RESEARCH & INNOVATION DIVISION
IJIST

An AI-Powered Browser Extension Using Roberta and XAI for Phishing Email Detection and Security Awareness

Ahmed Murtaza, Ayesha Shahid, Safeer-ul-Hassan, Syed Masood Umar Rizvi, Saima Siraj
Department of Information Technology Quaid-e-Awam University of Engineering, Science and Technology Nawabshah, Pakistan
*Correspondence: ahmeddmurtazaa@gmail.com, ayeshashahidkhoja@gmail.com, safeerhassan123459@gmail.com, syedmasoodumar7@gmail.com, saimasiraj@quest.edu.pk

Citation | Murtaza. A, Shahid. A, Hassan. S. U, Rizvi. S. M. U, Siraj. S, "An AI-Powered Browser Extension Using RoBERTa and XAI for Phishing Email Detection and Security Awareness", IJIST, Vol. 07 Special Issue. pp 97-106, May 2025
Received| April 13, 2025 **Revised|** May 11, 2025 **Accepted|** May 12, 2025 **Published|** May 15, 2025.

Phishing attacks are a common and serious cybersecurity threat today. They exploit human weaknesses by stealing sensitive information by sending fake emails and harmful links. Traditional email filtering systems like rule-based methods and black-box models, struggle to detect phishing. Rule-based filters fail when attackers use new tricks, and black-box models lack transparency, which limits user awareness.

This work introduces a smart browser extension that uses deep learning and Explainable AI (XAI) for phishing detection. We use a transformer-based model, Roberta, trained on a large email dataset, achieving 98.12% accuracy in classifying email content. For checking URLs, we use VirusTotal, which gathers threat intelligence from multiple sources. We also apply XAI tools to highlight key parts of the text that contributed to the classification of the email content, and a large language model (LLM) to provide simple explanations about phishing.

Our hybrid approach combines explainable deep learning with multi-source URL verification. This helps users understand phishing threats better and improves their ability to spot attacks on their own.

Keywords: Phishing Detection, Explainable AI, Cybersecurity Awareness, Deep Learning, Roberta

Introduction:

Phishing remains one of the most pervasive and dangerous forms of cybercrime, responsible for millions of successful attacks targeting individuals, businesses, and government organizations every year. In a typical phishing attack, malicious actors impersonate trusted entities to deceive victims into revealing sensitive information such as login credentials, personal data, or financial details. The consequences of these attacks include financial losses, data breaches, and compromised organizational security.

Traditional phishing detection techniques, including blocklists, rule-based filters, and heuristic-based approaches, have been moderately effective against straightforward phishing campaigns. These systems typically rely on identifying known malicious URLs, domains, or commonly used suspicious keywords. However, the growing sophistication of phishing tactics such as involving AI-generated emails, obfuscated URLs, personalized content, and social engineering techniques has rendered these static methods increasingly inadequate [1].

In response to these challenges, recent advances in machine learning (ML) and deep learning (DL) have significantly improved phishing detection performance by analyzing email content, headers, and embedded URLs more intelligently [2]. Among these, transformer-based models like BERT and Roberta have shown exceptional promise due to their ability to capture complex language patterns and contextual relationships. Despite their effectiveness, many of these AI-based systems operate as "black boxes," offering little to no insight into their decision-making processes [3]. This lack of transparency contributes to low user trust in automated cybersecurity tools and limits opportunities for user education and awareness.

Given the escalating nature of phishing threats and the limitations of existing solutions, there is an urgent need for phishing detection systems that are not only accurate but also interpretable and user-focused. Effective cybersecurity tools should empower users by providing understandable feedback about detected threats, fostering awareness, and promoting safer online behaviors.

In this study, we propose an advanced phishing detection system that addresses both technical accuracy and user interpretability. The system leverages the RoBERTa transformer model for email content classification and integrates Explainable AI (XAI) techniques and a Large Language Model (LLM) to provide clear, user-friendly explanations of detection outcomes. Additionally, it incorporates the VirusTotal API for real-time URL safety verification, checking against over eighty sources and security vendors. The system utilizes a browser extension as the user interface to deliver real-time phishing detection and educational feedback directly within the user's browsing environment. Achieving a detection accuracy of **98.12%**, the proposed system enhances phishing detection capability while also serving as a practical tool for raising cybersecurity awareness.

Objectives:

1. To train a RoBERTa transformer model on a labeled email dataset for classifying phishing and legitimate emails based on content.
2. To determine the safety of embedded URLs by extracting them from emails and verifying their status using the VirusTotal API.
3. To integrate Explainable AI (XAI) techniques and a large language model (LLM) to highlight phishing indicators, explain the reasoning behind detection outcomes in accessible, non-technical terms, and provide actionable cybersecurity guidance to users.
4. To develop a browser extension that offers users real-time phishing detection, detailed explanations, and educational feedback to improve email safety awareness.

Novelty Statement:

This study introduces a novel, hybrid phishing detection framework that combines the high accuracy of transformer-based models with real-time, explainable feedback mechanisms for end users. Unlike traditional spam filters and conventional detection systems that silently

block threats, the proposed solution actively educates users by translating complex model outputs into clear, actionable guidance.

The research uniquely integrates a RoBERTa-based detection model to classify email content, utilizing Explainable AI (XAI) and large language models (LLMs) to deliver transparent, trustworthy explanations for detection outcomes. Additionally, by incorporating VirusTotal API-based URL verification and deploying the system through a freely accessible browser extension, this study bridges the gap between cutting-edge AI research and practical, user-friendly cybersecurity tools.

By shifting the focus from merely detecting phishing attempts to empowering users with knowledge and awareness, this work represents a stepping stone toward the development of future AI-powered cybersecurity systems that prioritize both technical performance and human-centered design.

Related Work:

Evolution of Phishing Detection Methods:

Phishing remains one of the top cybersecurity threats. Attackers use fake emails, websites, and malicious links to steal sensitive information. Traditional anti-phishing methods like blocklists and rule-based systems struggle to keep up with constantly changing phishing tactics [1]. As cybercriminals become more advanced, machine learning (ML) and deep learning (DL) have shown greater effectiveness in detecting phishing attempts with higher accuracy [2].

Traditional Phishing Detection Approaches:

In the past, static techniques like keyword filtering and blocklists were commonly used. However, attackers can easily bypass them by changing URLs or using trusted platforms [4]. We compare ML models such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Neural Networks (NNet) using ROC curves to predict whether a webpage is phishing. An experiment using 2,889 phishing and legitimate emails with forty-three features revealed that ML methods consistently outperform static techniques [5]. ML techniques have also been explored for phishing website detection. A feature-based approach was used to train classifiers on phishing datasets, showing the strength of Decision Trees and Random Forest. However, these models are still vulnerable to evasive and adversarial phishing tactics, highlighting the need for more robust deep learning-based solutions [6].

Deep Learning for Phishing Detection:

Deep learning enables more advanced text analysis, improving phishing detection accuracy. Taxonomies have been used to classify and assess different phishing detection methods based on their strengths and limitations. However, deep learning models still face challenges such as the need for manual parameter tuning, long training times, and inefficiencies [3]. Recent studies compare models like Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and hybrid CNN-LSTM models. In one study, the CNN model achieved 97.6% accuracy, the LSTM model 96.8%, and the CNN-LSTM hybrid reached 99.2% [7]. More recent methods combine multiple deep-learning models for even better results. One such framework uses ResNeXt and Gated Recurrent Unit (GRU) models, combining GRU's sequence modeling with ResNeXt's feature extraction. This hybrid approach strengthens phishing detection and performs well against complex threats using commercially available resources [8].

The Role of Explainable AI (XAI) in Cybersecurity:

Explainable AI (XAI) improves transparency and helps users understand how phishing is detected. While XAI increases user trust and system security, it also adds computational overhead and can be vulnerable to adversarial attacks. Future research must focus on balancing detection efficiency and security [9].

LLMs and Transformer Models for Phishing Detection:

Large Language Models (LLMs) like Roberta enhance phishing detection by capturing advanced language patterns. They also help improve user awareness by identifying complex phishing tactics. Integrating LLMs with XAI increases interpretability and makes AI-based detection systems more transparent and trustworthy [10].

Transformer Models in Cybersecurity:

Traditional ML models often miss deeper context, but Transformer models like BERT and RoBERTa outperform them in phishing detection. Roberta, in particular, is highly effective due to its optimized training methods and dynamic masking, making it ideal for classifying phishing emails [11].

Explainable Transformer Models for Phishing Detection:

One study fine-tuned a DistilBERT model and combined it with LLMs and XAI for phishing email detection. The authors addressed class imbalance in the dataset and used interpretation methods like LIME and Transformer visualization tools to improve model transparency. While similar to our approach, this work used DistilBERT instead of RoBERTa and focused more on implementation [12].

Vulnerabilities in Explainable AI (XAI) to Adversarial Attacks:

While XAI adds transparency and trust, recent research shows it can be manipulated. Adversarial attacks may trick XAI models into producing misleading explanations, allowing attackers to bypass detection. These findings raise concerns about the reliability of XAI-based solutions and stress the need for strong defense mechanisms that maintain interpretability without weakening security [13].

The Impact of User Awareness on Phishing Detection:

User education plays a key role in improving phishing detection. One study highlighted how training programs help users recognize phishing attempts, significantly reducing attack success rates. Awareness efforts empower users to detect and avoid threats more effectively [14].

Summary:

Despite advancements, phishing detection still faces challenges. Traditional methods are outdated against modern tactics, and many AI-based systems lack explainability, which weakens user trust. Most detection tools focus only on identifying phishing and overlook the importance of user awareness. Additionally, reliance on static features limits their ability to detect AI-generated phishing. Overcoming these challenges is vital to improving cybersecurity defenses.

System Architecture and Design:

The proposed phishing detection system uses a multi-component architecture designed to provide strong email classification and enhance user education, as illustrated in Figure. 1.

The Key Components of the Proposed System Include:

1. **Email Data Processing Module:** Cleans and preprocesses raw email content, removing irrelevant information while preserving key phishing indicators.
2. **Roberta-based Classification Engine:** Tokenizes email text and accurately classifies messages as either phishing or legitimate.
3. **URL Verification Module:** Extracts and evaluates embedded URLs using the VirusTotal API to assess potential threats.
4. **Explainable AI (XAI) Framework:** Provides clear, understandable justifications for the classification results generated by the Roberta model.
5. **LLM-Based User Feedback System:** Delivers interpretable explanations to help users understand phishing risks and improve cybersecurity awareness.

6. **Browser Extension Interface:** Enables real-time scanning and phishing detection directly within the user's browser as new emails arrive.

The system employs a hybrid approach, combining advanced text classification with external threat intelligence to enhance the accuracy and reliability of phishing detection.

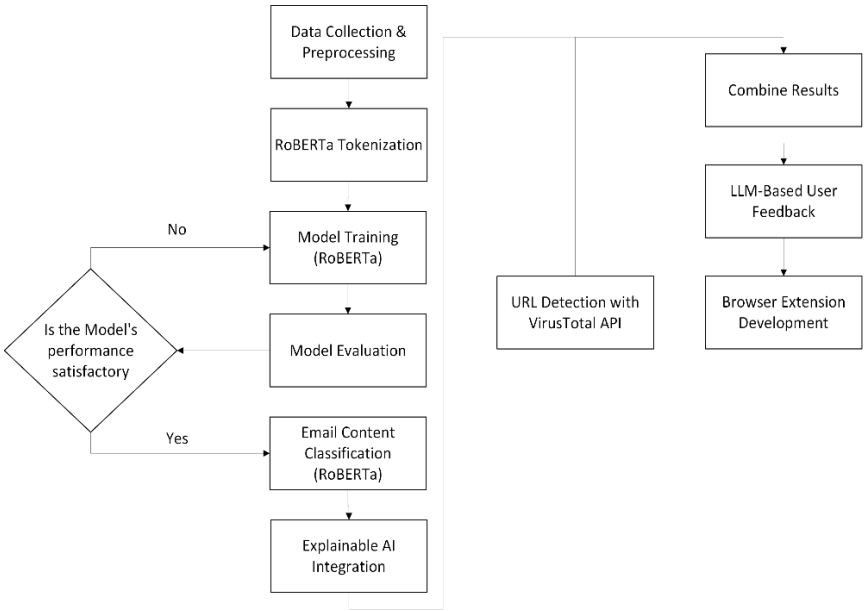

Figure 1. Methodology

Methodology:

1) **Data Collection and Preprocessing:** The dataset comprises over 18,000 emails obtained from the Enron Corporation via Kaggle [15]. The email text undergoes preprocessing steps such as normalization and duplicate removal. However, stop words and special characters are retained to preserve patterns that may indicate phishing. The cleaned data is then prepared for the next stage—tokenization.

2) **Roberta Tokenization:** The RoBERTa tokenizer transforms the preprocessed email text into numerical vectors, maintaining the contextual meaning of the words. These tokenized inputs are then fed into the classification model to enable accurate phishing detection.

Table 1. Training loss across epochs for each fold

Fold	Epoch one	Epoch two	Epoch three
1	0.1251	0.0478	0.0416
2	0.0408	0.0310	0.0297
3	0.0349	0.0263	0.0266
4	0.0277	0.0258	0.0223
5	0.0299	0.0243	0.0235

3) **Model Training:**

The RoBERTa model is fine-tuned on the preprocessed dataset using 5-fold cross-validation to ensure robust evaluation. The model is trained for 3 epochs in each fold, with one-fold used as the validation set and the remaining four for training. The AdamW optimizer with learning rate scheduling improves convergence during training. Table 1 shows the loss values across epochs for each fold, indicating a consistent decline, evidence of stable and effective learning. Additionally, hyperparameter tuning is performed to enhance classification accuracy and reduce the risk of overfitting.

4) **Model Evaluation:**

The model's performance is evaluated using accuracy, precision, recall, and F1-score to provide a comprehensive assessment of its classification effectiveness (see Figure. 2).

Confusion matrices and classification reports are also examined to identify patterns of misclassification, offering valuable insights for further model optimization and improvement (see Figure. 2 and Figure.3).

5) **Email Content Classification:** The trained RoBERTa model analyzes tokenized emails to detect linguistic patterns, deceptive phrasing, and contextual anomalies, classifying each email as either phishing or legitimate.

6) **Explainable AI (XAI) Integration:** XAI techniques are used to highlight key terms and provide reasoning behind phishing classifications, enhancing model transparency and building user trust.

7) **Combining Results:** A weighted scoring mechanism integrates the outcomes of email content classification and URL risk analysis, ensuring a more comprehensive and accurate phishing detection process.

8) **LLM-Based User Feedback:** A large language model (LLM) generates user-friendly explanations, pointing out phishing indicators and offering insights to improve cybersecurity awareness.

9) **Browser Extension Development:** The final system includes a browser extension that combines phishing classification, URL verification, and easy-to-understand explanations to deliver real-time protection for users.

Results:

The performance of the proposed phishing detection model was thoroughly evaluated using multiple performance metrics and cross-validation. On the validation set, the model achieved a high accuracy of 98.12%, with a precision of 98%, recall of 98%, and an F1-score of 98% (see Figure. 2). These metrics indicate strong and consistent detection capabilities for both phishing and legitimate instances. The confusion matrix further revealed a notably low false positive rate (see Figure. 3), reducing the risk of legitimate emails being flagged incorrectly, which is crucial for maintaining a positive user experience in practical applications.

Throughout the training process, loss values consistently decreased over successive epochs for all five folds, confirming the model's stable convergence and effective learning behavior. The final average training loss achieved was 0.0287 and loss remained low and stable across folds, indicating strong generalization to unseen data (Table I). No signs of overfitting or performance degradation were observed during training, further supporting the model's robustness.

To evaluate the effectiveness of the proposed model relative to traditional and deep learning approaches, its performance was compared against several baseline classifiers, including a Support Vector Machine (SVM), and a standard BERT (see Figure 4). The RoBERTa-model outperformed all baseline methods across all evaluated metrics. It achieved higher accuracy and precision while maintaining a lower false positive rate, reinforcing the advantage of transformer-based models for handling complex phishing content with nuanced linguistic patterns (see Figure. 2 and Figure. 3).

The incorporation of the Explainable AI (XAI) method provided valuable interpretability to the model's decision-making process. Using feature attribution and language pattern analysis, the system identified key phishing indicators, such as urgency-driven phrasing significant contributors to phishing classifications. Through XAI the model's classification decisions were explained and then transferred to a Large Language Model (LLM) to provide user-friendly explanations and actionable tips. These explanations were delivered to users via the browser extension in real time, increasing transparency and offering immediate insights into detection decisions. Preliminary qualitative feedback suggested that users appreciated having visibility into why specific emails or web content were flagged, which could enhance long-term cybersecurity awareness.

In addition to the Roberta model's content classification, the integration of the VirusTotal API enhanced detection reliability. Through VirusTotal our system cross-references URLs against an up-to-date threat database and checks against over eighty sources which enhances credibility. This integration complemented the model's prediction capabilities by adding a robust, external verification step for potentially malicious URLs attached to the email.

While incorporating RoBERTa and XAI components introduced some additional computational overhead, the system's responsiveness remained within acceptable limits for real-time deployment. This minor increase in processing time was considered a worthwhile trade-off for the added interpretability and user education benefits.

This study primarily aimed to achieve high detection accuracy and lay the groundwork for a platform that promotes user education in email phishing awareness. While a full-scale user engagement analysis is planned for future work, integrating explainable feedback into detection systems can positively influence cybersecurity awareness and behavior.

Final Test Set Evaluation:
Accuracy: 98.12%

Classification Report:

	precision	recall	f1-score	support
0	0.98	0.98	0.98	2214
1	0.98	0.99	0.98	2315
accuracy			0.98	4529
macro avg	0.98	0.98	0.98	4529
weighted avg	0.98	0.98	0.98	4529

Figure 2. Classification Report of Fine-Tuned Roberta Model

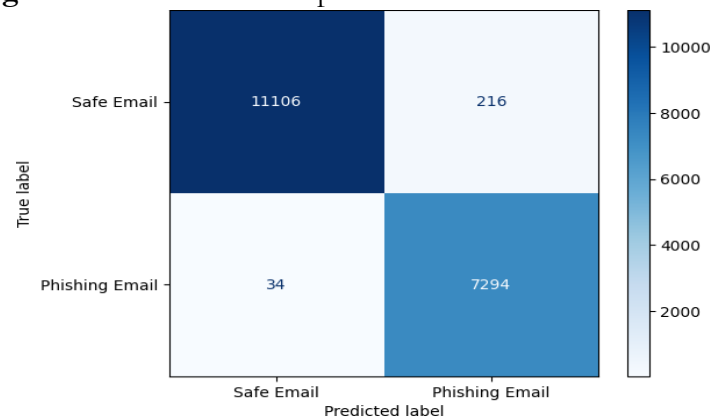

Figure 3. Confusion Matrix for Fine-Tuned Roberta Model

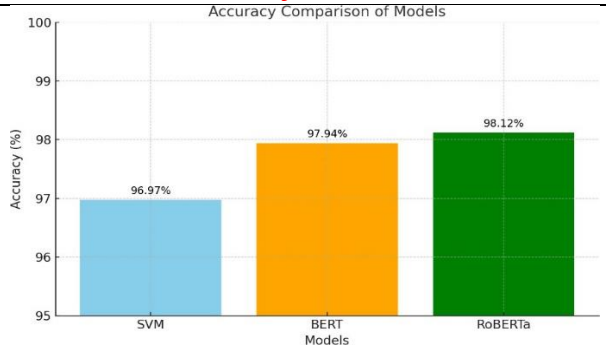

Figure 4. Accuracy Comparison of Models

Discussion:

The high accuracy and precision achieved by our model on the validation set (see Figure. 2) reaffirm the growing consensus in recent literature that machine learning (ML) and deep learning (DL) approaches significantly outperform traditional anti-phishing methods. Our results align with the findings of [2] and [5], where ML models consistently surpassed static techniques like blocklists and keyword filters. This continued validation across studies highlights the increasing relevance of data-driven approaches in countering dynamic phishing tactics.

Compared to existing deep learning-based phishing detection methods, our model demonstrated competitive performance. Prior studies such as [7] reported an accuracy of 97.6% with CNN models and up to 99.2% with CNN-LSTM hybrids and our model's high accuracy of 98.12% positions it among these state-of-the-art solutions while incorporating unique elements like XAI integration and real-time browser protection.

One distinguishing aspect of our approach is the integration of Explainable AI (XAI) and Large Language Models (LLMs), particularly Roberta, to improve both system transparency and user awareness. Previous works like [10], [11], and [12] explored LLMs and transformer models for phishing detection, demonstrating superior performance over traditional ML models. Our results complement these findings, reinforcing the effectiveness of transformer-based models in capturing advanced phishing language patterns.

The addition of XAI and LLM addresses a growing need for interpretability in AI-based cybersecurity tools. Studies like [9] and [13] highlighted both the benefits and challenges of integrating XAI, particularly its potential computational overhead and vulnerability to adversarial manipulation. Our implementation showed that while computational demands slightly increased, the trade-off was justified by enhanced detection transparency and improved user trust, aligning with the conclusions of these earlier works.

Another crucial consideration is user education and awareness. Our system incorporates real-time feedback through a browser extension, supporting previous research findings that awareness campaigns significantly reduce phishing attack success rates. While our current study primarily prioritized accuracy as a performance benchmark, future research will focus on assessing user engagement and behavioral outcomes.

Overall, the combination of high accuracy, explainability, and real-time user interaction forms a comprehensive phishing detection framework that not only strengthens system security but also educates users which is an aspect often overlooked in earlier studies.

Conclusion and Future Work:

This research presents an advanced phishing detection system that integrates deep learning, XAI, and LLMs to enhance cybersecurity awareness. The fine-tuned Roberta model achieved an accuracy of 98.12% in email classification, while the VirusTotal API improves detection reliability by analyzing URLs embedded within emails. The system's browser

extension ensures real-time protection, and the XAI and LLM-generated feedback further elevates user awareness.

Future work will focus on incorporating additional features for improved classification, enhancing cybersecurity awareness through new techniques, expanding the dataset for greater diversity, and integrating more threat intelligence sources to strengthen phishing detection capabilities.

Acknowledgment:

We thank Dr. Saima Siraj Soomro for her invaluable support, continuous assistance, and guidance throughout this research, from project selection to proposal development. Their insightful guidance and encouragement were instrumental in our decision-making throughout the research. We also acknowledge Quaid-e-Awam University of Engineering, Science and Technology Nawabshah for supporting this research and providing resources for development.

References:

[1] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, pp. 345–357, Mar. 2019, doi: 10.1016/J.ESWA.2018.09.029.

[2] P. H. Kyaw, J. Gutierrez, and A. Ghobakhlou, "A Systematic Review of Deep Learning Techniques for Phishing Email Detection," *Electronics*, vol. 13, no. 19, p. 3823, Sep. 2024, doi: 10.3390/ELECTRONICS13193823.

[3] N. Q. Do, A. Selamat, O. Krejcar, E. Herrera-Viedma, and H. Fujita, "Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions," *IEEE Access*, vol. 10, pp. 36429–36463, 2022, doi: 10.1109/ACCESS.2022.3151903.

[4] P. Prakash, M. Kumar, R. Rao Kompella, and M. Gupta, "PhishNet: Predictive blacklisting to detect phishing attacks," *Proc. - IEEE INFOCOM*, 2010, doi: 10.1109/INFCOM.2010.5462216.

[5] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," *ACM Int. Conf. Proceeding Ser.*, vol. 269, pp. 60–69, 2007, doi: 10.1145/1299015.1299021.

[6] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing Detection Using Machine Learning Technique," *Proc. - 2020 1st Int. Conf. Smart Syst. Emerg. Technol. SMART-TECH 2020*, pp. 43–46, Nov. 2020, doi: 10.1109/SMART-TECH49988.2020.00026.

[7] Z. Alshingiti, R. Alaqel, J. Al-Muhtadi, Q. E. U. Haq, K. Saleem, and M. H. Faheem, "A Deep Learning-Based Phishing Detection System Using CNN, LSTM, and LSTM-CNN," *Electronics*, vol. 12, no. 1, pp. 232–232, Jan. 2023, doi: 10.3390/ELECTRONICS12010232.

[8] F. S. Alsubaei, A. A. Almazroi, and N. Ayub, "Enhancing Phishing Detection: A Novel Hybrid Deep Learning Framework for Cybercrime Forensics," *IEEE Access*, vol. 12, pp. 8373–8389, 2024, doi: 10.1109/ACCESS.2024.3351946.

[9] V. L. and C. S. N. Capuano, G. Fenza, "Explainable Artificial Intelligence in CyberSecurity: A Survey," *IEEE Access*, vol. 10, pp. 93575–93600, 2022, doi: 10.1109/ACCESS.2022.3204171.

[10] T. Koide, N. Fukushi, H. Nakano, and D. Chiba, "ChatSpamDetector: Leveraging Large Language Models for Effective Phishing Email Detection," Feb. 2024, Accessed: Apr. 20, 2025. [Online]. Available: https://arxiv.org/abs/2402.18093v2

[11] R. Meléndez, M. Ptaszynski, and F. Masui, "Comparative Investigation of Traditional Machine-Learning Models and Transformer Models for Phishing Email Detection," *Electron. 2024, Vol. 13, Page 4877*, vol. 13, no. 24, p. 4877, Dec. 2024, doi: 10.3390/ELECTRONICS13244877.

[12] S. Jamal, H. Wimmer, and I. H. Sarker, "An Improved Transformer-based Model for

Detecting Phishing, Spam, and Ham: A Large Language Model Approach," Nov. 2023, Accessed: Apr. 20, 2025. [Online]. Available: https://arxiv.org/abs/2311.04913v2

[13] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," *Inf. Fusion*, vol. 107, p. 102303, Jul. 2024, doi: 10.1016/J.INFFUS.2024.102303.

[14] K. Patil and S. R. Arra, "Detection of Phishing and User Awareness Training in Information Security: A Systematic Literature Review," *Proc. 2nd Int. Conf. Innov. Pract. Technol. Manag. ICIPTM 2022*, pp. 780–786, 2022, doi: 10.1109/ICIPTM54933.2022.9753912.

[15] S. Chakraborty, "Phishing email detection," *Kaggle*, 2023.