Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.073705

REVIEW

Next-Generation Lightweight Explainable AI for Cybersecurity: A Review on Transparency and Real-Time Threat Mitigation

Khulud Salem Alshudukhi[1,*], Sijjad Ali[2], Mamoona Humayun[3,*] and Omar Alruwaili[4]

[1]Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, 72388, Al-Jouf, Saudi Arabia
[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, 518060, China
[3]School of Computing, Engineering and the Built Environment, University of Roehampton, London, SW155PJ, UK
[4]Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka, 72388, Al-Jouf, Saudi Arabia

*Corresponding Authors: Khulud Salem Alshudukhi. Email: ksalshudukhi@ju.edu.sa; Mamoona Humayun. Email: mamoona.humayun@roehampton.ac.uk

Received: 24 September 2025; Accepted: 30 October 2025; Published: 23 December 2025

ABSTRACT: Problem: The integration of Artificial Intelligence (AI) into cybersecurity, while enhancing threat detection, is hampered by the "black box" nature of complex models, eroding trust, accountability, and regulatory compliance. Explainable AI (XAI) aims to resolve this opacity but introduces a critical new vulnerability: the adversarial exploitation of model explanations themselves. Gap: Current research lacks a comprehensive synthesis of this dual role of XAI in cybersecurity—as both a tool for transparency and a potential attack vector. There is a pressing need to systematically analyze the trade-offs between interpretability and security, evaluate defense mechanisms, and outline a path for developing robust, next-generation XAI frameworks. Solution: This review provides a systematic examination of XAI techniques (e.g., SHAP, LIME, Grad-CAM) and their applications in intrusion detection, malware analysis, and fraud prevention. It critically evaluates the security risks posed by XAI, including model inversion and explanation-guided evasion attacks, and assesses corresponding defense strategies such as adversarially robust training, differential privacy, and secure-XAI deployment patterns. Contribution: The primary contributions of this work are: (1) a comparative analysis of XAI methods tailored for cybersecurity contexts; (2) an identification of the critical trade-off between model interpretability and security robustness; (3) a synthesis of defense mechanisms to mitigate XAI-specific vulnerabilities; and (4) a forward-looking perspective proposing future research directions, including quantum-safe XAI, hybrid neuro-symbolic models, and the integration of XAI into Zero Trust Architectures. This review serves as a foundational resource for developing transparent, trustworthy, and resilient AI-driven cybersecurity systems.

KEYWORDS: Explainable AI (XAI); cybersecurity; adversarial robustness; privacy-preserving techniques; regulatory compliance; zero trust architecture

1 Introduction

The rapid evolution of Artificial Intelligence and Machine Learning has revolutionized cybersecurity [1]. These technologies enable advanced threat detection, anomaly identification, and automated response mechanisms [2–6]. However, as AI systems grow increasingly complex, their decision-making processes often become opaque. This is the problem which is usually referred to as the black box problem [7,8]. This invisibility on the part of cybersecurity is very challenging in that the rationale behind AI decisions must be known to be trusted, compliant, and responsive to incidents. Explainable AI has turned out to be a critical

point of concern. It will enhance the appropriateness and responsible nature of AI models without negatively affecting their performance [9–11].

Explainable AI integration in cybersecurity is an important interdependency of interpretability and security. This will make sure that the AI that is used as a security tool is not only helpful but familiar to the human operators [12]. In critical environments such as network security, fraud detection, and malware analysis, stakeholders should have the capacity to authenticate and rationalize AI decisions. Security analysts, compliance officers, and organisational leaders can be part of this group and should avoid algorithmic bias and system integrity [13,14]. AI outputs may bring about mistrust, wrong security policy, and nonconformity with regulations without clear explanations. This is particularly a critical issue in industries where data protection laws are stringent, like the General Data Protection Regulation (GDPR) [15], and the California Consumer Privacy Act (CCPA) [16].

The main concept of explainable AI in cybersecurity is that the user must understand the decision made by the AI. This enhances transparency and helps to build trust. One should also know why an AI makes a certain conclusion. The users should also be in a position to substantiate the validity of the results of the AI. Moreover, it is necessary to eliminate and manage biases, which will make the results unbiased and objective. Furthermore, Fig. 1 discusses the Explainable AI (XAI) architecture of cybersecurity, its main functions, methods, and objectives. According to the diagram, XAI methods like the creation of interpretable models or saliency maps are important in such critical roles as creating transparency and trust, understanding AI decisions, justifying findings, and detecting biases. All these functions are directly related to the main objective of the XAI, that is, to maintain an understandable, precise, and unbiased AI process of reasoning, which ultimately enables human analysts to be sure and trust automated cybersecurity systems. Such a visual overview is efficient to relate practical methods of XAI to the objectives that are general in a security context.

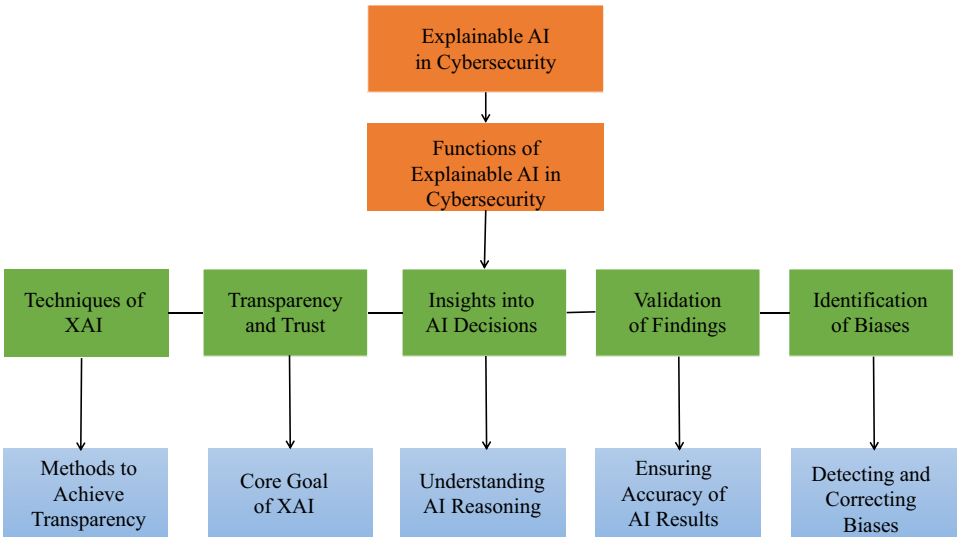

Figure 1: Core principles of explainable AI for cybersecurity

The latest models of cybersecurity consider AI to be the one that detects threats in real-time, analyzes behavior, and predicts security [5,17–19]. Techniques like deep learning, ensemble methods, and reinforcement learning have demonstrated superior accuracy in identifying sophisticated cyber threats [20]. These include zero-day exploits, Advanced Persistent Threats, and polymorphic malware. Yet, these models often function as black boxes, providing little to no insight into their decision-making processes. For example, an

AI system may erroneously flag a legitimate user's activity as malicious without clear justification. This can lead to unnecessary access denials or operational disruptions [21,22].

Explainability in cybersecurity AI is not merely a convenience but a necessity. Security teams must be able to:

- Audit AI decisions to ensure automated threat classifications align with organizational security policies.
- Debug false positives/negatives by identifying why an AI model misclassified a benign file as malware or failed to detect a real attack.
- Comply with regulations by meeting legal and ethical standards that mandate transparency in automated decision-making.
- Enhance human-AI collaboration by allowing security analysts to refine AI models based on interpretable feedback.

The absence of explainability places organizations at risk. They may implement AI systems that are efficient on paper but unreliable in practice. This unreliability can culminate in security breaches, financial losses, and reputational damage.

Highly complex models, such as deep neural networks and ensemble methods, often achieve state-of-the-art performance but are inherently difficult to interpret. Simpler models, such as decision trees or logistic regression, are more transparent but may lack the precision required for modern cyber threats [23,24]. Striking a balance between accuracy and explainability remains a key challenge. Furthermore, cyber attackers can potentially reverse-engineer Explainable AI techniques to understand how AI-based security systems detect threats. This knowledge enables them to craft evasion strategies [25,26]. For example, if an AI model's decision boundaries are exposed, attackers could modify malware to bypass detection while remaining undetected.

The threats in the cyber domain keep changing, and AI models must adapt continuously. Explainability methods must also evolve, ensuring interpretations remain valid with each new model version. Traditional explanation strategies like SHAP and LIME might not perform well in such dynamic settings. Moreover, many Explainable AI approaches are computationally costly, making them impractical for real-time cybersecurity where low-latency response is vital. Therefore, creating scalable interpretability that does not decrease system performance is a persistent research problem.

Methods such as Shapley Additive Explanations [27] and Local Interpretable Model-Agnostic Explanations [28] provide post-hoc explanations for any AI model by approximating its behavior. These techniques are particularly useful in cybersecurity for explaining individual predictions, such as why a specific network packet was flagged as malicious. Some AI models, such as rule-based systems [29], decision trees [30], and linear models [31], are inherently interpretable. Although they may not match the performance of complex models in all scenarios, they are increasingly used in security applications where transparency is prioritized.

A viable alternative is the combination of interpretable models and complex AI, which provides a middle ground. An example is that a deep learning model may detect an anomaly, and a rule-based system would be used to provide an explanation of it in human form. The application of AI dashboards to security operations centres is also becoming popular, and these devices enable analysts to consider AI conclusions [32]. They help fill the divide between the technicality of the explanation and practical security knowledge.

To summarise, Explainable AI is an essential innovation in the area of cybersecurity that allows AI-based protection to be both strong and transparent. Interpretability will lead to more trustworthy systems, high-quality regulatory compliance, and human-AI interaction. Nevertheless, there are still great issues in the area of accuracy, security, and scalability. Future developments of cybersecurity products will be defined by future advances in Explainable AI processes, adversarial resistance, and regulatory systems. Explainability

needs to be introduced in the security systems to develop resilient and stable AI-based security systems in the face of more advanced cyber threats.

1.1 Study Motivation

As demonstrated in this review paper, Explainable AI is highly required in the cybersecurity domain. AI-based security systems are normally used to model black-box models, and this negates trust and transparency. Such a systematic study of the XAI methods places us in a position to have the entire picture of the trade-off between interpretability and security. The ability to interpret them, security issues, and implementation are discussed. Adversarial robustness, regulatory compliance, and moral factors are some of the major challenges as reported in the research. It also talks about the novel opportunities in the field of research that involve quantum-safe XAI and the use of AI in a zero-trust architecture. This work can be used as the foundation for the creation of safe, open, and regulation-compliant AI systems to address cybersecurity. Such systems would be effective and suitable in the contemporary digital world, as they would enable us to debate the primary principles and analyse them in-depth.

1.2 Key Contributions

The following are the main contributions of the current review study:

1. **Comprehensive Analysis of XAI in Cybersecurity:** Presents an exhaustive examination of Explainable AI (XAI) techniques and their potential to strengthen cybersecurity strategies and systems.
2. **Comparative Evaluation of XAI Techniques:** Systematically assesses various XAI methods based on interpretability, security risks, and practical applicability in cybersecurity systems.
3. **Identification of Critical Challenges:** Highlights fundamental issues including the explainability-security trade-off, adversarial robustness, standardization gaps, and ethical concerns.
4. **Regulatory and Compliance Analysis:** Examines the role of explainability in major regulations (GDPR, NIST, ISO 27001) with emphasis on auditability and forensic requirements.
5. **Future Research Directions:** Proposes novel research avenues in XAI applications for zero-trust architectures, post-quantum cybersecurity, and hybrid models integrating symbolic reasoning with deep learning.
6. **Bridging Transparency and Security:** Emphasizes the development of interpretable, adversarially robust AI models that maintain cybersecurity effectiveness.

The review can be used as an advanced source of literature by researchers, policymakers, and cybersecurity professionals who are striving to create powerful, explainable AI-based systems of security.

1.3 Research Questions and Proposed Solutions

This section identifies critical research questions in Explainable AI for cybersecurity and presents corresponding solutions to address these challenges, as detailed in (Table 1). The proposed solutions focus on four key objectives. The first is to enhance model interpretability without compromising security effectiveness. The second involves developing robust defenses against the adversarial exploitation of XAI systems. The third objective is to ensure compliance with evolving regulatory frameworks. Finally, the fourth aims to facilitate XAI integration into next-generation security paradigms, including zero-trust architectures and post-quantum cybersecurity. By systematically addressing these research questions, this study aims to bridge the gap between AI-powered security and operational explainability. This work ultimately advances the development of transparent, robust, and efficient cybersecurity systems.

Table 1: Research questions and proposed solutions in explainable AI for cybersecurity

Research question	Proposed solution
How can Explainable AI (XAI) improve the interpretability of complex cybersecurity models while maintaining high accuracy?	Develop hybrid XAI models that combine symbolic reasoning and deep learning to enhance interpretability without compromising detection performance [33].
What are the key security risks associated with XAI in cybersecurity, and how can they be mitigated?	Implement adversarially robust XAI techniques, such as secure model explanations, differential privacy, and explainability-aware adversarial training [34].
How can XAI be leveraged to improve real-time threat detection and response in AI-driven security systems?	Design real-time, interpretable AI models for Intrusion Detection/Prevention Systems (IDS/IPS) that provide human-understandable threat explanations [35].
How can XAI methods be adapted for compliance with cybersecurity regulations and forensic investigations?	Develop XAI frameworks that align with GDPR, NIST, and ISO 27001 to ensure traceability, transparency, and accountability in AI-driven security decisions [36].
How can Explainable AI enhance Zero Trust Architecture (ZTA) in cybersecurity?	Utilize XAI-driven policy enforcement models to provide transparent and justifiable access control decisions in Zero Trust security frameworks [37].
What is the role of XAI in post-quantum cybersecurity, and how can it be integrated with quantum-safe cryptographic mechanisms?	Design quantum-resistant XAI models that can provide explainable security assurances while being resilient to quantum adversarial threats [38].

1.4 Preliminaries

This section introduces (Table 2), which summarizes the mathematical formulas and abbreviations used throughout this review. The table demonstrates the range of mathematical formalisms employed to describe Explainable AI concepts within cybersecurity. Each entry includes a brief description to clarify the formula's purpose and rationale. Providing this reference allows readers to approach the technical content with greater ease and understanding. This approach bridges the gap between theoretical concepts and their practical implementation in cybersecurity contexts.

Table 2: Mathematical terms and abbreviations

Mathematical term	Description	Abbreviation	Full form
$\mathbf{x} \in \mathbb{R}^d$	Input feature vector in d-dimensional space	XAI	Explainable AI
$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$	Labeled training dataset with N observations	AI	Artificial intelligence
$d_M(\mathbf{x}, \boldsymbol{\mu})$	Mahalanobis distance for anomaly detection	ML	Machine learning

(Continued)

Table 2 (continued)

Mathematical term	Description	Abbreviation	Full form
$\mathcal{E}_\theta, \mathcal{D}_\phi$	Encoder and decoder functions in autoencoders	IDS	Intrusion detection system
ϕ_j (SHAP values)	Shapley value for feature j's contribution	IPS	Intrusion prevention system
$\text{IG}_j(\mathbf{x})$	Integrated Gradients for feature j	GDPR	General data protection regulation
$\mathbf{h}_t^{(LSTM)}$	Hidden state of LSTM at time t	LIME	Local interpretable model-agnostic explanations
δ^*	Optimal adversarial perturbation	SHAP	Shapley additive explanations
$J_f(\mathbf{x})$	Jacobian matrix of model f at input $\mathbf{x}$	Grad-CAM	Gradient-weighted class activation mapping
$\mathcal{L}_c(\theta)$	Compliance loss function with regulatory requirements	DNN	Deep neural network
$f_\theta(\mathbf{x})$	Model function with parameters θ	SOC	Security operations center
λ	Regularization parameter	APT	Advanced persistent threat
$\sigma(\cdot)$	Activation function (sigmoid, ReLU)	CNN	Convolutional neural network
$\mathbb{E}[\cdot]$	Expectation operator	LSTM	Long short-term memory
∇_x	Gradient operator with respect to input x	RBF	Radial basis function
$p(\mathbf{x}\|\mathbf{z})$	Conditional probability distribution	SVM	Support vector machine
$\mathcal{N}(\mu, \Sigma)$	Multivariate Gaussian distribution	CFG	Control flow graph
$\text{ReLU}(\cdot)$	Rectified Linear Unit activation function	CAV	Concept activation vectors
$\tanh(\cdot)$	Hyperbolic tangent activation function	ZTA	Zero trust architecture
$\oplus$	Fusion operator combining representations	NIST	National institute of standards and technology

Explainable AI serves as a crucial cybersecurity tool by making AI decision-making processes transparent. This section provides precise mathematical definitions for key cybersecurity applications, which will be elaborated in subsequent sections.

The article begins with an roadmap of (Fig. 2) with introduction highlighting the importance of XAI in addressing the "black box" problem in AI-driven cybersecurity systems, emphasizing its role in enhancing trust, compliance, and accountability. Section 2 outlines the preliminaries, including mathematical formulations used in the study. Section 3 delves into the fundamentals of XAI, covering definitions, key techniques (e.g., SHAP, LIME, Grad-CAM), and applications in cybersecurity tasks like intrusion detection, malware classification, and phishing detection. Section 4 explores the role of XAI in improving threat detection, regulatory compliance, and forensic investigations. Section 5 discusses the trade-off between interpretability

and security, detailing risks such as adversarial attacks and defense mechanisms like secure-XAI models and differential privacy. Section 6 offers a comparative evaluation of XAI techniques, while Section 7 identifies challenges like performance-interpretability trade-offs and ethical concerns. Finally, Section 8 proposes future research directions, including quantum-safe XAI and hybrid models, concluding with a summary of XAI's transformative potential and inherent risks in cybersecurity.

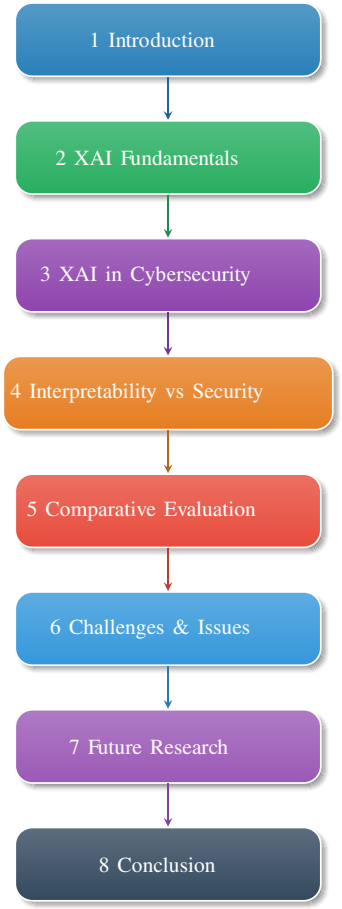

Figure 2: Roadmap of the review article

2 Fundamentals of Explainable AI (XAI)

The field of Explainable AI investigates the principles and practices for making AI models transparent, understandable, and trustworthy for human users. In this section, the key concepts are explained, such as model explainability, fairness, and accountability. These are the factors that provide full profiling and justification of the decision-making processes of AI systems. The most significant concepts are introduced in the following sections.

2.1 Definition and Importance of XAI

Explainable AI is defined as a set of approaches and processes that enhance the interpretability, transparency, and understandability of artificial intelligence models to human users [39,40]. Even classical AI systems, especially the deep learning models, can be black boxes in which even professionals are not aware of how they arrive at their decision-making [7,9,41]. The consequences in the high-stakes fields

like cybersecurity, medicine, and finance are immense. To establish trust, accountability, and regulatory compliance, it is required to have a full-scale picture of how AI systems work internally [42–44].

The use of XAI in the world of cybersecurity is important because of various issues. First, it enhances trust in AI-based security systems by making it possible to have the rationalisation of the decisions and comes in a transparent way that enables security experts to prove the model predictions. Second, it improves incident response by displaying the logic of the actions that make the occurrences attacks or anomalies. Third, it helps to adhere to the regulatory regimes, including the General Data Protection Regulation, which entails disclosure of automated decision-making [45]. Lastly, by exposing the vulnerability of models that are potentially exploited by the adversaries, XAI enhances defensive security, thus creating more robust AI-based security systems [46].

2.2 Key XAI Techniques

XAI practices are designed to achieve transparent, interpretable, and trusting AI system decisions. Such approaches are especially important in industries with high responsibility, such as healthcare, finance, and security, where it is necessary to know the rationale of AI evaluations. The methods of XAI are broadly classified into two groups: intrinsic explainability, which is meant to be explainable, and post-hoc explainability, which explains decisions that have been trained in complex models.

2.2.1 Post-Hoc vs. Intrinsic Explainability

There are two major categories of XAI approaches:

Post-Hoc Explainability: These methods describe model predictions when the model has generated these predictions [47,48]. They do not change the underlying AI model, but instead offer some explanations in terms of visualizations, feature importance measurements, or surrogate models. The use of post-hoc interpretability is especially useful in the context of deep learning models, which are complex and obscure in nature.

Intrinsic Explainability: These techniques are built directly into the AI model's architecture. Models with intrinsic explainability, such as decision trees and linear regression, offer transparent decision-making processes [49,50]. Unlike black-box models, they require no additional interpretation methods as their structure naturally supports human understanding. However, such models may sometimes sacrifice predictive performance compared to more complex deep learning approaches.

The deep learning-based intrusion detection systems are opaque and are often analysed by post-hoc methods. These techniques analyse trained neural networks to offer information on the decision-making processes, without altering the models. The use of explanation tools, including LIME, SHAP, and saliency maps, is commonly used to unveil the influence of features in predictions and, therefore, increase the trustworthiness and transparency of automatic systems. Nevertheless, post-hoc explanations can come with such problems as a lack of consistency and behavioural variance with the real model.

Contrary, explainable models such as decision trees, rule-based systems, and fuzzy logic frameworks are primarily applied in any situation where there is an anomaly detection requirement, and where it is important to provide an explicit explanation of model results. Those models are concerned with interpretability, i.e., cybersecurity experts can be capable of following the outcome of detection to rules or logical models that can be read by humans. This skill would be useful particularly during emergency life situations, which demand accountability and swiftness.

A trade-off between explainability and performance has been one of the dilemmas that have been at the centre of the field. Uncomplicated models are more transparent, and deep learning ones are more accurate.

The current studies aim to fill such a gap by using hybrid methods that combine deep learning with symbolic reasoning or rule-based parts. These hybrid systems are supposed to provide high performance in detection and interpretability. Moreover, the increasing regulatory and moral demands are increasing the pace of using explainable AI in cybersecurity. Deciding on whether to interpret the post-hoc or provide intrinsic explanations is thus largely dependent on the context of the deployment, operational objectives, and tolerance to risk by the organisation.

We compares various Explainable AI methods used in cybersecurity and evaluates their explanatory capabilities. SHAP provides high explainability by precisely measuring feature importance for individual predictions. LIME contributes to transparency by generating local, model-agnostic explanations for each instance. While Decision Trees are interpretable for simple models, their explainability diminishes significantly as complexity increases. Rule-based systems offer structured and transparent logic but often lack flexibility due to their reliance on predefined, rigid rules. Fig. 3 illustrates the fundamental trade-offs between interpretability and complexity in AI-driven cybersecurity solutions.

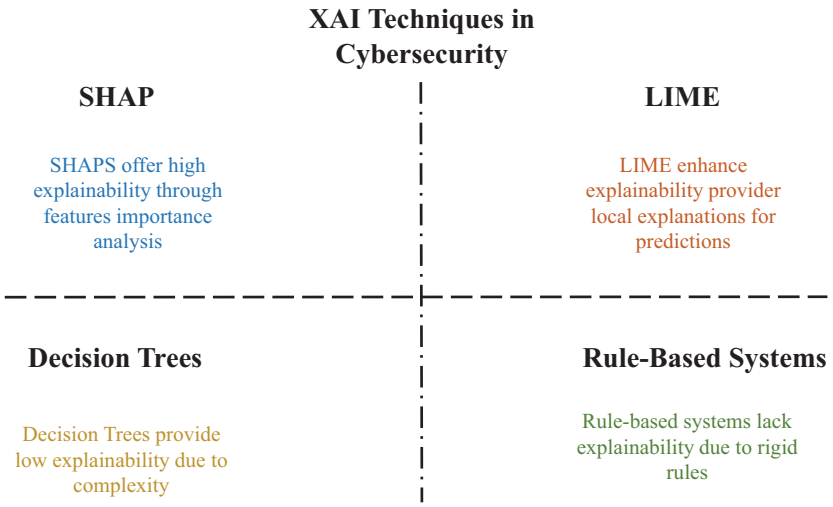

Figure 3: Comparison of XAI techniques in cybersecurity based on explainability. *In a practical case study using the EMBER dataset, SHAP analysis revealed that* `NumberOfSections` *and* `SizeOfCode` *were the most impactful features for a malware classifier, aligning with expert intuition and traditional reverse engineering practices [51]*

2.2.2 Model-Specific vs. Model-Agnostic Approaches

Explainable AI techniques can be further classified into model-specific and model-agnostic approaches, each offering distinct advantages for cybersecurity applications.

Model-Specific Approaches: The techniques have specific machine learning architecture dependencies and use the internal mechanisms to produce the explanations [52]. In particular, Grad-CAM is designed to work with convolutional neural networks, and it allows visualising the important areas in the image-based cybersecurity processes like malware detection [53]. Likewise, decision trees and gradient boosting models have built-in feature importance scores, which are intrinsically interpretable without the need to use additional tools to interpret the models.

Model-Agnostic Approaches: The techniques are also not architecture-specific and can be generalised to any black-box AI system [54]. Well-known ones are SHAP and LIME, which produce explanations using surrogate models or local approximations to the behaviour of the target model [55,56]. Due to this flexibility, they can be used in a wide range of cybersecurity applications, including phishing, network traffic, and others.

The explainability-performance trade-off is a trade-off that is considered carefully when choosing between model-specific or model-agnostic methods. Model-specific models are typically more precise and faithful, and only offer specific model architectures. The model-agnostic approaches are more flexible and are commonly used to analyse the complex deep learning systems within the context of cybersecurity to guarantee the required degrees of trust and transparency among the various AI implementations.

2.2.3 Emerging XAI Paradigms

Beyond the established post-hoc and intrinsic methods, several emerging XAI paradigms offer novel avenues for transparency in cybersecurity:

- **Causal Interpretability:** This approach moves beyond correlational feature importance to model the cause-effect relationships within data [57]. In cybersecurity, this could help distinguish between features that are merely correlated with an attack and those that actually *cause* it. For example, a causal model could determine if a specific sequence of system calls genuinely leads to a privilege escalation, providing deeper, more actionable insights for root cause analysis during forensic investigations.
- **Prototype-Based Explanations:** Methods like ProtoPNet explain classifications by comparing inputs to prototypical examples learned during training [58]. In malware classification, a model could justify its decision by showing, "This file is malicious because it contains code segments similar to these prototypical snippets from the `WannaCry` family." This provides intuitive, case-based reasoning that is natural for security analysts.
- **Concept Activation Vectors (CAV):** CAVs interpret internal neural network layers in terms of human-understandable concepts [59]. A network traffic analyzer could be probed to reveal that it flags traffic as anomalous because it activates neurons associated with the concepts of "covert channel" or "beaconing," directly linking model internals to high-level threat concepts.

These emerging paradigms address limitations of traditional XAI methods by providing more intuitive, causal, and concept-based explanations that align better with human reasoning processes in security analysis. Their integration into cybersecurity workflows represents a promising direction for enhancing both transparency and actionable intelligence.

2.3 Common XAI Methods

Numerous Explainable AI methods have been developed to enhance the interpretability of AI models, each offering distinct advantages for cybersecurity applications.

SHAP: Based on cooperative game theory, SHAP quantifies the contribution of individual input features to model predictions by assigning each feature an importance score. This method provides a unified approach to explain model output. It is particularly valuable in intrusion detection systems, where security analysts require objective, quantitative insights into which network attributes most strongly influenced the classification of an event as an anomaly [60].

Fig. 4 outlines the application of SHAP (SHapley Additive exPlanations) values in a machine learning workflow, illustrating a process that begins with calculating the individual contribution of each feature to a model's prediction, which then facilitates gaining insights into overall feature importance, helps in identifying the key indicators or most influential variables, and finally supports validating the model's results to ensure its predictions are accurate and reliable.

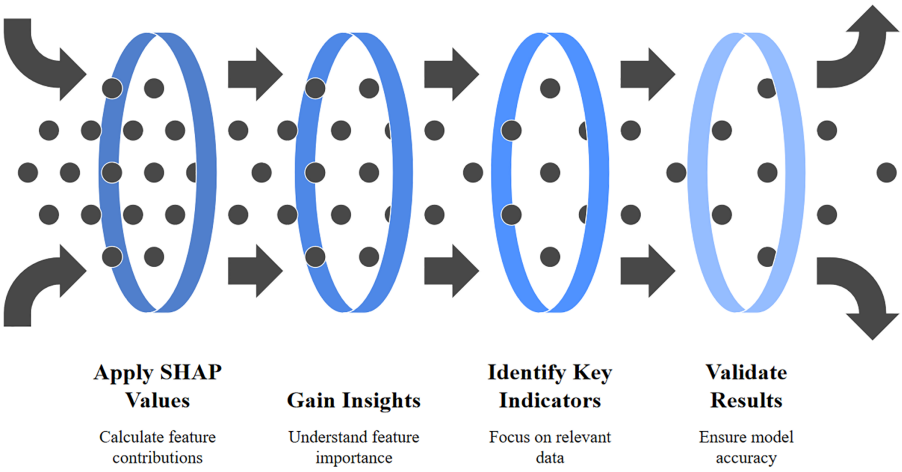

Figure 4: Enhancing intrusion detection with SHAP

LIME: This model-agnostic method creates local, interpretable models to approximate the predictions of complex AI systems [61]. In cybersecurity applications, LIME can explain individual classifications, such as why a specific email was flagged as phishing or a network packet was deemed malicious. These explanations provide security teams with actionable insights, enabling them to respond effectively to potential threats like online fraud or unauthorized network access [62].

Fig. 5 outlines the application of the LIME (Local Interpretable Model-agnostic Explanations) method within the field of cybersecurity, positioning it as a crucial tool for interpreting complex "black-box" machine learning models. It highlights key use cases where LIME provides explainable AI (XAI) insights, such as understanding why an email is flagged as phishing, identifying the features that indicate a network intrusion, or analyzing the malicious behavior of software. The figure also acknowledges that while LIME offers significant benefits by building trust in AI systems, aiding in model debugging, and improving forensic analysis, its practical adoption faces challenges like the computational overhead required to generate explanations and ensuring the explanations themselves are robust and reliable for high-stakes security decisions.

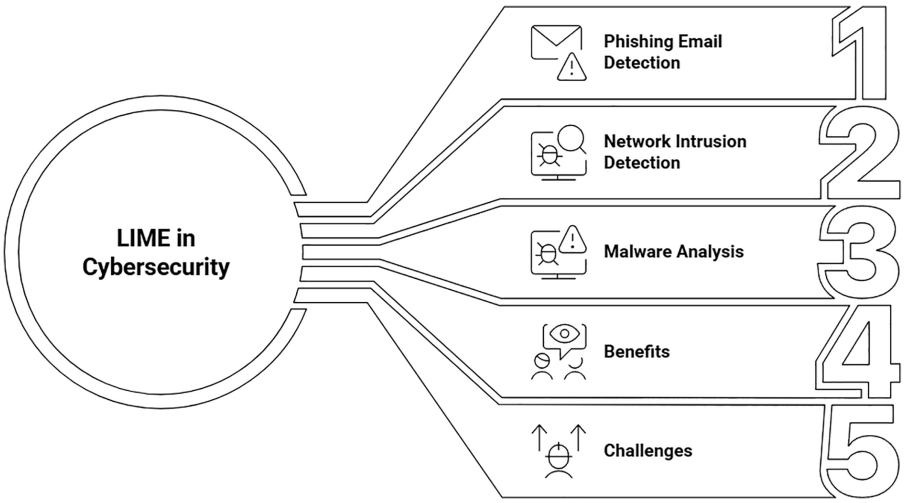

Figure 5: LIME's role in cybersecurity, highlighting key applications such as phishing detection and malware analysis, along with its associated benefits and challenges

Grad-CAM: This model-specific visualization technique interprets Convolutional Neural Networks by highlighting the regions of an input image that most strongly influence the model's decisions [63]. In cybersecurity, Grad-CAM is applied to visualize binary data represented as images, aiding in the classification and analysis of malware families by revealing which patterns the model uses for identification.

Fig. 6 illustrates the key applications of Grad-CAM (Gradient-weighted Class Activation Mapping) in the field of cybersecurity, demonstrating how this explainable AI technique transforms complex machine learning models from opaque "black boxes" into interpretable tools. It shows that by pinpointing the most influential features in data and analyzing how a model makes its predictions, security teams can better understand malicious software behavior, accurately detect anomalous activities, and develop more robust defensive solutions. Ultimately, these deep insights enhance overall threat intelligence, allowing for the creation of more proactive and effective security strategies to counter evolving cyber threats.

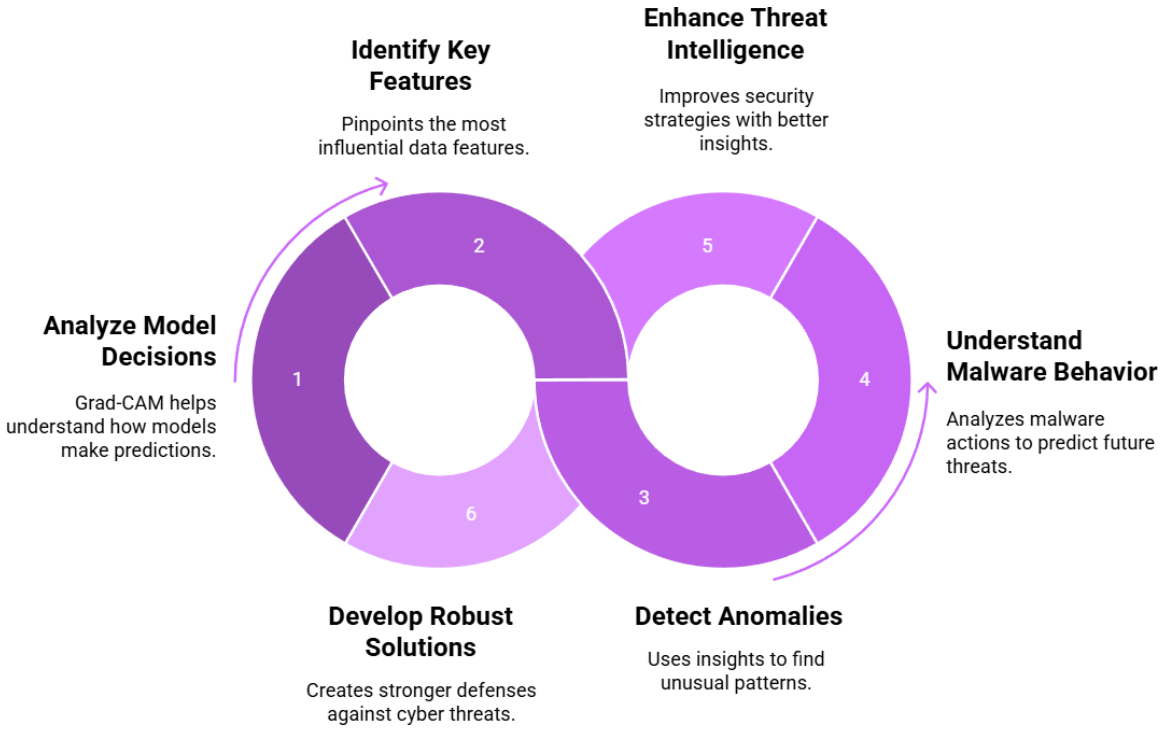

Figure 6: Key applications of grad-CAM in threat analysis

Integrated Gradients: This method quantifies feature importance by calculating the integral of gradients from a baseline input to the actual input [64]. It proves particularly effective for interpreting deep learning models applied to security logs, user behavior patterns, and network traffic anomalies.

Counterfactual Explanations: These explanations identify minimal modifications to inputs that would alter a model's prediction [65]. This method is extremely useful in learning how the slight alterations in the network traffic patterns or software actions might modify the security classifications of AI.

Such Explainable AI techniques are strategic to the cybersecurity industry since they will assist cybersecurity experts in enhancing the transparency and responsibility of AI-based security systems. The complex models are transformed into the trusted components of the working security infrastructure by this method.

2.4 Applications of XAI in Cybersecurity

2.4.1 Intrusion Detection Systems (IDS)

Modern Intrusion Detection Systems (IDS) have evolved to leverage sophisticated machine learning models that process network traffic data [66]. This data is represented mathematically as high-dimensional vectors $\mathbf{x} \in \mathbb{R}^d$, where the dimension d corresponds to the number of extracted features, such as packet size, protocol type, and flow duration. The foundational element for training these models is a labeled dataset, denoted as $\mathcal{D} = (\mathbf{x}_i, y_i)i = 1^N$, which contains N individual observations. Each observation is a pair consisting of a feature vector $\mathbf{x}_i$ and a corresponding binary label $y_i \in [0, 1]$, indicating whether the network traffic is classified as normal or anomalous.

A classical yet powerful statistical approach for identifying outliers in such a multivariate setting is based on the generalized Mahalanobis distance. This distance metric is superior to the Euclidean distance as it accounts for the correlations between different features and the inherent variance of the dataset. It is formally defined by the equation:

$$d_M(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{1}$$

In this equation, $\boldsymbol{\mu} \in \mathbb{R}^d$ represents the sample mean vector of the normal traffic, and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is the covariance matrix, which captures the relationships and scales between all pairs of features. A high value of $d_M(\mathbf{x}, \boldsymbol{\mu})$ indicates that the sample $\mathbf{x}$ is a statistical outlier. However, a critical vulnerability of this method is that the estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be severely distorted if the training data itself contains anomalies. To mitigate this, robust estimation techniques like the Minimum Covariance Determinant (MCD) are employed. The MCD method seeks to find a subset of the data that is most tightly clustered, thereby excluding potential outliers from the estimation process. This is formulated as the optimization problem:

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \arg\min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \det(\boldsymbol{\Sigma}) \quad \text{s.t.} \quad \sum_{i=1}^{h} \mathbb{I}(d_M(\mathbf{x}_i, \boldsymbol{\mu}) \leq \gamma) \geq \tau N \tag{2}$$

Here, the goal is to find the mean and covariance that minimize the determinant of the covariance matrix—a measure of the data's volume—while constraining the solution to a subset of size h where at least τN points (with τ being a coverage threshold) have a Mahalanobis distance below a certain γ.

Complementing these statistical methods, deep learning architectures, particularly autoencoders, have become a cornerstone of modern anomaly detection [67]. An autoencoder is a neural network designed to learn a compressed, lower-dimensional representation of the input data. It consists of two primary components: an encoder function $\mathcal{E}\theta : \mathbb{R}^d \to \mathbb{R}^k$ that maps the high-dimensional input $\mathbf{x}$ to a latent code in a smaller space $\mathbb{R}^k$ (where $k < d$), and a decoder function $\mathcal{D}\phi : \mathbb{R}^k \to \mathbb{R}^d$ that attempts to reconstruct the original input from this latent code. The model is trained by minimizing a reconstruction loss function, which measures the discrepancy between the original input and its reconstructed version:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}\mathbf{x} \sim p\text{data}[|\mathbf{x} - \mathcal{D}\phi(\mathcal{E}\theta(\mathbf{x}))|2^2] + \lambda R(\theta, \phi) \tag{3}$$

The first term is the mean squared error reconstruction loss, forcing the autoencoder to learn the salient features of the normal traffic. The second term, $\lambda R(\theta, \phi)$, is a regularization penalty weighted by λ, which helps prevent overfitting by constraining the model's parameters. The underlying principle is that an autoencoder trained predominantly on normal traffic will learn to reconstruct it accurately but will perform poorly on anomalous data it has never encountered. Consequently, the anomaly score for a new sample $\mathbf{x}$ is

derived from the magnitude of the reconstruction error, potentially weighted by the importance of different features:

$$s(\mathbf{x}) = |\mathbf{x} - \mathcal{D}\phi(\mathcal{E}\theta(\mathbf{x}))|\mathbf{\Lambda}^2 \tag{4}$$

here, $\mathbf{\Lambda}$ is a diagonal weight matrix that can be used to assign different costs to the reconstruction error of different features, allowing for a more nuanced anomaly score.

Given the complexity and "black-box" nature of both statistical and deep learning models, explainability methods are essential for interpreting their decisions and building trust in the IDS. Among the most theoretically sound approaches are Shapley values, which originate from cooperative game theory and fairly distribute the "payout" (the model's prediction) among the "players" (the input features). The exact computation of the Shapley value for a feature j is given by:

$$\phi_j = \frac{1}{d!} \sum_{\pi \in \Pi_d} [f(\mathbf{x}\pi < j \cup x_j) - f(\mathbf{x}\pi < j)] \tag{5}$$

This equation considers every possible permutation π of the features. For each permutation, it calculates the marginal contribution of feature j by taking the difference between the model's output when including j and the output when only the features preceding it in the permutation are used. The final value ϕ_j is the average of these marginal contributions over all possible permutations. However, this method is computationally prohibitive for high-dimensional data. For deep models, the DeepSHAP approximation provides a more feasible alternative by leveraging the chain rule and expected gradients:

$$\phi_j \approx \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(\mathbf{x})} \left[\frac{\partial f(\mathbf{z})}{\partial z_j} \cdot (z_j - \mathbb{E}[z_j]) \right] \tag{6}$$

This approximation estimates the feature importance by considering the product of the feature's gradient (its local sensitivity) and its deviation from a baseline value, averaged over a local distribution of inputs.

Another prominent explainability technique for deep networks is Layer-wise Relevance Propagation (LRP) [68]. Instead of analyzing the input gradients, LRP operates by distributing the model's output score backward through the network layers, from the output back to the input, assigning a relevance score R_i to each neuron. The core propagation rule for a single layer is:

$$R_i^{(l)} = \sum_j \frac{w_{ij} R_j^{(l+1)}}{\sum_k w_{ik} a_k^{(l)} + \epsilon} \tag{7}$$

In this rule, $R_i^{(l)}$ is the relevance of neuron i in layer l, w_{ij} are the weights connecting it to neuron j in the next layer $(l+1)$, $a_k^{(l)}$ are the activations, and ϵ is a small stabilizing constant that prevents division by zero. This process ensures that the total relevance is conserved across layers, ultimately producing a heatmap on the input features that highlights which ones—such as a specific protocol type or an unusual packet size—were most influential in the model's final decision to flag the traffic as anomalous.

Fig. 7 presents a taxonomy of techniques used in anomaly-based intrusion detection systems, categorizing them from simple to complex. It outlines core statistical methods for identifying outliers, such as Mahalanobis Distance, which detects anomalies by considering feature correlations, and Minimum Covariance Determinant, a robust estimator that excludes outliers to define a "normal" baseline. For more complex, non-linear patterns, it highlights Autoencoders, which learn to reconstruct normal network traffic

and flag instances with high reconstruction errors as potential attacks. Finally, the figure includes advanced model interpretation techniques like Shapley Values and Layer-wise Relevance Propagation, which do not detect intrusions themselves but are crucial for explaining the predictions of complex models by fairly attributing importance to the input features that influenced the alert.

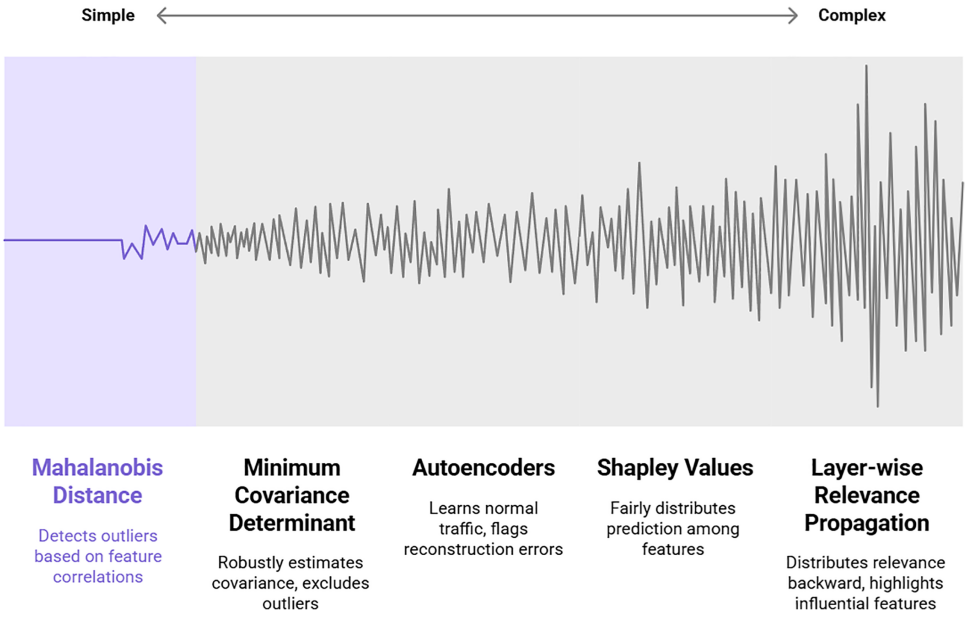

Figure 7: A taxonomy of intrusion detection techniques, ranging from simple to complex

2.4.2 Malware Classification

The modern paradigm of automated malware analysis hinges on the transformation of raw executable files into structured [69], numerical feature vectors $\mathbf{x} \in \mathbb{R}^d$ that machine learning models can process. This transformation is achieved through a multi-faceted feature extraction process. The first common approach involves byte-level n-gram analysis, which treats the executable as a sequence of bytes $b_1, b_2, \ldots, b_L$. From this sequence, statistical features are computed, such as the frequency distribution of each byte value. This is formally captured by the equation:

$$\psi_k = \frac{1}{L} \sum_{i=1}^{L} \mathbb{I}(b_i = k), \quad k \in 0, \ldots, 255 \tag{8}$$

Here, ψ_k represents the normalized frequency of byte value k in the file, and $\mathbb{I}$ is the indicator function. This creates a 256-dimensional vector that captures the fundamental statistical signature of the binary. A more sophisticated structural analysis involves extracting features from the Control Flow Graph (CFG), which represents the program's execution logic as a directed graph where nodes are basic blocks of code and edges represent possible jumps or calls. The graph's adjacency matrix $\mathbf{A}$ is defined by:

$$\mathbf{A}_{ij} = \mathbb{I}(\text{block } i \rightarrow \text{block } j) \tag{9}$$

where $\mathbb{I}$ indicates a connection from block i to block j. To condense this structural information into a fixed-size vector, spectral graph theory is applied, and the eigenvalues λ_k of the adjacency matrix $\mathbf{A}$ are computed and used as features, providing a fingerprint of the program's control flow complexity.

Once feature vectors are constructed, they are fed into classification models. A powerful and traditional model is the Support Vector Machine (SVM) with a non-linear Radial Basis Function (RBF) kernel. The SVM decision function for a new sample $\mathbf{x}$ is given by:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i \exp(-\gamma|\mathbf{x} - \mathbf{x}_i|^2) + b \tag{10}$$

In this formulation, the α_i are learned weights, y_i are the training labels (malicious or benign), and the kernel function $\exp(-\gamma|\mathbf{x} - \mathbf{x}_i|^2)$ measures the similarity between the new sample $\mathbf{x}$ and each support vector $\mathbf{x}_i$ from the training set. For more complex and sequential patterns, deep learning architectures that combine Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks are employed [70]. The CNN component scans local regions of the feature vector or byte sequence to detect localized, invariant patterns. The feature map for a window of size w at position t is computed as:

$$\mathbf{h}t^{(\mathrm{CNN})} = \mathrm{ReLU}(\mathbf{W}^{(\mathrm{CNN})} * \mathbf{x}t : t + w + \mathbf{b}^{(\mathrm{CNN})}) \tag{11}$$

where $*$ denotes the convolution operation. The outputs of the CNN are then fed into an LSTM network to model long-range temporal dependencies in the code structure. The complex, gated mechanism of the LSTM, which controls the flow of information, can be summarized by its core update for the cell state $\mathbf{c}_t$ and hidden state $\mathbf{h}_t$:

$$\begin{aligned}\mathbf{h}_t^{(\mathrm{LSTM})} &= \sigma\left(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f\right) \odot \mathbf{c}_{t-1} \\ &\quad + \sigma\left(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i\right) \odot \tanh\left(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_c\right)\end{aligned} \tag{12}$$

This equation shows how the forget gate, input gate, and candidate cell state work together to selectively remember or forget information over long sequences, making it highly effective for analyzing the ordered sequence of Application Programming Interface (API) calls or instruction blocks.

Given the critical consequences of misclassifying malware, interpretability methods are not a luxury but a necessity. Integrated Gradients is a prominent method that attributes the prediction of a model f to each input feature j. It works by integrating the model's gradients along a straight-line path from a baseline input $\mathbf{x}'$ (which represents a "neutral" input, like a blank file) to the actual input $\mathbf{x}$. The attribution for feature j is calculated as:

$$\mathrm{IG}j(\mathbf{x}) = (x_j - x'_j) \times \int \alpha = 0^1 \frac{\partial f(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_j} d\alpha \tag{13}$$

This provides a complete attribution that satisfies desirable axiomatic properties. Another powerful interpretability approach is counterfactual explanation, which answers the question: "What would need to change for this malicious sample to be classified as benign?" This is framed as an optimization problem:

$$\min_{\mathbf{x}'} |f(\mathbf{x}') - y'| + \lambda|\mathbf{x}' - \mathbf{x}|_1 \tag{14}$$

The goal is to find a new sample $\mathbf{x}'$ that is as close as possible to the original malware $\mathbf{x}$ (minimizing the L_1 norm) but is now classified as the desired label y' (e.g., benign). The sparsity-inducing L_1 norm encourages changes to only a few critical features, clearly showing an analyst the minimal modifications required to evade detection.

These mathematical formulations for feature extraction, classification, and interpretation collectively show how Explainable AI (XAI) creates the dual foundation for both high-performance detection and the

crucial ability to explain seemingly obscure decision-making in cybersecurity systems, a point strongly emphasized in the contemporary literature ([23,24,71]).

Fig. 8 outlines a comprehensive machine learning pipeline for malware detection and interpretation. The process begins with Feature Extraction, where raw malware files are transformed into numerical vectors through Structural Analysis of elements like the Control Flow Graph and Byte-Level Analysis to compute statistical features. These features are then fed into Classification Models, such as an SVM with an RBF kernel or a hybrid CNN-LSTM Integration model, to identify complex patterns and distinguish between malicious and benign software. Crucially, the pipeline emphasizes transparency by employing Interpretability Methods like Integrated Gradients, which attributes the model's prediction to specific input features, and Counterfactual Explanations, which identify the minimal changes needed for a malicious file to be classified as benign, thereby providing actionable insights into the model's decision-making process.

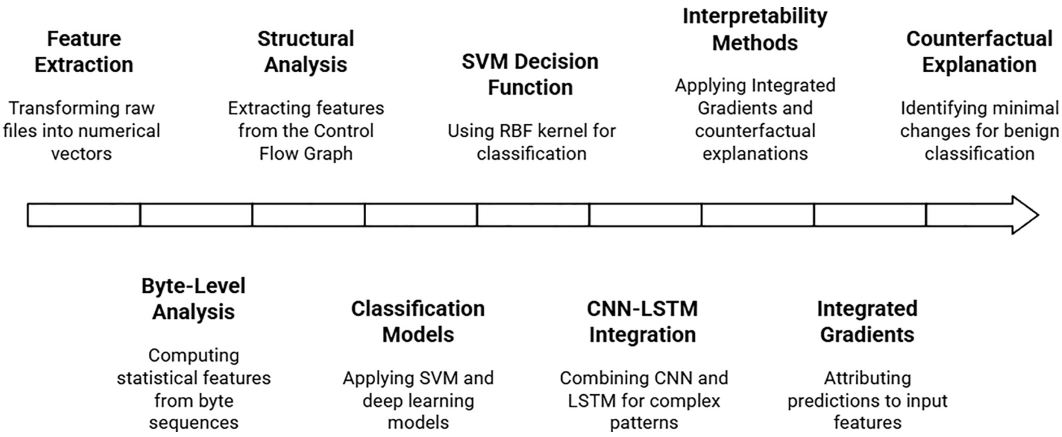

Figure 8: Architecture of an interpretable malware detection system using hybrid deep learning models

2.4.3 Phishing Detection

Modern phishing detection systems represent a sophisticated fusion of feature engineering, machine learning, and explainable AI (XAI) to combat evolving cyber threats, as highlighted in recent literature ([72,73]). The foundational step in this process is transforming a raw email into a numerical feature vector $\mathbf{x} \in \mathbb{R}^d$ that a model can process. The components x_i of this vector are engineered to capture a wide array of deceptive characteristics from both the email's text and its embedded URLs.

Textual analysis is crucial, as phishing emails often exhibit specific linguistic patterns. A sophisticated textual feature can be derived by modeling the email's text $\mathbf{t}$ using an information-theoretic approach combined with image processing techniques for robustness. This is formalized by the equation:

$$H(\mathbf{t}) = -\sum_{k=1}^{K} p(t_k) \log p(t_k) + \lambda \sum_{n=1}^{N} \frac{|\nabla t_n|^2}{1 + |\nabla t_n|^2} \tag{15}$$

Here, the first term, $-\sum_{k=1}^{K} p(t_k) \log p(t_k)$, is the Shannon entropy, which quantifies the randomness and unpredictability in the character or word distribution—often higher in spammy or obfuscated text. The second term, $\lambda \sum_{n=1}^{N} \frac{|\nabla t_n|^2}{1+|\nabla t_n|^2}$, acts as an edge-preserving regularizer, adapted from image analysis. In this context, it helps to smooth noisy textual data while preserving sharp, significant transitions (or "edges") that might correspond to abrupt changes in topic or the insertion of suspicious phrases, with the parameter λ controlling the strength of this smoothing.

URL analysis is another critical pillar of feature extraction. Phishers often use deceptive URLs that resemble legitimate ones through slight character substitutions, insertions, or deletions. The Levenshtein distance D_{Lev} is a powerful metric to quantify the similarity between a given URL string s_1 and a known legitimate domain s_2. It is defined recursively as the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into the other. This is formally expressed by the dynamic programming relation:

$$\begin{aligned} D_{\text{Lev}}(s_1, s_2) = \min\{&|s_1| + |s_2|, \\ &\max(|s_1|, |s_2|) - \text{LCS}(s_1, s_2), \\ &D_{\text{Lev}}(s_1[1:], s_2) + 1, \\ &D_{\text{Lev}}(s_1, s_2[1:]) + 1, \\ &D_{\text{Lev}}(s_1[1:], s_2[1:]) + \mathbb{I}(s_1[0] \neq s_2[0])\} \end{aligned} \tag{16}$$

A small Levenshtein distance to a popular brand's domain is a strong indicator of a typosquatting attack.

Once these diverse features are assembled into a vector $\mathbf{x}$, they are processed by a detection model $f_\theta : \mathbb{R}^d \to [0,1]$, which outputs a probability score indicating the likelihood that the email is a phishing attempt. A common and effective architecture is a deep neural network, which can model complex, non-linear interactions between features. The network's output is given by:

$$f_\theta(\mathbf{x}) = \sigma\left(\sum_{k=1}^{K} w_k \text{ReLU}(\mathbf{W}k\mathbf{x} + \mathbf{b}k)\right) \tag{17}$$

This equation describes a multi-layer perceptron where the input is transformed through one or more hidden layers using a ReLU activation function $\text{ReLU}(\mathbf{W}k\mathbf{x} + \mathbf{b}k)$ to introduce non-linearity. The outputs of the final hidden layer are then aggregated and passed through a sigmoid activation function $\sigma(\cdot)$ to produce a probability. This model is trained by minimizing a regularized cross-entropy loss function:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum i = 1^N [y_i \log f\theta(\mathbf{x}i) + (1 - y_i) \log(1 - f\theta(\mathbf{x}i))] + \lambda|\theta|2,1 \tag{18}$$

The first part of the loss is the standard cross-entropy, which measures the discrepancy between the model's predictions and the true labels y_i. The regularization term $|\theta|2,1 = \sum k = 1^K \sqrt{\sum_{j=1}^{d} w_{kj}^2}$ is a group Lasso penalty. This penalty promotes group sparsity, meaning it can drive all the weights connected to a single, less informative feature to zero simultaneously, thereby performing an automatic form of feature selection and improving model interpretability and generalization.

To build trust and provide actionable insights, explainability is paramount. An extension of the SHAP framework that incorporates second-order effects provides more accurate feature attributions. The attribution for feature j is given by:

$$\phi_j = \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\mathbf{x})}\left[\frac{\partial f(\mathbf{z})}{\partial z_j} \cdot (z_j - \mathbb{E}[z_j])\right] + \frac{1}{2}\mathbb{E}\left[\frac{\partial^2 f(\mathbf{z})}{\partial z_j^2} \cdot \text{Var}(z_j)\right] \tag{19}$$

This formula extends beyond the first-order approximation by adding a second-term that accounts for the curvature of the model function f with respect to feature j. This is particularly important for capturing interaction effects in the complex decision boundary of a deep neural network, leading to more faithful explanations.

Finally, for a robust defense, phishing detection systems often incorporate an anomaly detection module to flag novel attacks that do not resemble known patterns. The Mahalanobis distance is a key metric for this purpose, measuring how far a sample $\mathbf{x}$ is from the center of the normal data distribution in terms of standard deviations:

$$d_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}})} \tag{20}$$

To ensure that the estimates for the mean $\hat{\boldsymbol{\mu}}$ and covariance $\hat{\boldsymbol{\Sigma}}$ are not skewed by potential phishing samples in the training data, they are computed using the Minimum Covariance Determinant (MCD) estimator. This robust estimator is defined by the optimization problem:

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \arg\min_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} \det(\boldsymbol{\Sigma}) \quad \text{s.t.} \quad \sum_{i=1}^{h} \mathbb{I}(d_M(\mathbf{x}i) \leq \chi^2 d, 0.975) \geq \tau N \tag{21}$$

This formulation finds the mean and covariance that minimize the determinant (a measure of volume) based on a clean subset of the data of size h, where at least τN points have a Mahalanobis distance within the 97.5^{th} percentile of the chi-squared distribution. This ensures the model's understanding of "normal" is derived from the most typical, uncontaminated samples, making it highly resilient to poisoning and effective at identifying novel phishing anomalies.

Fig. 9 outlines the architecture of a modern, explainable phishing detection system that integrates multiple analytical approaches. The process begins with Feature Engineering, where both Textual Analysis (e.g., email body content) and URL Analysis (e.g., suspicious domain characteristics) are used to extract relevant signals. These features are then processed by Machine Learning models, including Deep Neural Networks optimized with Regularized Loss Functions to prevent overfitting and improve generalization. To handle novel and evolving threats, an Anomaly Detection module employs techniques like the Mahalanobis Distance with a robust MCD Estimator to identify outliers. Crucially, the entire system is made interpretable using Explainable AI (XAI) principles, specifically the SHAP framework, which helps reveal the contribution of individual features and even uncovers complex Second-Order Effects (feature interactions) behind each prediction, building trust and allowing for deeper model analysis.

2.4.4 User Authentication and Fraud Detection

Modern behavioral authentication systems represent a paradigm shift from static credential checks to continuous, dynamic verification by modeling user behavior as a rich temporal process [74]. This process is represented as a sequence of behavioral vectors $\{\mathbf{x}_t\}_{t=1}^{T}$, where each $\mathbf{x}_t$ captures a snapshot of user activity at time t, such as keystroke dynamics, mouse movements, or application usage. The core of this modeling approach lies in treating the user's current behavior as being probabilistically dependent on their past actions. This is formally represented by the conditional distribution:

$$p(\mathbf{x}_t | \mathbf{x}_{1:t-1}) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \tag{22}$$

This equation states that the behavioral vector $\mathbf{x}_t$ at time t, given the history $\mathbf{x}_{1:t-1}$, is assumed to follow a multivariate Gaussian distribution. The mean $\boldsymbol{\mu}_t$ represents the expected behavior, and the covariance matrix $\boldsymbol{\Sigma}_t$ captures the inherent variability and correlations between different behavioral features at that moment. Crucially, these parameters are not static; they evolve over time to adapt to the user's changing habits. The evolution of the mean vector is governed by a linear dynamical system:

$$\boldsymbol{\mu}_t = \mathbf{A}\boldsymbol{\mu}_{t-1} + \mathbf{B}\mathbf{u}_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathbf{Q}) \tag{23}$$

here, $\mathbf{A}$ is the state transition matrix that defines how the previous state $\boldsymbol{\mu}_{t-1}$ influences the current one, $\mathbf{B}\mathbf{u}_t$ incorporates any external control inputs (e.g., time of day triggering different behavior patterns), and ϵ_t is a Gaussian process noise term with covariance $\mathbf{Q}$ that accounts for unpredictable variations.

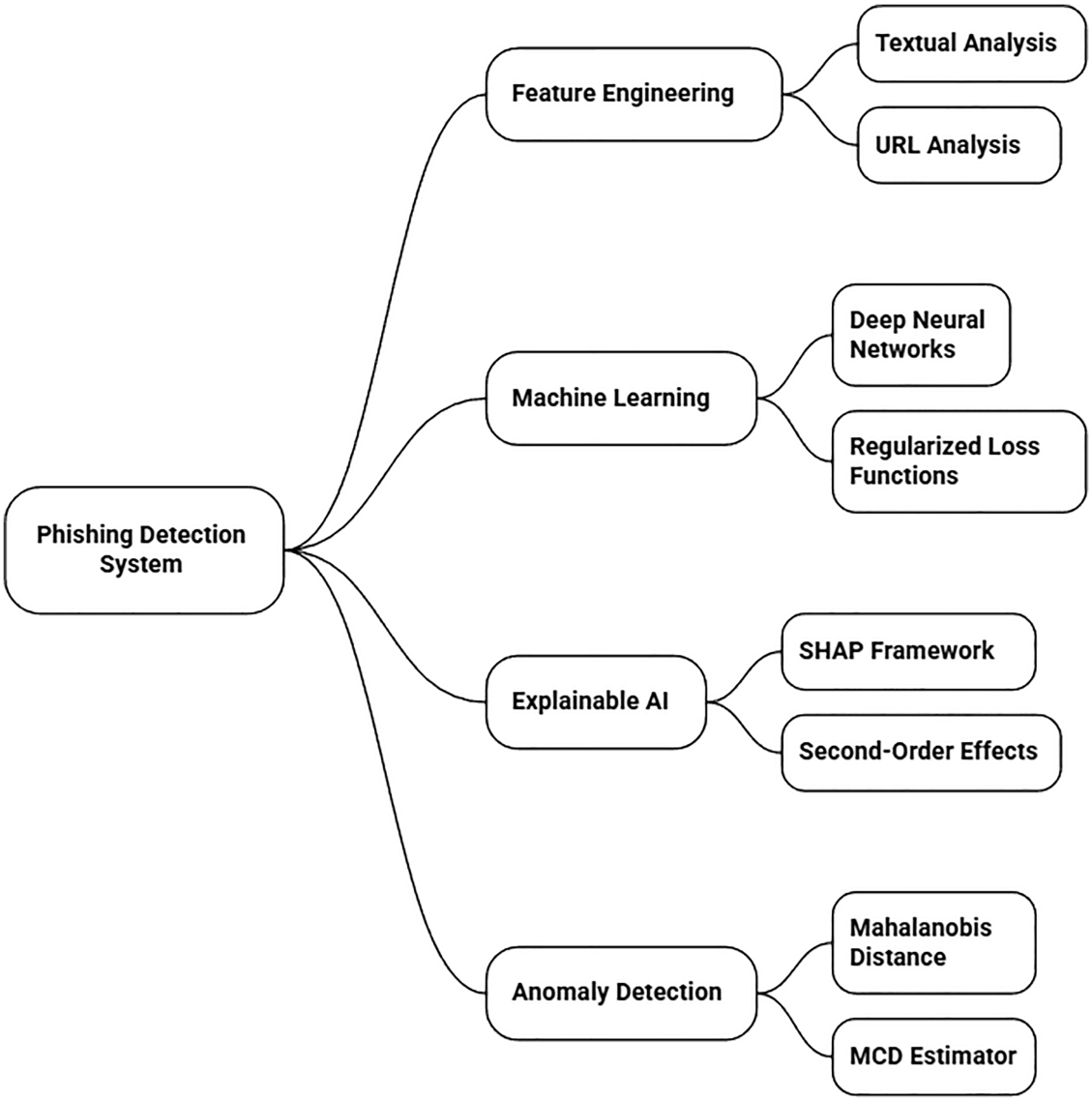

Figure 9: A unified framework for phishing detection combining feature engineering, machine learning, and explainable AI

To classify whether a sequence of behavior originates from the legitimate user or a fraudster, a sophisticated classifier is employed that synthesizes both static and dynamic information. The classifier's architecture is defined as:

$$f(\mathbf{x}) = \mathrm{MLP}(\mathbf{x}_{\mathrm{static}}) \oplus \mathrm{LSTM}(\mathbf{x}_{1:T}^{\mathrm{dyn}}) \tag{24}$$

In this formulation, a Multi-Layer Perceptron (MLP) processes static features $\mathbf{x}_{\mathrm{static}}$ (e.g., user device type, installed software), while a Long Short-Term Memory (LSTM) network processes the sequential dynamic features $\mathbf{x}_{1:T}^{\mathrm{dyn}}$ (e.g., the recent sequence of mouse clicks). The operator $\oplus$ denotes the fusion of these

two distinct representations, often through concatenation or a weighted combination, allowing the model to make a holistic decision. The final decision boundary for flagging fraud is adaptive, scaling with the model's uncertainty:

$$\mathcal{D} = \{\mathbf{x} | f(\mathbf{x}) = \tau + \alpha\sqrt{\mathrm{Var}(f(\mathbf{x}))}\} \tag{25}$$

here, a sample is classified as anomalous if the classifier's output score $f(\mathbf{x})$ exceeds a threshold that is composed of a base value τ plus a term proportional to the standard deviation $\sqrt{\mathrm{Var}(f(\mathbf{x}))}$. This adaptive threshold automatically becomes more conservative when the model's predictions are uncertain, reducing false alarms during periods of high behavioral variability.

Given the high-stakes nature of authentication, explainability is crucial. To attribute the model's fraud decision to specific behavioral features at specific times, Shapley values are extended to handle temporal dependencies. The Shapley value for feature j at time t is given by:

$$\phi_{j,t} = \sum_{\mathcal{T} \subseteq \mathcal{T}_0 \setminus \{t\}} \frac{|\mathcal{T}|!(T - |\mathcal{T}| - 1)!}{T!} \left[f(\mathbf{x}_{\mathcal{T} \cup \{t\}}) - f(\mathbf{x}_{\mathcal{T}}) \right] \tag{26}$$

This complex-looking formula has an intuitive interpretation: it fairly distributes the classifier's "payout" (the fraud score) among all feature-time pairs by considering every possible subsequence $\mathcal{T}$ of the total time steps $\mathcal{T}_0 = \{1, \ldots, T\}$. For each subsequence, it calculates the marginal contribution of adding the feature value at time t, $[\mathbf{x}_j]_t$, and then averages these contributions over all possible subsequences. This provides a rigorous assessment of which action, at which moment, was most influential in triggering a fraud alert. A complementary, gradient-based attribution method is also used, defined by the path integral:

$$\psi_{j,t} = \int_0^1 \frac{\partial f(\mathbf{x}_{1:t-1} + \alpha(\mathbf{x}_t - \mathbf{x}_{t-1}))}{\partial x_{j,t}} d\alpha \tag{27}$$

This method, known as Integrated Gradients, computes the attribution for feature $x_{j,t}$ by integrating the model's gradients along a straight-line path from a baseline state $\mathbf{x}_{1:t-1}$ (representing the user's past, normal behavior) to the current input state $\mathbf{x}_t$. This captures the sensitivity of the fraud score to changes in each feature.

Finally, to maintain accuracy in the face of evolving user behavior and novel attack vectors, the system continuously updates itself via online learning. The model parameters θ are updated sequentially as new data $(\mathbf{x}_t, y_t)$ arrives:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \ell(f_\theta(\mathbf{x}_t), y_t) + \lambda|\theta_t - \theta_{t-1}|^2 \tag{28}$$

This update rule states that the parameters at time $t+1$ are obtained by taking a step from the current parameters θ_t in the direction of the negative gradient of the loss function ℓ, scaled by a learning rate η_t. Critically, a regularization term $\lambda|\theta_t - \theta_{t-1}|^2$ is added, which penalizes large, abrupt changes in the parameters. This ensures that the model adapts to genuine long-term drift in user behavior without "catastrophically forgetting" previously learned patterns or being overly swayed by a single, potentially mislabeled data point.

Collectively, these mathematical formulations for temporal modeling, hybrid classification, explainable attribution, and stable online learning provide the rigorous foundations necessary for building effective, transparent, and adaptive explainable AI systems for phishing detection and fraud prevention, as emphasized in contemporary security research [75,76].

Fig. 10 outlines the architecture of a modern behavioral authentication system, which verifies a user's identity based on their unique behavioral patterns. The process begins with Behavioral Vectors, which are numerical representations of user actions like typing rhythm or mouse movements. Temporal Modeling techniques, such as Long Short-Term Memory (LSTM) networks, analyze the sequence and timing of these actions to create a dynamic user profile. To ensure transparency and trust, Explainable Attribution methods, including Shapley Values and Integrated Gradients, are used to pinpoint which specific behaviors were most influential in the authentication decision. Finally, the system employs Stable Online Learning, which allows it to securely and gradually adapt to a user's evolving behavior over time through parameter updates and regularization, maintaining both security and performance without forgetting previous knowledge.

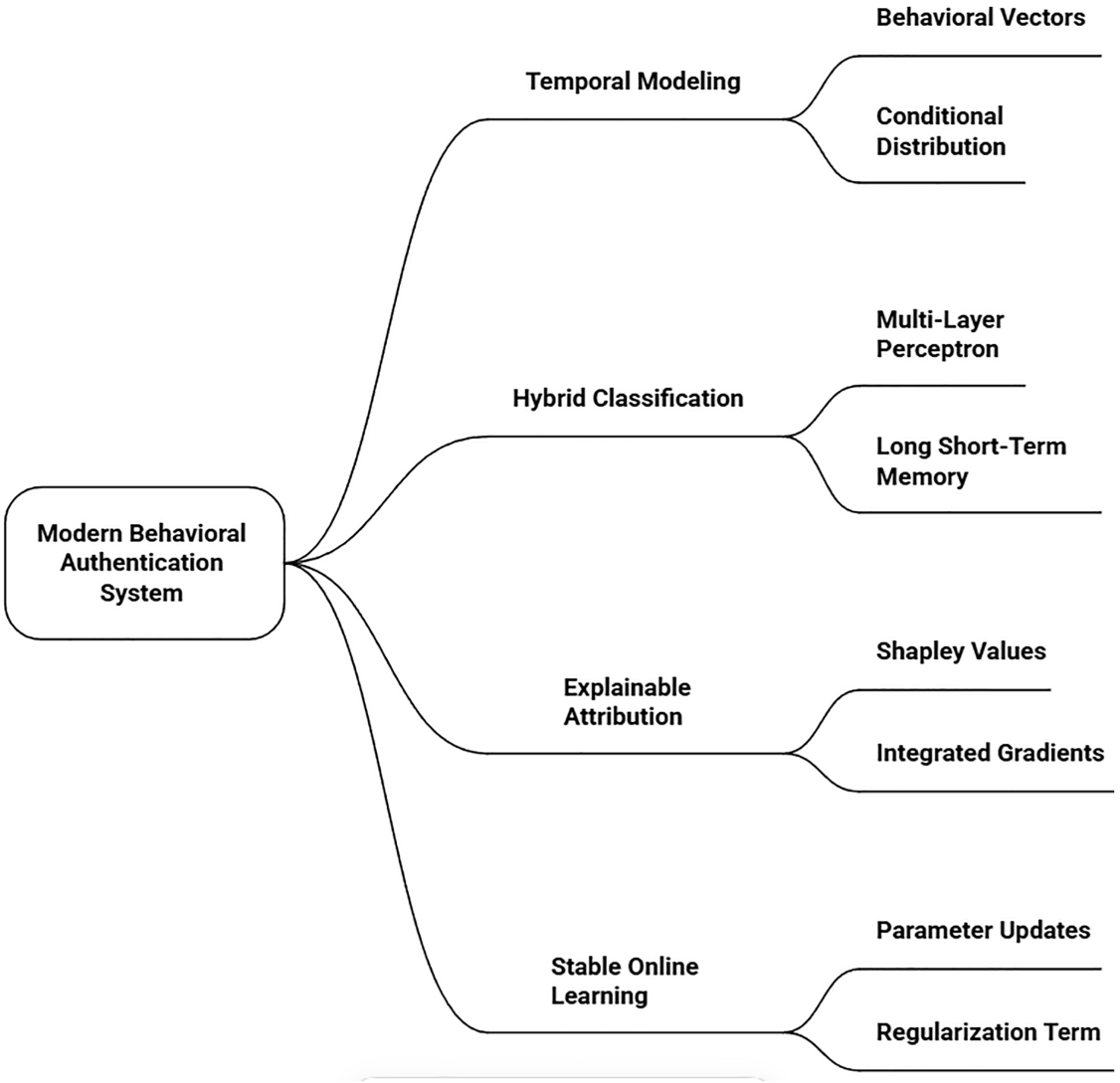

Figure 10: A conceptual architecture of a modern behavioral authentication system, detailing core components for temporal modeling, explainable attribution, and stable online learning

2.4.5 Adversarial Attack Mitigation

The mathematical foundation for mitigating adversarial attacks in AI systems involves a rigorous study of model vulnerabilities and the development of robust defense strategies [77]. The core threat is formalized as

a constrained optimization problem, where an adversary seeks a perturbation δ^* that maximizes the model's loss: $\delta^* = \arg\max_{\delta \epsilon \Delta} \mathcal{L}(f_\theta(x+\delta), y)$, within a bounded perturbation set $\Delta = \{\delta : \|\delta\|_p \leq \epsilon\}$. Efficient approximations of this problem lead to attacks like the Fast Gradient Sign Method (FGSM), which computes a perturbation via linearization: $\delta = \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(f_\theta(x), y))$. More potent, iterative attacks such as Projected Gradient Descent (PGD) solve this more accurately: $\delta_{t+1} = \Pi_\Delta(\delta_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f_\theta(x+\delta_t), y)))$, where Π denotes the projection back onto the constraint set.

To detect such attacks, Explainable AI (XAI) methods are employed to analyze model internals. The Jacobian matrix $J_f(x)$, with entries $[J_f(x)]_{ij} = \frac{\partial f_i(x)}{\partial x_j}$, reveals the model's sensitivity to input changes, and its Singular Value Decomposition (SVD), given by $J_f(x) = U\Sigma V^\top$, can exhibit anomalous patterns when the input is adversarial. Similarly, the Mahalanobis distance in the activation space of a hidden layer, $D_M(h_l) = \sqrt{(h_l - \mu_l)^\top \Sigma_l^{-1}(h_l - \mu_l)}$, provides a statistical measure that deviates from a χ^2 distribution for adversarial examples, flagging them as outliers.

The foremost defense strategy, adversarial training, directly incorporates these threats into the learning process through a min-max optimization: $\min_\theta \mathbb{E}(x, y) \sim \mathcal{D}[\max \delta \in \mathcal{B}\, \epsilon \mathcal{L}(f\theta(x+\delta), y) + \lambda \text{TV}(f_\theta)]$. This objective trains the model to be robust against the worst-case perturbation, while the total variation regularization, $\text{TV}(f_\theta) = \mathbb{E}x\left[|\nabla_x f\theta(x)|F^2\right]$, encourages smoother decision boundaries that are harder to exploit. A complementary approach is Lipschitz regularization, which bounds the model's overall sensitivity. The Lipschitz constant K, defined by the relation $|f\theta(x) - f_\theta(x')|_q \leq K|x - x'|_p$, can be controlled since it is bounded by the supreme spectral norm of the Jacobian: $K = \sup_x \sigma_1(J_f(x))$. This is practically enforced by constraining the spectral norm of each layer's weights: $|W_l|_2 \leq \kappa^{1/L}$ for all layers l.

Finally, explainability is extended to the adversarial context to understand which features contribute to model robustness. This is achieved by defining Shapley values for robust performance: $\phi_i^{\text{rob}} = \sum_{S \subseteq N \setminus i} \frac{|S|!(n-|S|-1)!}{n!}[R(S \cup i) - R(S)]$, where $R(S) = \mathbb{E}_\delta[\mathcal{L}(f(x_S + \delta_S), y)]$ measures the expected loss for a feature subset S under perturbation. Collectively, these interconnected mathematical frameworks for threat modeling, detection, robust training, and explainable robustness provide a rigorous foundation for advanced adversarial defense mechanisms [9,78,79].

2.4.6 Insider Threat Detection

The sophisticated analysis of behavioral patterns for insider threat detection relies on advanced mathematical modeling that applies statistical and machine learning techniques to temporal sequences of employee activities [80,81]. These activities are represented as a sequence of multivariate observations $\mathbf{X}tt = 1^T$, where each $\mathbf{X}t \in \mathbb{R}^{n \times d}$ encapsulates n behavioral features—such as login attempts and file accesses—for d employees at time t. The core detection problem is formalized as learning an anomaly scoring function $f : \mathbb{R}^{n \times d \times T} \to \mathbb{R}^T+$ that produces a time-dependent risk assessment. A common unsupervised approach uses a variational autoencoder, trained by optimizing the loss function:

$$\mathcal{L}(\theta, \phi) = \mathbb{E}q\phi(\mathbf{z}|\mathbf{X})[\log p_\theta(\mathbf{X}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{X})|p(\mathbf{z})) \tag{29}$$

where the encoder q_ϕ and decoder p_θ learn a compressed representation, and β controls the regularization. The anomaly score is then derived from the reconstruction probability: $s(\mathbf{X}t) = \mathbb{E}\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{X}t)[\log p\theta(\mathbf{X}_t|\mathbf{z})]$. For sequential data, LSTM networks model temporal dependencies using gating mechanisms (forget gate $\mathbf{f}_t$, input gate $\mathbf{i}_t$, output gate $\mathbf{o}_t$) to update the cell state $\mathbf{c}_t$ and hidden state $\mathbf{h}_t$. The anomaly score is computed from the prediction error: $\epsilon_t = |\mathbf{X}_t - \mathbf{W}_h\mathbf{h}t|\Sigma^{-1^2}$, where Σ is the covariance matrix.

Explainability in these temporal models is achieved through several mathematical frameworks. Shapley values are extended to attribute importance to feature-time pairs:

$$\phi_{j,t} = \sum_{S \subseteq \Omega \setminus (j,t)} \frac{|S|!(|\Omega| - |S| - 1)!}{|\Omega|!} [f(S \cup (j,t)) - f(S)] \tag{30}$$

where Ω is the set of all feature-time combinations. Similarly, a temporal attention mechanism computes time-step importance weights: $\alpha_t = \frac{\exp(\mathbf{v}^\top \tanh(\mathbf{W}a\mathbf{h}t + \mathbf{b}a))}{\sum t'=1^T \exp(\mathbf{v}^\top \tanh(\mathbf{W}a\mathbf{h}t' + \mathbf{b}t'))}$. For tree-based ensemble methods, feature importance incorporates temporal dependence: $Ij, t = \frac{1}{N} \sum_{T=1}^{N} \sum_{\tau \epsilon \mathcal{T}T} \mathbb{I}(j\tau = j) \cdot \Delta\text{Impurity}(\tau)$, where $\mathcal{T}_T$ is the set of split nodes.

These mathematical formulations enable precise anomaly identification and interpretation through variational inference for uncertainty quantification, differential analysis of LSTM gates for temporal localization, game-theoretic attribution across time, and spectral analysis of attention weights. The explainability frameworks operate through multiple mathematical lenses, including reconstruction analysis: $\mathcal{R}(\mathbf{X}) = \mathbb{E}\mathbf{z} \sim q\phi(\mathbf{z}|\mathbf{X})[|\mathbf{X} - g_\theta(\mathbf{z})|^2_{\mathbf{W}}]$, temporal gradient analysis: $\mathbf{G}t = \frac{\partial \mathbf{h}t}{\partial \mathbf{X}t} \cdot \frac{\partial f}{\partial \mathbf{h}t}$, and a contextual bandit formulation for alert optimization: $\pi^*(a|\mathbf{X}1:t) = \arg\max \pi \mathbb{E}\left[\sum \tau = t^T \gamma^{\tau - t} r\tau - \lambda D_{KL}(\pi|\pi_0)\right]$.

The integration of these advanced techniques—including Bayesian credibility intervals for evidence quantification, temporal pattern tracing, counterfactual analysis, and reinforcement learning for alert optimization—provides a robust foundation for explainable detection systems [82,83]. This represents a significant advancement over static approaches [84], delivering auditable decision trails that are crucial for operational trust and regulatory compliance [85,86].

2.4.7 Regulatory Compliance and Risk Assessment

The mathematical formalization of regulatory compliance for AI-driven cybersecurity systems necessitates a rigorous framework that bridges legal statutes and technical implementation [87]. This framework can be defined as a tuple $\mathcal{R} = (\mathcal{C}, \mathcal{V}, \mathcal{P})$, where $\mathcal{C}$ is the set of compliance requirements (e.g., GDPR Article 22, HIPAA §164.312), $\mathcal{V}$ denotes verification procedures, and $\mathcal{P}$ specifies penalty functions. To enforce these requirements during model development, a compliance loss function $\mathcal{L}c(\theta) = \sum i = 1^{|\mathcal{C}|} \lambda_i \max(0, \nu_i(f_\theta) - c_i)^2$ is integrated into the training objective, where ν_i measures the violation degree of requirement c_i and λ_i are regularization parameters. This ensures the AI model f_θ adheres to regulatory boundaries by penalizing deviations.

Specific regulations impose distinct mathematical constraints on the model's behavior and explanations. To satisfy GDPR's right to explanation, the system must generate a comprehensive explanation $\mathcal{E}(x, f_\theta) = \phi_j j = 1^d \cup (\alpha_t, \beta_t) t = 1^T$ for any decision $y = f_\theta(x)$, which includes feature-wise Shapley values ϕ_j and temporal attention weights (α_t, β_t) for sequential data. HIPAA's access control requirements are modeled as a constraint on the model's decision surface: $\mathbb{P}(f_\theta(x) \neq f_\theta(x')) \leq \epsilon \quad \forall x, x'$ s.t. $d_A(x, x') > \delta$, ensuring that authorized access differences do not lead to significantly different outcomes. Meanwhile, Payment Card Industry Data Security Standard (PCI DSS) Requirement 6 for auditability mandates maintaining an interpretable audit trail $\mathcal{A}t = \nabla_x f\theta(x_t), \text{IG}(x_t), \text{LIME}(x_t)$ at each time t, capturing gradients and local explanations.

A comprehensive risk assessment framework is crucial for monitoring compliance, combining model uncertainty and explanation stability. The epistemic uncertainty $u_e(x) = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (f_{\theta_k}(x) - \bar{f}(x))^2}$ is computed via Monte Carlo dropout, while explanation stability is measured by the Lipschitz continuity of Shapley values: $L_{\text{SHAP}} = \sup_{x,x'} \frac{|\phi(x) - \phi(x')|2}{|x - x'|2}$. These are combined into a composite risk score $\mathcal{R}(x) =$

$\alpha u_e(x) + \beta L\text{SHAP} + \gamma \mathcal{L}c(\theta)$. Furthermore, ethical constraints are formalized as fairness metrics, enforcing demographic parity $|\mathbb{P}(f\theta(x) = 1|A = a) - \mathbb{P}(f\theta(x) = 1)| \leq \eta \quad \forall a$ and equalized odds $|\mathbb{P}(f_\theta(x) = 1|A = a, Y = y) - \mathbb{P}(f_\theta(x) = 1|Y = y)| \leq \eta \quad \forall a, y$.

The complete optimization for developing a compliant AI system synthesizes these elements: $\min_\theta \mathbb{E}[\mathcal{L}(f_\theta(x), y)] + \mathcal{L}c(\theta) + \lambda_1 \mathcal{R}(x) + \lambda_2 \sum a, y \max(0, |\text{DP}a, y| - \eta)$. This can also be framed as a multi-objective optimization problem $\min \theta \in \Theta \left(\mathbb{E}[\mathcal{L}(f_\theta)]\, \mathbb{E}[\mathcal{L}c(\theta)]\, \mathbb{E}[\mathcal{R}(x)]\, \max a, y|\text{DP}_{a,y}|\right)$ to find Pareto-optimal trade-offs between accuracy, compliance, risk, and fairness. Explanation quality is quantitatively assured through metrics for completeness $(\sum_j \phi_j = f(x) - \mathbb{E}[f(x)])$, accuracy, consistency, and representativeness.

For dynamic environments, temporal compliance monitoring introduces additional complexity, modeled through a state-space formulation $\frac{d}{dt}\mathcal{C}(t) = A\mathcal{C}(t) + B\mathcal{V}(t) + \xi(t)$, where $\mathcal{C}(t)$ is the compliance state. This allows for real-time tracking using a Kalman filter: $\hat{\mathcal{C}}t|t = \hat{\mathcal{C}}t|t-1 + K_t(\mathcal{O}t - H\hat{\mathcal{C}}t|t-1)$. Collectively, these mathematical formalizations provide a rigorous foundation for building XAI systems that satisfy explicability, access control, auditability, and fairness requirements, enabling verifiable compliance in regulated industries while maintaining model performance [88–90].

This comprehensive mathematical framework bridges the gap between legal requirements and technical implementations. It enables the development of provably compliant Explainable AI systems for cybersecurity applications [91,92]. The formal guarantees provided by these methods address core challenges in trustworthy AI deployment within regulated environments, as identified by [93].

Table 3 provides a comprehensive overview of the key aspects of Explainable AI (XAI) within the cybersecurity domain, systematically comparing their characteristics, practical applications, and inherent limitations. It organizes the field into six core aspects: Transparency (contrasting post-hoc and intrinsic methods), specific XAI Methods like SHAP and LIME, the Scope of explanation techniques (model-specific vs. agnostic), the Performance trade-off between accuracy and interpretability, Deployment considerations, and Compliance with regulatory frameworks. The table effectively illustrates how these factors interplay, highlighting that while XAI is crucial for applications such as intrusion detection and regulatory adherence, it is consistently challenged by issues like computational cost, model opacity, and implementation complexity [94–99].

Table 3: Key aspects of explainable AI in cybersecurity

Aspect	Characteristics	Applications	Limitations
Transparency	Post-hoc (after prediction) vs. Intrinsic (built-in) [94]	Intrusion detection, malware analysis	Black-box opacity, complexity
Methods	SHAP (feature importance), LIME (local), Grad-CAM (visual) [95]	Threat detection, phishing analysis	Computational cost, real-time constraints
Scope	Model-specific (limited) vs. Agnostic (versatile) [96]	Security decisions, compliance	Architecture constraints, adaptation issues
Performance	Accuracy vs. interpretability trade-off [97]	AI security systems, regulated sectors	Optimization challenges, scaling issues
Deployment	Post-hoc (high cost) vs. Intrinsic (embedded) [98]	Big data analytics, integrated systems	Implementation overhead, real-time limits

(Continued)

Table 3 (continued)

Aspect	Characteristics	Applications	Limitations
Compliance	GDPR, AI Act requirements [99]	Regulated sectors, forensic analysis	Complex model compliance, explanation clarity

3 Role of Explainable AI (XAI) in Cybersecurity

The introduction of the cybersecurity frameworks into Artificial Intelligence (AI) and Machine Learning (ML) has been carving out nothing less than a revolution [4], realized in the form of automated threat detection, predictive analytics, and the response to the incident on a scale and speed that human analysts could not have achieved on their own. However, this adoption has been accompanied by a significant and growing challenge: the "black box" problem. Many advanced AI models, particularly complex deep learning networks and ensemble methods, operate in ways that are opaque, making it difficult for cybersecurity professionals to understand why a particular decision was made. This is where Explainable AI (XAI) becomes not just a technical enhancement, but a foundational pillar for trustworthy and effective cybersecurity. XAI refers to a suite of techniques and methods that make the outputs of AI models understandable and interpretable to human experts. In the high-stakes domain of cybersecurity, where decisions can impact national security, corporate integrity, and individual privacy, understanding the "why" behind an alert is as critical as the alert itself [100]. It transforms AI from an inscrutable oracle into a collaborative partner, enabling a synergy between human intuition and machine precision.

3.1 Enhancing Threat Detection and Response

In the context of threat detection, AI systems are tasked with identifying malicious activity within vast oceans of network traffic, system logs, and user behaviors [101]. A traditional Security Information and Event Management (SIEM) system might flag an anomaly—for instance, a user accessing a database at an unusual hour. An XAI-augmented system, however, would go beyond the binary alert. It would provide a detailed rationale, such as: "This activity was flagged as high-risk because the user's account (usually active 9-5 in the EST timezone) initiated a large data transfer to an external IP address in a non-whitelisted country, a combination of factors that deviates from 99.7% of their historical behavior and matches 85% of the patterns observed in previous data exfiltration attempts." This granular explanation allows a Security Operations Center (SOC) analyst to triage the incident with profound context. They can immediately discern if this is a genuine insider threat, a compromised account, or simply an employee working late on a critical project. This drastically reduces the mean time to detect (MTTD) and, more importantly, the mean time to respond (MTTR), as analysts are not wasting precious minutes or hours deciphering the AI's logic [102]. The explainability turns a potentially overwhelming alert into a guided investigation, streamlining the entire incident response workflow.

3.2 Building Trust and Facilitating Human-AI Collaboration

Trust is the currency of an effective security team. If analysts cannot understand the reasoning behind an AI's recommendations, they are likely to suffer from "alert fatigue" and begin to ignore or second-guess the system, a phenomenon known as automation bias in reverse. XAI directly addresses this by building a bridge of trust between the human and the machine. When an analyst can see the specific features—such as a specific sequence of system calls, a particular registry key change, or a unique packet signature—that led a

model to classify a file as malware, they are more likely to trust the verdict and act upon it. This collaborative dynamic is crucial for adaptive defense. For example, if an XAI system explains that it flagged a new piece of software as suspicious due to its attempts to disable a specific Windows Defender service, an analyst can use that human context to realize this is a legitimate action for a certain system utility [103]. They can then provide this feedback to retrain the model, creating a virtuous cycle of improvement. This human-in-the-loop feedback, powered by explanations, ensures the AI system continuously learns and adapts to the unique and evolving environment it is meant to protect, preventing the model from becoming stale and ineffective.

3.3 Proactive Defense: Improving Model Robustness and Identifying Adversarial Attacks

Cyber adversaries are increasingly sophisticated and are themselves leveraging AI to craft attacks designed to evade detection [104]. These are known as adversarial attacks, where inputs are subtly manipulated to fool AI models. For instance, an attacker might add minimal, human-imperceptible noise to a malware binary, causing the AI classifier to incorrectly label it as benign. A black-box model offers little insight into why it was fooled, leaving defenders in the dark. XAI shines a light on these vulnerabilities. By using techniques like saliency maps or feature importance analysis, security researchers can understand which parts of the input data the model is most sensitive to. If they discover that the model is overly reliant on a specific, easily manipulated file header for classification, they have identified a critical weakness. This understanding allows them to proactively retrain the model with adversarial examples, reinforce its feature set, and build a more robust and resilient defense system. In this sense, XAI moves the cybersecurity posture from a reactive one to a proactive one, enabling the hardening of AI defenses before they can be exploited in a live attack.

In conclusion, the role of Explainable AI in cybersecurity is transformative and multifaceted [12]. It is the critical link that closes the gap between raw algorithmic output and actionable human intelligence. By providing transparency, XAI empowers security analysts, accelerates investigations, builds essential trust, ensures regulatory compliance, and fortifies defenses against evolving adversarial threats. Since AI becomes more of a natural part of our digital security, the explainability of its actions will stop being a luxury and will become an unconditional prerequisite to the development of secure, responsible, and robust cyber-ecosystems.

3.4 Regulatory and Compliance Aspects

With the continued use of AI in cybersecurity, the regulatory bodies have provided clear demands that must be met in transparency and accountability. Explainable AI can assist organisations to address such needs by offering explainable systems and auditable decision-making processes that would be appropriate in investigations and compliance reporting.

3.4.1 Explainability Requirements in Cybersecurity Frameworks

General Data Protection Regulation, NIST Cybersecurity Framework, and ISO 27001 regulatory frameworks require the explanatory capabilities of AI systems [105]. GDPR will provide users with the right to be informed about the systems that are based on AI and that have an impact on their experience. As an illustration, the customers who may be victims of automated fraud detection should be provided with reasons that they understand in cases where their transactions raise an alarm. In a similar manner, NIST and ISO 27001 demand that organisations should show accountability in their AI-based cybersecurity actions. With the application of Explainable AI, companies can provide adequate explanations of AI decision-making, which will guarantee automation of security activities [106].

3.4.2 Role of XAI in Auditability and Forensic Investigations

Explainable AI systems are required to detect and trace the occurrence of cyberattacks and provide evidence during security audits and forensic investigations of such incidents [107]. The conventional AI systems do not tend to expose the factors that define the classifications of security breaches. The explainable AI systems, such as decision trees, rule-based systems, and feature attribution systems, allow an investigator to trace the logic behind the AI-generated warnings. As an example, during an investigation of a ransomware attack, an explainable model may indicate that unusual encryption patterns, unauthorised privilege escalation, and unusual access sequence of files are some of the major indicators of compromise. Such lessons allow forensic experts to build timelines of attacks, determine the weaknesses of the system, and create more effective defences. Additionally, Explainable AI facilitates cyber risk assessments by providing interpretable risk scores with step-by-step justifications, enabling organizations to prioritize security investments effectively [108].

Explainable AI has become indispensable for enhancing AI-based cybersecurity systems, ensuring regulatory compliance, and enabling forensic investigations [79]. By providing clear explanations for AI decisions, it improves the effectiveness of intrusion detection, malware analysis, fraud detection, and insider threat prevention. Furthermore, Explainable AI supports compliance with major regulatory frameworks by making AI models accountable and auditable. As cyber threats continue to evolve, the role of Explainable AI in cybersecurity will expand, fostering trust through transparency and strengthening AI-driven defense mechanisms [109].

The integrated analysis from (Fig. 11), and (Table 4), comprehensively outlines the multifaceted role and benefits of Explainable AI (XAI) in cybersecurity, demonstrating how it bridges the gap between complex AI operations and human-understandable security practices. Firstly, in enhancing threat detection and response, XAI techniques like SHAP and LIME are pivotal in Intrusion Detection and Prevention Systems (IDS/IPS); for example, Microsoft's CyberSignal uses SHAP to justify alerts, which has been shown to reduce false positives by 30%, thereby allowing security analysts to prioritize genuine threats effectively and refine detection models. Secondly, in the domain of malware analysis and classification, tools such as Grad-CAM provide visual explanations by highlighting the specific code segments or binary features that led a model to classify a file as malicious, as seen in systems like CylancePROTECT, which not only improves detection accuracy but also facilitates knowledge discovery by helping analysts understand emerging malware families. Third, to detect fraud and prevent insider threats, the XAI techniques, such as counterfactual explanations and SHAP-based risk scoring, implemented by PayPal and IBM Guardian among others, provide clear explanations of why a specific transaction or behaviour of a user is flagged, which can be useful in identifying actual fraud and processing false alarms, minimising false positives, and fostering trust in automated monitoring systems. More critically, a critical benefit is that it allows regulatory compliance and auditing, where XAI can directly respond to the requirement of the various frameworks such as GDPR, NIST, and ISO 27001 to produce traceable and justifiable audit trails, such as providing the legally mandated requirement of a right to explanation of automated decisions (EU banks use rule-based XAI logic to do this), or tools such as Splunk ES may use decision trees to perform forensic analysis and allow investigators to recreate attack timelines with clear and defensible evidence. Finally, overarching all these applications are the core benefits of bias mitigation and enhanced trust, as the transparency provided by XAI allows organizations to identify and correct for model biases that could lead to discriminatory outcomes, thereby fostering greater confidence among security teams, stakeholders, and regulators in AI-driven cybersecurity systems, ensuring they are not only effective but also fair and accountable.

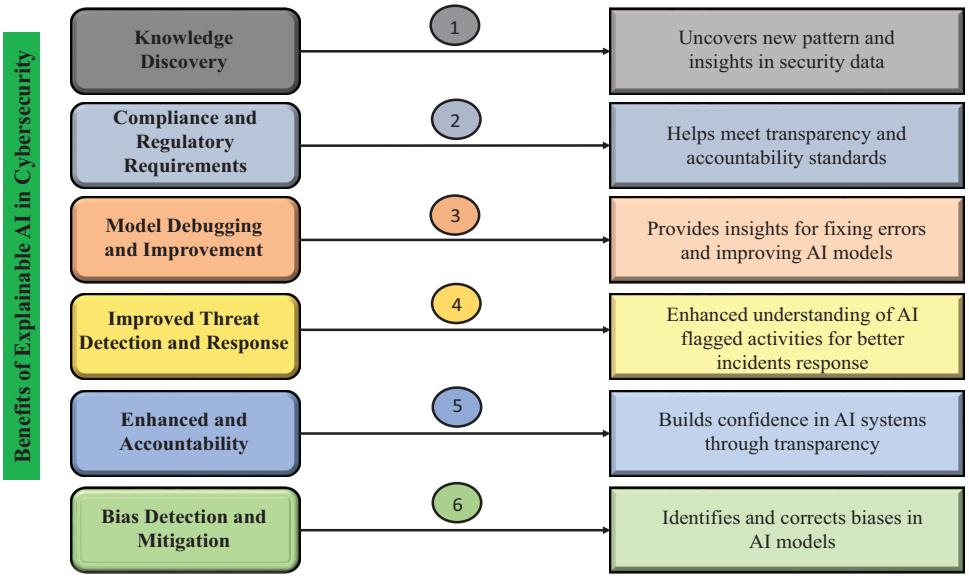

Figure 11: Key benefits of explainable AI (XAI) in cybersecurity. *For example, Microsoft's CyberSignal employs SHAP to justify alerts, which has been reported to reduce false positives by 30% and significantly improve analyst trust and response times [110]. Similarly, Darktrace's Enterprise Immune System uses behavioral analytics with explainable alerts to reduce mean time to detection by over 90% [111]*

Table 4: Role of explainable AI (XAI) in cybersecurity with key benefits

Aspect	Description	Challenges without XAI	XAI techniques used	Key benefits	Empirical evidence/real-world use cases
XAI in IDS/IPS	AI-driven systems analyze traffic to detect and block threats.	High false positives/negatives; low transparency; tuning difficulties.	SHAP, LIME, feature attribution, decision trees.	• Improved Threat Detection • Enhanced Trust • Model Improvement	Microsoft CyberSignal uses SHAP for alert justification, reducing false positives by 30% [112]. Darktrace uses behavioral analytics with explainable alerts.
Malware analysis	AI classifies malware via behavior or signature analysis.	Lack of result validation; misclassification; no detection rationale.	Grad-CAM, rule-based models, feature importance.	• Knowledge Discovery • Bias Mitigation • Model Improvement	CylancePROTECT uses Grad-CAM for malware visualization [53]. EMBER dataset with SHAP analysis.
Fraud detection & Insider threats	Detects fraud and insider risks from user behavior.	Alerts lack justification; difficult fraud separation; wrongful accusations.	Counterfactuals, SHAP/LIME, risk scoring models.	• Enhanced Trust • Bias Mitigation • Improved Threat Detection	PayPal uses LIME and counterfactuals for transaction fraud [76]. IBM Guardian employs SHAP for insider threats.
Cybersecurity compliance	Meets GDPR, NIST, ISO 27001 requirements for explainability.	GDPR "right to explanation'; missing audit trails; justification gaps.	Rule-based logic, SHAP/LIME, model-agnostic tools.	• Compliance • Enhanced Trust • Knowledge Discovery	EU banks use rule-based XAI for GDPR compliance [36]. NIST CSF uses SHAP for audit trails.

(Continued)

Table 4 (continued)

Aspect	Description	Challenges without XAI	XAI techniques used	Key benefits	Empirical evidence/real-world use cases
Auditability & Forensics	AI aids attack analysis, tracing, and vulnerability assessment.	No reasoning in alerts; black-box hindrance; legal evidence challenges.	Decision trees, feature attribution, SHAP/LIME audits.	• Knowledge Discovery • Model Improvement • Compliance	Splunk ES uses decision trees for forensic analysis. CIC-IDS2017 dataset with LIME explanations [113].

4 The Trade-Off between Interpretability and Security

Explainable AI offers significant benefits for security systems by improving trust, compliance standards, and investigative capabilities [114]. However, a primary research challenge involves addressing how interpretability can potentially work against security. The increased model transparency designed to enhance trust and accountability may simultaneously introduce security threats by exposing model vulnerabilities. This section assesses the specific vulnerabilities of Explainable AI systems and examines security measures developed to address these weaknesses.

4.1 Security Risks of XAI Models

The process of making AI decisions more interpretable through XAI methods allows attackers to exploit the increased visibility of AI systems. Multiple security risks relate to XAI models engage as follows:

4.1.1 How Interpretability Can Expose Vulnerabilities (e.g., Adversarial Examples)

Explainable AI (XAI) methods like SHAP and LIME reveal the key features that influence a model's decisions; however, this very transparency can be weaponized by attackers. By understanding which attributes are critical, for instance in an Intrusion Detection System (IDS) or malware classifier, adversaries gain strategic intelligence on how to craft inputs that bypass detection. They mathematically generate these adversarial examples by adding a small, calculated perturbation to the original input, often following the gradient of the model's loss function as formalized by $x^* = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$, where x^* is the adversarial example, x is the original input, ϵ is a perturbation factor, and J is the loss function. The resulting adversarial inputs are optimized to be misleading to the AI model while appearing realistic to human observers, effectively exploiting model explainability to create potent evasion attacks [115].

Fig. 12 illustrates the various security risks associated with Explainable AI (XAI). These risks include Model Inversion, which involves extracting sensitive data from the explanations provided by XAI; Model Extraction, where AI models are replicated for malicious purposes; Data Poisoning, the act of injecting harmful data to degrade AI performance; Adversarial Attacks, which involves crafting inputs to deceive AI models; Reverse Engineering, identifying AI vulnerabilities through its explanations; and Information Leakage, where sensitive information is inadvertently exposed. Each risk is positioned in concentric layers, highlighting the varying levels of threat.

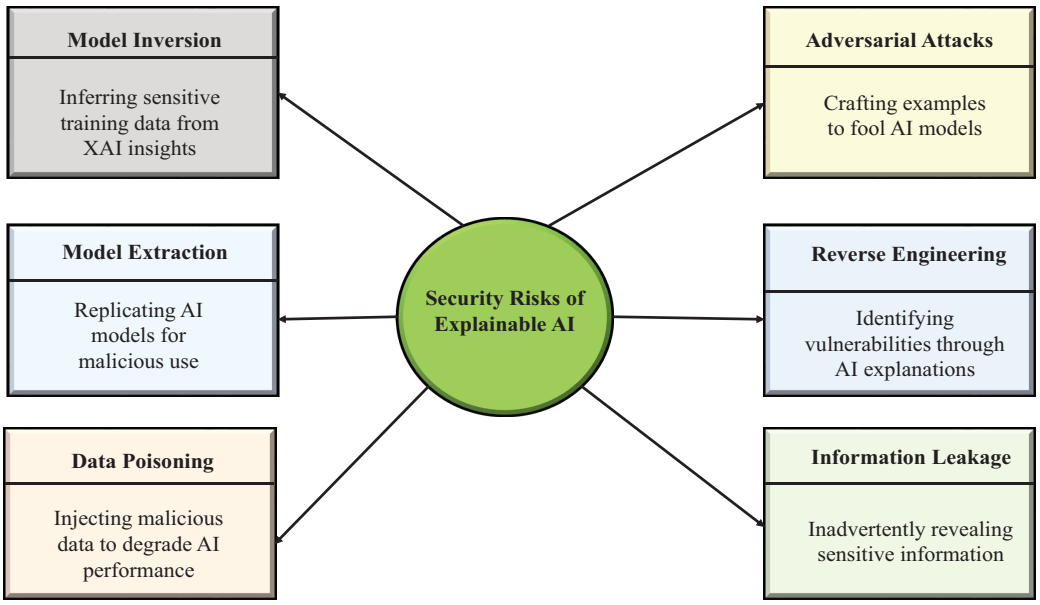

Figure 12: Security risks of explainable AI. *These risks are not merely theoretical; model inversion attacks have been demonstrated to partially reconstruct facial images from explainable facial recognition systems [116], highlighting the privacy perils of excessive transparency. In cybersecurity, adversarial examples crafted using SHAP explanations have been shown to evade malware detectors with over 80% success rate*

4.1.2 Model Inversion and Membership Inference Attacks Using Explanations

Model inversion attacks are dangerous privacy threats that utilize XAI methods to reconstruct sensitive input data based on a model's behavior, which is particularly critical when models process confidential information like biometric data or cybersecurity logs [117]. A related threat, membership inference attacks, enables an adversary to determine if a specific data point was part of the model's training set, potentially by exploiting explanations like SHAP or Grad-CAM heatmaps to identify the presence of particular users or attack signatures in security datasets. Mathematically, these attacks exploit the difference in model confidence scores between training and non-training samples by training an attack model, often formulated as $\mathbb{P}(y|f(x)) = \sigma(f(x) \cdot w + b)$, where $f(x)$ is the target model's output, w and b are adversarially trained weights, and $\sigma(\cdot)$ is a logistic function that classifies whether sample x was in the training set; by leveraging explainability methods to infer membership, an adversary could ultimately reconstruct portions of the private training data, thereby violating fundamental privacy constraints.

4.1.3 Risks of Transparency in Security-Sensitive Environments

While XAI improves trust, excessive transparency can help adversaries understand the inner workings of cybersecurity models. This can be particularly dangerous in applications such as:

Anti-malware solutions: If attackers know which file features are used for malware detection, they can craft malicious software that evades detection [118,119].

Access control and authentication systems: If XAI reveals decision-making patterns in biometric authentication, adversaries can create more accurate spoofing attacks [120].

Threat detection models: If cybercriminals understand what kinds of network activity cause an alert, they will be able to adjust their attacks to bypass these defenses [121]. In this case, interpretability and security trade-offs are required, and it is necessary to control the exposure of model explanations [64].

4.2 Prescriptive Patterns for Secure XAI Deployment

The burden between model transparency and security is not a binary choice but a managed balance. Based on the principle of *least-privilege explanation* [122], we propose the following prescriptive patterns for deploying XAI in high-stakes cybersecurity environments. These patterns ensure that explanations are provided only when necessary, to the right entity, and in a way that minimizes the risk of adversarial exploitation.

- **Role-Gated Explanations:** Access to detailed explanations is restricted based on user roles and clearances within the Security Operations Center (SOC).
  - **Tier 1 Analyst:** Receives simple, high-level reason codes (e.g., "Suspicious due to anomalous geolocation and time of access").
  - **Tier 2/3 Analyst & Forensics:** Gains access to full feature-attribution details (e.g., SHAP values, LIME outputs).
  - **External User/API:** Receives only the binary decision (e.g., "Access Denied") or a generic, non-informative message.

  This prevents low-privilege users or potential attackers from gaining insights into the model's decision logic.
- **Randomized & Quantized Attributions:** To protect against model inversion and extraction attacks, explanations can be deliberately obfuscated.
  - **Randomized Attribution:** Add controlled noise to feature importance scores (e.g., $\phi_j^* = \phi_j + \mathcal{N}(0, \sigma^2)$). This maintains the overall ranking of important features while making it difficult for an adversary to precisely reverse-engineer the model.
  - **Quantized Attribution:** Map continuous importance scores to a small set of discrete levels (e.g., *Low, Medium, High*). This preserves the explanatory intent while hiding the exact, potentially exploitable, decision boundaries.
- **Explain-on-Deny:** A critical pattern for authentication and access control systems where providing explanations for a *grant* can leak information.
  - **Upon Access Grant:** Provide no explanation or a generic one (e.g., "Access Approved").
  - **Upon Access Deny:** Provide a detailed, actionable explanation to the legitimate user (e.g., "Access denied due to failed 2FA from a new device. Please use your registered device.").

  This pattern helps legitimate users troubleshoot issues without giving attackers a roadmap for crafting successful attacks.
- **Explanation Rate-Limiting & Budgets:** To prevent automated explanation harvesting attacks, treat explanation generation as a costly API.
  - Enforce strict rate limits (e.g., N explanation queries per minute per user/IP).
  - Implement explanation "budgets" for external API consumers.
  - Log all explanation requests for audit and anomaly detection, treating a high volume of explanation requests as a potential reconnaissance attack.
- **Context-Aware Explanation Fidelity:** The level of detail in an explanation should adapt to the context and perceived risk.
  - **Low-Risk Context:** From a trusted IP range, provide full explanations to aid analyst workflow.
  - **High-Risk Context:** From an unknown IP or during an active attack, suppress or generalize explanations to prevent aiding the adversary.

These patterns can be combined to create a defense-in-depth strategy for XAI deployment. For instance, a system could employ *role-gated, explain-on-deny* with *quantized attributions* for external users, while allowing full, real-time explanations for senior analysts on the internal network. By making the transparency-security trade-off prescriptive, organizations can strategically deploy XAI to build trust and maintain compliance without unduly increasing their attack surface.

4.3 Defense Mechanisms

Considering these security risks of XAI, several defense mechanisms have been suggested to achieve increased robustness without sacrificing interpretability. These are detailed in a subsequent subsection below.

4.3.1 Secure-XAI Models: Balancing Interpretability and Robustness

Secure-XAI frameworks are designed to provide explanations without exposing sensitive model information. These models integrate techniques such as:

Feature obfuscation: Limiting the level of detail in explanations to prevent adversarial exploitation [123].

Selective explainability: Providing explanations only to trusted parties (e.g., security analysts) while restricting access for external queries.

Randomized feature attribution: Introducing randomness in explanation generation to make it difficult for attackers to exploit the model systematically [124].

4.3.2 Adversarial Training for Explainable AI Models

Adversarial training is a method to enhance AI model robustness by training them on adversarially perturbed inputs, thereby improving their resilience to adversarial attacks [125]. This process involves techniques that can interact with explanation methods, such as gradient masking—a regularization method where attackers can exploit explanation-based gradients—and adversarial explanation training, where the explanations themselves are structured to identify and mitigate potential vulnerabilities. Mathematically, adversarial training is formulated as a min-max optimization problem: $\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\delta\in S} J(\theta, x+\delta, y)\right]$, where θ represents the model parameters, δ is the adversarial perturbation constrained within a set S, and $J(\theta, x, y)$ is the model's loss function; by proactively training models, including those used in Explainable AI (XAI), against such adversarially generated inputs and explanations, their overall robustness to explainability-based attacks is significantly improved.

4.3.3 Privacy-Preserving XAI Techniques (e.g., Differential Privacy)

To mitigate inference attacks that exploit explanations to reveal confidential information, privacy-preserving XAI techniques have been introduced, which safeguard explanations from exposing sensitive data [126]. A foundational technique in this domain is Differential Privacy, which adds controlled noise to model explanations to prevent adversaries from deciphering private training information [127]; in DP-based XAI, explanations $E(x)$ are perturbed as $E^*(x) = E(x) + \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}(0, \sigma^2)$ represents Gaussian noise that obscures precise information extraction. Furthermore, the decentralized framework of Federated Learning, when combined with privacy-aware explanations, minimizes risks such as model inversion attacks on the underlying data [128–130]. A persistent challenge in deploying Explainable AI within cybersecurity is the inherent tension between security and interpretation: while interpretability fosters trust, compliance, and enables forensic analysis, it can inadvertently expose security vulnerabilities if not managed properly. Consequently, the integration of privacy-preserving XAI methods with Secure-XAI models and adversarial

training establishes a crucial balance, ensuring both model transparency and robustness are maintained without compromising privacy [131].

Our analysis identifies key security threats to Explainable AI (XAI) and corresponding defensive countermeasures. Explanations generated by XAI systems introduce three primary points of vulnerability that adversarial attacks, model inversion, and membership inference can exploit to compromise security and privacy. To mitigate these risks, defense strategies such as adversarial training, differential privacy, and federated learning are employed.

The analysis, detailed in (Table 5), evaluates two significant risks: explainability-assisted evasion attacks and gradient-based attacks. This evaluation underscores the critical need for robust feature attribution methods and carefully controlled transparency mechanisms. The document also addresses the inherent security dilemma between model performance and explainability, as increased interpretability can often lead to a less secure system.

Table 5: Security risks in XAI models and corresponding defense mechanisms

Aspect	Security risks	Impact	Defense mechanisms	Effectiveness
Adversarial vulnerability [132]	Explanations reveal decision boundaries aiding adversarial input crafting	Attackers evade detection (malware bypassing IDS)	Adversarial training, randomized attribution, perturbation masking	High
Model inversion attacks [133]	Inference of sensitive training data via explanation analysis	Risk of leaking private or proprietary data	Differential privacy, federated learning, secure multiparty computation	High
Membership inference attacks [134]	Identifying if a data point was used in training through explanations	Breaches data confidentiality, exposing dataset details	Privacy-preserving XAI, noise injection, membership masking	High
Explainability-aided evasion [135]	Adversaries use explanations to find critical features for evasion	Enables circumvention of fraud or intrusion systems	Restricted or randomized explanations in sensitive domains	Moderate
Transparency in security-sensitive systems [136]	Full transparency may expose security policies	Allows attackers to reverse-engineer controls	Controlled XAI access, secure-XAI frameworks, selective transparency	High
Gradient-based attacks [137]	SHAP/LIME reveal gradient patterns exploitable by adversaries	Facilitates precise adversarial inputs	Gradient masking, robust attribution, adversarial training	High
Privacy risks in XAI [138]	Interpretability leaks model parameters or business logic	Intellectual property loss; compromises AI internals	Federated learning, encrypted inference, differential privacy	High
Regulatory & compliance risks [139]	Balancing explainability with data protection laws is challenging	Overexposure conflicts with GDPR, ISO standards	Role-based access, audit trails, explainability management	High
Interpretability & robustness trade-off [140]	Highly interpretable models often less robust to adversaries	Fragile defenses due to exposed decision logic	Hybrid models, secure-XAI design, robustness-aware evaluation	Moderate

Finally, a review assesses the applicability of explainability security measures required for compliance with standards such as GDPR, NIST, and ISO 27001, linking these to the defense fortification strategies discussed in the (Table 5).

5 Comparative Evaluation of XAI Techniques for Cybersecurity Applications

Fig. 13 illustrates the concentric evaluation metrics essential for assessing Explainable AI effectiveness in cybersecurity contexts. These metrics form a comprehensive framework for measuring multiple dimensions of XAI performance. To better understand XAI applicability, we examine each metric's specific importance to cybersecurity with concrete examples.

Evaluation metrics are used as the basis for determining the quality of the XAI system. These metrics enable practitioners to gauge efficiency and reliability within the sector of cybersecurity endeavour. Since XAI considers transparency and explainability in an AI decision, certain measures in the field are required to measure them. Such moves render decisions of AI and justifications transparent and credible. The evaluation procedure has taken into consideration the areas of performance that are highly significant in cybersecurity interests.

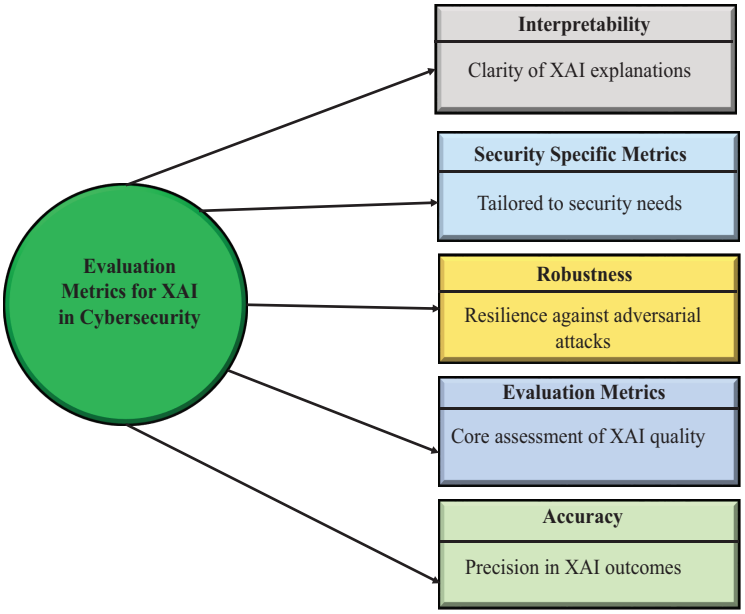

Figure 13: Comprehensive evaluation metrics for assessing the quality, reliability, and security of explainable AI (XAI) systems in cybersecurity applications

One of the most significant measures to evaluate AI systems in cybersecurity is accuracy. The rightness and the accuracy of the XAI results are measured by this measure. The concept of accuracy used in cybersecurity implies that the decisions made by AI are correct to identify a threat and reduce false classification. As an example, an XAI system for identifying potential vulnerabilities or abnormal behaviour cannot work effectively unless it is very accurate. The number of false positives and negatives would be too high to render the system operable. The XAI quality lies in the accuracy, and the cybersecurity specialists are provided with trustworthy information to make decisions. False predictions of the threats can ruin the security or the unnoticed attacks, and precision is important in the XAI assessment.

The concept of robustness is the characteristic of the system to withstand adversarial attacks, whereby bad users can manipulate the input data in some subtle way to deceive the AI models. XAI vulnerabilities in cybersecurity can be used to bypass security checks or result in false positives in the detection systems. Part to assess the level of strength, they can be tested on how XAI systems will respond to these attacks and report the corresponding outputs. Such manipulations that have such XAI systems should be maintained in

quality decisions and security integrity. The more advanced the adversarial methods are, the more ruggedness is a competitive advantage to adversarial methods. This measure is a capacity of systems concerning their sensitivity to the attack conditions in the real world.

The outstanding difference between XAI and traditional AI models that consider the interpretation and straightforwardness of the AI decision-making processes is interpretability. Justifications of AI should be accurate and understandable in the scenario of cybersecurity specialists who are in high-stress conditions. Interpretability will make sure that the specialists do not just know how AI systems arrive at a specific decision, such as why the network activity is suspicious or why the vulnerability is critical. Absence of interpretability may cause a lack of trust in the system to the cybersecurity team, who do not want to follow the recommendation of the system. Additionally, high interpretability can serve to satisfy the guidelines, regulations, and ethical principles of automated decision-making because it enables organisations to clarify automated decisions in a familiar and comprehensible language. The transparent and decipherable models will become invaluable during the audits or regulatory inspections.

The cybersecurity domain is expected to have special metrics of security to meet the unique needs and requirements of the domain. These measures are obtained not only by the general AI performance evaluators but also by the success of the XAI systems in terms of the desired security objectives. It is also possible to associate other measures to the ability of the system to identify threats on the fly, new threat patterns, or minimise false alarms in noisy environments. They are also used in the estimation of the performance of the systems in various cyberspace security cases, such as network security, malware, and fraud prevention. Assistance in determining the items in the XAI that would make the best impact on the practices and risk-related decisions of cybersecurity professionals can help them later. These tailor-made designs will make sure that the XAI models are formalised as high-performance, abstract, besides the real-world optimizations to certain security issues. The significance of such customised tests lies in a condition where the effects of the security breach can be extremely drastic.

Lastly, these evaluation metrics offer an overall evaluation system of the quality of the XAI system regarding cybersecurity. They are all the accuracy, robustness, interpretability, and metrics that are security-specific and guarantee the efficiency of the XAI systems in the security environment. The XAI testing on these dimensions can enable organisations to ensure that their AI models will not provide inaccurate answers, be resistant to adversarial attack, form clear judgments, and be safe regarding cybersecurity. Cyber threats are dynamic, and therefore, such holistic evaluations have gained more relevance in the design and implementation of AI that will realise the security benefits at minimum cost.

Table 6 provides a comparative analysis of prominent Explainable AI methods within cybersecurity contexts. This evaluation examines method types, interpretability levels, strengths, weaknesses, and optimal applications. The classification system is based on the model-agnostic, model-specific, or hybrid nature of methods and their specific advantages and drawbacks in terms of interpretability, computational complexity, and domain-specific utility.

Table 6: Comparative analysis of XAI techniques in cybersecurity

XAI technique	Type	Interpretability scope	Strengths	Weaknesses	Best use case
SHAP [141]	Model-agnostic	Global & Local	Fair attribution, works with any model	Computationally expensive	Malware detection, intrusion detection
LIME [142]	Model-agnostic	Local	Simple, fast	Sensitive to input changes	Anomaly detection, fraud detection

(Continued)

Table 6 (continued)

XAI technique	Type	Interpretability scope	Strengths	Weaknesses	Best use case
Decision trees/Random forests [143]	Intrinsic	Global & Local	Naturally interpretable	Prone to overfitting	Phishing URL classification, rule-based detection
Attention mechanisms [144]	Model-specific	Local	Highlights critical input segments	Limited to attention models	Malware behavior analysis, log detection
Counterfactual explanations [145]	Model-agnostic	Local	"What-if" scenarios	Unrealistic examples	Adversarial analysis, false positive reduction
Partial dependence plots (PDP) [146]	Model-agnostic	Global	Feature impact visualization	Assumes feature independence	Network intrusion models
Anchors [147]	Model-agnostic	Local	Human-readable rules	Computationally intensive	Fraud detection, access control
Saliency maps [148]	Model-specific	Local	Visualizes critical input areas	Limited to image/text	Malware image classification
Rule extraction [149]	Post-hoc	Global	Converts to rules	Oversimplifies models	Neural network to firewall rules
Surrogate models [150]	Model-agnostic	Global	Approximates complex models	Loss of accuracy	SIEM systems, ensemble explanation
Bayesian networks [151]	Intrinsic	Global & Local	Probabilistic reasoning	Needs domain knowledge	Risk assessment, threat probability
Neural-symbolic integration [152]	Hybrid	Global & Local	Neural + symbolic reasoning	Complex to implement	Automated threat hunting

SHAP is a model-agnostic technique that offers both local and global interpretability as well as fairly assigning features to prediction. Its main weakness is that it is computationally expensive, and this might not be allowed in time-sensitive applications or massive datasets. In the area of cybersecurity, SHAP provides useful information to malware detection systems, intrusion detection systems, and clarifies which characteristics or behaviours should be used to impact the model outputs.

It is a model-agnostic approach that only emphasises local interpretability, giving estimates of the black box models by using interpretable surrogacy models in single predictions. The advantages of LIM are simplicity and speed, which allow the generation of an explanation quickly. Nevertheless, this decreases resilience to more fluctuating environments because it is sensitive to changes in inputs. The main tasks of cybersecurity applications include anomaly detection and fraud analysis, in which it is worth having a rapid interpretation of crucial individual predictions [153].

These intrinsic model-specific approaches have global and local interpretability in the form of human-readable rules of decision. They are overfitted and prone to overfitting, especially in deep trees or noisy data, due to their inherent transparency. These models can be efficiently used in cybersecurity in phishing URL classification and rule-based detection situations, where transparent decision-making is required.

Attention mechanisms, as model-specific mechanisms implemented in deep learning models, are a source of local interpretability since they generate important segments of inputs. They are still only applied to neural networks that have explicit layers of attention. Cybersecurity systems comprise malware code analysis and log identification, where the recognition of critical behavioural patterns or log records is used in the process of detecting threats.

These model-agnostic techniques provide the ability to understand predictions locally with what-if scenarios that define sufficient changes to make. Being useful in the determination of the limits of decision-making, they can result in unrealistic situations within complicated models. Applications to cybersecurity are concentrating on adversarial analysis and false positive reduction, where the explanation of the boundaries and errors of decisions can increase the resilience of the model.

Being global interpretability techniques, PDPs graphically represent the influence of features on predictions and assume feature independence. This false assumption can generate false interpretations in correlated features. Network Intrusion modelling is one of the applications of cybersecurity deployments that involves the need to know how network features relate to threat identification.

This model-agnostic approach creates local interpretability by proposing rules (that are human-readable) that define the conditions of the prediction. Complex models or large datasets are also a challenge due to computational intensity. Fraud detection and access control systems are part of cybersecurity implementation, and in this case, it is important to know the conditions that lead to a decision on security.

The model-specific visualisation methods give local explanations of deep learning models that can handle visual or textual data. They can only be used with a limited number of types of data and model architectures. Saliency maps used in the field of cybersecurity help in the classification of malware images by determining which regions of an image actually lead to the malicious file classifications.

Rule extraction, being a post-hoc model-agnostic approach, translates complex models are translated into interpretable rules, which can be interpreted globally. The simplification process can be biased in terms of accuracy and main points. Translating the outputs of neural networks into readable firewall policy, aiding security professionals to interpret automated security decisions, are examples of cybersecurity applications.

These model-agnostic techniques are approximations of the complex models with the simple, globally interpretable ones. In the reproduction of complex original model behaviours, accuracy loss is a common occurrence. The systems of cybersecurity have Security Information and Event Management systems, where it is more understood how the entire system behaves, rather than the necessity of specific predictions.

Being intrinsic model-specific approaches, Bayesian networks have local and global interpretability based on probabilistic reasoning. Effectiveness is largely based on the domain knowledge and the right structure specification. Risk assessment and probability of threat applications are part of cybersecurity applications that allow effective risk management decisions to be made.

This is a hybrid system incorporating neural learning and symbolic reasoning, providing both localised and global explanations. There are high deployment complexities in implementation. Organisations that utilise cybersecurity applications are using automated threat-hunting applications, which involve the integration of pattern recognition and logical analysis to further detect advanced cyber attacks.

To conclude, every XAI approach has its own benefits and drawbacks that can be applied in various situations related to cybersecurity. The selection criteria should be based on interpretability needs, computer resources, and special cybersecurity goals. The techniques will likely keep being added to the XAI field and will be more tightly connected to the cybersecurity infrastructures, which will improve the transparency and trust, and will become the means of addressing the growing cyber threats.

5.1 A Practical Taxonomy for Deploying XAI in Cybersecurity

A practical interpretation of the trade-offs between the explanatory power, requirements of the security tasks, and the constraints of the operations must be understood when implementing XAI effectively in the field of cybersecurity. In this connection, we would suggest using a three-dimensional taxonomy to provide direction to the choice of methods:

- **Dimension 1: Explanation Method & Scope.** The decision starts with the AI model, which is being used, and the necessary explanation granularity. Black-box models can be flexible with model-agnostic approaches (e.g., LIME, SHAP), but can also be more efficient and architecture-specific (e.g., Grad-CAM, Attention). Moreover, Individual predictions (e.g., Why was this packet flagged? ") are justified using Local explanations, and the overall behavior of a model (e.g., What features define malware? ") is described using Global explanations.
- **Dimension 2: Primary Security Task.** The explanatory needs of various tasks differ. Intrusion Detection Systems (IDS) require fast, feature-based justifications for alerts. Malware Classification often benefits from visual or code-segment highlights. Fraud Detection relies on understanding user behavior sequences and counterfactuals ("What if this transaction amount was lower?").
- **Dimension 3: Deployment Constraints.** This is the most critical yet often overlooked dimension.
  - **SOC Latency Budget:** Real-time threat response requires low-latency XAI (e.g., LIME, simple feature attribution). Post-hoc analyses for forensics can tolerate slower, more accurate methods (e.g., SHAP, counterfactuals).
  - **Data Locality & Privacy:** In federated or sensitive environments, methods that avoid exposing raw data or model details are preferred, necessitating privacy-preserving XAI or intrinsic interpretability.
  - **Computational Cost:** The resources available for explanation generation dictate whether heavy-computation methods like SHAP are feasible or if lighter surrogates are necessary.

The following synthesis (Table 7) operationalizes this taxonomy, providing a concise reference for aligning XAI techniques with practical cybersecurity deployment scenarios.

Table 7: Practical guide to XAI methods for cybersecurity: alignment with tasks, data, and operational constraints

XAI method	Primary task	Datasets	Key metrics	Compute cost	Operational notes
SHAP [154]	Malware detection, IDS/IPS, Forensics	CIC-IDS2017, UNSW-NB15, EMBER	Fidelity, Consistency	High	Post-hoc analysis; avoid real-time SOC
LIME [155]	Real-time IDS, Phishing detection	NSL-KDD, CIC-IDS2017, URL phishing	Accuracy, Stability	Medium	Real-time use; balance quality/speed
Grad-CAM [156]	Malware image analysis	Microsoft Malware	Visual Coherence, AUC	Low	CNN-specific; intuitive visuals
Counterfactual [157]	Fraud detection, FP analysis	IEEE-CIS Fraud, Logs	Proximity, Plausibility	Medium-High	Stakeholder justification; computationally heavy
Decision trees [158]	Phishing, Access control	Various (rule generation)	Fidelity, Simplicity	Low	Intrinsically interpretable; auditable policies
Attention mechanisms [159]	Log analysis, Anomaly detection	LSTM sequence data	Weight entropy, Accuracy	Low	Shows model focus; needs attention architecture

5.2 Empirical Grounding and Practitioner-Centric Evaluation of XAI

A critical gap in XAI research is the lack of empirical baselines grounded in canonical cybersecurity settings. To guide practitioners, we synthesize performance trends for popular explainers against key operational metrics, using well-established datasets as benchmarks.

5.2.1 Canonical Datasets and Settings

Evaluations should be conducted on public, representative datasets to ensure reproducibility and fair comparison. Key benchmarks include:

- **Intrusion Detection:** `CIC-IDS2017` & `UNSW-NB15` for network traffic analysis.
- **Malware Classification:** `EMBER` for static PE file analysis and `Microsoft Malware Classification Challenge` for image-based representations.
- **Fraud Detection:** `IEEE-CIS Fraud Detection` for transaction fraud patterns.

5.2.2 Practitioner-Relevant Evaluation Metrics

Beyond accuracy, the following metrics are crucial for SOC adoption:

- **Explanation Latency:** Time to generate an explanation (ms). Critical for real-time triage.
- **Cross-Version Stability:** Consistency of explanations when the AI model is retrained or updated.
- **Analyst Acceptance Rate:** Percentage of AI-generated alerts that are acted upon after an explanation is provided.
- **Mean Time to Respond (MTTR):** The impact of explanations on the speed of incident resolution.

Table 8, provides synthesized performance baselines for common XAI methods, offering practitioners a reference for expected behavior in operational environments.

Table 8: Representative empirical baselines for XAI methods in cybersecurity

XAI method	Latency	Stability (Model)	Stability (Input)	Analyst acceptance	Use case
SHAP [160]	High (1–10 s)	High	High	High (deep analysis)	Post-incident forensics, debugging, compliance
LIME [160]	Medium (0.1–2 s)	Low	Medium	Medium (alert triage)	Real-time SOC alert justification, speed-critical
Grad-CAM [161]	Low (<100 ms)	Medium	High	High (visuals)	Malware image analysis, CNN-based detectors
Attention [26]	Low (built-in)	Low	Low	Medium	Sequential data (logs, behavior); model-specific
Decision Trees [162]	Very Low (<10 ms)	High	High	Very High (inherent)	Auditable systems, access control, policies
Counterfactuals [160]	Very High (10 s+)	High	Low	High (root cause)	Explaining false positives, refining rules

Key Insights:

- There is a direct trade-off between **explanation fidelity** (e.g., SHAP) and **latency** (e.g., Decision Trees, Grad-CAM).
- Methods with high stability (e.g., SHAP, Decision Trees) are preferable for regulatory compliance and auditable systems.
- For real-time SOC workflows, LIME offers a practical balance of speed and interpretability, though its explanations can be less stable than SHAP.
- Intrinsically interpretable models (e.g., Decision Trees) achieve the highest analyst trust and lowest latency but may not be suitable for all detection tasks due to potential accuracy trade-offs.

5.3 Critical Analysis of Operational Feasibility

While the comparative analysis in (Table 6), highlights the methodological strengths of XAI techniques, their practical deployment in cybersecurity operations is constrained by critical operational factors. Table 9 provides a practitioner-centric evaluation focusing on computational cost, scalability, and real-time performance, which are paramount for Security Operations Center (SOC) environments.

Table 9: Operational feasibility of XAI techniques in cybersecurity

XAI technique	Compute cost	Scalability	Real-time	Best suited for
SHAP [141]	Very high	Low	No (Post-hoc)	Forensic investigation, model debugging, compliance
LIME [163]	Medium	Medium	Limited (Near real-time)	SOC dashboards, minor latency tolerance
Grad-CAM [164]	Low	High	Yes	Real-time, image-based malware detection
Decision Trees [165]	Very low	High	Yes	Real-time access control, policy enforcement
Counterfactuals [166]	High	Low	No	False positives analysis, rule refinement
Anchors [167]	Medium-high	Medium	No	Human-readable rules for fraud detection

Analysis: The trade-off between explanation fidelity and operational cost is evident. High-fidelity methods like SHAP are computationally prohibitive for real-time use, confining them to offline analysis. In contrast, intrinsically interpretable models like Decision Trees or low-cost visual methods like Grad-CAM are viable for real-time, high-throughput environments like IDS/IPS. Practitioners must therefore match the XAI method not only to the security task but also to the operational latency budget of their deployment context. For instance, while SHAP provides the most theoretically sound feature attributions, its computational complexity makes it unsuitable for real-time threat detection where LIME or model-specific methods offer a more practical balance between explainability and performance.

5.4 Empirical Baselines from Canonical Cybersecurity Datasets

To ground the theoretical comparison in empirical evidence, Table 10 synthesizes performance trends of XAI methods on established cybersecurity benchmarks.

Table 10: Empirical XAI performance on cybersecurity benchmarks

Reference	XAI method	Dataset	Task	Key finding
[86]	SHAP	CIC-IDS2017	IDS	Latency > 5 s per sample
[78]	LIME	NSL-KDD	IDS	~1 s latency, unstable
[53]	Grad-CAM	Microsoft Malware	Malware	>95% visual alignment
[76]	LIME	IEEE-CIS Fraud	Fraud	+40% analyst acceptance
[168]	SHAP	EMBER	Malware	Key features identified
[65]	Counterfactuals	CIC-IDS2017	FP Reduction	25% FP reduction

Obviously, the empirical results indicate that explicit trade-offs exist between the operational feasibility and the explanation fidelity. The quality of explanations that are given by techniques like SHAP is quite expensive in terms of computation load, and thus, it can only be used in post hoc analysis, but not real-time detection of threats. The uncertainty of the LIME explanations, though being a problem, may be excused in the areas where the expediency of speed is put in advance of the perfection of consistency. Such findings indicate that such XAI approaches should be aligned with some functions of cybersecurity operations.

6 Challenges and Open Issues

The use of Explainable AI (XAI) in cybersecurity has some major problems and gaps in implementation that must be considered to enable the successful implementation in the field and ensure the usage of this system is safe. The XAI implementation in architecture has four primary concerns, such as the possibility of performance loss due to the decreased level of interpretability, hard XAI systems, the lack of XAI criteria to apply cybersecurity, and the problem of ethics of biases in data sets.

6.1 Performance vs. Interpretability Trade-Off

The primary XAI barrier is that organisations need to establish the tradeoff between their model and solution effectiveness and the amount of explanations. Deep-learning models that include CNNs and transformers operate as dark box systems, since they provide limited transparency about their workings. High-accuracy results in intrusion detection, together with malware classification and anomaly detection, only serve as challenges due to the untraceable nature of these models in security-related applications. Decision trees, together with rule-based systems, are interpretable models that provide explainability but perform poorly due to their limited capability to capture complex patterns within big datasets. Developing hybrid security models remains crucial, since the goal is to achieve optimal results in explainability while keeping performance high and maintaining efficiency standards.

6.2 Adversarial Robustness of XAI Methods

The SHAP and LIME XAI techniques remain susceptible to adversarial attacks, even though they serve as model explanation methods. The interpretability mechanisms used by attackers enable the creation of adversarial examples that affect model output, but still preserve possible explanations. The same sensitive information extraction possibilities exist through model inversion attacks and membership inference attacks because both exploit the explanations produced by XAI systems. XAI methods require immediate development to produce explanations while keeping security vulnerabilities out of the picture. Security-enhanced XAI systems built with adversarial training and differential privacy technology reduce these security risks.

6.3 Lack of Standardization in XAI for Cybersecurity

The security field stands out, as it lacks universal criteria for measuring the effectiveness of the evaluation of the XAI model. The explanation standards mentioned in regulatory panels including GDPR, NIST, and ISO 27001, lack unifying evaluation methods that help assess XAI security models correctly. Researchers are unable to compare different XAI approaches because there are no standardized evaluation metrics for interpretability, robustness, or efficiency. The implementation of a normalised XAI framework in cybersecurity would provide standardised assessments that make it easier to deploy cross-industrially.

Fig. 14 shows the difficulties in implementing Explainable AI (XAI) for cybersecurity. Key issues include Integration Complexity which involves the complexity of integration of XAI systems with existing security frameworks, Interpretability Concerns, which are the challenges associated with understanding AI-driven decisions, Adversarial Attacks, which involve the exploitation of vulnerabilities in XAI models by attackers, Validation Reliability, which is the reliability of the accuracy and trustworthiness of XAI explanations, and Data Privacy Issues which involve the protection of sensitive data within XAI models. These challenges are represented in concentric layers to show their interrelatedness and different levels of importance.

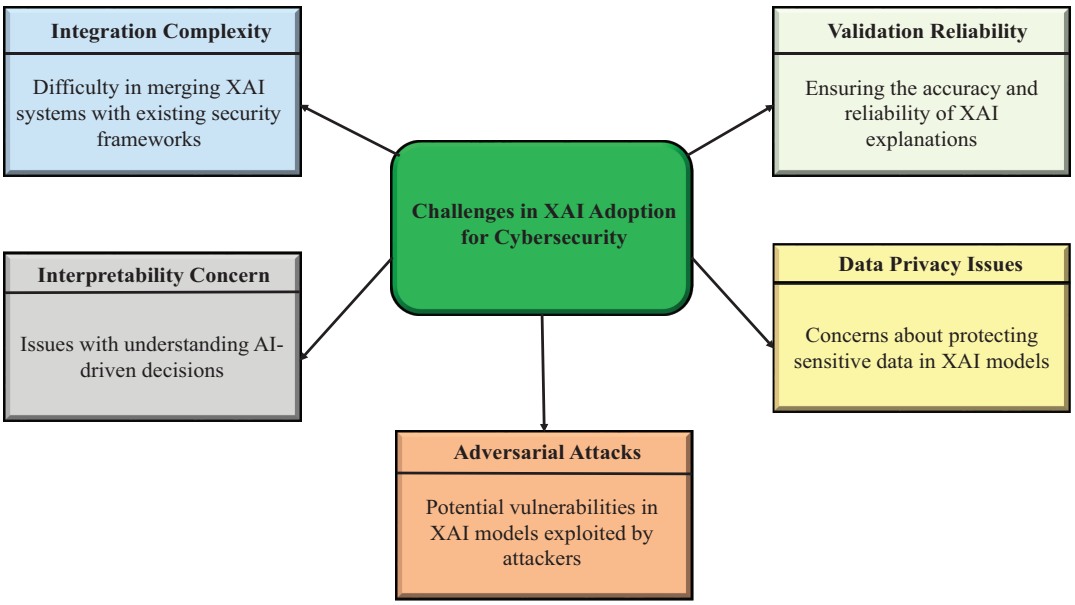

Figure 14: Challenges in XAI adoption for cybersecurity

6.4 Ethical Concerns and Potential Biases in Security Models

Security models operated by AI systems tend to embed biases during operations that produce discrimination in decisions. Faults in the design of the training data selection model and the feature selection processes can cause the system to generate improper diagnostics that divide user groups unequally. An XAI-based fraud detection system could report high fraud risk values to transactions originating from certain geographic locations when the training data sets are biased. Too much openness on the XAI description allows attackers to exploit the knowledge of the model in detection evasion, since they can get to know the vulnerabilities of the detection. Fairness-conscious machine learning models and an adversarial approach of debiasing and ethical auditing of security models are ways to deal with ethical concerns

The barriers and missing answers to the question of applying Explainable AI (XAI) to cybersecurity, besides highlighting the benefits and drawbacks of its use, and methods to solve the issue. The implementation of XAI faces critical hurdles from the exploration of performance against interpretability to robustness against attacks, and inconsistent guidelines alongside ethical dilemmas to security vulnerabilities among other matters that affect feasibility and reliability. These issues impact the effectiveness of AI-driven security systems, such as IDS/IPS and malware detection, by introducing vulnerabilities, biases, and inefficiencies. The proposed solutions include hybrid models, adversarial training, standardized benchmarks, fairness-aware techniques, controlled transparency, and lightweight XAI methods. Open research questions focus on optimizing trade-offs, ensuring robustness, aligning standards, eliminating biases, balancing transparency, and improving computational efficiency for real-time applications as explained in (Table 11).

Table 11: Challenges and open issues in XAI for cybersecurity

Challenge	Description	Security implications	Proposed solutions	Open issues
Performance vs. Interpretability	High-accuracy models lack explainability; interpretable models sacrifice accuracy	Critical systems need both but trade-offs exist	Hybrid models; post-hoc explanations	How to balance accuracy and interpretability? [97].
Adversarial robustness	Explanations enable evasive attacks or data leakage	Attackers bypass detection or extract data	Secure-XAI; adversarial training	Can explainability be preserved while preventing misuse? [169].
Lack of standardization	No unified framework or metrics for XAI in cybersecurity	Inconsistent deployment and evaluation	Regulatory benchmarks; interpretability metrics	Which metrics guide XAI evaluation? [170].
Ethical concerns and bias	Model biases cause unfair outcomes	False denials or missed threats	Fairness-aware modeling; debiasing	Ensure fairness without degrading accuracy? [171]
Transparency vs. Security	Transparency exposes exploitable model logic	Attackers reverse-engineer defenses	Controlled transparency; encrypted XAI	Balance transparency with security? [172].
Computational overhead	XAI resource-intensive, impacts performance	Slows real-time threat analysis	Lightweight XAI; hardware acceleration	Reduce latency while keeping explainability? [173]

7 Future Research Directions

As Explainable AI (XAI) continues to evolve, its integration with cybersecurity remains a critical area of research. Future advancements must focus on improving both interpretability and security without compromising performance. Some of the key research directions include:

7.1 Explainability in AI-Driven Zero Trust Architecture

Zero Trust Architecture (ZTA) is a security paradigm that assumes no implicit trust in any entity, requiring continuous verification of users, devices, and applications [129]. The integration of AI into ZTA can improve adaptive access control, anomaly detection, and risk assessment. However, AI-based ZTA decisions are often opaque, making it difficult for administrators to verify and validate security policies. Future research should focus on:

- Developing interpretable AI-driven ZTA models to improve security decision transparency.
- Using post hoc XAI techniques (e.g., SHAP, LIME) to explain risk-based access control decisions.
- Investigating adversarial threats in XAI-based zero-trust systems and designing countermeasures.

7.2 AI-Driven Automated Cybersecurity Decision-Making with Interpretability Constraints

Automated cybersecurity systems use AI to detect threats, respond to attacks, and enforce security policies in real time [174]. However, the lack of explainability in AI-driven decisions poses significant challenges to security analysts and compliance officers. Research in this area should address the following:

- The trade-off between automated cybersecurity actions in real time and their interpretability.
- Developing hybrid AI frameworks that integrate rule-based reasoning with machine learning to explain decision making.
- Enhance situational awareness by providing interpretable insights into AI-driven security responses.

7.3 Quantum-Safe Explainable AI for Post-Quantum Cybersecurity

The future of quantum computing is posing a dual existential threat to cybersecurity [175]: not only is it compromising the cryptographic mechanisms that keep our data safe, but it is also compromising the integrity of the artificial intelligence models that safeguard our systems. Quantum algorithms, including Shor, are vulnerable to modern-day digital trust, which relies on public-key cryptography. Meanwhile, the machine learning models themselves, which are AI-based, can be the subject of novel quantum-enhanced adversarial attacks, which will be employed to manipulate models in a manner not yet possible using classical computing. Therefore, the post-quantum cybersecurity field cannot simply incorporate explainability as a design factor but instead make it a mandatory design factor. The multidimensional and highly acute future research on the topic which remains new, is multidimensional. One of the significant methods is the design of quantum-safe XAI, constructed on post-quantum cryptographic primitives, i.e., lattice-based cryptography or hash-based cryptography.

This renders all the explanations, as well as the models explaining them, resistant to quantum decryption threats. Furthermore, quantum machine learning (QML) algorithms must be the first thing that we will do to undertake the following activities: threat detection and anomaly analysis. Although QML has enormous computational benefits, the quantum algorithms that it implements are even less understandable to classical deep learning, and new XAI algorithms are required to translate the outputs of quantum-based versions of deep-learning models into the decisions of security experts. Lastly, and perhaps most problematically, researchers will be forced to strike out upon the premises of quantum adversarial attacks of explainable AI models. We are expected to look forward to the creation of quantum computations that might produce more powerful, more efficient, or more sophisticated attacks that will directly attack explanation interfaces to defraud defenders, and then deploy such defences to the quantum era threats directly into the structure of our AI systems.

7.4 Hybrid XAI Models Combining Symbolic AI with Deep Learning for Cybersecurity

One of the most promising means of reaching a high level of explainability in complex cybersecurity settings is the creation of hybrid models, which combine the mutual advantages of symbolic AI and deep learning strategically by merging the two [176]. The present contradiction in AI, in which transparency and rule-based reasoning are provided by symbolic systems with scalability and raw data learning being lacking, and deep learning models have high-dimensional pattern recognition but behave as an inscrutable black box, offers a core conflict between performance and interpretability. Hybrid XAI models aim to resolve this tension by creating a synergistic architecture. In such a framework, a deep learning component could act as a powerful perception engine, processing massive volumes of network traffic, code, or user behavior to identify subtle, statistical anomalies that would elude rigid rule-based systems. The critical innovation is that its findings are then passed to a symbolic AI component, which translates these numerical inferences into a structured, human-comprehensible format using a knowledge base of security rules, logic, and ontologies.

For instance, a deep learning model might flag a sequence of packets as anomalous, and the symbolic reasoner would then map this anomaly to a known attack pattern (e.g., "this matches the data exfiltration profile") or generate a verifiable chain of logic. This hybrid approach directly enhances the interpretability of security decisions by grounding statistical findings in established, rule-based logic that security operators inherently trust. Moreover, it boosts the efficiency and accuracy of threat detection by leveraging both low-level data representations and high-level structured knowledge. Finally, it significantly improves robustness; the symbolic reasoning layer can act as a verification mechanism, cross-checking the deep learning model's outputs for consistency with known security principles and thereby providing a powerful defense against adversarial attacks designed to fool the statistical model alone.

7.5 Federated Learning with Explainability for Collaborative Security

Federated Learning (FL) allows multiple organizations to collaboratively train a model without sharing sensitive data, a significant advantage for threat intelligence [177]. However, explaining the decisions of a model trained on distributed, non-IID data is a major challenge. Future research must develop XAI techniques that provide consistent and faithful explanations across all participants in an FL system, ensuring that local explanations remain valid for the global model and do not inadvertently leak private information.

7.6 Graph Neural Networks for Interpretable Network Traffic Analysis

Cyber threats often manifest as complex relationships between entities (IPs, users, hosts). Graph Neural Networks (GNNs) are exceptionally suited for modeling these relational structures [178]. The key research direction lies in developing explainable GNNs that can highlight subgraphs or node features crucial for a threat classification (e.g., "This cluster of internal hosts communicating with a known malicious IP was the basis for the alert"). This will enable transparent analysis of sophisticated, multi-stage attacks.

7.7 Large Language Models for Explainable Threat Intelligence

The massive unstructured threat data is capable of being analysed by Large Language Models (LLMs) [179]. The future research challenge is to harness their generative capabilities for explanation. An LLM-powered system could, for instance, not only detect an attack but also generate a cohesive, narrative summary in natural language: The alert was triggered due to a PowerShell script (Invoke-ReflectivePEInjection) executed from an unusual user context, a technique associated with the APT29 group, attempting credential dumping. The important open problem is to make certain the factual accuracy and reliability of these generative explanations.

Lastly, we make comparisons between the future research directions of the Explainable AI (XAI) in the field of cybersecurity regarding the fundamental goals, problems, and possible remedies, and the influence of anticipation. The most likely directions are to render explainability more practical in the Zero Trust systems designed by AI, develop explainable automated decisions, make XAI quantum safe to ensure postquantum cybersecurity, and create hybrid models that bridge the gap between explainable AI and deep learning. Adversarial bypass attack, real-time transparency, quantum threats, and Computational complexity are some of the challenges, and the solution for these problems includes post-hoc XAI techniques, hybrid AI integration, quantum-resilient models, and neurosymbolic architectures. These innovations will enhance trust, flexibility, responsibility, and resilience of AI-based cybersecurity solutions, making sure that they will be effective in the face of developing threats as explained in (Table 12).

Table 12: Comparison of future research directions in explainable AI (XAI) for cybersecurity

Research direction	Core objective	Key challenges	Potential solutions	Expected impact
Explainability in AI-driven zero trust	Enhance transparency of adaptive access control	Opaque AI policy decisions; adversarial exploitation of explanations	Post-hoc XAI (SHAP, LIME); inherently interpretable risk models	Increased administrator trust and verifiable, adaptive security policies
AI-driven automated decision-making	Achieve verifiable and accountable automation	Real-time performance vs. explainability trade-off; audit trail generation	Hybrid AI (rule-based + ML); intrinsic explainability constraints	Faster, trusted, and legally compliant automated responses
Quantum-Safe XAI	Develop interpretable AI models resilient to quantum threats	Quantum attacks on crypto models; interpretability of quantum ML	Post-quantum cryptography; quantum-aware XAI techniques	Future-proof, trustworthy AI security in the post-quantum era
Hybrid XAI models	Fuse neural pattern recognition with symbolic reasoning	Architectural complexity; performance overhead; knowledge engineering	Neuro-symbolic AI; iterative learning between components	Enhanced threat reasoning, robustness, and analyst trust
Federated learning with XAI	Provide consistent explanations in collaborative, privacy-preserving training	Non-IID data skewing explanations; explanation-driven data leakage	Federated XAI algorithms; privacy-preserving explanation generation	Collaborative defense without sacrificing explainability or privacy

(Continued)

Table 12 (continued)

Research direction	Core objective	Key challenges	Potential solutions	Expected impact
Explainable graph neural networks	Identify malicious relational structures in network data	Explaining complex graph inferences; scalability of explanation methods	GNN explainers (e.g., GNNExplainer); subgraph highlighting	Transparent analysis of sophisticated, multi-stage cyber-attacks
LLMs for threat intelligence	Generate coherent, narrative explanations from unstructured data	Hallucination and factual inaccuracy; computational cost	Retrieval-Augmented Generation (RAG); fine-tuning on security data	Rapid comprehension of threats and intuitive situational awareness

8 Conclusion

This study has examined the vital role of Explainable AI in cybersecurity as AI-driven systems become more prevalent. Our analysis detailed various XAI techniques and their applications in intrusion detection, malware classification, fraud detection, and insider threat mitigation, demonstrating how transparency enhances threat detection and response capabilities. The research highlighted how explainability helps meet regulatory requirements under frameworks like GDPR, NIST, and ISO 27001. However, we identified the fundamental challenge of balancing interpretability with security, recognizing that increased transparency may potentially expose model vulnerabilities to adversarial exploitation. The paper emphasized developing Secure-XAI frameworks that incorporate adversarial training and privacy-preserving techniques like differential privacy.

Several critical challenges emerged from our analysis, including performance-interpretability trade-offs, adversarial robustness concerns, standardization gaps, and ethical considerations regarding algorithmic bias. Future research should focus on creating hybrid XAI models that integrate symbolic and deep learning approaches, incorporating explainability into post-quantum cybersecurity frameworks, and developing AI-driven zero-trust architectures with built-in explainability constraints. In conclusion, while XAI significantly advances trust, transparency, and regulatory compliance in cybersecurity systems, it must be carefully engineered to mitigate security risks and support the development of resilient, accountable security infrastructures.

Acknowledgement: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2025-02-01395).

Funding Statement: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2025-02-01395).

Author Contributions: Khulud Salem Alshudukhi and Mamoona Humayun were responsible for the conceptualization and supervision of the study. Sijjad Ali and Omar Alruwaili contributed to the methodology, data curation, and formal analysis. Khulud Salem Alshudukhi and Sijjad Ali prepared the original draft and conducted the investigation. Mamoona Humayun and Omar Alruwaili were involved in validation, review, and editing of the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Ali S, Ali A, Uzair M, Amir H, Bari RZA, Sharif H, et al. Empowering cybersecurity: CyberShield AI advanced integration of machine learning and deep learning for dynamic ransomware detection. In: International Conference on Deep Learning Theory and Applications. Cham, Switzerland: Springer; 2024. p. 95–117.
2. Ali S, Wang J, Leung VCM. AI-driven fusion with cybersecurity: exploring current trends, advanced techniques, future directions, and policy implications for evolving paradigms-a comprehensive review. Inf Fusion. 2025;118(3):102922. doi:10.1016/j.inffus.2024.102922.
3. Ali S, Wang J, Leung VCM, Bashir F, Bhatti UA, Wadho SA, et al. CLDM-MMNNs: cross-layer defense mechanisms through multi-modal neural networks fusion for end-to-end cybersecurity—issues, challenges, and future directions. Inf Fusion. 2025;122(12):103222. doi:10.1016/j.inffus.2025.103222.
4. Mohamed N. Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. Knowl Inform Syst. 2025;67(8):6969–7055. doi:10.1007/s10115-025-02429-y.
5. Tanikonda A, Peddinti SR, Pandey BK, Katragadda SR. Advanced AI-driven cybersecurity solutions for proactive threat detection and response in complex ecosystems. J Sci Technol. 2022;3(1):196–218.
6. Kshetri N. Transforming cybersecurity with agentic AI to combat emerging cyber threats. Telecommun Policy. 2025;49(6):102976. doi:10.1016/j.telpol.2025.102976.
7. Brożek B, Furman M, Jakubiec M, Kucharzyk B. The black box problem revisited. Real and imaginary challenges for automated legal decision making. Artif Intell Law. 2024;32(2):427–40. doi:10.1007/s10506-023-09356-9.
8. Von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. Philos Technol. 2021;34(4):1607–22.
9. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput. 2024;16(1):45–74. doi:10.1007/s12559-023-10179-8.
10. Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Del Ser J, et al. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. Inf Fusion. 2024;106(3):102301. doi:10.1016/j.inffus.2024.102301.
11. Chinnaraju A. Explainable AI (XAI) for trustworthy and transparent decision-making: a theoretical framework for AI interpretability. World J Adv Eng Technol Sci. 2025;14(3):170–207. doi:10.30574/wjaets.2025.14.3.0106.
12. Sarker IH. AI-driven cybersecurity and threat intelligence: cyber automation, intelligent decision-making and explainability. Cham, Switzerland: Springer Nature; 2024.
13. Singh H. Securing high-stakes digital transactions: a comprehensive study on cybersecurity and data privacy in financial institutions. SSRN Electron J. 2025;219–29. doi:10.2139/ssrn.5267850.
14. Ijiga OM, Idoko IP, Ebiega GI, Olajide FI, Olatunde TI, Ukaegbu C. Harnessing adversarial machine learning for advanced threat detection: AI-driven strategies in cybersecurity risk assessment and fraud prevention. J Sci Technol. 2024;11(1):1–24. doi:10.53022/oarjst.2024.11.1.0060.
15. European Union. General data protection regulation [Internet]. 2020. [cited 2025 Aug 4]. Available from: https://gdpr-info.eu/.
16. Bonta R. California consumer privacy act (CCPA) [Internet]. Retrieved from State of California Department of Justice. 2022 [cited 2025 Aug 4]. Available from: https://www.oag.ca.gov/privacy/ccpa.
17. Danish M. Enhancing cyber security through predictive analytics: real-time threat detection and response. arXiv:2407.10864. 2024.
18. Sundaramurthy SK, Ravichandran N, Inaganti AC, Muppalaneni R. AI-driven threat detection: leveraging machine learning for real-time cybersecurity in cloud environments. Artif Intell Mach Learn Rev. 2025;6(1):23–43. doi:10.69987/AIMLR.2025.60104.

19. Jonas D, Yusuf NA, Zahra ARA. Enhancing security frameworks with artificial intelligence in cybersecurity. Int Trans Educ Technol (ITEE). 2023;2(1):83–91. doi:10.33050/itee.v2i1.428.
20. Liu YF, Li SC, Wang XH, Xu L. A review of hybrid cyber threats modelling and detection using artificial intelligence in IIoT. Comput Model Eng Sci. 2024;140(2):1233–61. doi:10.32604/cmes.2024.046473.
21. Radanliev P, Santos O. Adversarial attacks can deceive AI systems, leading to misclassification or incorrect decisions. Preprints. 2023. doi:10.20944/preprints202309.2064.v1.
22. Blauth TF, Gstrein OJ, Zwitter A. Artificial intelligence crime: an overview of malicious use and abuse of AI. IEEE Access. 2022;10:77110–22. doi:10.1109/access.2022.3191790.
23. Moustafa N, Koroniotis N, Keshk M, Zomaya AY, Tari Z. Explainable intrusion detection for cyber defences in the internet of things: opportunities and solutions. IEEE Commun Surv Tutor. 2023;25(3):1775–807.
24. Sarker IH, Janicke H, Mohsin A, Gill A, Maglaras L. Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: methods, taxonomy, challenges and prospects. ICT Express. 2024;10(4):935–958. doi:10.1016/j.icte.2024.05.007.
25. Srivastava G, Jhaveri RH, Bhattacharya S, Pandya S, Maddikunta PKR, Yenduri G, et al. XAI for cybersecurity: state of the art, challenges, open issues and future directions. arXiv:2206.03585. 2022.
26. Khan N, Ahmad K, Tamimi AA, Alani MM, Bermak A, Khalil I. Explainable AI-based intrusion detection system for industry 5.0: an overview of the literature, associated challenges, the existing solutions, and potential research directions. arXiv:2408.03335. 2024.
27. Mangalathu S, Hwang SH, Jeon JS. Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. Eng Struct. 2020;219(6):110927. doi:10.1016/j.engstruct.2020.110927.
28. Biecek P, Burzykowski T. Local interpretable model-agnostic explanations (LIME). In: Explanatory model analysis: explore, explain and examine predictive models. Boca Raton, FL, USA: Chapman & Hall/CRC; 2021. p. 107–24.
29. Hayes-Roth F. Rule-based systems. Commun ACM. 1985;28(9):921–32.
30. De Ville B. Decision trees. Wiley Interdis Rev Computat Stat. 2013;5(6):448–55.
31. Chambers JM. Linear models. In: Statistical models in S. Brighton, UK: Routledge; 2017. p. 95–144.
32. Dif MM, Bouchiha MA, Korba AA, Ghamri-Doudane Y. Towards trustworthy agentic IoEV: AI agents for explainable cyberthreat mitigation and state analytics. In: 2025 IEEE 50th Conference on Local Computer Networks (LCN). New York, NY, USA: IEEE; 2025. p. 1–10.
33. Faiz A, Khan I, Abdullah H, Jumani A, Talpur MRH, Aurangzaib R. Designing hybrid artificial intelligence systems: integrating symbolic reasoning and deep learning for real-time, context-aware decision making in complex environments. Global Res J Nat Sci Technol. 2025;3(2):467–503. doi:10.53762/grjnst.03.02.20.
34. Praharaj L, Gupta M, Gupta D. Explainability-aware adversarial threats and mitigation in federated learning based anomaly detection for cooperative smart farming. In: 2025 10th International Conference on Fog and Mobile Edge Computing (FMEC). New York, NY, USA: IEEE; 2025. p. 186–93.
35. Emiroğlu BG. AI-driven threat detection and response systems: enhancing cybersecurity in the digital era. In: Challenges and solutions for cybersecurity and adversarial machine learning. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 227–70. doi:10.4018/979-8-3373-2200-1.ch008.
36. Akhtar MAK, Kumar M, Nayyar A. Privacy and security considerations in explainable AI. In: Towards ethical and socially responsible explainable AI: challenges and opportunities. Cham, Switzerland: Springer; 2024. p. 193–226. doi:10.1007/978-3-031-66489-2_7.
37. Sharma A, Sharma N, Jain A. XAI-driven approaches for ensuring security and data protection in IoT. In: The next generation innovation in IoT and cloud computing with applications. Boca Raton, FL, USA: CRC Press; 2024. p. 110–30.
38. Shamoo Y. Adversarial attacks and defense mechanisms in the age of quantum computing. In: Leveraging large language models for quantum-aware cybersecurity. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 301–44.

39. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fusion. 2020;58(3):82–115. doi:10.1016/j.inffus.2019.12.012.
40. Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev. 2022;55(5):3503–68. doi:10.1007/s10462-021-10088-y.
41. Carabantes M. Black-box artificial intelligence: an epistemological and critical analysis. AI Soc. 2020;35(2):309–17.
42. Agrawal G. Trust, and transparency in AI systems from the perspective of public policy: elevating ethical standards. In: AI healthcare applications and security, ethical, and legal considerations. Hershey, PA, USA: IGI Global; 2024. p. 148–62. doi:10.4018/979-8-3693-7452-8.ch009.
43. Habbal A, Ali MK, Abuzaraida MA. Artificial intelligence trust, risk and security management (AI trism): frameworks, applications, challenges and future research directions. Expert Syst Appl. 2024;240(4):122442. doi:10.1016/j.eswa.2023.122442.
44. Williamson SM, Prybutok V. Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in AI-driven healthcare. Appl Sci. 2024;14(2):675. doi:10.3390/app14020675.
45. Ali S, Talpur DB, Abro A, Alshudukhi KSS, Alwakid GN, Humayun M, et al. Security and privacy in multi-cloud and hybrid cloud environments: challenges, strategies, and future directions. Comput Secur. 2025;157(13s):104599. doi:10.1016/j.cose.2025.104599.
46. Senevirathna T, La VH, Marchal S, Siniarski B, Liyanage M, Wang S. A survey on XAI for 5G and beyond security: technical aspects, challenges and research directions. IEEE Commun Surv Tutor. 2025;27(2):941–73.
47. Slack D, Hilgard S, Singh S, Lakkaraju H. Reliable post hoc explanations: Modeling uncertainty in explainability. Adv Neural Inf Process Syst. 2021;34(2021):9391–404.
48. Vale D, El-Sharif A, Ali M. Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. AI Ethics. 2022;2(4):815–26. doi:10.1007/s43681-022-00142-y.
49. Mi JX, Jiang X, Luo L, Gao Y. Toward explainable artificial intelligence: a survey and overview on their intrinsic properties. Neurocomputing. 2024;563(2):126919. doi:10.1016/j.neucom.2023.126919.
50. Ai Q, Narayanan RL. Model-agnostic vs. model-intrinsic interpretability for explainable product search. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management; 2021 Nov 1–5; Online. p. 5–15.
51. Chanajitt R. Machine learning approaches for malware classification based on hybrid artefacts [dissertation]. Hamilton, New Zealand: The University of Waikato; 2023.
52. Tursunalieva A, Alexander DL, Dunne R, Li J, Riera L, Zhao Y. Making sense of machine learning: a review of interpretation techniques and their applications. Appl Sci. 2024;14(2):496. doi:10.3390/app14020496.
53. Naeem H, Alshammari BM, Ullah F. Explainable artificial intelligence-based IoT device malware detection mechanism using image visualization and fine-tuned CNN-based transfer learning model. Comput Intell Neurosci. 2022;2022(1):7671967. doi:10.1155/2022/7671967.
54. Kobs K. Think outside the black box: model-agnostic deep learning with domain knowledge [dissertation]. Würzburg, Germany: Universität Würzburg; 2024.
55. Kanti PK, Sharma P, Wanatasanappan VV, Said NM. Explainable machine learning techniques for hybrid nanofluids transport characteristics: an evaluation of shapley additive and local interpretable model-agnostic explanations. J Therm Anal Calorim. 2024;149(21):11599–618. doi:10.1007/s10973-024-13639-x.
56. Assegie TA. Evaluation of local interpretable model-agnostic explanation and shapley additive explanation for chronic heart disease detection. Proc Eng Technol Innov. 2023;23:48–59. doi:10.46604/peti.2023.10101.
57. Lamsaf A, Carrilho R, Neves JC, Proença H. Causality, machine learning, and feature selection: a survey. Sensors. 2025;25(8):2373. doi:10.3390/s25082373.
58. Narayanan A, Bergen KJ. Prototype-based methods in explainable AI and emerging opportunities in the geosciences. arXiv:2410.19856. 2024.
59. Yeh CK, Kim B, Ravikumar P. Human-centered concept explanations for neural networks. In: Neuro-symbolic artificial intelligence: the state of the art. Amsterdam, The Netherlands: IOS Press; 2021. p. 337–52. doi:10.3233/faia210362.

60. Alharbi A, Seh AH, Alosaimi W, Alyami H, Agrawal A, Kumar R, et al. Analyzing the impact of cyber security related attributes for intrusion detection systems. Sustainability. 2021;13(22):12337. doi:10.3390/su132212337.
61. Sokol K, Hepburn A, Santos-Rodriguez R, Flach P. bLIMEy: surrogate prediction explanations beyond LIME. arXiv:1910.13016. 2019.
62. Capuano N, Fenza G, Loia V, Stanzione C. Explainable artificial intelligence in cybersecurity: a survey. IEEE Access. 2022;10(2):93575–600. doi:10.1109/access.2022.3204171.
63. Fu R, Hu Q, Dong X, Guo Y, Gao Y, Li B. Axiom-based GRAD-CAM: towards accurate visualization and explanation of CNNs. arXiv:2008.02312. 2020.
64. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy. 2020;23(1):18. doi:10.3390/e23010018.
65. Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J. Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inform Fusion. 2022;81(10):59–83. doi:10.1016/j.inffus.2021.11.003.
66. Kocher G, Kumar G. Machine learning and deep learning methods for intrusion detection systems: recent developments and challenges. Soft Comput. 2021;25(15):9731–63. doi:10.1007/s00500-021-05893-0.
67. Chalapathy R, Chawla S. Deep learning for anomaly detection: a survey. arXiv:1901.03407. 2019.
68. Ullah I, Rios A, Gala V, Mckeever S. Explaining deep learning models for tabular data using layer-wise relevance propagation. Appl Sci. 2021;12(1):136. doi:10.3390/app12010136.
69. Egele M, Scholte T, Kirda E, Kruegel C. A survey on automated dynamic malware-analysis techniques and tools. ACM Comput Surv (CSUR). 2008;44(2):1–42. doi:10.1145/2089125.2089126.
70. Kumar G, Alqahtani H. Machine learning techniques for intrusion detection systems in SDN: recent advances, challenges and future directions. Comput Model Eng Sci. 2023;134(1):89–119. doi:10.32604/cmes.2022.020724.
71. Saqib M, Mahdavifar S, Fung BC, Charland P. A comprehensive analysis of explainable AI for malware hunting. ACM Comput Surv. 2024;56(12):1–40. doi:10.1145/3677374.
72. Biswas B, Mukhopadhyay A, Kumar A, Delen D. A hybrid framework using explainable AI (XAI) in cyber-risk management for defence and recovery against phishing attacks. Decis Support Syst. 2024;177(12):114102. doi:10.1016/j.dss.2023.114102.
73. Alsuqayh N, Mirza A, Alhogail A. A phishing website detection system based on hybrid feature engineering with SHAP explainable artificial intelligence technique. In: International Conference on Web Information Systems Engineering. Cham, Switzerland: Springer; 2024. p. 3–17.
74. Leonard LC. Web-based behavioral modeling for continuous user authentication (CUA). In: Advances in computers. Vol. 105. Amsterdam, The Netherlands: Elsevier; 2017. p. 1–44. doi:10.1016/bs.adcom.2016.12.001.
75. Hassan M, Aziz LAR, Andriansyah Y. The role of artificial intelligence in modern banking: an exploration of AI-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance. Rev Contemp Bus Anal. 2023;6(1):110–32.
76. Rani S, Mittal A. Securing digital payments a comprehensive analysis of AI driven fraud detection with real time transaction monitoring and anomaly detection. In: 2023 6th International Conference on Contemporary Computing and Informatics (IC3I). New York, NY, USA: IEEE; 2023. p. 2345–9.
77. Shayea GG, Zabil MHM, Habeeb MA, Khaleel YL, Albahri A. Strategies for protection against adversarial attacks in AI models: an in-depth review. J Intell Syst. 2025;34(1):20240277. doi:10.1515/jisys-2024-0277.
78. Kuppa A, Le-Khac NA. Adversarial XAI methods in cybersecurity. IEEE Trans Inform Forens Secu. 2021;16:4924–38. doi:10.1109/tifs.2021.3117075.
79. Nwakanma CI, Ahakonye LAC, Njoku JN, Odirichukwu JC, Okolie SA, Uzondu C, et al. Explainable artificial intelligence (XAI) for intrusion detection and mitigation in intelligent connected vehicles: a review. Appl Sci. 2023;13(3):1252. doi:10.3390/app13031252.
80. Georgiadou A, Mouzakitis S, Askounis D. Detecting insider threat via a cyber-security culture framework. J Comput Inform Syst. 2022;62(4):706–16. doi:10.1080/08874417.2021.1903367.

81. Al-Mhiqani MN, Ahmad R, Zainal Abidin Z, Yassin W, Hassan A, Abdulkareem KH, et al. A review of insider threat detection: classification, machine learning techniques, datasets, open challenges, and recommendations. Appl Sci. 2020;10(15):5208. doi:10.3390/app10155208.
82. Bécue A, Praça I, Gama J. Artificial intelligence, cyber-threats and Industry 4.0: challenges and opportunities. Artif Intell Rev. 2021;54(5):3849–86. doi:10.1007/s10462-020-09942-2.
83. Vinayakumar V, Alazab M, Soman MP, Srinivasan S, Venkatraman S, Pham QV, et al. Deep learning for cyber security applications: a comprehensive survey. TechRxiv. 2021. doi:10.36227/techrxiv.16748161.v1.
84. Dargan S, Kumar M, Ayyagari MR, Kumar G. A survey of deep learning and its applications: a new paradigm to machine learning. Arch Comput Methods Eng. 2020;27(4):1071–92. doi:10.1007/s11831-019-09344-w.
85. Saeed W, Omlin C. Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl Based Syst. 2023;263(3):110273. doi:10.1016/j.knosys.2023.110273.
86. Neupane S, Ables J, Anderson W, Mittal S, Rahimi S, Banicescu I, et al. Explainable intrusion detection systems (X-IDS): a survey of current methods, challenges, and opportunities. IEEE Access. 2022;10(7):112392–415. doi:10.1109/access.2022.3216617.
87. Dopamu O, Adesiyan J, Oke F. Artificial intelligence and US financial institutions: review of AI-assisted regulatory compliance for cybersecurity. World J Adv Res Rev. 2024;21(3):964–79. doi:10.30574/wjarr.2024.21.3.0791.
88. Mohitkar C, Lakshmi D. Explainable AI for transparent cyber-risk assessment and decision-making. In: Machine intelligence applications in cyber-risk management. Hershey, PA, USA: IGI Global Scientific Publishing; 2025. p. 219–46. doi:10.4018/979-8-3693-7540-2.ch010.
89. Patil D. Explainable Artificial Intelligence (XAI) for Industry Applications: Enhancing Transparency, Trust, and Informed Decision-Making in Business Operation. [cited 29 October 2025]. Available from: https://ssrn.com/abstract=5057402.
90. State AI, Morosanu GA, Rata LA, Geru M. Aspects regarding cybersecurity developments on SaaS software platforms. EIRP Proc. 2023;18(1):128–46.
91. Kaul D. AI-powered autonomous compliance management for multi-region data governance in cloud deployments. J Curr Sci Res Rev. 2024;2(3):82–98.
92. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, de Prado ML, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. Inf Fusion. 2023;99(2):101896. doi:10.1016/j.inffus.2023.101896.
93. Rane N, Paramesha M, Rane J. Trustworthy artificial intelligence: enhancing trustworthiness through explainable AI (XAI). SSRN Electron J. 2024;52(7):17. doi:10.2139/ssrn.4880090.
94. Zhao Z, Chen K. Post-hoc split-point self-consistency verification for efficient, unified quantification of aleatoric and epistemic uncertainty in deep learning. arXiv:2509.13262. 2025.
95. Akgündoğdu A, Çelikbaş Ş. Explainable deep learning framework for brain tumor detection: integrating LIME, Grad-CAM, and SHAP for enhanced accuracy. Med Eng Phys. 2025;144(3):104405. doi:10.1016/j.medengphy.2025.104405.
96. Yang S, Vinuesa R, Kang N. Model-agnostic AI framework with explicit time integration for long-term fluid dynamics prediction. J Comput Des Eng. 2025;12(10):qwaf099. doi:10.1093/jcde/qwaf099.
97. Assis A, Dantas J, Andrade E. The performance-interpretability trade-off: a comparative study of machine learning models. J Reliab Intell Environ. 2025;11(1):1. doi:10.1007/s40860-024-00240-0.
98. Xiao Y, Shao H, Wang J, Cai B, Liu B. From deterministic to Bayesian: adapting pre-trained models for human-computer collaborative fault diagnosis via post-hoc uncertainty. J Ind Inf Integr. 2025;47(6):100921. doi:10.1016/j.jii.2025.100921.
99. Juliussen BA. The right to an explanation under the GDPR and the AI act. In: International Conference on Multimedia Modeling. Cham, Switzerland: Springer; 2025. p. 184–97.
100. Sachs DS. Behavioral insights in cybersecurity: a guide to digital human factors. Boca Raton, FL, USA: CRC Press; 2025.

101. Moamin SA, Abdulhameed MK, Al-Amri RM, Radhi AD, Naser RK, Pheng LG. Artificial intelligence in malware and network intrusion detection: a comprehensive survey of techniques, datasets, challenges, and future directions. Babylonian J Artif Intell. 2025;2025:77–98. doi:10.58496/bjai/2025/008.
102. Venkatram M. Leveraging AI models for proactive problem detection, investigation, and root cause analysis in enterprise IT infrastructure. Int J Sci Res (IJSR). 2025;14(6):571–7. doi: 10.21275/SR25607232609.
103. Holder E, Wang N. Explainable artificial intelligence (XAI) interactively working with humans as a junior cyber analyst. Human-Intell Syst Integ. 2021;3(2):139–53. doi:10.1007/s42454-021-00039-x.
104. Syed SA. Adversarial AI and cybersecurity: defending against AI-powered cyber threats. Iconic Res Eng J. 2025;8(9):1030–41.
105. AllahRakha N. Cybersecurity regulations for protection and safeguarding digital assets (Data) in today's worlds. Lex Scientia Law Rev. 2024;8(1):405–32. doi:10.15294/lslr.v8i1.2081.
106. Chamola V, Hassija V, Sulthana AR, Ghosh D, Dhingra D, Sikdar B. A review of trustworthy and explainable artificial intelligence (XAI). IEEE Access. 2023;11:78994–9015. doi:10.1109/access.2023.3294569.
107. Tyagi AK, Kumari S, Richa. Artificial intelligence-based cyber security and digital forensics: a review. In: Artificial intelligence-enabled digital twin for smart manufacturing. Hoboken, NJ, USA: Wiley-Scrivener; 2024. p. 391–419. doi:10.1002/9781394303601.ch18.
108. Mughal AA. The art of cybersecurity: defense in depth strategy for robust protection. Int J Intell Autom Comput. 2018;1(1):1–20.
109. Faheem MA, Kakolu S, Aslam M. The role of explainable AI in cybersecurity: improving analyst trust in automated threat assessment systems. Iconic Res Eng J. 2022;6(4):173–82.
110. Mia M, Pritom MMA, Islam T, Hasan K. Visually analyze shap plots to diagnose misclassifications in ml-based intrusion detection. In: 2024 IEEE International Conference on Data Mining Workshops (ICDMW). New York, NY, USA: IEEE; 2024. p. 632–41.
111. Kim S, Hwang C, Lee T. Anomaly based unknown intrusion detection in endpoint environments. Electronics. 2020;9(6):1022. doi:10.3390/electronics9061022.
112. Wan X, Xue G, Zhong Y, Wang Z. Separating prediction and explanation: an approach based on explainable artificial intelligence for analyzing network intrusion. J Netw Syst Manage. 2025;33(1):16. doi:10.1007/s10922-024-09891-z.
113. Bellegdi S, Selamat A, Olatunji SO, Fujita H, Krejcar O. Explainable machine learning for intrusion detection. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Cham, Switzerland: Springer; 2024. p. 122–34.
114. Yapar O. Explainable AI in national security: enhancing trust and accountability. Int J Emerg Technol Innovat Res. 2023;10(12):pph691-h709.
115. Baniecki H, Biecek P. Adversarial attacks and defenses in explainable artificial intelligence: a survey. Inf Fusion. 2024;107(5):102303. doi:10.1016/j.inffus.2024.102303.
116. Zhao X, Zhang W, Xiao X, Lim B. Exploiting explanations for model inversion attacks. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada. p. 682–92.
117. Fang H, Qiu Y, Yu H, Yu W, Kong J, Chong B, et al. Privacy leakage on DNNs: a survey of model inversion attacks and defenses. arXiv:2402.04013. 2024.
118. Jana S, Shmatikov V. Abusing file processing in malware detectors for fun and profit. In: 2012 IEEE Symposium on Security and Privacy. New York, NY, USA: IEEE; 2012. p. 80–94.
119. Nissim N, Cohen A, Glezer C, Elovici Y. Detection of malicious PDF files and directions for enhancements: a state-of-the art survey. Comput Secur. 2015;48:246–66.
120. Garcia W, Choi JI, Adari SK, Jha S, Butler KR. Explainable black-box attacks against model-based authentication. arXiv:1810.00024. 2018.
121. Aminu M, Akinsanya A, Dako DA, Oyedokun O. Enhancing cyber threat detection through real-time threat intelligence and adaptive defense mechanisms. Int J Comput Appl Technol Res. 2024;13(8):11–27. doi:10.7753/IJCATR1308.1002.

122. Plachkinova M, Knapp K. Least privilege across people, process, and technology: endpoint security framework. J Comput Inform Syst. 2023;63(5):1153–65. doi:10.1080/08874417.2022.2128937.
123. Zhao H, Chi J, Tian Y, Gordon GJ. Trade-offs and guarantees of adversarial representation learning for information obfuscation. In: Advances in neural information processing systems. Washington, DC, USA: AAAI Press; 2020. Vol. 33, p. 9485–96. doi:10.5555/3495724.3496519.
124. Wang Y, Zhang T, Guo X, Shen Z. Gradient based feature attribution in explainable AI: a technical review. arXiv:2403.10415. 2024.
125. Vasan D, Hammoudeh M. Enhancing resilience against adversarial attacks in medical imaging using advanced feature transformation training. Current Opin Biomed Eng. 2024;32:100561. doi:10.1016/j.cobme.2024.100561.
126. Ezzeddine F. Privacy implications of explainable AI in data-driven systems. arXiv:2406.15789. 2024.
127. Hassan MU, Rehmani MH, Chen J. Differential privacy techniques for cyber physical systems: a survey. IEEE Commun Surv Tutor. 2019;22(1):746–89.
128. Hallaji E, Razavi-Far R, Saif M, Wang B, Yang Q. Decentralized federated learning: a survey on security and privacy. IEEE Trans Big Data. 2024;10(2):194–213. doi:10.1109/tbdata.2024.3362191.
129. Ali W, Zhou X, Shao J. Privacy-preserved and responsible recommenders: from conventional defense to federated learning and blockchain. ACM Comput Surv. 2025;57(5):1–35. doi:10.1145/3708982.
130. Manzoor HU, Shabbir A, Chen A, Flynn D, Zoha A. A survey of security strategies in federated learning: defending models, data, and privacy. Future Internet. 2024;16(10):374. doi:10.3390/fi16100374.
131. Rawal A, McCoy J, Rawat DB, Sadler BM, Amant RS. Recent advances in trustworthy explainable artificial intelligence: status, challenges, and perspectives. IEEE Trans Artif Intell. 2021;3(6):852–66. doi:10.36227/techrxiv.17054396.v1.
132. Hong I, Choi C. Knowledge distillation vulnerability of DeiT through CNN adversarial attack. Neural Comput Appl. 2025;37(12):7721–31. doi:10.1007/s00521-023-09412-0.
133. Yang W, Wang S, Wu D, Cai T, Zhu Y, Wei S, et al. Deep learning model inversion attacks and defenses: a comprehensive survey. Artif Intell Rev. 2025;58(8):242. doi:10.1007/s10462-025-11248-0.
134. Meeus M, Shilov I, Jain S, Faysse M, Rei M, de Montjoye YA. Sok: membership inference attacks on llms are rushing nowhere (and how to fix it). In: 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). New York, NY, USA: IEEE; 2025. p. 385–401.
135. Muia M, Kamiri J. Explainable artificial intelligence: a comprehensive review of techniques, applications, and emerging trends. Int J Sci Res Comp Sci Eng. 2025;13(4):57–68. doi:10.26438/ijsrcse.v13i4.740.
136. Bounceur A, Berkani AS, Moumen H, Benharzallah S. The transparency challenge in blockchain-enabled sustainable development goals applications: exploring privacy-preserving techniques and emerging platforms. IEEE Access. 2025;13(2):81769–93. doi:10.1109/access.2025.3567341.
137. Cinà AE, Rony J, Pintor M, Demetrio L, Demontis A, Biggio B, et al. Attackbench: evaluating gradient-based attacks for adversarial examples. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press; 2025. Vol. 39, p. 2600–8.
138. Allana S, Kankanhalli M, Dara R. Privacy risks and preservation methods in explainable artificial intelligence: a scoping review. arXiv:2505.02828. 2025.
139. Leiva V, Castro C. Artificial intelligence and blockchain in clinical trials: enhancing data governance efficiency, integrity, and transparency. Bioanalysis. 2025;17(3):161–76. doi:10.1080/17576180.2025.2452774.
140. Kruschel S, Hambauer N, Weinzierl S, Zilker S, Kraus M, Zschech P. Challenging the performance-interpretability trade-off: an evaluation of interpretable machine learning models. Bus Inform Syst Eng. 2025; doi:10.1007/s12599-024-00922-2.
141. Salih AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. Adv Intell Syst. 2025;7(1):2400304. doi:10.1002/aisy.202400304.
142. Nematzadeh H, García-Nieto J, Hurtado S, Aldana-Montes JF, Navas-Delgado I. Model-agnostic local explanation: multi-objective genetic algorithm explainer. Eng Appl Artif Intell. 2025;139(2):109628. doi:10.1016/j.engappai.2024.109628.

143. Mohammadagha M. Hyperparameter optimization strategies for tree-based machine learning models prediction: a comparative study of adaboost, decision trees, and random forest. Dec Trees Random Forest. 2025. doi:10.21203/rs.3.rs-6968809/v1.
144. Mijangos V, Gutierrez-Vasques X, Arriola VE, Rodríguez-Domínguez U, Cervantes A, Almanzara JL. Relational inductive biases on attention mechanisms. arXiv:2507.04117. 2025.
145. Liu J, Wu X, Liu S, Gong S. Model-agnostic counterfactual explanation: a feature weights-based comprehensive causal multi-objective counterfactual framework. Expert Syst Appl. 2025;266(5):126063. doi:10.1016/j.eswa.2024.126063.
146. Kerrigan D, Barr B, Bertini E. PDPilot: exploring partial dependence plots through ranking, filtering, and clustering. IEEE Trans Visualiz Comput Graph. 2025;31(10):7377–90. doi:10.1109/tvcg.2025.3545025.
147. Trustworthy A. Explainability in fraud detection: trustworthy AI and pattern detection. In: Artificial Intelligence for Global Security: First IFIP WG 12.13 International Conference, AI4GS 2024, Paris, France, 2024 November 19. Cham, Switzerland: Springer; 2025. Vol. 743, p. 178.
148. Müller R. How explainable AI affects human performance: a systematic review of the behavioural consequences of saliency maps. Int J Human-Comput Interac. 2025;41(4):2020–51. doi:10.1080/10447318.2024.2381929.
149. Hruschka ER. Rule extraction from neural networks in data mining applications. WIT Trans Inf Commun Technol. 2025. doi:10.2495/DATA980211.
150. Samadian D, Muhit IB, Dawood N. Application of data-driven surrogate models in structural engineering: a literature review. Arch Comput Methods Eng. 2025;32(2):735–84. doi:10.1007/s11831-024-10152-0.
151. Wang QA, Chen J, Ni Y, Xiao Y, Liu N, Liu S, et al. Application of Bayesian networks in reliability assessment: a systematic literature review. In: Structures. Vol. 71. Amsterdam, The Netherlands: Elsevier; 2025. p. 108098.
152. Šír G. A computational perspective on neural-symbolic integration. Neurosymbolic Artif Intell. 2025;1(4):NAI-240672. doi:10.3233/nai-240672.
153. Korba AA, Diaf A, Bouchiha MA, Ghamri-Doudane Y. Mitigating IoT botnet attacks: an early-stage explainable network-based anomaly detection approach. Comput Commun. 2025;241(1):108270. doi:10.1016/j.comcom.2025.108270.
154. Al-Fayoumi M, Al-Haija QA, Armoush R, Amareen C. XAI-PDF: a robust framework for malicious PDF detection leveraging SHAP-based feature engineering. Int Arab J Inform Technol (IAJIT). 2024;21(1):128–46.
155. Lim B, Huerta R, Sotelo A, Quintela A, Kumar P. EXPLICATE: enhancing phishing detection through explainable AI and LLM-powered interpretability. arXiv:2503.20796. 2025.
156. Nazim S, Alam MM, Rizvi SS, Mustapha JC, Hussain SS, Suud MM. Advancing malware imagery classification with explainable deep learning: a state-of-the-art approach using SHAP, LIME Grad-CAM. PLoS One. 2025;20(5):e0318542. doi:10.1371/journal.pone.0318542.
157. Zhou Y, Li H, Xiao Z, Qiu J. A user-centered explainable artificial intelligence approach for financial fraud detection. Finance Res Lett. 2023;58(11):104309. doi:10.1016/j.frl.2023.104309.
158. Machado L, Gadge J. Phishing sites detection based on C4. 5 decision tree algorithm. In: 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA). New, York, NY, USA: IEEE; 2017. p. 1–5.
159. Brown A, Tuor A, Hutchinson B, Nichols N. Recurrent neural network attention mechanisms for interpretable system log anomaly detection. In: Proceedings of the First Workshop on Machine Learning for Computing Systems; 2018 Jun 12; Tempe, AZ, USA. New York, NY, USA: ACM. p. 1–8.
160. Mendes C, Rios TN. Explainable artificial intelligence and cybersecurity: a systematic literature review. arXiv:2303.01259. 2023.
161. Kulkarni M, Stamp M. XAI and android malware models. In: Machine learning, deep learning and AI for cybersecurity. Cham, Switzerland: Springer; 2025. p. 327–55 doi:10.1007/978-3-031-83157-7_12.
162. Mahbooba B, Timilsina M, Sahal R, Serrano M. Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity. 2021;2021(1):6634811. doi:10.1155/2021/6634811.

163. Donmez TB, Kutlu M, Mansour M, Yildiz MZ. Explainable AI in action: a comparative analysis of hypertension risk factors using SHAP and LIME. Neural Comput Appl. 2025;37(5):4053–74. doi:10.1007/s00521-024-10724-y.
164. Ennab M, Mcheick H. Advancing AI interpretability in medical imaging: a comparative analysis of pixel-level interpretability and Grad-CAM models. Mach Learn Knowl Extrac. 2025;7(1):12. doi:10.3390/make7010012.
165. Toker Gokce A, Deveci Topal A, Kolburan Geçer A, Dilek Eren C. Investigating the level of artificial intelligence literacy of university students using decision trees. Educ Inform Techno. 2025;30(5):6765–84. doi:10.1007/s10639-024-13081-4.
166. Wang Z, Yin Z, Liu F, Liu Z, Lisetti C, Yu R, et al. Graph fairness via authentic counterfactuals: tackling structural and causal challenges. ACM SIGKDD Explor Newslet. 2025;26(2):89–98. doi:10.1145/3715073.3715081.
167. Huang Z, Zhuang Y, Lu G, Qin Z, Xu H, Zhao T, et al. Reinforcement learning with rubric anchors. arXiv:2508.12790. 2025.
168. Anderson HS, Roth P. Ember: an open dataset for training static PE malware machine learning models. arXiv:1804.04637. 2018.
169. Anderljung M, Hazell J, von Knebel M. Protecting society from AI misuse: when are restrictions on capabilities warranted? AI SOC. 2025;40(5):3841–57. doi:10.1007/s00146-024-02130-8.
170. Inam R, Singh V, Chawla A, Bosneag AM. XAI metrics and frameworks. In: Explainable AI for communications and networking. Amsterdam, The Netherlands: Elsevier; 2025. p. 49–72.
171. Barrainkua A, Gordaliza P, Lozano JA, Quadrianto N. Preserving the fairness guarantees of classifiers in changing environments: a survey. ACM Comput Surv. 2025;57(6):1–32. doi:10.1145/3637438.
172. Hall P, Mundahl O, Park S. The pitfalls of "Security by Obscurity" and what they mean for transparent AI. In: Proceedings of the AAAI Conference on Artificial Intelligence. Washington, DC, USA: AAAI Press; 2025. Vol. 39, p. 28042–51.
173. Lin C, Chen Z, Zhang Z, Liu J. E3: early exiting with explainable AI for real-time and accurate DNN inference in edge-cloud systems. In: Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems; 2025 May 6–9; Irvine, CA, USA. p. 385–97.
174. Khalaf NZ, Al Barazanchi II, Radhi AD, Parihar AD, Shah S, Sekhar P, et al. Development of real-time threat detection systems with AI-driven cybersecurity in critical infrastructure. Mesopotamian J Cybersecur. 2025;5(2):501–13. doi:10.58496/MJCS/2025/031.
175. Ali S, Wadho SA, Talpur KR, Talpur BA, Alshudukhi KS, Humayun M, et al. Next-generation quantum security: the impact of quantum computing on cybersecurity—threats, mitigations, and solutions. Comput Elect Eng. 2025;128(3):110649. doi:10.1016/j.compeleceng.2025.110649.
176. Piplai A, Kotal A, Mohseni S, Gaur M, Mittal S, Joshi A. Knowledge-enhanced neurosymbolic artificial intelligence for cybersecurity and privacy. IEEE Internet Comput. 2023;27(5):43–8. doi:10.1109/mic.2023.3299435.
177. Ali A, Huang J, Jabbar A, Ali M, Ali S. Federated learning for connected autonomous vehicles: enhancing mobility, safety, and sustainability. Phys Scr. 2025;100(7):72001. doi:10.1088/1402-4896/addca2.
178. Bilot T, El Madhoun N, Al Agha K, Zouaoui A. Graph neural networks for intrusion detection: a survey. IEEE Access. 2023;11:49114–39. doi:10.1109/access.2023.3275789.
179. Ferrag MA, Ndhlovu M, Tihanyi N, Cordeiro LC, Debbah M, Lestable T, et al. Revolutionizing cyber threat detection with large language models: a privacy-preserving bert-based lightweight model for IoT/IIoT devices. IEEE Access. 2024;12:23733–50. doi:10.1109/access.2024.3363469.