ID 368

Review on Feasibility of Building An Explainable Artificial Intelligence Model for Anti-Phishing Detection

YLDH Yakandawala[1#], MKP Madushanka[2]

Department of Computer Science, Faculty of Computing General Sir John Kotelawala Defence University Ratmalana, Sri Lanka [1#,2]

38-cs-6229@kdu.ac.lk[1#]

***Abstract*— The viability of developing an Explainable Artificial Intelligence (XAI) model for anti-phishing detection is examined in this review. The significance of Explainable Artificial Intelligence (XAI), it's principles, methods/types, challenges, ethical issues, vulnerability aspects are discussed. The areas of machine learning for phishing detection, XAI models for phishing detection, developing appropriate explanation messages for warnings, feasibility issues, and a comparison with conventional approaches are all covered. The importance of XAI in enhancing the clarity and interpretability of AI models are further emphasized in the paper. It shows different XAI techniques, difficulties in striking a balance between explainability and performance, and XAI ethics. The evaluation looks at phishing scams, machine learning detection methods, and the advantages of XAI models. It suggests a thorough strategy for conveying explanatory messages and examines the viability of creating XAI models. In highlighting the promise of XAI to improve transparency and interpretability, the research also acknowledges the difficulties that must be overcome in order to create scalable and reliable XAI models for anti-phishing detection.**

Keywords—XAI, Phishing, Anti-Phishing, Detection, Cyber Security, Threats

I. INTRODUCTION

However, the traditional machine learning models used in this field have a lack of transparency and interpretability, which raises concerns about their reliability and their ability to explain their decisions. To address this issue, Explainable Artificial Intelligence (XAI) techniques have emerged as a promising solution. They aim to make machine learning models more transparent and interpretable.

The main objective of this review paper is to assess the possibility of developing an XAI model for anti-phishing detection. By using the interpretability features of XAI, such a model can not only accurately detect phishing attacks but also provide clear explanations for its decisions. This has the potential to greatly improve user trust, facilitate a better understanding of the patterns used for detection, and enable the identification and mitigation of false positives and false negatives.

To achieve this objective, we will conduct a comprehensive review of the existing literature and state-of-the-art techniques in anti-phishing detection. We will analyze the limitations of current approaches, with a particular emphasis on the need for improved interpretability. Additionally, we will explore the available XAI techniques and assess their suitability for building a transparent and interpretable model for anti-phishing detection.

In this review, we will also address several important considerations that affect the feasibility of constructing an XAI model for anti-phishing detection. These considerations include the availability of labeled data specifically annotated for phishing attacks, the development of relevant and comprehensive features, the selection of appropriate machine learning or deep learning algorithms, and the application of interpretable techniques to explain the model's predictions.

By critically examining existing research, identifying challenges, and exploring potential opportunities, this review paper aims to contribute to our understanding of the feasibility and potential benefits of building an XAI model for anti-phishing detection. Ultimately, our goal is to provide insights into the advancements, limitations, and future directions in this important research area, paving the way for more transparent and effective anti-phishing detection systems.

II. LITERATURE REVIEW

To evaluate the viability of constructing an Explainable Artificial Intelligence (XAI) model for anti-phishing detection, we adopted a systematic literature review methodology. The following steps were undertaken to gather and analyze relevant research papers:

A. Research Objective Refinement:

The research objectives were refined to ensure clarity and focus. The specific research questions to be addressed were identified, including the availability of labeled data, development of features, model selection, and interpretability techniques.

B. Search Strategy:

A comprehensive search strategy was formulated to retrieve relevant literature. Major academic databases, such as IEEE Xplore, Research Gate, and Google Scholar, were utilized. Search terms included variations of "anti-phishing detection," "explainable artificial intelligence," "XAI," "interpretability," and related keywords.

C. Screening and Selection:

Initially, titles and abstracts of the retrieved papers were screened to identify potentially relevant papers. Full-text articles were then thoroughly reviewed based on the inclusion and exclusion criteria. Any disagreements during the

screening process were resolved through discussion and consensus.

D. Data Extraction:

Relevant data from selected papers were extracted and the papers were added to Zotero for reference management.

E. Result Reporting:

The findings were organized and reported in a clear and concise manner. The results were presented through summaries, and visual representations, where appropriate.

F. Limitations and Bias:

The limitations of the methodology were acknowledged, including potential publication bias and the possibility of overlooking relevant papers despite efforts to be comprehensive. Steps were taken to minimize bias by following a systematic approach and including diverse sources in the search strategy.

By employing this systematic methodology, the review paper aimed to gather a comprehensive set of relevant research papers and provide an objective assessment of the feasibility of building an XAI model for anti-phishing detection. The methodology ensured transparency, rigor, and replicability in the literature review process.

III. REVIEW FINDINGS

The comprehensive review of the literature on the feasibility of building an Explainable Artificial Intelligence (XAI) model for anti-phishing detection yielded significant findings and insights. The key findings are as follows:

A. XAI and its Importance(Reddy and Kumar, 2023)

Explainable AI (XAI) addresses concerns about the opacity of AI models by making them transparent and interpretable. It aims to improve reliability, trustworthiness, and accountability. The rise of complex machine learning models has made transparency more important, leading to increased interest in XAI from academia and industry. In the context of Industry 4.0, lack of transparency in technologies like AI, robotics, and IoT poses challenges for functionality and security of IoT devices and networks.

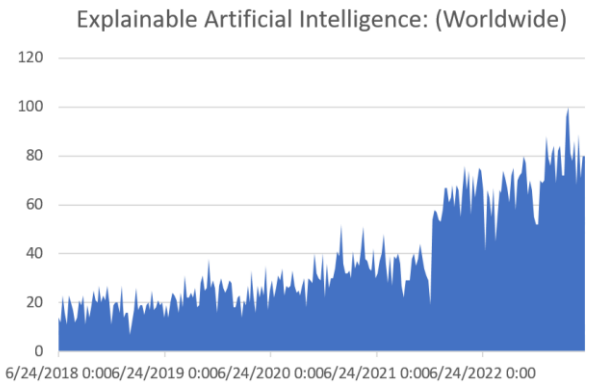

Figure 1. Google trends statistics for Explainable AI from 2018 to 2022 (Source: Google Trends)

The popularity of blackbox models has increased rapidly, but the emphasis has been on improving accuracy rather than prioritizing explainability. While explainability may currently be seen as optional, it is expected to become a necessary requirement as AI decision-making systems expand. Transparency throughout the implementation of blackbox models will inevitably require explainability in the future to ensure accountability and understanding.

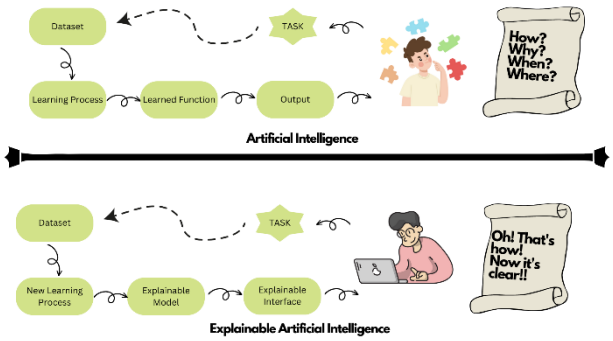

Figure 2. AI Processes vs XAI Processes (Source: Original content)

B. XAI Principles(Phillips et al., *2021)*

This section presents four fundamental principles for explainable AI systems that focus on the interaction between the system and its human recipients. The principles cover the delivery of explanations, ensuring their meaningfulness, accuracy, and adherence to knowledge limits. They apply to various AI techniques and emphasize the importance of providing explanations for a system to be considered explainable.

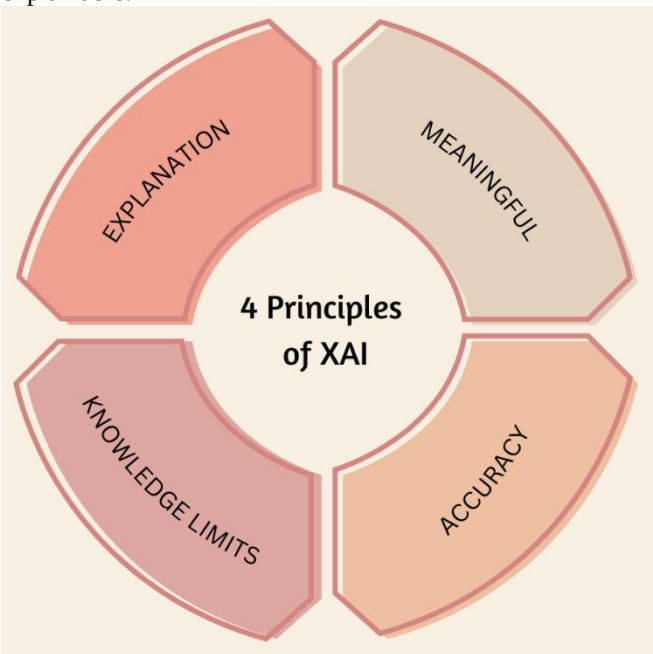

Figure 3. Principles of XAI

1) Explanation

The Explanation principle states that an explainable AI system should provide evidence, support, or reasoning related to its outcomes or processes. The principle focuses on the provision of explanations rather than their quality, correctness, or intelligibility. The quality and meaningfulness of explanations are addressed by other principles. The implementation of explanations can vary based on the system and scenario, allowing for flexibility in their execution and integration. A broad definition of explanation is adopted to accommodate different applications and use cases.

2) Meaningful

The Meaningful principle emphasizes that an explanation provided by an AI system should be understandable to its intended recipient. Factors such as the audience's prior knowledge, experiences, and psychological differences influence what they consider to be a "good" explanation. Developers need to consider the needs and desires of different groups of people who interact with the system, including developers and end-users. The meaning of an explanation can vary based on its purpose and the specific context in which it is provided. Achieving meaningful explanations requires understanding the audience's needs, expertise, and the relevance of the explanation to the question or query at hand. Measuring the meaningfulness of explanations is an ongoing challenge, and developing adaptable measurement protocols is crucial. The context of an explanation plays a vital role in assessing the quality of AI explanations and guiding their execution in a goal-oriented and meaningful manner.

3) Explanation Accuracy

The Explanation Accuracy principle in explainable AI emphasizes the importance of ensuring that the explanations provided by a system accurately reflect the process by which the system generates its output. While the Explanation and Meaningful principles focus on intelligibility and understanding, the Explanation Accuracy principle adds the requirement of veracity to the explanations.

Accuracy of explanations is distinct from decision accuracy, which refers to the correctness of the system's judgments. Researchers have established metrics for decision accuracy, but metrics for explanation accuracy are still being developed. The level of detail in the explanation is also a factor to consider. Simple explanations may be sufficient for certain audiences or purposes, while more detailed explanations may be necessary for experts or specific contexts.

The trade-off between explanation accuracy and meaningfulness is acknowledged. Detailed explanations may accurately represent the system's process but might be less accessible to some audiences. Conversely, simple explanations may be highly understandable but may not fully capture the complexity of the system. Flexibility in defining and measuring explanation accuracy is necessary to account for these considerations.

4) Knowledge Limits

The Knowledge Limits principle in explainable AI focuses on the importance of systems recognizing and acknowledging their limitations. This principle ensures that systems are aware of cases where they are not designed or authorized to operate, or when their answers may not be reliable. By declaring these knowledge limits, systems can prevent providing misleading, dangerous, or unjust outputs, thus increasing trust in their results.

There are two main ways in which a system can encounter its knowledge limits. Firstly, when the system is presented with an operation or query that falls outside its domain or area of expertise, it should appropriately respond by indicating that it cannot provide an answer. This serves as both an answer and an explanation to the user. For example, if a system designed to classify bird species is given an image of an apple, it should inform the user that it cannot find any birds in the input image.

Secondly, a system may reach its knowledge limits when the confidence in its most likely answer is too low, based on an internal confidence threshold. In such cases, even if the system recognizes the input as pertaining to its domain (e.g., a blurry image of a bird), it can indicate that the image quality is too low to identify the species. This type of explanation would inform the user that a bird was detected but the system lacks sufficient information to provide a precise answer.

By adhering to the Knowledge Limits principle, AI systems can ensure responsible and accurate outcomes by avoiding inappropriate or unreliable judgments.

C. XAI Methods / Types(Reddy and Kumar, 2023)

1) Feature visualization

Feature visualization is a technique used to generate visualizations of the features learned by an AI model. It helps to understand what the model is looking for in the input data and can assist in identifying potential biases or errors. For instance, in image recognition tasks, feature visualization can generate images that activate specific neurons in the model to understand which parts of the input image are crucial for the model's decision. This technique is commonly used with deep neural networks, and Google's DeepDream method, introduced in 2015, is a popular raw example that modifies input images to maximize activation of certain neurons, revealing visual patterns associated with specific features and providing insights into the network's functioning.

2) Saliency Mapping

Saliency mapping is an interpretable artificial intelligence (AI) technique that produces heat maps, which indicate the significant regions in the input data that influence the decision-making process of the model. By using saliency maps, biases and errors can be identified and addressed, resulting in improved accuracy and fairness. These maps provide concise explanations for the model's decisions, enhancing their comprehensibility to human users. Various approaches can be employed to generate saliency maps, such as calculating gradients of the network's output concerning the input image or utilizing gradients between the output and feature maps in a specific layer of the network, as demonstrated in the case of Grad-CAM.

3) Model Interpretation:

Model interpretation plays a crucial role in Explainable Artificial Intelligence (XAI) by improving transparency and interpretability. Model-agnostic techniques, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations), are particularly valuable as they analyze the input-output relationship without relying on specific model intricacies. LIME accomplishes this by generating surrogate models that approximate individual predictions, providing local interpretability. On the other hand, SHAP decomposes predictions to identify the contributions of different features, facilitating a global understanding of the model's behavior.

These techniques are particularly useful for complex models, as they enable a transparent decision-making process. Moreover, they can be combined with feature visualization and saliency mapping to gain further insights into the model's behavior and to detect biases or errors. By leveraging these interpretability techniques, researchers and practitioners can enhance their understanding of how the model arrives at its predictions, promoting trust, accountability, and fairness in AI systems.

a) LIME:

LIME is an explainability method that offers specific, comprehensible explanations for specific predictions. It was first introduced in 2016. By altering the initial input data and watching how this affects the black-box model's output, it produces interpretable features. This could involve altering words in text inputs, while introducing or removing minor quantities of noise in numerical data. LIME builds a more straightforward, understandable model that describes the behavior of the original model in a particular area by determining the most crucial elements there.

b) SHAP:

c) Using the SHAP approach, every AI model's output is explained by calculating the impact of each input characteristic on the prediction. It measures the average marginal contributions of each feature using Shapley values and cooperative game theory. For particular cases, SHAP offers local feature importance and a feature ranking at the global level. For a variety of applications, it provides precise and understandable explanations.

d) EBM:

The Explainable Boosting Machine (EBM) is a transparent model that offers comparable performance to advanced models like Random Forest Boosted Tree while maintaining high interpretability and explainability. It is an improved version of the Generalized Additive Model and utilizes bagging and boosting techniques. One key advantage of EBMs is their intelligibility, as the contribution of each feature can be visualized and understood by analyzing the corresponding functions.

D. XAI CHALLENGES (Reddy and Kumar, 2023)

1) Explainability vs Performance:

Balancing explainability and performance in machine learning models is a trade-off. Future XAI research should focus on developing advanced techniques that strike a balance between these objectives to achieve both high performance and explainability.

Trust in Non-Explainable Systems: Trust in non-explainable AI systems is questioned, as public backlash against AI errors and biases indicates a preference for human expertise. Independent validation is proposed for trustworthy AI.

Challenges in Deep Learning: Deep learning systems, while highly accurate, require extensive training data, expensive hardware, and face robustness issues. Ensuring accuracy and robustness for future data is a complex task.

Narrow Focus on Deep Learning: The current debate on explainability vs. accuracy is predominantly centered around deep learning imaging applications, overlooking other powerful AI algorithms with better explainability.

Achieving Explainability without Sacrificing Accuracy: In feature-based AI systems, it is possible to eliminate low-discriminating features, improving explainability without compromising accuracy.

2) Human Factors:

Understanding how people interpret XAI explanations is a challenge. Research should explore the human factors involved and develop techniques tailored to different user groups to enhance trust and usability.

3) Universal Standard:

The lack of a universal standard or framework hinders the development and evaluation of XAI techniques. Future research should aim to establish a standardized framework that can be widely adopted across domains and applications.

4) Bias:

Addressing biases and ensuring fairness is crucial in XAI. Techniques should be designed to identify and mitigate biases in machine learning models, promoting fair and ethical decision-making.

5) Evaluation:

Evaluating the effectiveness of XAI techniques is challenging due to the absence of consensus on evaluation criteria. Future research should prioritize the development of standardized metrics and benchmarks for evaluating XAI techniques' accuracy and practicality.

E. Ethical Principles in XAI(Hanif, Zhang and Wood, 2021)

Ethical Principles in Explainable Artificial Intelligence (XAI) include:

- Accountability: AI systems should explain and justify their decisions.
- Responsibility: Individuals and AI systems should take responsibility for mistakes.
- Transparency: Users have the right to understand decisions in clear terms.
- Fidelity: Explanations should influence the decision-making process.
- Bias: Measures should be taken to prevent biased perspectives in AI systems.
- Causality: The model should provide insights and uncover the decision-making process.
- Fairness: Decisions should be unaffected by data limitations and assessed for fairness.

- Safety: Users should trust AI system choices, even without full transparency.

These principles ensure responsible, transparent, and fair AI systems.

*F. Phishing Attacks(*Phishing : ENISA Threat Landscape, *no date)*

Phishing is a fraudulent technique where attackers attempt to deceive users and steal their sensitive information through methods like fake emails or websites. Spear phishing, a targeted form of phishing, involves researching victims to make the scam more convincing. Emotional responses often lead people to fall for phishing attempts, making it essential to train users to recognize and avoid such attacks. Domain-based email authentication standards like DMARC help block fraudulent emails. While email remains a popular phishing method, attackers are increasingly using social media messaging platforms. The use of adversarial AI to create and send sophisticated phishing messages is expected to rise. Phishing and spear phishing also serve as major attack vectors for other threats like unintentional insider threats.

Phishing attacks have increasingly targeted webmail and software-as-a-service (SaaS) providers, surpassing attacks against payment services. In Q1 2019, they accounted for 36% of all phishing attacks, with Microsoft 365 being a primary target for phishers.

BEC attacks remained a significant threat, with 88% of organizations worldwide experiencing spear phishing attacks and 86% encountering BEC attacks. Microsoft 365 was targeted for credential harvesting, allowing attackers to access sensitive data and potentially launch spear-phishing attacks. BEC attacks saw a 120% increase in Q1 2019, resulting in substantial financial losses reaching up to US $26.2 billion.

The number of phishing sites using HTTPS has seen a significant increase, with 74% of phishing sites using HTTPS in Q4 2019 compared to 32% two years earlier. However, the presence of HTTPS and SSL may create a false sense of trust, and threat actors can also exploit legitimate hacked websites to host phishing content.

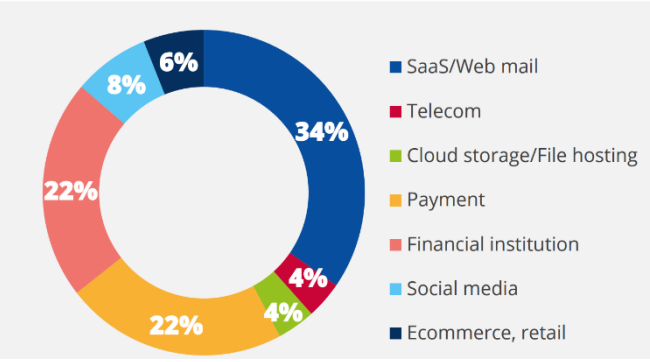

Figure 4. Phishing target attacks. Source: *(*Phishing : ENISA Threat Landscape, *no date)*

G. Phishing Vulnerability(Abroshan et al., *2021)*

Users' response to phishing attacks can be influenced by their risk-taking behavior.

Gender is a predictor of clicking on a phishing link, with women being slightly more susceptible.

Risk-taking behavior and gender influence users' likelihood of becoming risky users in the second step of the phishing process.

Risk-taking behavior and decision-making style do not significantly affect opening phishing emails, clicking on phishing links, or submitting sensitive data on phishing websites.

Technical solutions and anti-spoofing measures can help prevent phishing emails from reaching users.

Psychotherapy techniques and focused training can be used to reduce users' likelihood of clicking on phishing links, particularly for individuals with high risk-taking scores.

Gender can have indirect effects on falling prey to a phishing attack, potentially influenced by factors like technical knowledge and training.

Phishing emails can contain infected attachments or links to malicious websites that can compromise personal and organizational data.

These facts provide insights into the relationship between user behavior, gender, and susceptibility to phishing attacks, as well as the importance of preventive measures and user education in mitigating phishing risks.

H. Motivation for Phishing(Yu, Nargundkar and Tiruthani, 2008)

The primary motivation behind phishing attacks is financial gain, as scammers aim to profit from their fraudulent activities. However, phishers may also be motivated by other factors such as identity theft, industrial espionage, and the distribution of malware. While financial gain remains the main driving force, these additional factors contribute to the overall motivation behind phishing attacks.

1) Financial gain:

Phishing is primarily motivated by the opportunity to gain easy money, particularly targeting the financial sector. By spoofing the brand of financial institutions and accessing victims' account details, phishers aim to exploit financial resources.

2) Identity theft:

Phishing enables identity theft, where stolen identities are used for financial gain, fraud, or launching further phishing attacks. Stolen identities may also be sold to interested parties, creating a demand in the online community.

3) Identity trafficking:

Phishers engage in identity theft and sell stolen identities on online forums for a premium. This activity can lead to fraud, criminal activities, and financial gain. Profits from identity trafficking may be used for criminal purposes and can be challenging to track internationally.

4) Industrial espionage:

Sophisticated phishing attacks are conducted to spy on victims and gather valuable information, such as browsing patterns and product loyalties. This information is either utilized directly or sold to interested parties. Industrial

espionage through phishing can result in significant monetary losses.

5) Malware distribution:

Phishing attacks may involve distributing malware, such as Trojans, keyloggers, or fake browsers. Unsolicited phishing emails with malware attachments are sent to infect victims' machines and turn them into zombies. The distributed malware can be used for harvesting information or conducting future scams.

6) Harvesting passwords:

Phishers employ various methods, including keyloggers and malware, to harvest user passwords. Harvested passwords are used for financial gain, fraud, identity theft, or sold to interested parties.

7) Fame and notoriety:

Some phishing attacks are motivated by the desire for recognition and notoriety within the online community. Phishers engage in phishing to gain attention and glory, rather than solely for financial gain.

8) Exploiting security holes:

Individuals may attempt to exploit security flaws in systems to launch phishing attacks or sell compromised systems to other phishers. This motivation is driven by financial gain and the prospect of gaining fame and notoriety.

I. Machine Learning for Phishing Detection(Galego Hernandes et al., *2021)*

Machine learning plays a crucial role in phishing detection, offering high performance in both supervised and unsupervised techniques. Ridor and eDRI techniques have been identified as effective methods with high accuracy rates. Natural Language Processing (NLP) has also been utilized to analyze the text in attack links, searching for malicious content based on specific characteristics. Studies have shown that machine learning algorithms, such as Logistic Regression and Random Forest, can achieve high accuracy in detecting phishing websites. Researchers have developed various datasets and employed classification algorithms and NLP-based features to build real-time anti-phishing systems, resulting in accuracy rates ranging from 93% to 98%. These approaches demonstrate the effectiveness of machine learning in combating phishing attacks.

J. XAI Models for Phishing Detection(Galego Hernandes et al., *2021)*

1) LIME:

LIME is a model that provides interpretability and transparency for black-box models by analyzing their generated data. The model was processed using Random Forest and SVM algorithms, achieving high accuracy rates above 97.9% for both. The ROC curve for LIME with Random Forest demonstrated excellent performance, with an AUC of 0.9955. (Results of experiment done in (Galego Hernandes *et al.*, 2021))

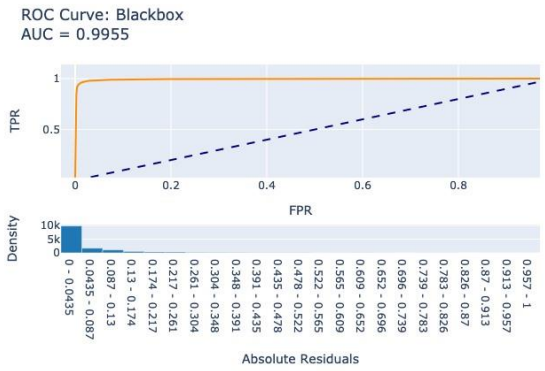

Figure 5. ROC Curve for LIME using Random Forest. (Source : (Galego Hernandes et al., 2021))

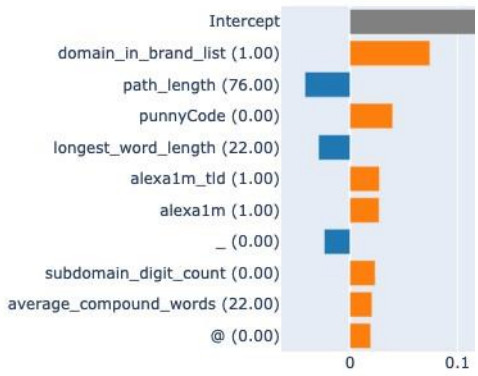

Figure 6. Explainable results presented by LIME. (Source: (Galego Hernandes et al., 2021))

Figure 6 illustrates the results of LIME, revealing the influential characteristics in the decision-making process of classifying URLs as phishing or legitimate. Each characteristic is represented by a bar, with its size indicating the weight of influence on the model's decision. The color of the bar (orange or blue) represents whether the characteristic contributes to identifying a legitimate URL or a phishing URL, respectively.

The analysis highlights the significance of the "domain in brand list" characteristic, indicating that URLs belonging to popular trademark domains have a strong influence. Another important characteristic is "punny code," which suggests ambiguity or double meaning in the URL composition. It's important to note that LIME provides local explanations rather than global explanations for classifications and results.

2) EBM:

EBM is a transparent and interpretable model that offers a desirable option over posthoc techniques, even if it means a potential loss in performance. However, EBM has demonstrated competitive performance, as indicated in Table I, where it is compared with other interpretable algorithms. The distinguishing feature of EBM is its ability to provide explanations at both the global and local levels.

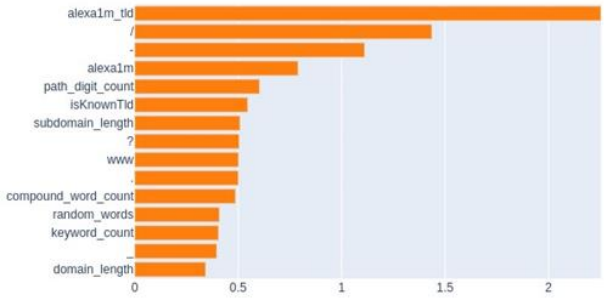

Figure 7. EBM overall global explanation (Source: (Galego Hernandes et al., 2021))

Figure 7 provides an overview of the most relevant characteristics influencing the model's decisions between phishing and legitimate samples. The characteristic "alexa1m tld," which represents the presence or absence of the Higher-Level Domain (TLD) of URLs in the most accessed sites according to Alexa, is identified as the most important for predictions.

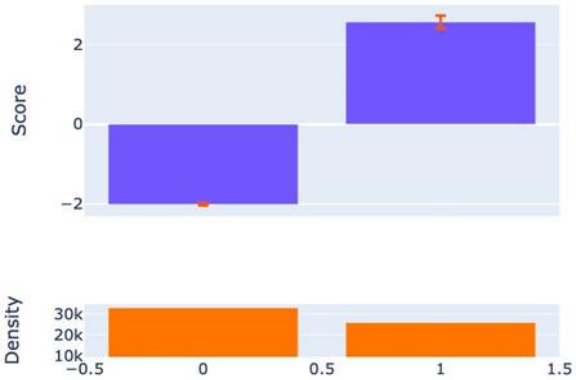

Figure 8. EBM global explanation for the presence of TLD in Alexa's top 1M sites. (Source: (Galego Hernandes et al., 2021))

Figure 8 further illustrates how this characteristic influences the algorithm's decisions, with true values favoring legitimate URLs and false values favoring phishing URLs.

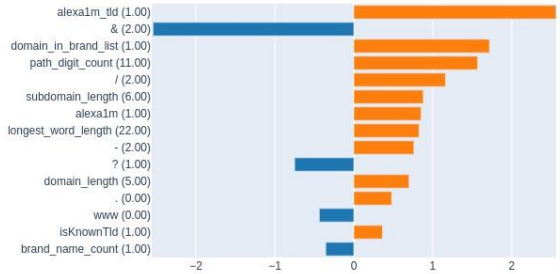

Figure 9. Local explanation of EBM (correct prediction of legitimate URL). (Source: (Galego Hernandes et al., 2021))

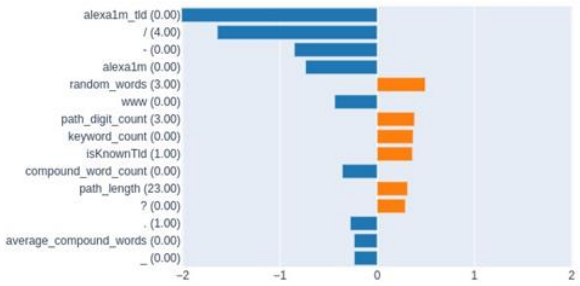

Figure 10. Local explanation of EBM (correct prediction of a Phishing URL). (Source: (Galego Hernandes et al., 2021))

Figures 9 and 10 present explanations for correct predictions made by the algorithm for legitimate and phishing URLs, respectively. The characteristics are represented with colors and horizontal positioning, with orange and right indicating relevance for legitimate URLs, while blue and left indicate relevance for phishing URLs.

The coherence between local and global explanations is reinforced by the significance of the "alexa1m tld" characteristic, as depicted in Figure 6. Such information can contribute to the development of more accurate models and enhance knowledge about the problem being addressed.

K. Creating Suitable Explanation Messages for Warnings(Greco, Desolda and Esposito, no date)

The delivery of the explanation plays a crucial role in building user trust and ensuring the user-friendliness of the XAI-based tool. Users need to understand and trust the explanations provided by the system to make informed decisions about potentially phishing emails. By employing design indications from the field of warnings for phishing attacks, such as the C-HIP model(Wogalter, 2018) and warning messages design guidelines(Bauer *et al.*, no date), the proposed message takes into account established best practices. This approach enhances the effectiveness of the warning system and increases user confidence in the tool's ability to detect phishing attempts.

The message itself is designed to have three distinct parts, each serving a specific purpose.

a) Detection Feature

Firstly, a short description is provided, highlighting the key features that the XAI model utilized to determine that the email is likely a phishing attempt. This empowers users with an understanding of the underlying factors contributing to the warning.

b) Hazard Explanation

Secondly, a concise explanation about the hazard associated with the suspicious email is included. This section aims to clearly communicate the potential risks and dangers involved in interacting with such emails.

c) Consequences

Lastly, the consequences of a successful phishing attack are outlined, emphasizing the potential harm that could occur if the user were to fall victim to the phishing attempt.

This comprehensive approach to message design ensures that users are well-informed about the nature of the threat and the importance of exercising caution.

By structuring the message in this manner, we aim to provide users with clear and meaningful information, enabling them to make informed decisions and take appropriate actions when encountering potentially malicious emails. The combination of feature description, hazard explanation, and consequences of a successful attack enhances user understanding, fosters trust in the XAI model, and ultimately promotes user safety in the face of phishing threats.

L. Considerations for Feasibility of XAI for Phishing Detection(Capuano et al., *2022)*

The current state of Explainable Artificial Intelligence (XAI) in phishing detection is limited, with only a few ad-hoc methods developed specifically for detecting phishing attacks. Phishing is a significant threat that affects anyone using technology, making AI-based prevention and detection crucial. It is important to have AI systems that not only make decisions but also provide explanations and awareness to ensure effective decision-making for both businesses and individual users. Users are more accepting of AI mistakes if they are explained and given an opportunity to improve, especially in cases of false negatives. In addition to XAI in cybersecurity, educating individuals using technological devices and browsing the internet about phishing is essential. Like training AI models, strategies should be devised to teach individuals how to avoid falling victim to phishing attacks, and these strategies must be explainable so that everyone can understand why certain decisions are made.

1) Availability of Labeled Data:

The availability of labeled data specifically annotated for phishing attacks emerged as a critical factor for building an XAI model. While several datasets exist for phishing detection, the availability of comprehensive and up-to-date datasets that include interpretability annotations is limited. This scarcity poses a challenge in training and evaluating XAI models for anti-phishing detection. Future efforts should focus on creating large-scale, diverse, and interpretable datasets to facilitate the development and evaluation of XAI models.

2) Feature Engineering:

Feature engineering plays a crucial role in identifying relevant characteristics of phishing attacks. The review identified a range of features commonly used for anti-phishing detection, including URL-based features (e.g., domain similarity, presence of suspicious keywords), content-based features (e.g., linguistic patterns, HTML tags), and behavior-based features (e.g., mouse movement, typing dynamics). However, choosing and creating features that accurately reflect the intricate and ever-evolving nature of phishing attempts continues to be a difficult research problem. Strong feature engineering methods are required in order to properly detect the subtle signs of phishing assaults and adapt to new phishing approaches.

3) Model Selection:

Anti-phishing detection has been used with a variety of machine learning and deep learning methods. Although these models have exhibited promising accuracy results, their interpretability is frequently constrained. For a XAI model for anti-phishing detection to be effective, accuracy and interpretability must be balanced. A potential answer might be provided by hybrid strategies that incorporate the advantages of various models, such as ensemble methods or rule-based classifiers. To find unique model architectures and strategies that can deliver high accuracy and insightful explanations, more study is required.

4) Interpretability Techniques:

The context of anti-phishing detection has led to the exploration of a number of interpretability strategies. Rules-based approaches, feature importance analysis (e.g., permutation feature importance, gradient-based approaches), local interpretability approaches (e.g., LIME, SHAP), and model-agnostic approaches (e.g., Layer-wise Relevance Propagation, Grad-CAM) are a few of these. These methods can offer justifications for the choices the model made, fostering greater openness and confidence. The use of these techniques for anti-phishing detection is still in its infancy, thus more investigation is required to determine how effective and scaleable they are in the face of phishing attempts. It may also be investigated to use specialized interpretability methods that can capture the unique traits of phishing indications.

5) Evaluation Metrics:

There are particular difficulties in evaluating XAI models for anti-phishing detection. Additional metrics that represent the interpretability and explainability features of the models need to be established. Traditional metrics like accuracy, precision, recall, and F1 score are frequently utilized. The review emphasized the requirement for cutting-edge evaluation measures that gauge the caliber and value of the justifications offered by XAI models. These metrics ought to take into account elements like explanation clarity, user comprehension, and influence on decision-making. Fair and thorough comparisons between various models and strategies would be made possible by the creation of standardized assessment frameworks and benchmark datasets tailored to XAI in anti-phishing detection.

Overall, the review findings indicate that building an XAI model for anti-phishing detection is a promising approach to enhance transparency and interpretability in this domain. However, challenges related to the availability of labeled data, feature engineering, model selection, and the development of effective interpretability techniques need to be addressed. Future research should focus on overcoming these challenges and developing robust and scalable XAI models that not only detect phishing attacks accurately but also provide meaningful and understandable explanations for their decisions. Such advancements will contribute to building more trustworthy and user-centric

IV. DISCUSSION AND ANALYSIS

The discussion and analysis of the feasibility of building an Explainable Artificial Intelligence (XAI) model for anti-

phishing detection revealed significant findings and insights. Here are the key points:

A. XAI and Its Importance:

Explainable AI (XAI) addresses concerns about the transparency and interpretability of AI models, improving reliability, trustworthiness, and accountability. The rise of complex machine learning models has made transparency more important, leading to increased interest in XAI. Lack of transparency in technologies like AI, robotics, and IoT poses challenges for functionality and security in Industry 4.0.

B. XAI Principles:

Explanation: AI systems should provide explanations related to their outcomes or processes.

Meaningful: Explanations should be understandable to recipients, considering their knowledge and needs.

Explanation Accuracy: Explanations should accurately reflect the system's process, balancing detail and comprehensibility.

Knowledge Limits: Systems should recognize and declare their limitations to avoid unreliable outputs.

C. XAI Methods/Types:

- Feature Visualization: It helps understand what an AI model looks for in input data and identify potential biases or errors. Deep neural networks commonly use this technique.

- Saliency Mapping: It generates heat maps highlighting important regions in the input data for the model's decision-making. It improves accuracy, fairness, and provides concise explanations.

- Model Interpretation: Model-agnostic techniques like LIME, SHAP, and EBM analyze the input-output relationship without relying on specific model details, promoting transparency.

D. XAI Challenges:

- Balancing Explainability and Performance: Achieving both high performance and explainability is a trade-off that requires the development of advanced techniques.

- Human Factors: Understanding how people interpret XAI explanations is a challenge, and research should focus on tailoring techniques to enhance trust and usability.

- Universal Standard: The lack of a universal standard or framework hinders the development and evaluation of XAI techniques.

- Bias: Addressing biases and ensuring fairness is crucial in XAI, requiring techniques to identify and mitigate biases in machine learning models.

- Evaluation: Standardized metrics and benchmarks are needed to evaluate the accuracy and practicality of XAI techniques effectively.

E. Ethical Principles in XAI:

Ethical principles in XAI include accountability, responsibility, transparency, fidelity, bias mitigation, causality understanding, fairness, and safety. These principles ensure responsible, transparent, and fair AI systems.

F. Phishing Attacks:

- Phishing is a fraudulent technique where attackers deceive users to steal sensitive information through fake emails or websites.
- Spear phishing is a targeted form of phishing that involves researching victims for more convincing scams.
- User training is crucial to recognize and avoid phishing attacks driven by emotional responses.
- Domain-based email authentication standards like DMARC help block fraudulent emails.
- Phishing attacks are expanding to social media platforms, and adversarial AI is expected to play a larger role.
- Phishing and spear phishing are major vectors for unintentional insider threats.
- Webmail and SaaS providers are increasingly targeted, surpassing attacks on payment services.
- Business Email Compromise (BEC) attacks, including spear phishing, pose significant threats, resulting in substantial financial losses.
- The number of phishing sites using HTTPS has risen, but HTTPS and SSL can create false trust, and legitimate hacked websites can host phishing content.

G. Phishing Vulnerability:

User behavior, gender, and risk-taking influence susceptibility to phishing attacks. Preventive measures, technical solutions, and user education are essential for mitigating phishing risks.

H. Machine Learning for Phishing Detection:

Machine learning algorithms, such as Logistic Regression and Random Forest, have shown high accuracy in detecting phishing websites. Natural Language Processing (NLP) is also used to analyze text in attack links.

I. XAI Models for Phishing Detection:

LIME and EBM are XAI models applied to anti-phishing detection. LIME provides interpretability for black-box models, and EBM offers transparency with competitive performance. They provide explanations for their decisions, enhancing transparency and trust.

J. Creating Suitable Explanation Messages for Warnings

The delivery of explanations is crucial for user trust and the user-friendliness of the XAI-based tool. The proposed message incorporates design indications from the field of phishing warnings, such as the C-HIP model and warning messages design guidelines. It consists of three parts: a description of the detection features, an explanation of the hazard, and the consequences of a successful attack. This comprehensive approach ensures that users are well-informed, fostering trust in the XAI model and promoting user safety against phishing threats.

K. Considerations for Feasibility of XAI for Phishing Detection:

Availability of labeled data, robust feature engineering, model selection balancing accuracy and interpretability, effective interpretability techniques, and development of

evaluation metrics are critical for building feasible XAI models for anti-phishing detection.

L. Comparison with Traditional Approaches:

Compare the performance and interpretability of XAI models with traditional machine learning models used in anti-phishing detection. Discuss the potential advantages of XAI models in terms of improved accuracy, user trust, and explainability. Analyze the trade-offs between the interpretability provided by XAI models and the performance achieved by traditional models. Highlight the potential of hybrid approaches that combine the strengths of both traditional and XAI models to achieve a balance between accuracy and interpretability.

In conclusion, building an XAI model for anti-phishing detection shows promise in enhancing transparency and interpretability. However, challenges related to data availability, feature engineering, model selection, interpretability techniques, and evaluation metrics need to be addressed for the development of robust and scalable XAI models in this domain.

V. FUTURE RESEARCH DIRECTIONS

Based on the findings and discussions presented in the review paper on the feasibility of building an XAI model for anti-phishing detection, several promising research directions and areas for future exploration can be identified:

A. Enhanced Interpretability Techniques:

Further research is needed to develop and refine interpretability techniques specifically tailored to anti-phishing detection. This includes investigating methods to generate more intuitive and understandable explanations for the decisions made by XAI models. Exploring visualization techniques, interactive interfaces, and natural language generation approaches can contribute to enhancing the interpretability and user-friendliness of XAI models in the context of anti-phishing detection.

B. Creation of Comprehensive and Interpretable Datasets:

To train and test XAI models, vast, varied, and interpretable datasets must be accessible. The compilation of comprehensive datasets with annotations that capture both the existence of phishing attacks and the interpretable features employed by the models should be the main goal of future research. Collaboration between researchers and business partners can make it easier to gather real-world data while maintaining data security and privacy.

C. Hybrid Approaches:

Look into hybrid strategies that combine the advantages of XAI models with conventional machine learning models. This may include creating ensemble approaches that combine the accuracy of conventional models with the interpretability of XAI models. Additionally, looking at ways to incorporate rule-based classifiers or expert systems with XAI models can produce anti-phishing detection systems that are comprehensible and interpretable.

D. User-Centric Design and User Studies:

To assess user views, preferences, and trust in XAI models for anti-phishing detection, conduct user research and gather feedback. This study can be used to pinpoint user requirements, design considerations, and usability issues. XAI models can be adapted to match user expectations and successfully help decision-making in anti-phishing scenarios by integrating user feedback into the development process.

E. Evaluation Metrics for Interpretability:

Create assessment metrics that particularly evaluate the parts of XAI models for anti-phishing detection that are interpretable and explainable. The effectiveness and value of the explanations given might not be fully captured by conventional criteria like correctness and precision. The influence of explanations on user comprehension and decision-making should be assessed using novel metrics that quantify explanation clarity, completeness, and impact. Additionally, analyzing the relationship between user trust and interpretability can reveal important information about how effective XAI models are.

F. Real-World Deployment and Integration:

Examine the difficulties and factors to be taken into account while implementing XAI models for anti-phishing detection in practical situations. Problems regarding scalability, computational effectiveness, and integration with current anti-phishing systems should be addressed. Look for strategies to adapt XAI models to changing phishing tactics and retrain them in dynamic environments to maintain their effectiveness.

G. Ethical and Privacy Considerations:

Investigate the ethical implications and privacy concerns associated with the deployment of XAI models for anti-phishing detection. Research should address issues such as the transparency of data usage, potential biases in model decision-making, and the secure handling of sensitive information. Developing guidelines and frameworks that ensure the responsible and ethical use of XAI models in anti-phishing scenarios is essential.

Future research directions can contribute to the development and application of XAI models for anti-phishing detection. By focusing on these directions, it is possible to achieve more transparent, comprehensible, and efficient systems to effectively combat phishing attempts.

VI. CONCLUSION

The feasibility of building an Explainable Artificial Intelligence (XAI) model for anti-phishing detection has been thoroughly examined through an extensive literature review. The review findings highlight the potential of XAI models in enhancing transparency, interpretability, and user trust in anti-phishing systems. However, several challenges must be addressed, including the scarcity of labeled data with interpretability annotations and the need for comprehensive and interpretable datasets specific to anti-phishing detection. Additionally, the development of enhanced interpretability techniques, hybrid approaches combining traditional machine learning models and XAI models, user-centric design, and evaluation metrics for interpretability are identified as promising future research directions. Ethical considerations and privacy concerns related to the deployment of XAI models for anti-phishing detection should also be carefully addressed. By addressing these challenges and pursuing the suggested research directions, transparent, interpretable, and

effective anti-phishing systems can be developed, empowering users to make informed decisions and fortify their defenses against phishing attacks. Continued research, collaboration, and user-centric design are crucial for realizing the full potential of XAI in anti-phishing detection and creating a safer digital environment.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to all the individuals and organizations who have contributed to this review paper on the feasibility of building an Explainable Artificial Intelligence (XAI) model for anti-phishing detection. I am truly thankful to the researchers and experts in the field whose valuable studies and insights have formed the foundation of this review. I also want to acknowledge the support and resources provided by organizations and institutions that have contributed to this research.

Furthermore, I would like to extend my deep appreciation to my supervisor and the academic staff of the KDU Faculty of Computing for their guidance, expertise, and continuous support throughout the research process and the writing of this review paper. Their knowledge and mentorship have played a crucial role in shaping my understanding and refining my ideas. I am grateful for their dedication to academic excellence and their commitment to fostering a research-oriented environment, which has provided me with valuable opportunities for growth and learning. Their patience, encouragement, and valuable feedback have significantly contributed to the development of this paper.

I am also inspired by their passion for advancing knowledge in the field of computing and their commitment to nurturing the next generation of researchers and professionals. Their dedication has motivated me to strive for excellence in my work.

REFERENCES

Abroshan, H. *et al.* (2021) 'Phishing Happens Beyond Technology: The Effects of Human Behaviors and Demographics on Each Step of a Phishing Process', *IEEE Access*, 9, pp. 44928–44949. Available at: https://doi.org/10.1109/ACCESS.2021.3066383.

Bauer, L. *et al.* (no date) 'Warning Design Guidelines'.

Capuano, N. *et al.* (2022) 'Explainable Artificial Intelligence in CyberSecurity: A Survey', *IEEE Access*, 10, pp. 93575–93600. Available at: https://doi.org/10.1109/ACCESS.2022.3204171.

Galego Hernandes, P.R. *et al.* (2021) 'Phishing Detection Using URL-based XAI Techniques', in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, Orlando, FL, USA: IEEE, pp. 01–06. Available at: https://doi.org/10.1109/SSCI50451.2021.9659981.

Greco, F., Desolda, G. and Esposito, A. (no date) 'Explaining Phishing Attacks: An XAI Approach to Enhance User Awareness and Trust'.

Hanif, A., Zhang, X. and Wood, S. (2021) 'A Survey on Explainable Artificial Intelligence Techniques and Challenges', in *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*. *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*, Gold Coast, Australia: IEEE, pp. 81–89. Available at: https://doi.org/10.1109/EDOCW52865.2021.00036.

Phillips, P.J. *et al.* (2021) *Four principles of explainable artificial intelligence*. NIST IR 8312. Gaithersburg, MD: National Institute of Standards and Technology (U.S.), p. NIST IR 8312. Available at: https://doi.org/10.6028/NIST.IR.8312.

Phishing : ENISA Threat Landscape (no date). enisa: European Union Agency for Cyber Security, p. 24. Available at: https://www.enisa.europa.eu/topics/cyber-threats/threats-and-trends/etl-review-folder/etl2020-phishing.

Reddy, G.P. and Kumar, Y.V.P. (2023) 'Explainable AI (XAI): Explained', in *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*. *2023 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Vilnius, Lithuania: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/eStream59056.2023.10134984.

Wogalter, M.S. (2018) 'Communication-Human Information Processing (C-HIP) Model', in M.S. Wogalter (ed.) *Forensic Human Factors and Ergonomics*. 1st edn. Boca Raton : Taylor & Francis, 2018. | Series: Human factors and ergonomics: CRC Press, pp. 33–49. Available at: https://doi.org/10.1201/9780429462269-3.

Yu, W.D., Nargundkar, S. and Tiruthani, N. (2008) 'A phishing vulnerability analysis of web based systems', in *2008 IEEE Symposium on Computers and Communications*. *2008 IEEE Symposium on Computers and Communications (ISCC)*, Marrakech: IEEE, pp. 326–331. Available at: https://doi.org/10.1109/ISCC.2008.4625681.