Explainable AI in Cybersecurity: Interpreting Machine Learning Models for Phishing Detection

Matthew McCarthy

Abstract—As phishing detection systems evolve from traditional classifiers to transformer-based architectures, understanding model behaviour has become a critical security priority. This systematic literature review investigates the application of Explainable AI (XAI) frameworks—specifically LIME and SHAP—in addressing transparency challenges in phishing detection. Synthesizing evidence from 12 primary experimental studies and 2 comprehensive surveys (2016–2025), the review assesses explanation fidelity, attribution stability, and human-centred utility. Findings indicate that while post-hoc surrogate explanations demonstrate strong fidelity for ensemble-based models, they exhibit instability and inconsistent token-level attributions in transformers such as BERT and RoBERTa. An evaluation gap is also identified, as many studies emphasize qualitative visual coherence over quantitative metrics of fidelity and robustness. Although XAI methods provide a theoretical foundation for transparency, their practical utility in professional security environments remains limited. Future research should establish standardized fidelity benchmarks and integrate human-in-the-loop evaluations to reduce analyst cognitive load and enable trustworthy deployment.

1 Introduction

Cybersecurity threats have increased substantially in both frequency and sophistication over the past decade, with phishing attacks remaining one of the most prevalent methods for compromising sensitive information. Phishing emails impersonate trusted entities through social engineering techniques that exploit urgency, fear, and authority to trick recipients into revealing credentials or downloading malware. These attacks are increasingly personalized through spear-phishing tactics, making automated detection progressively more challenging [2].

To counter this threat, machine learning (ML) has become central to modern phishing detection systems. Traditional approaches based on blacklists and rule-based heuristics proved inadequate against rapidly evolving attack strategies. Modern ML models—ranging from Support Vector Machines to transformer-based architectures like BERT—have demonstrated improved detection performance by analysing email content, metadata, and linguistic patterns [9, 7]. However, these advances have introduced a critical limitation: most high-performing models operate as "black boxes," providing minimal insight into their decision-making processes. This opacity challenges security professionals who must validate alerts and investigate false positives without understanding model reasoning [1].

Explainable Artificial Intelligence (XAI) addresses this limitation by providing human-interpretable explanations for model predictions. Among XAI techniques, Local Interpretable Model-agnostic Explanations (LIME) [11] and SHapley Additive exPlanations (SHAP) [6] have gained prominence due to their model-agnostic nature. These methods explain predictions across diverse architectures, from traditional machine learning to deep learning and transformers. This flexibility makes LIME and SHAP particularly valuable for phishing detection interpretability [10, 13].

This literature review systematically examines the application of LIME and SHAP to phishing detection systems. The review makes three key contributions: (1) it analyzes how LIME and SHAP perform across traditional ML, deep learning, and transformer-based detection models; (2) it critically evaluates the practical limitations of these techniques, including computational costs and explanation consistency; and (3) it identifies methodological gaps in current research, particularly regarding explanation quality evaluation and real-world deployment considerations.

To achieve these objectives, this review addresses three research questions:

1) **RQ1:** How effective are LIME and SHAP in providing meaningful explanations for phishing detection across different model architectures?
2) **RQ2:** What are the practical limitations of LIME and SHAP in cybersecurity contexts?
3) **RQ3:** What critical research gaps exist in applying explainable AI (XAI) to phishing detection?

The remainder of this paper is structured as follows. Section 2 describes the systematic review methodology. Section 3 provides background on phishing detection evolution and foundational XAI concepts. Section 4 examines specific applications of LIME and SHAP to phishing detection. Section 5 offers critical analysis of current approaches and their limitations. Section 6 identifies research gaps and proposes future directions. Finally, Section 7 concludes with implications for practice and research.

2 Methodology

This literature review employed a systematic multi-stage search strategy across academic databases to identify studies on LIME and SHAP applications in phishing detection. Searches were conducted in November–December 2025, targeting publications from 2020–2025.

2.1 Search Strategy

An initial scoping search was conducted in Google Scholar using the query "Phishing detection using machine learning

and explainable AI" (2020–2025), yielding 17,200 results. Title screening of the first 50 results identified key papers and keywords (LIME, SHAP, XAI) to inform subsequent targeted searches. Google Scholar served as a scoping exercise to establish the research landscape; all papers were subsequently verified through systematic searches in peer-reviewed academic databases.

Based on this scoping exercise, the following targeted database searches were then conducted using refined search strings. IEEE Xplore advanced search queries used can be seen in Listing 1 and Listing 2.

Listing 1. IEEE Xplore advanced search query

```
("All Metadata":Phishing detection)
AND ("All Metadata":machine learning)
AND ("All Metadata":LIME)
AND ("All Metadata":SHAP)
AND ("All Metadata":Explainable AI)
```

Listing 2. IEEE Xplore second advanced search query

```
("LIME" OR "SHAP"
OR "Explainable AI" OR "XAI")
AND ("phishing detection" OR "malicious email"
OR "URL classification")
```

Snowballing identified additional relevant papers by examining the reference lists (backward snowballing) and citation records (forward snowballing) of primary studies, which may include papers published in academic databases such as ACM and MDPI.

2.2 Screening Process

A three-stage screening was applied:

1) **Title screening** for keyword presence (phishing detection, machine learning, LIME/SHAP, explainable AI).
2) **Abstract review** for methodological relevance to phishing detection.
3) **Full-text assessment** using keyword searches and manual review to verify LIME/SHAP implementation and experimental results.

- **Inclusion:** Peer-reviewed papers implementing machine learning for phishing detection, including those utilizing LIME and/or SHAP for model interpretability.
  Note: Original LIME (2016) and SHAP (2017) works are included for methodological foundation.
- **Exclusion:** Non-peer-reviewed works such as pre prints from databases like ArXiv, papers published pre-2020, and studies lacking a primary focus on phishing detection.

2.3 Final Corpus

Following the removal of duplicates and the application of screening criteria, a final corpus of **14 papers** was selected for detailed review. This selection process followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to ensure methodological transparency and rigor. Papers identified during the initial Google Scholar scope were re-retrieved via their primary publishers or indexed sources (e.g., IEEE Xplore and ScienceDirect) to ensure metadata accuracy and version control. Consequently, these are accounted for within their respective database totals in Table 1.

TABLE 1
Systematic Literature Search Results

Database/Source	Initial Results	Final Selected
Google Scholar (Scoping)	17,200	–[a]
IEEE Xplore	16	6
ScienceDirect	83	3
Other Sources (Snowballing)	–	5
Total	**17,299**	**14**

[a]Papers re-verified via primary databases to ensure peer-review status.

3 Background & Related Work

Phishing detection has evolved from static rule-based systems to sophisticated machine learning (ML) and deep learning models. While these computational methods have significantly improved detection accuracy, their internal complexity often results in a "black box" nature that limits trust and transparency, a critical issue in cybersecurity contexts [14, 5]. This section outlines the evolution of phishing detection methodologies and the subsequent shift toward Explainable AI (XAI) as a necessary solution to the interpretability gap.

3.1 The Evolution of Phishing Detection

Traditional phishing detection relied heavily on rule-based filters, blacklists, and signature matching [1, 3]. While computationally efficient, these static methods frequently failed against novel zero-day attacks or obfuscation techniques [14]. To address these limitations, classical Machine Learning (ML) models, such as Support Vector Machines (SVM) and Random Forests, were introduced to improve adaptability. However, these approaches typically required extensive manual feature engineering and struggled to detect sophisticated social engineering tactics embedded within email content [7, 3].

More recently, Deep Learning (DL) architectures, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have automated the feature extraction process, with some reported accuracy rates ranging between 92% and 97% [3]. State-of-the-art Transformer-based models, such as BERT and RoBERTa, have further pushed performance boundaries, with some studies reporting accuracies as high as 99.43% [7]. Despite these performance gains, the non-linear complexity of DL models obscures their decision-making logic. In cybersecurity, where automated decisions directly impact user safety and data integrity, understanding *why* a model flags a specific email or URL is essential for alert validation, analyst trust, and regulatory compliance [1].

3.2 Explainable AI (XAI) Frameworks

XAI addresses the opacity of complex algorithms by generating human-understandable interpretations of model predictions. Within the cybersecurity domain, XAI techniques are generally categorized along two primary dimensions:

- **Model-Agnostic vs. Model-Specific:** Model-agnostic methods, such as LIME [11] and SHAP [6], can interpret any classifier regardless of its internal architecture. This flexibility makes them particularly valuable for retrofitting interpretability onto high performing "black box models" without altering their underlying structure. Conversely, model-specific methods rely on the intrinsic properties of a specific architecture, such as attention mechanism visualization in transformers.
- **Local vs. Global Interpretability:** Local explanations clarify the reasoning behind a *single* prediction (e.g., why a specific URL was flagged), which is crucial for security analysts investigating individual alerts [10]. Global explanations summarize the model's behaviour across an entire dataset, helping researchers validate that the model relies on legitimate security indicators rather than spurious correlations or bias.

3.3 Key Post-Hoc Techniques: LIME and SHAP

Two model-agnostic, post-hoc techniques are commonly used in the literature for explaining phishing detection:

- **LIME (Local Interpretable Model-agnostic Explanations):** LIME generates explanations by fitting an interpretable surrogate model locally around a specific prediction. For a "black-box" model f and an input x, LIME seeks a surrogate $g \in G$ that minimizes the objective:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g), \tag{1}$$

  where $\mathcal{L}(f, g, \pi_x)$ measures the fidelity of g in approximating f within the local neighbourhood π_x of x, and $\Omega(g)$ penalizes the complexity of g to ensure interpretability [11]. LIME generates perturbed versions of x, obtains the corresponding predictions from f, and fits g to these perturbed samples weighted by proximity to x. The resulting surrogate identifies the features that most influence the model's prediction locally.
- **SHAP (SHapley Additive exPlanations):** SHAP provides a theoretically principled approach to feature attribution, grounded in cooperative game theory. For a model f and input x, it computes the Shapley value ϕ_i for each feature i, which represents the average marginal contribution of that feature across all subsets $S \subseteq F \setminus \{i\}$:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{2}$$

  To compute ϕ_i, SHAP evaluates the model's output for all combinations of feature subsets and measures the marginal contribution of adding feature i. Unlike LIME, SHAP satisfies *local accuracy*, *missingness*, and *consistency*, ensuring a mathematically robust and consistent attribution of feature importance [6].

These frameworks provide the methodological foundation for the applied studies reviewed in Section 4.

4 Applications of XAI in Phishing Detection

The systematic selection process yielded a corpus of studies characterizing the current state of explainability in phishing mitigation. While most works focus on established machine learning (ML) classifiers or state-of-the-art transformer architectures, only a subset employ post-hoc frameworks to demystify automated decision-making. The following analysis synthesizes these findings across three thematic dimensions: the technical fidelity of explanations across different model architectures, the specific linguistic and structural features identified as malicious indicators, and the practical utility of XAI visualizations for human intervention.

4.1 Performance and Fidelity Across Architectures

The reviewed literature reveals a technical divergence in how explainability is applied across varying model complexities, directly addressing RQ1. Traditional machine learning (ML) ensembles, such as Random Forest and XGBoost, remain prevalent due to their high performance and ease of interpretation using post-hoc methods. For instance, Al-Subaiey et al. [13] and Pavani, Mahitha, and Maheswari [10] demonstrate that LIME and SHAP provide stable local explanations when paired with traditional ML, though they often rely on qualitative visual alignment rather than quantitative fidelity metrics such as deletion/insertion tests. In contrast, research into deep learning (DL) and transformer architectures highlights a critical "black box" limitation. While models such as DistilBERT and RoBERTa achieve state-of-the-art accuracies [7], their internal operations remain opaque. Authors like Aldoufani, Eleyan, and Bejaoui [1] have begun addressing this by integrating LIME to provide token-level explanations for transformers, yet a significant gap persists in measuring the mathematical consistency or "stability" of these explanations across diverse datasets.

4.2 Thematic Feature Attribution

A consensus exists across the corpus regarding the primary indicators of malicious intent, though the focus shifts between URL-based and content-based features. Studies focusing on URL structure consistently identify features such as `punny_code`, `domain_in_brand_list`, and the presence of external hyperlinks as the most influential indicators [5]. Mia, Derakhshan, and Pritom [8] further refine this by showing that lexical features like URL length and special character frequency are most influential. Conversely, when analysing email bodies, XAI reveals that models prioritize semantic "urgency" tokens such as "verify," "suspension," and "password" [13]. Notably, Mia, Derakhshan, and Pritom [8] challenge the assumption of feature universality, demonstrating that feature importance can vary significantly across datasets, which underscores the risk of dataset-specific bias in phishing detection models.

4.3 Visualization and Human-Centric Utility

While technical interpretability is well-documented, practical human-centric utility remains an area of significant controversy, providing critical insights into the practical limitations of current frameworks (RQ2). Many studies utilize bar charts and colour-coded text to present feature importance, assuming these formats reduce the cognitive load for security analysts. Some researchers have moved toward practical deployment, such as Al-Subaiey et al. [13] and Aldoufani, Eleyan, and Bejaoui [1], who implemented web-based platforms and Chrome extensions to provide real-time explanations. However, a critical methodological flaw identified across the reviewed papers is the near-total absence of empirical user validation. Without human-in-the-loop testing, it remains unproven whether these XAI visualizations actually improve the speed or accuracy of a human analyst's response to phishing threats [14].

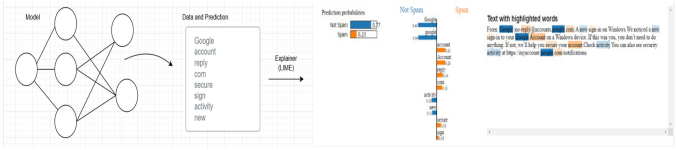

Fig. 1. Conceptual XAI pipeline (left) alongside a LIME explanation output (right) for a misclassified phishing email, showing token-level feature importance [13].

As illustrated in Figure 1, these visualizations aim to bridge the gap between complex model logic and actionable security insights by highlighting high-influence tokens directly within the email body [13].

5 Discussion and Critical Analysis

The systematic review of the selected corpus reveals that while Explainable AI (XAI) significantly enhances the transparency of phishing detection, several critical methodological and practical gaps remain. This section evaluates the trade-offs between model performance and interpretability, the lack of empirical human validation, and the limitations regarding explanation stability. By identifying these discrepancies, this section directly addresses the research gaps outlined in RQ3.

5.1 The Stability-Fidelity Gap in Phishing Explanations

A significant methodological oversight identified across the corpus is the lack of quantitative metrics for explanation stability and fidelity. Most studies, including those by Al-Subaiey et al. [13] and Pavani, Mahitha, and Maheswari [10], rely on qualitative visual alignment to justify the accuracy of LIME and SHAP outputs. However, without formal fidelity scores or deletion/insertion tests, the degree to which an explanation accurately reflects the model's underlying logic remains unverified [8]. Furthermore, while Lundberg and Lee [6] emphasizes that SHAP provides unique solutions satisfying local accuracy and consistency, perturbation-based methods such as LIME are known to exhibit stochastic instability [1] that seems to be underexamined in the literature. In a high-stakes cybersecurity environment, an unstable explanation that produces varying feature importance for the same input could undermine analyst trust and lead to inconsistent incident response.

5.2 The Evaluation Crisis: Absence of Human-Centric Validation

Despite the stated goal of XAI being "human understandable," there is a pervasive "human-out-of-the-loop" trend in current phishing research. Out of the reviewed corpus, only the foundational works by Ribeiro, Singh, and Guestrin [11] and Lundberg and Lee [6] incorporate empirical human subject experiments to validate interpretability. While recent work by Fan et al. [4] successfully applies SHAP to model human susceptibility factors—demonstrating the value of XAI in understanding user behaviour—this human-centric focus has not yet translated to the evaluation of detection tool interfaces. Modern applied studies in phishing detection, such as Aldoufani, Eleyan, and Bejaoui [1] and Shafin [12], present sophisticated visualizations like color-coded tokens and SHAP waterfall plots but do not conduct user studies to measure their actual utility in professional environments. [5, 10]. This creates an "interpretability paradox" where explanations are technically sound but potentially too complex for real-time human intervention. Without measuring metrics like *time-to-decision* or *cognitive load*, the practical value of these frameworks remains theoretical.

TABLE 2
The Interpretability Paradox in Phishing Detection

Dimension	Technical Success	Practical Limitation
Cognitive Load	High-fidelity feature attribution (e.g., SHAP).	Complex plots may exceed analyst processing capacity in real-time.
Trust	"Persuasive" visual highlights (e.g., LIME).	Risk of automation bias; users stop verifying the underlying threat.
Decision Accuracy	High model confidence and precision.	Explanations do not currently guarantee faster or better human decisions.

As summarized in Table 2, the gap between algorithmic transparency and human utility suggests that current XAI implementations may inadvertently hinder rather than help a security analyst's performance under pressure.

5.3 Dataset Bias and the Generalizability of Explanations

The findings of Mia, Derakhshan, and Pritom [8] represent a critical pivot in the literature by demonstrating that feature importance is not universal across datasets. This suggests that XAI may highlight dataset-specific biases or "shortcuts" learned by the model (*Clever Hans effect*) instead of true security indicators. For instance, a model might achieve high accuracy by over-relying on a specific sender domain present in the training set, which would fail in a real-world zero-day scenario. This critique is particularly relevant for high-performing Transformer models like RoBERTa and BERT [7], which achieve near-perfect accuracy but offer no

inherent mechanism to distinguish between genuine semantic learning and spurious correlations within the training data. Addressing this gap requires a shift toward cross-dataset validation of XAI explanations to ensure their generalizability across diverse phishing landscapes.

6 Research Gaps & Future Directions

Based on the critical analysis of the selected corpus, three primary gaps necessitate immediate research attention to transition XAI from a theoretical enhancement to a practical security tool.

6.1 Standardization of Fidelity Metrics

Current literature predominantly relies on qualitative visual checks to validate explanations, a method deemed insufficient for high-stakes cybersecurity environments. A major gap exists in the use of rigorous, quantitative metrics—such as local fidelity scores, the consistency of feature importance, and deletion/insertion tests on phishing detection models. Future research should establish a standardized "Phishing XAI Benchmark" that mandates these mathematical validations before visualizations are presented to analysts. This would prevent the deployment of visually plausible but quantitatively unverified explanations, addressing a critical risk in current phishing XAI research [8].

6.2 Operational Human-Centric Validation

While Al-Subaiey et al. [13] and Aldoufani, Eleyan, and Bejaoui [1] have successfully integrated XAI into web-based detection platforms, the actual impact of these tools on human decision-making remains largely unexplored. Empirical studies are needed to measure *Time-to-Decision* (TTD) and *False Positive Assessment Accuracy* in operational contexts. Future work must move beyond algorithmic performance metrics to determine whether XAI visualizations genuinely improve user understanding or inadvertently contribute to information overload [14].

6.3 Cross-Dataset Generalizability

The demonstrated instability of feature importance across datasets, as highlighted by Mia, Derakhshan, and Pritom [8], exposes a vulnerability in current model training paradigms. XAI is currently utilized primarily for post-hoc justification rather than model debugging. A critical future direction involves using XAI insights during the training phase to identify and penalize dataset-specific artifacts (e.g., specific sender domains) that do not generalize. Developing "explanation-guided training" protocols could help align high-performance Transformer models with robust, universal security indicators.

7 Conclusion

This systematic review evaluated the efficacy and limitations of LIME and SHAP in deciphering black-box phishing detection models. In addressing the defined research questions, the analysis yields three concluding insights.

Regarding **RQ1 (Effectiveness)**, LIME and SHAP prove highly effective for interpreting traditional ML ensembles, providing clear feature attribution for URL and metadata characteristics. However, their application to Transformer-based architectures remains computationally expensive and susceptible to stability issues, often failing to capture the contextual nuance of deep learning models [1, 13].

Regarding **RQ2 (Limitations)**, the review identifies a significant "Human-out-of-the-loop" paradox. While technical implementations of XAI have advanced, the absence of empirical user validation means that the practical utility of these explanations for security professionals remains theoretical rather than proven [12].

Regarding **RQ3 (Gaps)**, the field is currently hindered by a lack of standardized fidelity metrics and cross-dataset validation. As demonstrated by Mia, Derakhshan, and Pritom [8], high accuracy does not equate to robust logic. Consequently, the future of phishing detection lies not merely in higher accuracy scores, but in the development of models that are demonstrably faithful, robust across diverse datasets, and empirically validated to enhance human operator performance.

References

[1] Youssuf Aldoufani, Amna Eleyan, and Tarek Bejaoui. "An Intelligent Phishing Detection System Using Deep Learning and Explainable AI". In: *2025 International Symposium on Networks, Computers and Communications (ISNCC)*. ISSN: 2768-0940. Oct. 2025, pp. 1–6. DOI: 10.1109/ISNCC66965.2025.11250409. URL: https://ieeexplore.ieee.org/document/11250409.

[2] Santosh Kumar Birthriya, Priyanka Ahlawat, and Ankit Kumar Jain. "Detection and prevention of spear phishing attacks: A comprehensive survey". In: *Computers & Security* 151 (Apr. 2025), p. 104317. ISSN: 0167-4048. DOI: 10.1016/j.cose.2025.104317. URL: https://www.sciencedirect.com/science/article/pii/S0167404825000069.

[3] Nguyet Quang Do et al. "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions". In: *IEEE Access* 10 (2022), pp. 36429–36463. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3151903. URL: https://ieeexplore.ieee.org/document/9716113.

[4] Zhengyang Fan et al. "Investigation of Phishing Susceptibility with Explainable Artificial Intelligence". In: *Future Internet* 16.1 (Jan. 2024), p. 31. ISSN: 1999-5903. DOI: 10.3390/fi16010031. URL: https://www.mdpi.com/1999-5903/16/1/31.

[5] Paulo R. Galego Hernandes et al. "Phishing Detection Using URL-based XAI Techniques". In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. Dec. 2021, pp. 01–06. DOI: 10.1109/SSCI50451.2021.9659981. URL: https://ieeexplore.ieee.org/document/9659981.

[6] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[7] René Meléndez, Michal Ptaszynski, and Fumito Masui. "Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection". In: *Electronics* 13.24 (Jan. 2024), p. 4877. ISSN: 2079-9292. DOI: 10.3390/electronics13244877. URL: https://www.mdpi.com/2079-9292/13/24/4877.

[8] Maraz Mia, Darius Derakhshan, and Mir Mehedi Ahsan Pritom. "Can Features for Phishing URL Detection Be Trusted Across Diverse Datasets? A Case Study with Explainable AI". In: *Proceedings of the 11th International Conference on Networking, Systems, and Security*. NSysS '24. New York, NY, USA: Association for Computing Machinery, Jan. 2025, pp. 137–145. ISBN: 979-8-4007-1158-9. DOI: 10.1145/3704522.3704532. URL: https://doi.org/10.1145/3704522.3704532.

[9] Denish Omondi Otieno, Akbar Siami Namin, and Keith S. Jones. "The Application of the BERT Transformer Model for Phishing Email Classification". In: *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. ISSN: 0730-3157. June 2023, pp. 1303–1310. DOI: 10.1109/COMPSAC57700.2023.00198. URL: https://ieeexplore.ieee.org/document/10197078.

[10] Bhupathi Vishva Pavani, Desham Mahitha, and B Uma Maheswari. "Enhancing Online Safety: Phishing URL Detection Using Machine Learning and Explainable AI". In: *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. ISSN: 2473-7674. June 2024, pp. 1–6. DOI: 10.1109/ICCCNT61001.2024.10723976. URL: https://ieeexplore.ieee.org/document/10723976.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. URL: https://dl.acm.org/doi/10.1145/2939672.2939778.

[12] Sakib Shahriar Shafin. "An explainable feature selection framework for web phishing detection with machine learning". In: *Data Science and Management* 8.2 (June 2025), pp. 127–136. ISSN: 2666-7649. DOI: 10.1016/j.dsm.2024.08.004. URL: https://www.sciencedirect.com/science/article/pii/S2666764924000419.

[13] Abdulla Al-Subaiey et al. "Novel interpretable and robust web-based AI platform for phishing email detection". In: *Computers and Electrical Engineering* 120 (Dec. 2024), p. 109625. ISSN: 0045-7906. DOI: 10.1016/j.compeleceng.2024.109625. URL: https://www.sciencedirect.com/science/article/pii/S0045790624005524.

[14] Y L D H Yakandawala, M K P Madushanka, and W M K S Ilmini. "Explainable AI for Transparent Phishing Email Detection". In: *2024 International Conference on Advances in Technology and Computing (ICATC)*. Dec. 2024, pp. 1–6. DOI: 10.1109/ICATC64549.2024.11025302. URL: https://ieeexplore.ieee.org/document/11025302.