Panamerican Mathematical Journal
ISSN: 1064-9735
Vol 35 No. 1s (2025)

Explainable Machine Learning Models for Real-Time Threat Detection in Cybersecurity

Bhagyashree D Shendkar[1], Dhanraj Dhotre[2], Sumit Arun Hirve[3], Palash Sontakke[4], Hyderali Hingoliwala[5], G B Sambare[6]

[1,2,3,5]Department of Computer Science and Engineering, MIT School of Computing, MIT Art Design and Technology University, Pune, India

[4]Department of Information Technology, MIT School of Computing, MIT Art Design and Technology University, Pune, India

[6]Pimpri Chinchwad College of Engineering, Pune, India

bhagyashree.shendkar@mituniversity.edu.in[1], dhanraj.dhotre@mituniversity.edu.in[2], sumit.hirve@gmail.com[3], palash.sontakke@mituniversity.edu.in[4], hyderali.hingoliwala@mituniversity.edu.in[5], santosh.sambare@pccoepune.org[6]

Article History:

Received: *30-08-2024*

Revised: *15-10-2024*

Accepted: *26-10-2024*

Abstract:

In the rapidly evolving landscape of cybersecurity, traditional machine learning models often operate as "black boxes," providing high accuracy but lacking transparency in decision-making. This lack of explainability poses challenges for trust and accountability, especially in critical areas like threat detection and incident response. Explainable machine learning models aim to address this by making the model's predictions more understandable and interpretable to users. This research integrates explainable machine learning models for real-time threat detection in cybersecurity. Data from multiple sources, including network traffic, system logs, and user behavior, undergo preprocessing such as cleaning, feature extraction, and normalization. The processed data is passed through various machine learning models, including traditional approaches like SVM and decision trees, as well as deep learning models like CNN and RNN. Explainability techniques such as LIME, SHAP, and attention mechanisms provide transparency, ensuring interpretable predictions. The explanations are delivered through a user interface that generates alerts, visualizations, and reports, facilitating effective threat assessment and incident response in decision support systems. This framework enhances model performance, trust, and reliability in complex cybersecurity scenarios.

Keywords: Explainable AI, Cybersecurity, Threat Detection, Machine Learning, SHAP, Decision Support System.

1. Introduction

In recent years, the reliance on machine learning (ML) in cybersecurity has grown significantly, driven by the increasing complexity and frequency of cyber threats. Traditional rule-based security systems have struggled to keep up with the rapidly evolving threat landscape, prompting a shift towards more adaptive, data-driven approaches[1]. Machine learning models, with their ability to process large volumes of data and identify patterns indicative of malicious activity, have become invaluable in various cybersecurity applications such as malware detection, phishing identification, intrusion detection, and anomaly detection. Even though these machine learning models have shown to be effective, cybersecurity is greatly hampered by their "black-box" design[2]. High accuracy but opaqueness characterizes most machine learning methods, particularly those involving sophisticated models like deep learning networks. This is especially troubling in the field of cybersecurity, where

prompt, dependable, and comprehensible judgments are essential for protecting against threats that arise in real time[3]. Cybersecurity experts may find it difficult to accept, analyze, or respond appropriately on the basis of model output if they do not know why the model identified a particular activity as harmful[4][5]. Moreover, explainability becomes crucial for compliance, auditing, and accountability in extremely sensitive sectors like finance, healthcare, or critical infrastructure.

Explainability in machine learning models is therefore not merely a technological problem but also a critical component of creating a transparent and trustworthy environment and facilitating efficient decision-making in high-stakes cybersecurity settings. To bridge the gap between sophisticated machine learning algorithms and the requirement for human-understandable reasoning, an explainable AI (XAI) method is becoming crucial, where machine learning models deliver interpretable insights.

The convergence of machine learning and cybersecurity, particularly in the field of real-time threat detection, is the main topic of this overview of the literature. The review's objectives are to investigate how explainability approaches are integrated to improve the transparency of machine learning models and how these models are currently utilized to identify hazards as they arise[6][7]. The paper will look at a variety of machine learning approaches used in real-time detection systems, ranging from more complex deep learning models (convolutional neural networks, recurrent neural networks) to more conventional models (decision trees, support vector machines). The evaluation will specifically look into how these models identify anomalies, malware, or illegal access in system logs, user activity, and network traffic. This review's emphasis on the explainability of these machine learning models is crucial[8]. The range of approaches covered will include several Explainable AI methods for interpreting, rationalizing, and supporting machine learning model decisions. The evaluation will also look at the difficulties and solutions associated with incorporating explainable models into real-time systems that must respond with accuracy, timeliness, and comprehensibility.

Its black-box nature makes deploying machine learning models for cybersecurity one of the biggest hurdles. Interpreting the conclusions made by many high-performing models—particularly deep learning algorithms—can be challenging. These models handle data in ways that are difficult for people to access, and they frequently produce judgments without giving an explanation[9][10]. In the field of cybersecurity, where knowing why a certain threat was discovered is just as crucial as identifying the threat itself, this lack of transparency presents a serious problem. This becomes even more urgent when considering real-time threat detection. For security professionals to properly minimize threats, choices must be made quickly and with confidence[11][12]. On the other hand, the security team may not fully believe the system's suggestions if an ML system detects an anomaly or possible attack but is unable to provide an explanation, which could cause hesitancy or mistakes in the response. Explainability also makes it harder to audit and confirm the system's actions, which is important in situations like critical infrastructure or financial systems where mistakes could have dire repercussions. Another issue is that most real-time detection systems in use today put interpretability last and speed and accuracy first. These systems frequently concentrate on swiftly identifying dangers; however, they infrequently offer clarifications that could aid cybersecurity teams in comprehending the nature of the issue and taking appropriate action[13][14]. The absence of interpretable output and the disparity between extremely fast and accurate detection methods pose a serious barrier to machine learning's wider application in cybersecurity.

Furthermore, the need for transparent decision-making in AI-based systems is growing due to regulatory compliance and accountability concerns. The General Data Protection Regulation (GDPR) in the European Union, for example, requires automated systems to provide an explanation to the individuals impacted by their decisions. This highlights the increasing need for explainable models in a variety of industries, including cybersecurity. Achieving a balance between interpretability and model performance is necessary to tackle the black-box issue in machine learning for cybersecurity. This

paper will look at the methods currently used to strike that balance and examine how Explainable AI (XAI) techniques can be used to improve the transparency and reliability of real-time cybersecurity systems.

2. Machine Learning in Cybersecurity: Current Approaches

2.1 Traditional ML Models for Cybersecurity

Traditional machine learning models have been widely applied in cybersecurity tasks due to their ability to recognize patterns and anomalies in data. Some of the most used models include:

- **Support Vector Machines (SVMs)**: SVMs are effective classifiers that are used to identify network breaches, malware, and spam. They work especially well on binary classification tasks and high-dimensional data[15]. SVMs function by determining the best hyperplane to divide data points into distinct groups, such harmful versus legitimate traffic.
- **Decision Trees:** Rule-based models such as decision trees provide an easily understood framework. The fact that they can deconstruct decision-making processes into comprehensible pathways makes them a popular component of intrusion detection systems (IDS)[16]. For example, using a sequence of rule-based judgments, a decision tree can determine whether incoming network packets are malicious or benign.
- **Random Forests**: Decision tree ensembles called random forests increase prediction accuracy and robustness[17]. They have been applied to cybersecurity to identify malware, spam, and phishing by combining predictions from several decision trees, which lowers the chance of overfitting and enhances generalization.
- **Neural Networks:** Early cybersecurity applications, such identifying network intrusions, also made use of shallow neural networks. By changing the network's weights to link inputs to outputs—for example, network logs to attack categories—these models are trained on data.

Applications of these models:

- **Malware Detection**: Traditional models like SVMs and random forests have been used to classify files as benign or malicious by analyzing features like file structure, permissions, and byte sequences.
- **Intrusion Detection Systems (IDS)**: Decision trees and random forests are widely employed in network-based intrusion detection to differentiate between normal and malicious traffic.
- **Phishing Detection**: Random forests and decision trees have been applied to email data to detect phishing attempts by analyzing features such as URL length, email subject, and content.
- **Anomaly Detection**: Unsupervised models like k-means clustering or autoencoders have been applied to detect unusual patterns in network traffic or user behavior, flagging potential security threats.

Successes and limitations of these models:

- **Successes**: Traditional machine learning models are well-understood, often easier to interpret, and relatively efficient in terms of computation, especially when compared to deep learning models. Their clear decision-making process (e.g., in decision trees) makes them more trustworthy for cybersecurity professionals who need to understand how a threat was detected.
- **Limitations**: These models often struggle with detecting **new or evolving threats**, such as zero-day attacks, because they rely heavily on pre-defined patterns or features. Additionally, they are less effective when working with large-scale or highly complex datasets, where the relationships between variables are non-linear.

2.2 Deep Learning in Cybersecurity

Deep learning has emerged as a powerful tool in cybersecurity due to its ability to analyze vast amounts of data and uncover intricate patterns that traditional models might miss. Well-known architectures in this subject include Long Short-Term Memory Networks (LSTMs) and Recurrent Neural Networks (RNNs), which include Convolutional Neural Networks (CNNs). By understanding binary data as images or sequences, CNNs—which were first created for image processing—have been repurposed for malware detection. Through their hierarchical feature learning, CNNs can identify complex and obfuscated malware variants with this approach. Analyzing sequential data, such network traffic and user behavior logs, is an area in which RNNs and LSTMs thrive[18]. Particularly LSTMs are skilled in spotting temporal patterns and anomalies, which is essential for quickly identifying anomalous user behavior or slow-moving attacks. Deep learning models have great performance, but they have a lot of cybersecurity challenges. Their "black-box" aspect, which obstructs transparency and interpretability, is a significant problem. In contrast to simpler models like decision trees, deep learning models frequently have an opaque decision-making process, which makes it challenging for cybersecurity experts to comprehend and rely on the model's results[19]. In situations with high stakes, where false positives or negatives can have dire repercussions, this lack of clarity can be problematic. In important security scenarios, the inability to provide an explanation for a particular threat's detection or decision-making process might impede effective incident response and lower the overall reliability of deep learning models.

2.3 Real-Time Threat Detection Systems

Real-time threat detection systems are designed to identify and mitigate threats as they occur, without significant delay. These systems often rely on machine learning models to analyze incoming network traffic, user behavior, and system activity in near-real-time, providing alerts or taking automated actions to prevent attacks. Key areas where machine learning-based real-time detection is employed include:

- **Intrusion Detection Systems (IDS)**: AI-powered real-time intrusion detection systems (IDSs) may identify unusual patterns in network traffic and indicate possible intrusions before they become visible[20]. These systems detect deviations from typical traffic behavior using both supervised (classification) and unsupervised (anomaly detection) learning techniques.

- **Malware detection:** By instantly recognizing a malware's signature or anomaly, ML models included into endpoint security solutions can scan files or processes for dangerous activity in real-time[21], assisting in the prevention of malware from starting to execute.

- **Anomaly Detection:** In real-time systems, unsupervised machine learning models are used to keep an eye on user behavior, network traffic[22], or system logs for any anomalous activity that might point to a malware infection, insider threat, or security breach.

Challenges of real-time threat detection:

- **Latency**: Latency is one of the main problems with real-time threat detection systems. To deliver fast answers to threats, machine learning models must digest vast volumes of data quickly—often in milliseconds. The computational intensity of deep learning models can cause severe delays.

- **Processing Speed:** To manage the enormous volume of data produced by contemporary networks, real-time systems need to analyze data quickly. Even deep learning models, which are scalable, may encounter bottlenecks in busy environments. Conventional machine learning models may not be tuned for real-time performance.

- **Accuracy vs. False Positives**: In real-time detection systems, achieving a balance between low false positive rates and high accuracy is crucial. An excessive number of false positives from a system can overwhelm security personnel and cause alert fatigue, which can result in the miss-identification of genuine threats. Conversely, low-sensitivity models may be unable to identify some types of attacks, which could result in breaches. One of the key goals of this field of study is to strike a compromise between minimizing false positives and ensuring prompt, accurate detection.

3. Explainability in Machine Learning Models

3.1 The Concept of Explainable AI (XAI)

To overcome the "black-box" aspect of conventional models, explainable AI (XAI) aims to make machine learning model outputs comprehensible and interpreted for users. In domains with significant risks, including healthcare, finance, and cybersecurity, comprehending the methods and rationale behind a model's conclusions is crucial for fostering confidence and guaranteeing its efficient implementation. Post-hoc approaches like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which offer insights into model predictions after the fact, are important XAI techniques. SHAP provides a game-theoretic method to ascribe each feature's impact on the final output, hence elucidating complex model decisions, whereas LIME uses specific prediction learning to mimic a black-box model locally[23]. Apart from ad hoc methods, models that are transparent by nature are intended to be transparent from the beginning. While logistic regression and other linear models provide explicit feature contributions through their coefficients, decision trees and rule-based systems provide straightforward, intelligible pathways for decision-making[24]. GAMs, or generalized additive models, combine transparency in feature contributions with non-linear interactions to further improve interpretability[25]. These techniques aim to strike a compromise between interpretability and model accuracy, which helps to lessen the opacity issue with sophisticated machine learning models and guarantees that their conclusions are trustworthy and understandable.

3.2 Importance of Explainability in Cybersecurity

Explainability is particularly crucial in cybersecurity due to the direct impact machine learning models have on protecting sensitive systems and data. For cybersecurity professionals, the ability to understand why a model has flagged a particular network activity, file, or user behavior as malicious is essential[26]. This clarity aids in making informed decisions during incident response, allowing professionals to validate predictions, prioritize threats, and determine appropriate countermeasures. Moreover, compliance with regulations like the General Data Protection Regulation (GDPR) mandates that automated decision-making systems provide explanations for their actions, especially when dealing with sensitive personal data. This requirement underscores the legal necessity of explainability in certain cybersecurity contexts. Additionally, building trust in AI-driven cybersecurity systems hinges on transparency. Security teams are more likely to rely on models that provide clear explanations for their recommendations, such as blocking a user or quarantining a file[27]. Explainable models foster trust by enabling users to understand and validate the AI's decisions, which is crucial for effective and confident use of these systems. For instance, in auditing and legal scenarios, having detailed explanations of model decisions is necessary for reviewing and defending actions taken by security systems. This is especially important in industries where incorrect model outputs could lead to compliance issues or impact sensitive data handling, making explainability a critical component for both operational and regulatory purposes.

3.3 Explainability vs. Performance Trade-offs

A major difficulty in machine learning, especially in cybersecurity, is balancing interpretability with model accuracy. By efficiently managing big information and identifying subtle patterns, ensemble techniques like random forests or complex models like deep neural networks frequently produce results

with great accuracy. Nevertheless, consumers often find it difficult to understand the underlying assumptions of these models, as they are often difficult to interpret. Simpler models, such as logistic regression or decision trees, on the other hand, provide more transparent and understandable decision routes but may not perform as well, particularly when dealing with unstructured data or subtle attack patterns[28]. This trade-off draws attention to a fundamental conundrum in the development of cybersecurity machine learning systems, where both high performance and comprehensible explanations are essential. Many interesting strategies are being investigated to tackle this problem. With the ability to produce predictions from a high-performing model and provide explanations for them from a simpler model, hybrid models combine the best features of both complicated and interpretable models[29]. Real-time, instance-based explanations are made possible by post-hoc explainability approaches like LIME and SHAP, which keep complex models operational while providing intelligible results for decisions. Furthermore, efforts to improve the interpretability of deep learning models by means of techniques such as feature visualization and attention mechanisms are advancing. By bridging the gap between high performance and explainability, these advancements may increase the transparency and accessibility of high-accuracy, real-time models for cybersecurity applications.

4. Current XAI Techniques Applied to Cybersecurity

4.1 Post-Hoc Explanation Techniques

LIME (Local Interpretable Model-Agnostic Explanations): LIME is a well-liked method that builds local, interpretable models around a specific instance of interest, explaining any machine learning model's predictions. LIME operates by introducing perturbations (e.g., changes to certain features) to the input data and tracking changes in the model's predictions. The behavior of the complex model in the immediate vicinity of the instance is then approximated by fitting a straightforward, understandable model, such as a linear regression[30]. This makes it possible for security analysts to comprehend the reasoning behind a machine learning model's prediction (such as labelling a file as malicious or raising a red flag for suspicious network activity). LIME has been used in the field of cybersecurity to describe machine learning models that are utilized for tasks like intrusion detection systems (IDS) and malware detection[31]. For instance, LIME can determine which aspects of a file—such as its structure, permissions, or byte patterns—were most important to the model when determining whether to classify it as harmful in the context of malware detection. When features like packet size, protocol type, or the source and destination of the traffic are highlighted in an IDS alert, LIME can assist security teams in understanding why a particular piece of network traffic was reported as suspicious. Analysts can promptly confirm the correctness of the discovery and choose the best course of action thanks to this localized information.

SHAP (SHapley Additive exPlanations): Based on cooperative game theory, SHAP is an explanation technique that quantifies each characteristic of a model by assigning a significance value (Shapley value) that indicates how much the feature contributes to a specific prediction. SHAP can offer global explanations, which give a summary of the key elements in all of the model's predictions, as well as local explanations, which explain specific forecasts. Because of its dual functionality, SHAP is particularly helpful in the field of cybersecurity, where it is essential to comprehend both broad trends (like typical signs of phishing attempts) and detailed choices (like the reason behind a particular email's flagging as a phishing effort). When it comes to phishing detection, SHAP can clarify why a machine learning model classified an email as phishing based on specific characteristics (such as the sender domain, quantity of links, or specific phrases)[32]. This makes it easier for cybersecurity experts to comprehend and evaluate the threats connected to each email that has been identified fast. When it comes to anomaly detection, SHAP can be used to describe the key elements of system logs or network traffic that led to the flagging of a particular anomaly, such as unusual login times, geographic

locations, or high volumes of network traffic. These clarifications are very helpful for real-time threat identification, where quick thinking and action are essential.

4.2 Inherently Interpretable Models

Because decision trees and rule-based models have easily comprehensible decision-making processes, they are frequently utilized in cybersecurity and are intrinsically interpretable. With each node representing a feature and each branch representing a decision rule, decision trees offer a simple framework that results in a classification at the leaf node. This approach is especially helpful for cybersecurity jobs where transparency is essential, including malware categorization and intrusion detection[33]. Security analysts can more easily comprehend and verify the judgments made by the model by using a decision tree, which, for example, can track the precise series of decisions that result in the classification of network traffic as benign or malicious. In a similar vein, rule-based systems work well for real-time threat detection since they apply predetermined criteria to incoming data. These systems can flag suspicious activity based on specific rules derived from historical data, such as unusual IP addresses or rare protocols, allowing for quick, interpretable responses to potential threats. An intermediate solution between interpretability and performance is provided by generalized additive models[34]. By giving each variable, a distinct, non-linear effect, GAMs model the link between features and outcomes while preserving some degree of transparency and ability to handle intricate patterns. In cybersecurity activities, where knowing feature contributions is crucial for efficient incident response, this balance is advantageous. While GAMs are not as popular in cybersecurity as other models, they are becoming more and more popular for applications like malware classification and network traffic analysis[35]. They are helpful in elucidating the reasons for the flagging of data points as suspicious since they model the contribution of each attribute to the prediction, yielding findings that are easy to understand. For real-time threat identification and analysis, GAMs are a viable solution because of their interpretability and reasonable prediction power.

4.3 Deep Learning and XAI

Deep learning models, which are typically recognized for being "black-box" systems, are becoming easier to understand thanks to new methods in Explainable AI (XAI). Saliency maps, which emphasize the most important input features, attention mechanisms, which indicate which portions of the input data the model concentrates on, and Grad-CAM, which creates heat maps that illustrate which input data areas contributed most to a model's prediction, are some notable innovations[36]. These methods are used in cybersecurity to analyze network traffic and detect malware, assisting in the recognition and comprehension of anomalous patterns or significant characteristics[37]. Deep learning models may maintain high performance while offering critical transparency by including these explainable methodologies, which enhances the efficacy and reliability of AI-driven cybersecurity systems.

5. Proposed Methodology

The figure 1 shows architecture for an explainable machine learning system used in real-time threat detection in cybersecurity.

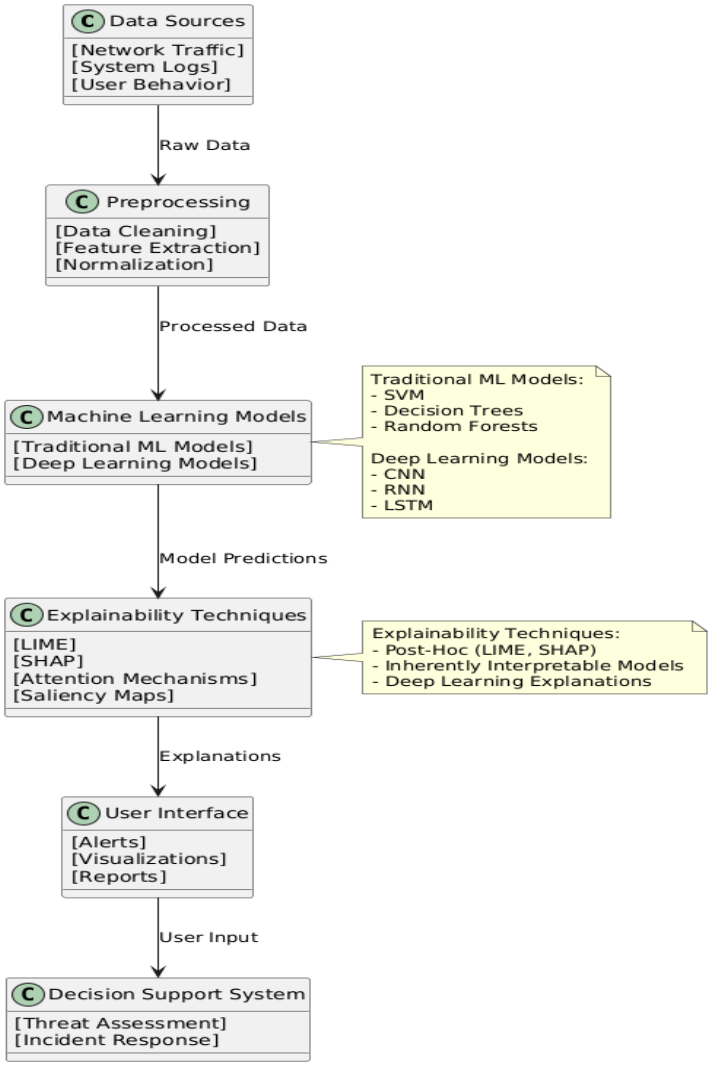

Figure 1: An architecture for an explainable machine learning system used in real-time threat detection in cybersecurity

Data Sources: This component represents the various inputs that the system analyses. **Network Traffic** includes data packets, protocols, and flow patterns from network communications. **System Logs** consist of records from operating systems, applications, and security tools that track user activity and system events. **User Behavior** data captures patterns in how users interact with systems, which can be critical for identifying suspicious activities or anomalies.

Preprocessing: Before feeding the data into machine learning models, it undergoes several preprocessing steps to ensure it is clean, relevant, and standardized. **Data Cleaning** involves removing noise and correcting errors in the data. **Feature Extraction** selects and transforms raw data into meaningful features that can enhance model performance. **Normalization** adjusts the scale of features to ensure consistent input, which helps in improving model accuracy and convergence.

Machine Learning Models: The heart of the system lies in this component, which applies various machine learning techniques to the pre-processed data. **Traditional ML Models** include algorithms like Support Vector Machines (SVMs), Decision Trees, and Random Forests. These models are well-established and often used for tasks such as malware detection and intrusion detection. **Deep Learning Models**, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs),

are used for more complex pattern recognition tasks. These models excel in analyzing large volumes of data and identifying subtle, intricate patterns that might indicate a threat.

Explainability Techniques: To address the opacity of machine learning models, this component focuses on providing understandable explanations for the model's predictions. **LIME (Local Interpretable Model-Agnostic Explanations)** and **SHAP (SHapley Additive exPlanations)** are post-hoc techniques that generate local and global explanations of model behavior, helping users understand why specific predictions were made. **Attention Mechanisms** and **Saliency Maps** are used for deep learning models to highlight which parts of the input data (e.g., specific network packets or code sections) were influential in the model's decision-making process.

User Interface: This component presents the results and explanations to cybersecurity professionals. **Alerts** notify users of detected threats or anomalies in real-time. **Visualizations** provide graphical representations of the data and model predictions, making complex information more accessible. **Reports** offer detailed summaries and analyses of threat detections, helping users understand and respond to security incidents effectively.

Decision Support System: Integrating with the user interface, this component helps in translating model predictions and explanations into actionable insights. **Threat Assessment** involves evaluating the severity and potential impact of detected threats. **Incident Response** guides the actions that security teams should take based on the assessment, such as isolating affected systems, applying patches, or notifying relevant stakeholders.

5. Challenges and Limitations of XAI in Real-Time Cybersecurity Applications

5.1 Latency and Computational Overhead

In real-time cybersecurity systems, explainability implementation frequently results in latency and extra processing overhead. Real-time threat detection systems may become slower because of XAI approaches like LIME and SHAP, which need additional processing to produce explanations. Research has indicated that this additional processing power may affect how quickly attacks are identified and addressed, which is a serious problem in situations where quick decisions are necessary to stop security breaches.

5.2 Balancing Explainability with Accuracy

Accuracy and explainability in real-time cybersecurity continue to be major challenges. Deep learning algorithms are among the many high-performing models that forgo interpretability in favor of greater predictive power. However, performance can suffer when explainability is prioritized, especially in difficult threat detection tasks. Research shows that more easily interpreted, more basic models—like decision trees—frequently miss subtle danger behaviors, which lowers the accuracy of detection.

5.3 Adversarial Attacks on XAI Systems

XAI techniques may unintentionally reveal security holes that adversaries could use to break into the system or launch hostile assaults. Attackers can circumvent detection by gaining insight into the model's decision-making process through a comprehension of the explanations given by XAI. Strong countermeasures are required to secure explainable models because research has shown how adversarial actors may utilize explanations to alter inputs and avoid detection.

5.4 Usability of Explanations for Cybersecurity Professionals

Making sure that the explanations produced by XAI techniques are both technically precise and simple for cybersecurity experts to understand is another difficulty. Many explanations have limitations in real-time circumstances due to their highly technical or abstract nature, especially those derived from

complicated models. Research indicates that there is a need for cybersecurity solutions with user-friendly interfaces and explanations since there is a gap between offering analysts who may not have extensive experience in machine learning precise model insights and making these insights practical.

6. Future Directions in Explainable AI for Cybersecurity

6.1 Advancements in Real-Time Explainable Models

To better serve real-time applications, explainable AI models' speed and efficiency are being improved through recent research. Work is being done to provide high-accuracy, low-latency methods that can quickly and accurately explain threat detection. The goal of this field's innovations is to reduce the computing overhead of XAI techniques while preserving their efficacy, enhancing real-time cybersecurity systems' responsiveness.

6.2 Human-in-the-Loop Approaches

Enhancing explainable models with human feedback appears to be a potential way to improve model performance and explanation relevance. Research is looking into the interactions and improvements that cybersecurity specialists make to explainable systems with the goal of developing better user-centered models that are in line with real-world requirements and decision-making procedures. For security experts, this integration facilitates the transition from technical explanations to practical insights.

6.3 Explainability for Deep Learning Models

Innovative methods are being created to increase cybersecurity deep learning models' explainability without sacrificing accuracy. Attention mechanisms and visualization tools designed for complex neural networks are examples of emerging trends. By shedding more light on the decision-making process of deep learning models, these developments hope to increase public confidence in and comprehension of AI-powered threat detection systems.

6.4 Combining XAI with Other Technologies

The future of XAI in cybersecurity may involve integrating explainable models with technologies like blockchain, homomorphic encryption, and federated learning. These integrations promise to enhance both the security and transparency of AI systems by combining robust privacy-preserving techniques with explainability, creating more secure and interpretable AI solutions for cybersecurity applications.

7. Conclusion

In conclusion, integrating explainable machine learning models into real-time cybersecurity systems significantly enhances both the transparency and effectiveness of threat detection. By combining traditional machine learning techniques with deep learning models, and leveraging explainability methods like LIME, SHAP, and attention mechanisms, this approach provides clear, interpretable insights into model predictions. The inclusion of a user interface that offers visualizations, alerts, and detailed reports allows for more informed and faster decision-making in threat assessment and incident response. This framework not only improves the accuracy and reliability of cybersecurity models but also builds user trust by making the decision-making process understandable. In complex cybersecurity environments, the balance between model performance and transparency is crucial, and this approach achieves that, paving the way for more secure, robust systems.

References

[1] H. I. Halim, “Deep Learning Methods in Web Intrusion Detection : A Systematic Review,” pp. 0–23, 2022.
[2] J. Ables *et al.*, “Creating an Explainable Intrusion Detection System Using Self Organizing Maps,” 2022, doi: 10.1109/SSCI51031.2022.10022255.
[3] F. Yan, S. Wen, S. Nepal, C. Paris, and Y. Xiang, “Explainable machine learning in cybersecurity: A survey,” *Int. J.*

Intell. Syst., vol. 37, no. 12, pp. 12305–12334, 2022, doi: 10.1002/int.23088.
[4] P. Chandre, P. Mahalle, and G. Shinde, "Intrusion prevention system using convolutional neural network for wireless sensor network," *IAES Int. J. Artif. Intell.*, vol. 11, no. 2, pp. 504–515, 2022, doi: 10.11591/ijai.v11.i2.pp504-515.
[5] H. Liu, C. Zhong, A. Alnusair, and S. R. Islam, "FAIXID: A Framework for Enhancing AI Explainability of Intrusion Detection Results Using Data Cleaning Techniques," *J. Netw. Syst. Manag.*, vol. 29, no. 4, pp. 1–30, 2021, doi: 10.1007/s10922-021-09606-8.
[6] S. Patil *et al.*, "Explainable Artificial Intelligence for Intrusion Detection System," *Electron.*, vol. 11, no. 19, 2022, doi: 10.3390/electronics11193079.
[7] S. S. Damre, B. D. Shendkar, N. Kulkarni, P. R. Chandre, and S. Deshmukh, "Smart Healthcare Wearable Device for Early Disease Detection Using Machine Learning," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 4s, pp. 158–166, 2024.
[8] A. Yayla, L. Haghnegahdar, and E. Dincelli, "Explainable Artificial Intelligence for Smart Grid Intrusion Detection Systems," *IT Prof.*, vol. 24, no. 5, pp. 18–24, 2022, doi: 10.1109/MITP.2022.3163731.
[9] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," *IEEE Access*, vol. 10, no. August, pp. 93104–93139, 2022, doi: 10.1109/ACCESS.2022.3204051.
[10] G. R. Pathak and S. H. Patil, "Mathematical Model of Security Framework for Routing Layer Protocol in Wireless Sensor Networks," *Phys. Procedia*, vol. 78, no. December 2015, pp. 579–586, 2016, doi: 10.1016/j.procs.2016.02.121.
[11] F. V. Farahani, K. Fiok, B. Lahijanian, W. Karwowski, and P. K. Douglas, "Explainable AI: A review of applications to neuroimaging data," *Front. Neurosci.*, vol. 16, 2022, doi: 10.3389/fnins.2022.906290.
[12] S. Makubhai, G. R. Pathak, and P. R. Chandre, "Comparative analysis of explainable artificial intelligence models for predicting lung cancer using diverse datasets," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, pp. 1978–1989, 2024, doi: 10.11591/ijai.v13.i2.pp1980-1991.
[13] S. Neupane *et al.*, "Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities," *IEEE Access*, vol. 10, pp. 112392–112415, 2022, doi: 10.1109/ACCESS.2022.3216617.
[14] Bhagyashree Pandurang Gadekar and Dr. Tryambak Hiwarkar, "A Conceptual Modeling Framework to Measure the Effectiveness using ML in Business Analytics," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 2, no. 1, pp. 399–406, 2022, doi: 10.48175/ijarsct-7703.
[15] E. Tcydenova, T. W. Kim, C. Lee, and J. H. Park, "Detection of Adversarial Attacks in AI-Based Intrusion Detection Systems Using Explainable AI," *Human-centric Comput. Inf. Sci.*, vol. 11, 2021, doi: 10.22967/HCIS.2021.11.035.
[16] P. R. Chandre, P. N. Mahalle, and G. R. Shinde, "Machine learning based novel approach for intrusion detection and prevention system: a tool based verification," in *2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN)*, Nov. 2018, pp. 135–140, doi: 10.1109/GCWCN.2018.8668618.
[17] L. Dhanabal and S. P. Shantharajah, "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 6, pp. 446–452, 2015, doi: 10.17148/IJARCCE.2015.4696.
[18] A. Gramegna and P. Giudici, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk," *Front. Artif. Intell.*, vol. 4, no. September, pp. 1–6, 2021, doi: 10.3389/frai.2021.752558.
[19] F. Charmet *et al.*, "Explainable artificial intelligence for cybersecurity: a literature survey," *Ann. des Telecommun. Telecommun.*, vol. 77, no. 11–12, pp. 789–812, 2022, doi: 10.1007/s12243-022-00926-7.
[20] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 35, no. 2, pp. 590–611, 2023, doi: 10.1016/j.jksuci.2023.01.004.
[21] S. Das Guptta, K. T. Shahriar, H. Alqahtani, D. Alsalman, and I. H. Sarker, "Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques," *Ann. Data Sci.*, vol. 11, no. 1, pp. 217–242, 2024, doi: 10.1007/s40745-022-00379-8.
[22] G. Palaniappan, S. Sangeetha, B. Rajendran, Sanjay, S. Goyal, and B. S. Bindhumadhava, "Malicious Domain Detection Using Machine Learning on Domain Name Features, Host-Based Features and Web-Based Features," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 654–661, 2020, doi: 10.1016/j.procs.2020.04.071.
[23] T. Sutter, A. S. Bozkir, B. Gehring, and P. Berlich, "Avoiding the Hook: Influential Factors of Phishing Awareness Training on Click-Rates and a Data-Driven Approach to Predict Email Difficulty Perception," *IEEE Access*, vol. 10, pp. 100540–100565, 2022, doi: 10.1109/ACCESS.2022.3207272.
[24] B. Gadekar and T. Hiwarkar, "International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Proposed Business Improvement Model Utilizing Machine Learning: Enhancing Decision-Making and Performance," *Orig. Res. Pap. Int. J. Intell. Syst. Appl. Eng. IJISAE*, vol. 2024, no. 1s, pp. 557–568, 2023, [Online]. Available: www.ijisae.org.
[25] M. A. Remmide, F. Boumahdi, N. Boustia, C. L. Feknous, and R. Della, "Detection of Phishing URLs Using Temporal Convolutional Network," *Procedia Comput. Sci.*, vol. 212, no. C, pp. 74–82, 2022, doi: 10.1016/j.procs.2022.10.209.

[26] A. Mahalakshmi, N. S. Goud, and G. V. Murthy, "A survey on phishing and it's detection techniques based on support vector method (Svm) and software defined networking(sdn)," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 2, pp. 498–503, 2018.
[27] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," *Front. Comput. Sci.*, vol. 3, no. March, pp. 1–23, 2021, doi: 10.3389/fcomp.2021.563060.
[28] S. Bojjagani, D. R. D. Brabin, and P. V. V. Rao, "PhishPreventer: A Secure Authentication Protocol for Prevention of Phishing Attacks in Mobile Environment with Formal Verification," *Procedia Comput. Sci.*, vol. 171, no. 2019, pp. 1110–1119, 2020, doi: 10.1016/j.procs.2020.04.119.
[29] S. Borde, P. Chandre, P. Mohd Shafi, and A. Jadhav, "Shaikh Ashfaq 2 Sonali A Patil Zero Trust Security Paradigm: A Comprehensive Survey and Research Analysis," *J. Electr. Syst.*, vol. 19, no. 2, pp. 28–37, 2023.
[30] M. Ahsan, K. E. Nygard, R. Gomes, M. M. Chowdhury, N. Rifat, and J. F. Connolly, "Cybersecurity Threats and Their Mitigation Approaches Using Machine Learning—A Review," *J. Cybersecurity Priv.*, vol. 2, no. 3, pp. 527–555, 2022, doi: 10.3390/jcp2030027.
[31] V. Bidve *et al.*, "Use of explainable AI to interpret the results of NLP models for sentimental analysis," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 35, no. 1, pp. 511–519, 2024, doi: 10.11591/ijeecs.v35.i1.pp511-519.
[32] E. S. Shombot, G. Dusserre, R. Bestak, and N. B. Ahmed, "An application for predicting phishing attacks: A case of implementing a support vector machine learning model," *Cyber Secur. Appl.*, vol. 2, no. November 2023, 2024, doi: 10.1016/j.csa.2024.100036.
[33] G. R. Pathak, M. S. G. Premi, and S. H. Patil, "LSSCW: A lightweight security scheme for cluster based Wireless Sensor Network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 10, pp. 448–460, 2019, doi: 10.14569/ijacsa.2019.0101062.
[34] P. M. Dinesh, M. Mukesh, B. Navaneethan, R. S. Sabeenian, M. E. Paramasivam, and A. Manjunathan, "Identification of Phishing Attacks using Machine Learning Algorithm," *E3S Web Conf.*, vol. 399, 2023, doi: 10.1051/e3sconf/202339904010.
[35] J. Kotwal, D. R. Kashyap, and D. S. Pathan, "Agricultural plant diseases identification: From traditional approach to deep learning," *Mater. Today Proc.*, vol. 80, no. xxxx, pp. 344–356, 2023, doi: 10.1016/j.matpr.2023.02.370.
[36] E. D. Frauenstein, "A Framework to Mitigate Phishing Threats A Framework to Mitigate Phishing Threats," no. October, 2014, [Online]. Available: https://www.researchgate.net/publication/267512601.
[37] V. S. Kore, B. A. Tidke, and P. Chandre, "Survey of Image Retrieval Techniques and Algorithms for Image-rich Information Networks," *Int. J. Comput. Appl.*, vol. 112, no. 6, pp. 39–42, 2015, [Online]. Available: https://www.ijcaonline.org/archives/volume112/number6/19674-1244%0Ahttp://research.ijcaonline.org/volume112/number6/pxc3901244.pdf.