: <https://www.clarix.com/whitepapers/best-practices-webinars-wp.pdf>.
- [13] G. H. Mohith, M. Adithya, G. P. S, and S. Vinay, "Development of anti-phishing browser based on random forest and rule of extraction framework," *Cybersecurity*, vol. 3, no. 1, Dec. 2020, doi: 10.1186/s42400-020-00059-1.
- [14] S. D. Gupta, K. T. Shahriar, H. Alqahtani, D. Alsaman, and I. H. Sarker, "Modeling hybrid feature-based phishing websites detection using machine learning techniques," *Annals of Data Science*, vol. 11, no. 1, pp. 217–242, Feb. 2024, doi: 10.1007/s40745-022-00379-8.
- [15] S. Paliath, M. A. Obeitah, and M. Aldwairi, "PhishOut: effective phishing detection using selected features," in *2020 27th International Conference on Telecommunications (ICT)*, IEEE, Oct. 2020, pp. 1–5, doi: 10.1109/ICT49546.2020.9239589.
- [16] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment," *Computer Communications*, vol. 175, pp. 47–57, Jul. 2021, doi: 10.1016/j.comcom.2021.04.023.
- [17] T. Sutter, A. S. Bozkir, B. Gehring, and P. Berlich, "Avoiding the hook: influential factors of phishing awareness training on click-rates and a data-driven approach to predict email difficulty perception," *IEEE Access*, vol. 10, pp. 100540–100565, 2022, doi: 10.1109/ACCESS.2022.3207272.
- [18] S. K. Punia, M. Kumar, T. Stephan, G. G. Deverajan, and R. Patan, "Performance analysis of machine learning algorithms for big data classification," *International Journal of E-Health and Medical Communications*, vol. 12, no. 4, pp. 60–75, Jul. 2021, doi: 10.4018/IJEHMC.20210701.oa4.
- [19] L. Liu, O. De Vel, Q.-L. Han, J. Zhang, and Y. Xiang, "Detecting and preventing cyber insider threats: a survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 1397–1417, 2018, doi: 10.1109/COMST.2018.2800740.
- [20] U. A. Butt, R. Amin, H. Aldabbas, S. Mohan, B. Alouffi, and A. Ahmadian, "Cloud-based email phishing attack using machine and deep learning algorithm," *Complex & Intelligent Systems*, vol. 9, no. 3, pp. 3043–3070, 2023, doi: 10.1007/s40747-022-00760-3.
- [21] L. Ribeiro, I. S. Guedes, and C. S. Cardoso, "Which factors predict susceptibility to phishing? an empirical study," *Computers & Security*, vol. 136, Jan. 2024, doi: 10.1016/j.cose.2023.103558.
- [22] A. A. Orunsolu, A. S. Sodiya, and A. T. Akinwale, "A predictive model for phishing detection," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 2, pp. 232–247, Feb. 2022, doi: 10.1016/j.jksuci.2019.12.005.
- [23] M. Sanchez-Paniagua, E. F. Fernandez, E. Alegre, W. Al-Nabki, and V. Gonzalez-Castro, "Phishing URL detection: a real-case scenario through login URLs," *IEEE Access*, vol. 10, pp. 42949–42960, 2022, doi: 10.1109/ACCESS.2022.3168681.
- [24] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 590–611, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.004.
- [25] G. Palaniappan, S. S, B. Rajendran, Sanjay, S. Goyal, and B. B S, "Malicious domain detection using machine learning on domain name features, host-based features and web-based features," *Procedia Computer Science*, vol. 171, pp. 654–661, 2020, doi: 10.1016/j.procs.2020.04.071.
- [26] M. A. Remmide, F. Boumahdi, N. Boustia, C. L. Feknous, and R. Della, "Detection of phishing URLs using temporal convolutional network," *Procedia Computer Science*, vol. 212, pp. 74–82, 2022, doi: 10.1016/j.procs.2022.10.209.
- [27] C. Opara, Y. Chen, and B. Wei, "Look before you leap: detecting phishing web pages by exploiting raw URL and HTML characteristics," *Expert Systems with Applications*, vol. 236, Feb. 2024, doi: 10.1016/j.eswa.2023.121183.
- [28] B. Gaddekar and T. Hiwarkar, "A proposed business improvement model utilizing machine learning: enhancing decision-making and performance," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 1s, pp. 557–568, 2023.
- [29] B. Gaddekar and T. Hiwarkar, "A critical evaluation of business improvement through machine learning: challenges, opportunities, and best practices," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 10s, pp. 264–276, Oct. 2023, doi: 10.17762/ijritcc.v11i10s.7627.
- [30] D. Dhotre, P. R. Chandre, A. Khandare, M. Patil, and G. S. Gawande, "The rise of crypto malware: leveraging machine learning techniques to understand the evolution, impact, and detection of cryptocurrency-related threats," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 7, pp. 215–222, Sep. 2023, doi: 10.17762/ijritcc.v11i7.7848.
- [31] J. G. Kotwal, R. Kashyap, and P. M. Shafi, "Artificial driving based efficientnet for automatic plant leaf disease classification," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 38209–38240, Oct. 2023, doi: 10.1007/s11042-023-16882-w.
- [32] A. Jawale, P. Warole, S. Bhandare, K. Bhat, and P. R. Chandre, "Jeevn-Net: brain tumor segmentation using cascaded U-Net & overall survival prediction," *International Research Journal of Engineering and Technology*, vol. 7, no. 1, 2020.
- [33] S. Ashfaq, S. A. Patil, S. Borde, P. Chandre, P. M. Shafi, and A. Jadhav, "Zero trust security paradigm: a comprehensive survey and research analysis," *Journal of Electrical Systems*, vol. 19, no. 2, pp. 28–37, Jan. 2024, doi: 10.52783/jes.688.
- [34] V. Bidve *et al.*, "Use of explainable AI to interpret the results of NLP models for sentimental analysis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 35, no. 1, Jul. 2024, doi: 10.11591/ijeecs.v35.i1.pp511-519.

BIOGRAPHIES OF AUTHORS






Dr. Pankaj Chandre    has obtained his B.E. degree in Information Technology from Sant Gadge Baba Amravati University, Amravati, India, M.E. degree in Computer Engineering from Mumbai University Maharashtra, India in the year 2011 and PhD in Computer Engineering from Savitribai Phule Pune University, Pune, India in the year 2021. He is currently working as an Associate Professor in Department of Computer Science and Engineering, MIT School of Computing, MIT ADT, Pune, India. He has published 60 plus papers at international journals and conferences. He is guiding 2 plus PhD research scholar at MIT Art Design and Technology University, Pune, India. He has guided more than 30 plus under-graduate students and 20 plus postgraduate students for projects. His research interests are network security and information security. He can be contacted at email: pankaj.chandre@mituniversity.edu.in or pankajchandre30@gmail.com.






Prof. Pallavi Bhujbal    is a seasoned academic and researcher with over 11+ years of experience in the field of Computer Science and Engineering. She holds a Bachelor's degree in Computer Science and Engineering from Bamu University, Aurangabad (2011) and an M.Tech. from Pune University (2014). She is currently pursuing her Ph.D. from the Indian Institute of Information Technology, Nagpur. At present, she is working at MIT School of Computing, where she guides both B.Tech. and M.Tech. students. She has published three international conference papers and two journal papers indexed in Scopus. Her research interests include network security, information security, and IoT. With a strong academic foundation and industry insight, she is committed to advancing knowledge and promoting innovation. She can be contacted at email: pallavi.bhujbal@mituniversity.edu.in.






Prof. Ashvini Jadhav    is a research scholar and a professional in the field of Computer Science and Engineering, specializing in computer networks. She obtained her Master's degree in Computer Science and Engineering (Computer Network) from G H Raisoni College of Engineering and Management, Wagholi, in 2013. Currently, she serves as a dedicated Assistant Professor in Department of Information Technology at the MIT School of Computing in Loni, Pune, where she shares her extensive knowledge and expertise with aspiring computer engineers. With more than 14 years of hands-on experience in computer engineering, she has honed her skills across various domains within the discipline, with a specific focus on computer networks, cyber security, and programming. Her dedication to advancing the field of computer science is underscored by her on-going pursuit of a Ph.D. at the MIT School of Computing, MIT ADT, Pune, India where she is actively involved in pioneering research endeavors. She can be contacted at email: ashvini.jadhav@mituniversity.edu.in.






Prof. Bhagyashree Dinesh Shendkar    has more than 16 years of teaching experience in the Computer Science Engineering and currently working as an Assistant Professor at Computer Science Engineering Department of MIT School of Computing, MIT Art Design and Technology University, Loni Kalbhor, Pune. She has received her B.E. Information Technology and M.E. Information Technology degree from Pune University. Currently, she is researching regarding conceptual modelling framework using machine learning as her work to receive her Ph.D. She authored four reference books. She published more than 43 research papers in national and international journals, conferences and presented 05 papers in national, international seminars and conferences. Her innovative work has also led to the filing of five patents, showcasing his commitment to advancing knowledge and technology. Her interested areas of academics are data structures, artificial intelligence, machine learning, and computer graphics. She is a life member of the Indian Society for Technical Education (ISTE). She can be contacted at email: bhagyashree.shendkar@mituniversity.edu.in.



Mrs. Aditi Wangikar    has obtained her B.E. degree in Information Technology from Dr. Babasaheb Marathwada University Sambhajinagar India, M.E. degree in Computer Science and Engineering from Dr. Babasaheb Ambedkar University Sambhajinagar, India in the year 2014. She is currently working as an Assistant Professor in Department of Computer Science and Engineering, MIT School of Computing, MIT ADT, Pune, India. She has guided more than 15 plus under-graduate students for projects. Her research interests are data analytics. She can be contacted at email: aditiwangikar@outlook.com.



Dr. Rajneeshkaur Sachdeo    is Dean Faculty of CS&IT, Professor, Department of Computer Science and Engineering, MIT School of Computing, MIT Art, Design and Technology University, Pune. She obtained Ph.D. (Computer Science and Engineering)- February 2017, SGBAU State University, Amravati, M.Tech. (Computer Engineering)- 2005, Government College of Engineering, Pune, B.E. Computer Science and Engineering - 1993, SGBAU State University Amravati. Her research areas include data security and privacy, natural language processing and linguistics, machine learning, data mining, and wireless network. She is member of IEEE, ISTE, IACSIT and CSI, received the MAEER's Trust "Staapna Diwas Award" in August 2010. She guided around 100 students for under-graduate projects and 18 students at the post-graduate level, 4 research scholar completed PhD, currently guiding 5 plus research scholars. She completed a research & development project as PI / Co-PI: "Resources for Internationalisation of Higher Education Institutions in India (RISHII)", Co-funded by ERASMUS+ Programme of the European Union "e-translator: Multilingual machine translation", funded by Board of College and University Development, University of Pune, "BookmyTool" App on Playstore - application development for SCHIMMEL Online Pvt. Ltd. Pune "ButterflyGarden" App for VEM Tooling Pvt. Ltd. Hongkong. She can be contacted at email: rajneeshkaur.sachdeo@mituniversity.edu.in.