Journal of Digital Science

ISSN 2686-8296

Volume 7 Issue 2 December 2025

Institute of Cited Scientists (ICS)

CONTENTS

Explainable AI Approaches for Detecting and Mitigating Phishing Attacks: A Review

Nipuna Sankalpa Thalpage [1,2][0009-0001-3374-1927], Eranga Jayarathne [3][0009-0004-8915-4467]

[1]Cardiff Metropolitan University, Wales, United Kingdom;
[2]Institute of Cited Scientists, Agia Napa, Cyprus
[3]University of Peradeniya, Sri Lanka

Received 01.12.2025/Revised 24.12.2025/Accepted 30.12.2025/Published 30.12.2025

https://doi.org/10.33847/2686-8296.7.2_4

Abstract. Phishing remains one of the most pervasive and sophisticated cybersecurity threats, increasingly leveraging social engineering, AI-driven content generation, and multi-vector delivery methods. While machine learning (ML) and deep learning (DL) models have significantly advanced phishing detection capabilities, their "black-box" nature often limits transparency, trust, and practical adoption in real-world security environments. Explainable Artificial Intelligence (XAI) offers a solution by providing interpretable insights into model decisions, enabling analysts and stakeholders to understand, validate, and act upon automated classifications. This semi-systematic review examines contemporary XAI techniques applied to phishing detection, focusing on studies published between 2017 and 2025. Searches conducted across Scopus, IEEE Xplore, and Google Scholar yielded peer-reviewed literature integrating explainability into ML/DL-based phishing detection. The selected studies were synthesized to identify the types of models used, the XAI methods employed, and their contributions to interpretability, operational value, and human–AI collaboration.

Findings show that feature attribution methods such as SHAP, LIME, and Integrated Gradients are the most widely adopted, offering both global and local explanations for text-based and URL-based phishing detection. Attention mechanisms and visualization techniques further enhance transparency in deep learning models, while interpretable models—such as decision trees and logistic regression, remain valuable for contexts requiring high clarity. However, gaps persist in real-world validation, dataset diversity, standard metrics for evaluating explanations, and deployment feasibility.

Overall, XAI strengthens phishing mitigation by improving user trust, supporting analyst decision-making, and enabling more accountable AI-driven security systems. The review highlights the need for scalable, human-centred, and adversarially robust XAI approaches to support the next generation of phishing detection frameworks.

Keywords: Phishing, Explainable AI, Cybersecurity.

1. Introduction

Phishing has become a significant global cybersecurity threat, affecting individuals, businesses, and governments across all sectors. As a form of social engineering, phishing attacks deceive users into disclosing sensitive information or performing harmful actions by impersonating trustworthy entities. With the rapid expansion of digital services, cloud platforms, mobile communication, and remote work environments, phishing incidents have increased in both frequency and

sophistication. Attackers now employ targeted spear-phishing, business email compromise (BEC), and multi-vector delivery methods that exploit human behavior rather than system vulnerabilities, making phishing one of the most challenging cybersecurity issues worldwide.

Despite continuous advancements in cybersecurity technologies, traditional phishing detection methods face several limitations. Signature-based filters struggle to identify newly crafted phishing emails, as attackers frequently modify URLs, payloads, and content to evade detection. Rule-based systems depend on predefined patterns and are often ineffective against zero-day attacks or dynamically generated phishing pages. Moreover, conventional spam filters can produce high false-positive and false-negative rates, reducing reliability and causing user fatigue. These limitations create a critical gap in existing defense mechanisms, allowing sophisticated phishing campaigns to bypass conventional security tools and compromise users [1].

In response to these challenges, Artificial Intelligence (AI) and Machine Learning (ML) have emerged as promising approaches for enhancing phishing detection. ML-based systems can automatically learn discriminative features from large datasets, enabling more accurate identification of malicious emails, URLs, and websites. AI-driven models such as deep learning, natural language processing (NLP), and anomaly detection techniques provide adaptive and scalable solutions capable of identifying previously unseen threats [2]. These intelligent systems can analyze behavioural, structural, and linguistic patterns, offering a more robust and proactive defense against evolving phishing tactics. As a result, AI/ML-based phishing detection has become a key focus in contemporary cybersecurity research.

1.1 Need for Explainability

The increasing reliance on machine learning for phishing detection has introduced new challenges related to the interpretability of these systems. Many state-of-the-art ML and deep learning models operate as "black boxes," providing high accuracy but offering little insight into how decisions are made. This opacity becomes problematic in cybersecurity applications, where understanding the rationale behind a model's classification is crucial. Without clear explanations, security analysts may struggle to validate alerts, investigate incidents, or identify model weaknesses, ultimately limiting the practical adoption and trustworthiness of such systems [3].

Transparency, trust, and accountability are therefore essential when deploying automated threat-detection tools. In high-risk environments, such as financial institutions, critical infrastructure, and governmental systems, stakeholders require assurance that AI-driven decisions are consistent, unbiased, and aligned with organizational policies. Explainable systems enable users to understand why an email or URL is classified as phishing, improving confidence in automated decisions and supporting compliance with regulatory standards. Additionally, explainability reduces the likelihood of misclassifications, facilitates human oversight, and enhances the collaborative interaction between analysts and intelligent detection systems.

The importance of explainable artificial intelligence extends beyond cybersecurity into broader digital transformation initiatives. Recent review studies highlight that integrating machine learning with explainable AI significantly enhances organizational trust, accountability, and decision-making across data-driven systems. These findings emphasize that explainability is a key enabler for the responsible adoption of AI technologies, reinforcing the relevance of XAI in high-risk domains such as phishing detection, where transparency and human oversight are critical [4].

Explainable Artificial Intelligence (XAI) plays a vital role in addressing these concerns within cybersecurity. The need for explainable models is further reinforced

by calls for a rigorous scientific foundation for interpretability, particularly in high-stakes decision-making domains such as cybersecurity [5].

XAI techniques aim to make ML model behavior more interpretable by identifying key features, patterns, or reasoning processes that influence predictions. In the context of phishing detection, XAI can highlight suspicious linguistic cues, structural anomalies in URLs, or abnormal sender behaviors that contribute to an alert. These insights help analysts refine detection strategies, understand emerging attack vectors, and improve model robustness against adversarial manipulation. Consequently, XAI has become a critical component of modern cybersecurity frameworks, enabling more secure, transparent, and reliable AI-driven defense mechanisms.

1.2 Purpose of the Review

The purpose of this review is to examine the role of Explainable Artificial Intelligence (XAI) in enhancing phishing detection systems. As phishing attacks continue to evolve in complexity, many organizations have turned to machine learning and deep learning models to strengthen their defensive capabilities. However, the opaque nature of these black-box models raises significant concerns regarding trust, interpretability, and operational reliability. By evaluating the current landscape of XAI techniques and their applications in phishing detection, this review aims to identify how explainability can address these limitations and support more informed, transparent decision-making in cybersecurity environments.

This review contributes to the body of knowledge in several ways. First, it synthesizes existing research on AI/ML-based phishing detection and highlights the specific challenges associated with model interpretability. Second, it provides a structured analysis of leading XAI methods, such as feature attribution, model-agnostic explanations, and visualization frameworks—and examines their effectiveness in cybersecurity contexts. Third, the review identifies gaps in current literature, including issues related to usability, scalability, adversarial robustness, and real-world deployment. Through this comprehensive evaluation, the review offers insights that can guide future research, support the development of more transparent detection systems, and promote the responsible integration of AI within cybersecurity practices.

In contrast to existing reviews that primarily focus on detection accuracy or algorithmic performance, this study enriches the subject area by foregrounding explainability as a central analytical lens. Specifically, it advances current literature by comparatively synthesising XAI techniques in phishing detection, examining their contribution to transparency, human–AI collaboration, and real-world deployment feasibility. By explicitly linking explainability methods to operational decision-making and trust, this review extends beyond technical surveys and provides interdisciplinary value relevant to cybersecurity, human-centered AI, and digital governance.

1.3 Research Questions

This review is guided by the following research questions, which aim to explore the application, effectiveness, and limitations of Explainable Artificial Intelligence (XAI) within phishing detection systems:

RQ1: What XAI methods have been applied in phishing detection?

This question seeks to identify and categorize the range of XAI techniques used in existing studies, including model-specific and model-agnostic approaches. It explores how various explainability methods, such as LIME, SHAP, attention mechanisms, and rule-based explanations, have been integrated into phishing detection models.

RQ2: How do XAI techniques enhance transparency and decision-making in phishing mitigation?

This question examines the role of explainability in improving analyst understanding, trust, and operational effectiveness. It investigates how XAI contributes to clearer interpretations of model outputs, better incident response, and more informed cybersecurity decision-making.

RQ3: What limitations and research gaps exist in current literature?

This question identifies shortcomings in present research, such as challenges in evaluating explanation quality, scalability concerns, adversarial vulnerabilities, dataset biases, and limited real-world deployment. It highlights areas where further investigation is needed to advance the development of robust, transparent AI-based phishing detection systems.

2. Methodology

This study employs a semi-systematic literature review to examine how Explainable Artificial Intelligence (XAI) techniques are applied in phishing detection and how interpretability contributes to cybersecurity decision-making. This review method was selected because it allows a structured yet flexible examination of emerging interdisciplinary research, particularly where technological, behavioural, and security aspects intersect.

A targeted search was conducted across three major academic databases: Scopus, IEEE Xplore, and Google Scholar. The search strategy used combinations of key terms such as: *"Explainable AI," "XAI," "phishing detection," "email phishing," "URL phishing," "interpretable machine learning,"* and *"explainability in cybersecurity."*

To ensure relevance and quality, the following inclusion criteria were applied:

- Peer-reviewed journal or conference publications
- Published between 2017 and 2025
- Focused on machine learning–based phishing detection with an XAI component
- Written in English

Studies were excluded if they: (1) Did not include any explainability method; (2) Focused solely on phishing awareness or user training, were non-peer-reviewed preprints, abstracts, or posters; (3) Used traditional rule-based detection without ML/XAI.

The screening procedure consisted of an initial title and abstract review to identify relevant papers, followed by a detailed full-text evaluation. To reduce subjectivity and strengthen reliability, all authors participated in the screening and data extraction process. This multidisciplinary collaboration ensured balanced interpretation by combining expertise from cybersecurity, AI/ML, and user-centered perspectives.

The final set of selected papers was examined using thematic analysis. Each study was reviewed to identify:

- The type of phishing detection model used (ML/DL),
- The XAI technique applied (e.g., SHAP, LIME, attention mechanisms),
- The nature of explanations provided (local/global),
- Key findings, strengths, and limitations related to interpretability in phishing detection.

This review process aligns with established practices for semi-systematic and mapping reviews, offering both methodological transparency and flexibility for exploring rapidly developing research areas [6].

Compared to purely narrative reviews, the semi-systematic approach adopted in this study provides improved methodological rigor by combining structured

database searches, explicit inclusion and exclusion criteria, and thematic synthesis. This methodology enables reproducibility while retaining flexibility to analyse emerging interdisciplinary research that spans explainable AI, cybersecurity, and human-centered design. By involving multiple authors in screening and synthesis, the approach also reduces selection bias and strengthens the reliability of the findings.

3. Background and Theoretical Foundation

3.1 Overview of Phishing Attacks

Phishing is a form of cyberattack in which adversaries deceive users into revealing confidential information or performing malicious actions by impersonating legitimate entities. Traditionally delivered through email, phishing has expanded into multiple channels, including fraudulent websites, SMS-based attacks (smishing), and voice-based scams (vishing). Spear-phishing—highly targeted attacks tailored to specific individuals or organizations, has become increasingly prevalent due to the availability of personal data on social platforms and public sources [7]. Large-scale empirical studies have shown that phishing campaigns continuously evolve in linguistic structure, delivery mechanisms, and visual deception strategies, complicating static detection approaches [8].

Modern phishing trends reflect a shift toward more sophisticated and evasive techniques. Attackers now use AI-generated content, dynamic phishing websites, URL obfuscation, and advanced social engineering strategies to bypass traditional security filters. Additionally, multi-stage and multi-vector phishing campaigns combine email, malicious links, and spoofed authentication pages to increase success rates. The rise of business email compromise (BEC), credential-harvesting kits, and adversarial tactics has made phishing a persistent and evolving cybersecurity threat [3].

3.2 ML Models in Phishing Detection

Machine learning (ML) has become a central mechanism for improving phishing detection by analyzing complex patterns in emails, URLs, and website structures. Feature-based models such as decision trees, random forests, support vector machines, and logistic regression rely on hand-crafted features, for example, URL length, domain age, sender attributes, or HTML structure, to differentiate phishing attempts from legitimate communication. These models often provide fast and explainable decisions but may struggle to adapt to rapidly evolving attack vectors [1]. Traditional supervised machine learning models using lexical and host-based features remain effective baselines for phishing detection and are frequently compared against deep learning approaches [9].

NLP-based approaches utilize natural language processing techniques to analyze the textual content of emails or messages. Models can detect linguistic anomalies, sentiment patterns, and semantic relationships indicative of phishing. Techniques such as TF-IDF, word embeddings, and transformer architectures have improved the ability to capture subtle cues in phishing messages [10].

Deep learning architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), LSTMs, and transformer-based models, have further advanced detection capabilities by automatically learning complex, high-dimensional representations. These models excel in classifying phishing websites, analyzing visual layouts, and processing unstructured textual data. However, their high accuracy often comes at the cost of interpretability, making them challenging to trust and validate in real-world settings [3].

Convolutional neural network architectures have been successfully applied to phishing website detection by learning discriminative visual and structural patterns directly from web content [11].

3.3 Explainable AI in Cybersecurity

Explainable Artificial Intelligence (XAI) aims to address the opacity of complex ML models by providing insights into how and why decisions are made. In cybersecurity, XAI is essential for supporting human analysts, ensuring transparency, and enabling reliable deployment of ML-driven detection systems.

One core distinction in XAI is between global and local explanations. Global explanations describe the overall behavior of a model, how features generally influence predictions, while local explanations focus on individual decisions, clarifying why a specific email or URL was flagged as phishing. Both perspectives are valuable for different security tasks, including model auditing and incident investigation [2].

Another key distinction lies between post-hoc and intrinsic interpretability. Post-hoc methods, such as LIME, SHAP, counterfactual explanations, and saliency maps, generate explanations after model training without altering the underlying architecture. In contrast, intrinsic interpretability involves building models, such as decision trees, rule-based models, or attention-enabled architectures, with explanations inherently embedded. While intrinsic models tend to be more transparent, post-hoc techniques allow high-performing deep learning models to be used without sacrificing interpretability [12, p. 2019].

XAI plays a critical role in increasing trust among users and security analysts. By revealing the reasoning behind detection outcomes, XAI improves confidence in automated decisions, supports regulatory compliance, and facilitates more effective threat investigation. It also enhances resilience against adversarial manipulation by allowing analysts to identify model weaknesses and potential exploitation paths. As phishing threats continue to grow, the integration of XAI into detection frameworks becomes essential for ensuring reliable, transparent, and actionable cybersecurity defenses [3].

4. XAI Techniques in Phishing Detection — Review Findings

The findings presented in Sections 4 and 5 represent the results of this semi-systematic review. Rather than reporting experimental outcomes, these sections synthesise evidence from existing studies to identify dominant XAI approaches, their interpretability characteristics, and their implications for phishing mitigation.

4.1 Feature Attribution Methods

Feature attribution methods were the most common XAI techniques found in phishing detection literature. These methods help identify which features—such as URL structure, token positions, domain characteristics, or suspicious keywords—contributed most to the model's prediction.

SHAP (Shapley Additive Explanations):

SHAP is widely used due to its strong theoretical foundation and ability to provide consistent global and local explanations. Studies showed that SHAP effectively identified high-impact phishing indicators such as URL length, abnormal characters, and deceptive lexical cues [13].

LIME (Local Interpretable Model-Agnostic Explanations):

LIME was applied to explain individual phishing samples, highlighting influential tokens or words in email bodies and URL segments. Although sometimes less stable than SHAP, it remained popular for its simplicity and intuitive visual outputs [14].

Permutation Feature Importance:

Permutation-based importance methods were used mainly in traditional ML models, such as Random Forests, to evaluate how shuffling a feature affects prediction performance. Features like "presence of IP in URL" and "domain age" commonly ranked high in phishing detection tasks.

Integrated Gradients:

Deep learning–based phishing detectors (e.g., LSTMs and transformers) applied integrated gradients to reveal token-level contributions across email content or URL structures. This method provided fine-grained explanations for complex neural models [15].

4.2 Interpretable Models

Some studies adopted models that are transparent by design, eliminating the need for post-hoc explainability tools.

Decision Trees:
Decision trees remained useful due to their rule-based, human-readable nature. They generated explicit decision paths such as "URL contains multiple subdomains → classify as phishing."

Rule-Based Models:
Rule-based systems offered deterministic explanations using if–then constructs, supporting environments where compliance and interpretability are essential.

Logistic Regression:

Logistic regression allowed direct interpretation of feature coefficients, enabling analysts to understand which features increased or decreased the likelihood of phishing.

While these models achieved strong interpretability, their performance often lagged behind deep learning models on sophisticated phishing attacks.

4.3 Visualization and Model Transparency Tools

Visualization-based explanation methods, including relevance heatmaps and saliency mapping, have been widely adopted to interpret deep learning decisions by highlighting influential input regions [16].

Heatmaps:
Heatmaps were used to emphasize influential regions in emails or webpage screenshots, highlighting suspicious forms, links, or layout structures. Fig. 1 illustrates a heatmap used against a phishing email.

Fig.1. Illustrative Heatmap Highlighting Suspicious Regions in a Phishing Email

Visualization-based XAI techniques provided clear insight into how DL models processed and classified phishing content.

Attention Mechanisms:

Transformer-based phishing detectors visualized attention weights to show which tokens or URL parts influenced classification outcomes most. Attention-based XAI is particularly effective for text-rich phishing emails [17].

Saliency Maps:

Saliency maps were employed in visual phishing detection to highlight critical UI elements, logos, login fields, and fake certificates that the model found suspicious.

These tools improved clarity for cybersecurity analysts, supporting human–AI collaboration in investigation workflows.

4.4. Hybrid and Emerging XAI Methods

Emerging studies explored hybrid approaches that combine symbolic reasoning, multiple explainers, or alternative forms of interpretability.

Neural-Symbolic XAI:
These approaches integrated rule-based logic with neural network architectures, providing a balance between accuracy and interpretability.

Ensemble Explainability:
Some works combined SHAP (global) with LIME (local) or added attention maps to produce multi-layered explanations, offering richer interpretive insights.

Counterfactual Explanations:
Counterfactual methods showed how small changes, such as modifying a suspicious keyword or altering a URL parameter, could shift a classification from phishing to legitimate. This method helped reveal model sensitivity and decision boundaries.

Although promising, hybrid methods require higher computational resources and remain less common in phishing-specific literature.

4.5. Comparative Synthesis

Table 1 presents a comparative synthesis of dominant XAI techniques applied in phishing detection, highlighting trade-offs between interpretability, performance, and real-world deployment feasibility.

Table 1. Comparative synthesis of XAI techniques applied in phishing detection

XAI Technique	Model Type	Explanation Level	Strengths	Limitations	Deployment Suitability
SHAP	ML / DL	Global & Local	Stable, theoretically grounded	Computational overhead	Medium
LIME	Model-agnostic	Local	Intuitive, flexible	Explanation instability	Medium
Attention Mechanisms	DL	Local	Context-aware explanations	Not causally grounded	High
Decision Trees	ML	Global	Fully interpretable	Lower detection accuracy	High
Counterfactuals	ML / DL	Local	Actionable explanations	Hard to generate	Low–Medium

A cross-study comparison reveals several important insights:

Best Performing XAI Methods

- **SHAP** provided the most reliable and widely applicable explanations across both ML and DL models.
- **Attention mechanisms** excelled for deep learning models processing text-based phishing content.

Most Interpretable Approaches

- **Decision trees** and **rule-based models** were easiest for analysts to interpret.
- **LIME** remained valuable for quick, instance-level explanations.

Key Trade-offs

- **Interpretable models offer transparency** but may sacrifice detection accuracy.
- **Deep models achieve higher accuracy** but rely heavily on post-hoc XAI.
- **Hybrid methods provide richer explanations** but at higher computational cost.

Overall, no single XAI technique is universally optimal. The literature suggests that combining feature attribution, visual interpretability, and interpretable model

design can provide the most balanced approach for real-world cybersecurity applications.

5. Impact of XAI in Phishing Mitigation – Review Findings

Explainable Artificial Intelligence (XAI) plays a critical role in enhancing the security, usability, and operational effectiveness of phishing detection systems. As phishing attacks continue to evolve in complexity, XAI enables both technical and non-technical stakeholders to understand the reasoning behind automated decisions, supporting better trust, coordination, and response mechanisms.

Further, Explainable AI methods can be broadly categorized into intrinsic and post-hoc approaches, offering different trade-offs between transparency, fidelity, and usability [18]. This chapter summarizes the impact of XAI on transparency, human–AI collaboration, and real-world deployment considerations.

5.1 Improving Transparency and Trust

One of the most significant contributions of XAI in phishing mitigation is its ability to improve transparency in automated decision-making. Explanations help analysts understand *why* a particular email or URL was classified as phishing, making model behaviour more predictable and interpretable. Human-centered explainability research indicates that explanations aligned with human reasoning significantly enhance trust, confidence, and decision quality in AI-assisted systems [19].

Influence on Analyst Decision-Making

XAI-generated explanations, such as SHAP plots, attention maps, or feature contribution scores, allow cybersecurity analysts to verify whether the model is focusing on meaningful indicators, such as deceptive keywords, suspicious sender domains, or structural anomalies in URLs. This validation helps analysts make confident decisions, reduce false alarms, and identify potential blind spots in the model's reasoning. Studies indicate that transparent ML models improve analysts' ability to detect misclassifications and refine incident response workflows [20].

User Trust Improvements

Trust is essential when deploying AI systems in high-risk cybersecurity environments. When end-users and analysts can clearly see why a system flagged content as malicious, they are more likely to accept and rely on model outputs. Evidence shows that interpretability increases perceived reliability and promotes responsible adoption of AI in security operations [21].

5.2 Human–AI Collaboration

XAI strengthens the interaction between human analysts and automated phishing detection systems, allowing both to work more effectively together.

Helping Cybersecurity Teams

By providing clear, interpretable insights into phishing indicators, XAI reduces cognitive load for analysts and accelerates triage processes. For example, attention-based explanations can highlight critical tokens or patterns that require immediate review, helping security teams prioritize alerts. This collaborative advantage enhances

overall detection performance and reduces manual analysis time.

Supporting End-User Training

XAI explanations can also be used in phishing awareness and training programs. Highlighting which elements of an email prompted its classification,such as unusual URL obfuscation, manipulated brand names, or mismatched sender information, helps educate users on real-world phishing characteristics. Prior work suggests that visual, example-based explanations improve user comprehension and retention [14]**.**

XAI therefore contributes not only to technical defense but also to behavioural and educational aspects of phishing mitigation.

5.3 Practical Deployment Considerations

While XAI provides substantial benefits, deploying explainable phishing detection systems in real-world environments introduces several operational challenges.

Real-Time Processing Constraints:

a) Generating explanations—especially for deep neural networks—can be computationally demanding. Methods like SHAP or Integrated Gradients may introduce latency that is incompatible with real-time email filtering or URL scanning. Lightweight explanation methods or approximation-based techniques are often necessary to maintain system responsiveness [15].

b) Explainability vs. Model Performance

5.4 A recurring challenge highlighted in the reviewed studies is the trade-off between interpretability and accuracy.

- Interpretable models (e.g., decision trees, logistic regression) are easier to understand but may fail to capture the complexity of sophisticated phishing attacks.
- High-performing deep learning models achieve superior detection rates but require post-hoc XAI to remain transparent.

Balancing these factors is crucial for building reliable, deployable phishing detection systems that meet both security and operational requirements.

6. Discussion

6.1 Discussion in Relation to RQ1: XAI Methods Applied in Phishing Detection

In relation to RQ1, the reviewed literature demonstrates that feature attribution–based XAI methods are the most widely adopted approaches in phishing detection systems. Techniques such as SHAP, LIME, Integrated Gradients, and attention mechanisms dominate existing research due to their flexibility and compatibility with both traditional machine learning and deep learning models. Among these, SHAP emerged as the most consistently applied method, offering both global and local explanations grounded in solid theoretical principles. Its ability to quantify feature contributions has been particularly effective in identifying phishing indicators such as URL obfuscation patterns, suspicious lexical tokens, and anomalous sender attributes.

Attention mechanisms were predominantly used in deep learning architectures, especially transformer-based models, to provide token-level explanations for text-based phishing detection. These mechanisms enable visualization of which parts of an email or URL the model prioritizes, thereby enhancing interpretability without significantly compromising detection performance. In contrast, interpretable models such as decision trees and logistic regression offer intrinsic transparency through rule-based reasoning and coefficient analysis but generally underperform when faced with sophisticated, evolving phishing attacks.

Overall, the findings indicate a clear research preference for post-hoc, model-agnostic XAI techniques that preserve high detection accuracy while providing meaningful explanations. However, the dominance of post-hoc methods also reflects the ongoing challenge of achieving both interpretability and performance within a single unified model architecture.

6.2 Discussion in Relation to RQ2: Enhancing Transparency and Decision-Making in Phishing Mitigation

Addressing RQ2, the review confirms that XAI techniques play a critical role in enhancing transparency, trust, and decision-making in phishing mitigation. Explainability mechanisms—such as feature importance plots, attention visualizations, saliency maps, and heatmaps—enable cybersecurity analysts to understand why a particular email, URL, or webpage is classified as phishing. This transparency allows analysts to validate model outputs, identify false positives, and assess whether the model's reasoning aligns with known phishing characteristics.

The reviewed studies highlight that XAI significantly improves human–AI collaboration by reducing cognitive load and supporting faster incident triage. Rather than treating AI systems as opaque decision-makers, analysts can interact with explanations to refine detection strategies and prioritize alerts more effectively. This is particularly important in operational security environments where analysts must justify actions, comply with regulatory requirements, and respond to threats in real time.

Beyond analyst support, XAI also contributes to user trust and awareness. Several studies suggest that explanation-driven insights can be integrated into phishing awareness and training programs, helping end-users recognize deceptive patterns in real-world attacks. By illustrating how specific elements—such as manipulated brand names, mismatched URLs, or urgent language—trigger detection, XAI enhances both technical defenses and human resilience against phishing.

Collectively, these findings demonstrate that explainability is not merely an auxiliary feature but a foundational requirement for responsible and effective AI-driven phishing mitigation.

6.3 Discussion in Relation to RQ3: Limitations and Research Gaps

In response to RQ3, the review identifies several persistent limitations and research gaps that constrain the practical impact of XAI-based phishing detection systems. The absence of standardized evaluation methodologies for explainability remains a critical challenge, particularly in high-risk domains such as cybersecurity [22] .A major concern is the overreliance on public benchmark datasets, such as PhishTank and static email corpora, which often fail to capture the diversity, context, and rapid evolution of real-world phishing attacks. This raises questions about the generalizability and robustness of both detection models and their explanations.

Another critical limitation is the lack of human-centered evaluation. Most studies assess XAI effectiveness using technical metrics or qualitative examples, with minimal empirical evaluation of how explanations are perceived and used by analysts

or end-users. Without systematic user studies, it remains unclear whether explanations genuinely improve understanding, trust, or decision quality in operational environments.

The absence of standardized metrics for explanation quality further complicates comparative evaluation across studies. Explanation clarity, usefulness, stability, and trustworthiness are often measured inconsistently or not at all, limiting the ability to draw strong conclusions about the relative effectiveness of different XAI approaches.

From a deployment perspective, computational overhead and latency remain significant challenges. Techniques such as SHAP and Integrated Gradients, while informative, may introduce delays incompatible with real-time email filtering or large-scale enterprise systems. Moreover, few studies address the adversarial robustness of XAI, despite the risk that attackers could manipulate inputs to exploit or mislead explanation mechanisms.

These limitations highlight the need for more holistic, deployment-aware, and user-focused research to advance explainable phishing detection beyond experimental settings.

6.4 Implications for Future Explainable Phishing Detection Research

The findings of this review suggest several important implications for future research. First, there is a growing need to develop intrinsically interpretable models that integrate transparency directly into the detection process while maintaining competitive accuracy. Second, human-centered evaluation frameworks should be prioritized to assess how explanations support real-world decision-making. Third, research should focus on lightweight and scalable XAI techniques suitable for real-time deployment. Finally, greater attention must be given to adversarial threats against explainable systems, ensuring that explanations enhance security rather than introduce new vulnerabilities.

7. Conclusion

This review demonstrates that Explainable Artificial Intelligence has substantially enriched phishing detection research by transforming opaque classification systems into transparent, accountable, and analyst-supportive tools. Evidence from the reviewed studies shows that XAI improves interpretability, supports informed decision-making, and strengthens trust in AI-driven cybersecurity systems. These findings directly address the central research objective by confirming that explainability is not a peripheral enhancement, but a foundational requirement for effective and responsible phishing mitigation.

The review examined the role of Explainable Artificial Intelligence (XAI) in phishing detection, highlighting how explainability enhances transparency, trust, and operational effectiveness in cybersecurity environments. Across the reviewed studies, it is evident that XAI contributes significantly to the interpretability of machine learning and deep learning models, offering insights into model behaviour that are essential for both analysts and end-users. Feature attribution methods, interpretable models, visualization tools, and emerging hybrid techniques each play a distinct role in supporting clearer, more accountable decision-making.

The findings demonstrate that while deep learning approaches continue to deliver strong predictive performance, their complexity requires robust explainability mechanisms to ensure they can be used responsibly and effectively in security operations. At the same time, interpretable models offer high levels of transparency, though often at the expense of detection accuracy. Hybrid frameworks provide a

promising balance, but require further refinement and computational optimisation before they can be widely adopted.

The review also revealed persistent challenges, including limited real-world evaluation, a lack of standardized metrics for explanation quality, reliance on static datasets, and constraints in deploying XAI-enabled models in real-time systems. Addressing these gaps will be crucial for advancing practical, trustworthy phishing detection solutions. Future research should focus on developing lightweight, scalable XAI methods; conducting human-centered usability studies; improving dataset diversity; and exploring adversarial robustness within explainable frameworks.

Overall, XAI represents a critical component of modern phishing mitigation strategies. As phishing attacks continue to evolve, integrating explainability into detection systems will not only strengthen cybersecurity defenses but also support more informed, collaborative, and trustworthy decision-making across both technical and non-technical users.

REFERENCES

1. N. Alsuqayh, A. Mirza and A. Alhogail, "Exploring feature engineering and explainable AI for phishing website detection: a systematic literature review," International Journal of Electrical and Computer Engineering (IJECE), pp. DOI: 10.11591/ijece.v15i6.pp5863-5878, 2025.
2. P.R. Chandre, P. Bhujbal, A. Jadhav, B. D. Shendkar, A. Wangikar and R. Sachdeo, "A comprehensive review of interpretable machine learning techniques for phishing attack detection," IAES International Journal of Artificial Intelligence, pp. DOI: 10.11591/ijai.v14.i4.pp3022-3032, 2025.
3. M. Tawfik, A. A. Abu-Ein, A. Abdelhaliem and S. FathiIslam, "Explainable few-shot learning with modern BERT for detecting emerging phishing attacks using XF PhishBERT: Explainable few-shot learning with modern BERT...M. Tawfik et al.," Scientific Reports, pp. DOI: 10.1038/s41598-025-27500-0, 2025.
4. Thalpage, N. The Integration of Machine Learning and Explainable AI in Business Digitization: Unleashing the Power of Data - A Review. Journal of Digital Science, 6(1), 2023. https://doi.org/10.33847/2686-8296.6.1_2 .
5. F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Computer Science, Philosophy, 2017. https://api.semanticscholar.org/CorpusID:11319376.
6. N. S. Thalpage and T. A. D. Nisansala, "Exploring the Opportunities of Applying Digital Twins for Intrusion Detection in Industrial Control Systems of Production and Manufacturing – A Systematic Review," Data Protection in a Post-Pandemic Society, pp. https://doi.org/10.1007/978-3-031-34006-2_4, 2023.
7. M. Mehdi, Y. Farzaneh, S. Farzaneh, S. Elham, S. Elham and H. Gharaee, "An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features," in Conference: 2019 5th International Conference on Web Research (ICWR), 2019. DOI: 10.1109/ICWR.2019.8765265
8. S. Baki and R. Verma, "Sixteen Years of Phishing User Studies: What Have We Learned?," IEEE Transactions on Dependable and Secure Computing, 2022, DOI: 10.1109/TDSC.2022.3151103
9. O. K. Sahingoz, E. Buber, O. Demir and B. Diri, "Machine learning based phishing detection from URLs," Expert Systems with Applications, 2019, https://doi.org/10.1016/j.eswa.2018.09.029.
10. R. Basnet, S. Mukkamala and A. H. Sung, "Detection of Phishing Attacks: A Machine Learning Approach," Studies in Fuzziness and Soft Computing, pp. DOI: 10.1007/978-3-540-77465-5_19, 2008.
11. M. Adebowale and K. Lwin, "Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection," in Deep Learning with Convolutional Neural Network and Long Short-Term Memory for Phishing Detection, 2019, DOI: 10.1109/SKIMA47702.2019.898242.
12. U. Bhatt, A. Xiang and P. Eckersley, "Explainable machine learning in deployment," in Computer Science Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, https://doi.org/10.1145/3351095.3375624 .

13. S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in NIPS, 2017, DOI: 10.48550/arXiv.1705.07874.
14. M. T. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in the 22nd ACM SIGKDD International Conference , 2016, https://doi.org/10.1145/2939672.2939778.
15. M. Sundararajan, A. TalyAnkur and T. Yan, "Axiomatic Attribution for Deep Networks," p. DOI: 10.48550/arXiv.1703.01365, 2017.
16. W. Samek, T. Wiegand, Klaus-Robert and M.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," 2017, DOI: 10.48550/arXiv.1708.0829.
17. Vaswani, N. Shazeer, N. Parmar and I. Polosukhin, "Attention Is All You Need," p. DOI: 10.48550/arXiv.1706.03762, 2017.
18. G. Riccardo, A. Monreale, F. Turini and F. Giannotti, "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys , p. DOI: 10.1145/3236009, 2018.
19. T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," p. DOI: 10.1016/j.artint.2018.07.007, 2017.
20. F. Doshi-Velez and B. Kim, "A Roadmap for a Rigorous Science of Interpretability," p. DOI: 10.48550/arXiv.1702.08608, 2017.
21. F. Charmet, H. Chandra, Tanuwidjaja, S. Ayoubi and Z. Zhang, "Explainable artificial intelligence for cybersecurity: a literature survey," annals of telecommunications - annales des télécommunications, pp. DOI: 10.1007/s12243-022-00926-7, 2022.

Aims and Objectives

Published online by Institute of Cited Scientists, Cyprus, two times a year, Journal of Digital Science (JDS) is an international peer-reviewed journal which aims at the latest ideas, innovations, trends, experiences and concerns in the field of digital science covering all areas of the scholarly literature of the sciences, social sciences and arts & humanities.

The principal ambition of this periodical is the efficacious propagation of original insights and outcomes derived from human cerebral activity and exemplified in scholarly treatises through the utilisation of contemporary information and digital technology.

The main topics currently covered include: Artificial Intelligence in Cybersecurity and Decision-making process, Cryptocurrency as Digital Assets, Digital decoding of language, Digital Technology in Health Care.

Contact information

Journal URL: https://ics.events/journal-of-digital-science/
Email: conf@ics.evnets
Printed online from the original layout under the imprint at:
1, Vlachou, Nicosia, The Republic of Cyprus

The picture on JDS cover was generated by Andrei Stepanov.