

EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR CYBERSECURITY: TECHNIQUES, APPLICATIONS, CHALLENGES, AND FUTURE RESEARCH DIRECTIONS.

Abstract

Artificial intelligence (AI) has become an important component of modern cybersecurity systems because of its ability to identify suspicious activities, recognize complex patterns, and support automated threat detection. Machine learning and deep learning algorithms are increasingly applied to network intrusion detection, malware classification, phishing detection, and anomaly identification. However, many advanced AI models operate as black boxes, making it difficult for security professionals to understand why a particular activity has been classified as malicious or legitimate. This lack of transparency can reduce trust in automated security systems and complicate the investigation of cybersecurity incidents. Explainable artificial intelligence (XAI) addresses this challenge by providing human-understandable explanations of model predictions. This paper presents a review of explainable AI techniques and their applications in cybersecurity. It examines interpretable machine learning models, feature importance analysis, Local Interpretable Model-Agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). The paper also discusses the benefits and limitations of integrating explainability into cybersecurity operations, including computational overhead, explanation reliability, adversarial manipulation, privacy concerns, and the balance between predictive performance and interpretability. Finally, it identifies research opportunities involving real-time explanation systems, privacy-preserving learning, human-centered security interfaces, and robust evaluation frameworks. The review indicates that explainability can support more transparent and accountable cybersecurity decisions, although explanations should be validated and used alongside conventional security controls rather than treated as proof of a model's correctness.

Keywords: Artificial Intelligence, Cybersecurity, Explainable AI, Machine Learning, Intrusion Detection, SHAP, LIME, Threat Detection.

Introduction

The rapid expansion of digital technologies has transformed communication, business operations, education, healthcare, banking, and public services. Organizations increasingly depend on interconnected computer systems, cloud infrastructure, mobile devices, and online platforms to store and exchange information. Although these technologies improve efficiency and accessibility, they also increase the number of potential entry points for cyberattacks. Protecting digital information has consequently become a major concern for researchers, software developers, network administrators, and security professionals.

Traditional cybersecurity systems commonly rely on predefined rules, signatures, and known attack patterns. These approaches can be effective against previously identified threats, but they may struggle to recognize unfamiliar attacks or malicious activities that differ from established patterns. Artificial intelligence provides additional capabilities by enabling computer systems to learn from historical data and identify suspicious behavior that may not be captured by conventional rules.

Machine learning algorithms can classify network traffic, detect unusual system activities, identify potentially malicious files, and estimate the likelihood that a digital event represents a security threat. Deep learning extends these capabilities by learning complex representations from large datasets. Nevertheless, the internal decision-making processes of sophisticated models can be difficult to understand. A model may identify a network connection as suspicious without providing a clear explanation of the characteristics that influenced its decision.

This limitation is particularly important in cybersecurity because automated predictions can influence incident investigations, access restrictions, system alerts, and the allocation of security resources. An unexplained alert may require additional investigation, while an incorrect prediction may result in unnecessary disruption or allow a genuine threat to remain undetected.

Explainable artificial intelligence has emerged as a research area that seeks to make AI decisions more understandable to human users. XAI techniques attempt to explain the factors influencing a prediction, identify important input features, and communicate the reasoning behind an algorithmic output. Existing research has explored these techniques in areas such as intrusion detection, malware classification, and network security [1–3].

This paper reviews the role of explainable AI in cybersecurity, describes representative techniques, examines practical applications, and identifies major implementation challenges. Its central argument is that cybersecurity systems should be evaluated not only by predictive performance but also by the quality, reliability, and usefulness of their explanations.

1.1. Research Objectives

The main objectives of this paper are:

1. To explain the importance of explainability in AI-based cybersecurity systems.
2. To examine major XAI techniques and their underlying principles.
3. To discuss applications of explainable AI in different cybersecurity tasks.
4. To identify technical, operational, and ethical challenges associated with XAI.
5. To propose future research directions for developing trustworthy cybersecurity solutions.

1.2. Research Questions

This review is guided by the following questions:

- How can explainable AI improve the interpretation of cybersecurity predictions?
- Which XAI techniques are suitable for different security applications?
- What limitations affect the reliability and practical usefulness of explanations?
- How can future systems combine predictive performance, transparency, and operational efficiency?

2. Background and Related Work

2.1. Artificial Intelligence in Cybersecurity

Artificial intelligence refers to computational techniques that enable systems to perform tasks associated with learning, classification, prediction, and decision support. In cybersecurity, AI is used to analyze large volumes of network traffic, system logs, user activities, and file characteristics.

Machine learning techniques can identify patterns associated with malicious behavior by learning from labeled or unlabeled data. Supervised learning uses examples with known labels, whereas unsupervised learning can identify unusual patterns without requiring every observation to have a predefined class. Deep learning uses multilayer neural networks to learn complex relationships within data.

These methods support several security functions, including intrusion detection, spam filtering, malware analysis, phishing identification, and behavioral anomaly detection. However, a high classification score alone does not necessarily explain why a system has reached a particular conclusion.

For example, an intrusion detection model might classify a connection as malicious because of unusual packet rates, unexpected protocol behavior, or an abnormal volume of transmitted data. If the model does not communicate these factors, a security analyst may find it difficult to assess the alert or determine an appropriate response.

2.2. Explainable Artificial Intelligence

Explainable artificial intelligence includes methods designed to make model behavior and predictions more understandable. The objective is not simply to produce a prediction but also to provide information that helps users interpret and evaluate that prediction.

XAI methods are commonly divided into two broad categories.

Intrinsic interpretability: These models are designed to be relatively understandable by their structure. Examples include decision trees, linear models, and rule-based classifiers. Their internal logic can often be examined directly.

Post-hoc explainability: These methods explain predictions generated by an already trained model. They can be applied to complex models whose internal operations are difficult to interpret directly. LIME and SHAP are widely discussed examples.

The appropriate method depends on the application, the complexity of the underlying model, the characteristics of the input data, and the explanation required by the intended user.

2.3. Related Research

Research on explainable AI for cybersecurity has expanded as machine learning has become more widely used in security operations. Zhang et al. [1] reviewed XAI applications in cybersecurity and examined their relevance to threat detection and security analysis. Their work highlighted the need to make AI-driven security decisions more transparent.

Rjoub et al. [2] surveyed explainable AI methods for cybersecurity and discussed their potential to improve the understanding of network security threats. The study also identified limitations that must be considered when integrating explainability into operational systems.

Arrieta et al. [3] presented a broader discussion of explainable artificial intelligence, including its concepts, methods, and challenges across different application areas. Their work provides a useful foundation for understanding the distinction between model interpretability and post-hoc explanations.

These studies indicate that explainability is an important component of trustworthy AI. However, cybersecurity introduces specific requirements: explanations should be generated quickly, remain useful as attack patterns change, and provide enough technical detail for analysts to investigate potentially malicious behavior.

3. Research Methodology

This paper adopts a narrative literature-review approach to examine explainable AI in cybersecurity. It synthesizes concepts, methods, and challenges discussed in established research publications concerning explainability, machine learning, and cyber threat detection.

The review focuses on three main areas: the technical foundations of XAI, its application to cybersecurity tasks, and the limitations affecting its deployment. Representative research on LIME, SHAP, and explainable machine learning is used to explain the principal methods and their relevance to security analysis.

The literature was organized thematically rather than through a quantitative meta-analysis. The paper does not claim to represent an exhaustive systematic review, and it does not report newly collected experimental data. Consequently, no original accuracy scores, experimental comparisons, or performance improvements are claimed.

This approach provides a conceptual foundation for understanding the opportunities and limitations of explainable AI while identifying areas requiring further empirical investigation.

4. Major Explainable AI Techniques

4.1. Decision Trees and Rule-Based Models

Decision trees represent classification decisions through a sequence of conditions. Each internal node evaluates a feature, each branch represents an outcome, and each terminal node produces a prediction.

In an intrusion detection application, a decision tree might evaluate whether a connection contains an unusual number of requests, an unexpected protocol combination, or an abnormal traffic duration. The resulting path through the tree can provide an understandable account of the classification.

Rule-based models operate similarly by applying explicit logical conditions. A simplified rule might classify an event for investigation when several independently defined indicators are present.

The primary advantage of these approaches is their relative transparency. Analysts can inspect the decision structure and determine which conditions contributed to a prediction.

However, interpretability may decrease when trees become excessively deep or rule sets become very large. Simple models may also fail to represent some complex relationships present in sophisticated cyber threats. Therefore, model simplicity must be balanced against predictive capability and the operational requirements of the security environment.

4.2. Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-Agnostic Explanations, commonly known as LIME, explain an individual prediction by approximating a complex model locally with a simpler interpretable model [4].

The central idea is that a complicated model may be difficult to understand globally, but its behavior around a particular prediction may be approximated using a simpler explanation.

For example, consider an AI-based intrusion detection system that flags a network connection as suspicious. LIME can help identify which input characteristics contributed most strongly to that specific classification. An explanation might highlight unusual connection duration, an unexpected traffic volume, or a high frequency of requests.

LIME is model-agnostic, meaning it can be applied to different types of predictive models without requiring direct access to their internal structures.

Despite these benefits, LIME has limitations. Its explanations depend on how nearby examples are generated, how features are represented, and how the local approximation is constructed. Different configurations may produce different explanations. In cybersecurity, where inputs can have complex dependencies, an explanation may also simplify relationships that are important to the actual prediction.

4.3. SHapley Additive exPlanations

SHapley Additive exPlanations, or SHAP, use concepts from cooperative game theory to estimate the contribution of individual features to a model prediction [5].

The method assigns feature contributions by considering how predictions change when features are included or excluded across different combinations. These contributions can help explain both individual predictions and broader model behavior.

In cybersecurity, SHAP may be used to identify the characteristics most influential in a malware classification or network intrusion alert. A local explanation can describe why a particular event was classified as suspicious, while an aggregate analysis can reveal which features frequently influence predictions across a dataset.

SHAP offers a consistent mathematical framework for feature attribution, but it does not automatically guarantee that an explanation is complete or causally correct. Results depend on assumptions about the model, the data distribution, and the method used to estimate feature contributions.

Highly correlated features can also complicate interpretation because their contributions may be distributed in ways that are not intuitive to users. For these reasons, SHAP results should be interpreted alongside domain knowledge and independent security evidence.

4.4. Feature Importance Analysis

Feature importance methods estimate the influence of different input variables on a model's predictive behavior. Depending on the method, importance may be assessed through changes in model performance, internal model properties, or the contribution of features to individual predictions.

For example, a security analyst may investigate whether unusual connection duration, repeated authentication failures, or abnormal request frequency is associated with a model's alerts.

Global feature importance helps identify broad patterns across a dataset. Local feature importance explains why a specific prediction was made.

These perspectives serve different purposes. Global analysis can support model auditing and feature selection, whereas local analysis can assist with incident investigation.

However, feature importance should not be confused with causation. A feature associated with a prediction does not necessarily cause the detected threat. It may reflect correlations, dataset artifacts, or indirect relationships. Reliable interpretation therefore requires careful validation.

4.5. Comparative Analysis of XAI Techniques

Technique	Main purpose	Advantages	Limitations
Decision trees	Explain predictions through decision paths	Directly interpretable	Large trees may become difficult to understand
Rule-based models	Express decisions as explicit rules	Transparent and easy to audit	Complex rule sets can become difficult to maintain
LIME	Explain individual predictions locally	Model-agnostic and flexible	Sensitive to sampling and configuration
SHAP	Estimate feature contributions	Consistent attribution framework	Computational cost and interpretation challenges
Feature importance	Identify influential input variables	Useful for model analysis	Does not establish causation

The comparison demonstrates that no single technique is ideal for every cybersecurity application. Intrinsically interpretable models may be preferable when transparency is essential, while post-hoc methods can help explain complex predictive models. In practice, a combination of approaches may provide a more useful understanding of model behavior.

5. Applications of Explainable AI in Cybersecurity

5.1. Network Intrusion Detection

Network intrusion detection systems monitor traffic to identify activities that may indicate unauthorized access, malicious communication, or exploitation attempts.

Machine learning models can process traffic features and classify connections as normal or suspicious. Explainability adds information about the characteristics that influenced these classifications.

For instance, an explanation may identify an unusual connection rate, unexpected protocol behavior, or a deviation from established traffic patterns. This information can help analysts prioritize alerts and determine which events require additional investigation.

Explanations can also support model auditing by revealing whether predictions depend on meaningful network characteristics or unreliable patterns in the training data.

Nevertheless, network traffic changes over time. New applications, legitimate changes in usage, and evolving attack strategies can alter data distributions. An explanation that was useful under one operating condition may become less informative when the underlying environment changes.

5.2. Malware Detection

Malware detection involves identifying software that may perform unauthorized or harmful activities. Machine learning models can analyze file metadata, behavioral observations, static features, and other representations to distinguish potentially malicious software from legitimate applications.

Explainable AI can help investigators understand which observed characteristics contributed to a malware prediction. Depending on the model and available data, these characteristics might include unusual permissions, suspicious behavior indicators, or combinations of file properties.

Such explanations can support malware triage and help researchers compare the characteristics of different samples.

However, a feature attribution should not be interpreted as definitive proof that a file is malicious. Legitimate applications may share characteristics with malware, and malicious software may deliberately change its observable properties. Explanations must therefore be considered alongside established malware-analysis procedures and additional evidence.

5.3. Phishing Detection

Phishing attacks use deceptive communications or websites to persuade individuals to reveal sensitive information or perform unsafe actions. Automated detection systems may analyze email text, sender information, URL characteristics, and webpage properties to identify potentially deceptive content.

Explainable models can highlight suspicious elements that contributed to a classification, such as inconsistencies in sender information, unusual URL structures, or language patterns associated with deceptive messages.

These explanations can help security teams understand why a communication was flagged and can support user education.

However, phishing methods evolve, and attackers may imitate legitimate communication styles. Explanations based on superficial indicators can become unreliable if attackers change those indicators. Systems should therefore combine multiple signals and be evaluated against varied and evolving examples.

5.4. Anomaly Detection

Anomaly detection identifies observations that differ significantly from expected behavior. It is useful in cybersecurity because some attacks may not match previously recorded signatures.

AI-based anomaly detection can examine login activity, network behavior, resource utilization, and other system events to identify unusual patterns.

Explainability can help analysts determine which characteristics made an observation unusual. For example, an alert might be associated with a login from an unfamiliar context combined with an unexpected sequence of system operations.

A major challenge is that unusual activity is not necessarily malicious. Legitimate administrative tasks, software updates, and changes in user behavior can produce anomalies.

Explanations should therefore provide context rather than automatically equating unusual behavior with an attack. Human review and additional verification can reduce unnecessary interventions.

5.5. Security Operations and Incident Response

Security operations teams frequently need to evaluate alerts from multiple monitoring systems. Excessive or poorly explained alerts can increase investigation workloads and make it harder to prioritize genuine threats.

Explainable AI can provide a concise summary of the factors supporting an alert, helping analysts understand the model's reasoning and decide whether further investigation is appropriate.

For example, an explanation interface might display the predicted threat category, the principal contributing features, the confidence estimate where properly calibrated, and the evidence that requires verification.

Such systems can also preserve explanation records for subsequent audits. These records may help organizations understand why an alert was prioritized or why a particular investigation was initiated.

Nevertheless, explanations should not replace independent verification. Automated decisions that block users, isolate systems, or modify access permissions should be governed by appropriate security policies, risk thresholds, and human oversight.

6. Challenges and Limitations

6.1. Trade-Off Between Accuracy and Interpretability

Complex models can capture relationships that are difficult to represent with simple decision structures. However, their complexity may make them harder to explain.

Conversely, an interpretable model may provide a clear decision path but fail to capture certain patterns present in complicated datasets.

This creates a practical trade-off between predictive performance and interpretability. The best solution depends on the consequences of incorrect predictions, the needs of security analysts, and the resources available for investigation.

Future systems should evaluate both predictive quality and explanation usefulness instead of assuming that one metric is sufficient.

6.2. Reliability of Explanations

An explanation is useful only when it accurately reflects the relevant behavior of the model. Some post-hoc methods approximate complex decision processes and may produce simplified representations that do not fully capture the original model.

Explanations may also vary when inputs are slightly changed or when a method uses different sampling procedures. Inconsistent explanations can make it difficult for analysts to develop reliable expectations about system behavior.

Researchers should therefore examine explanation stability, fidelity to the original model, consistency across similar examples, and usefulness in realistic security investigations.

6.3. Computational Overhead

Cybersecurity systems may process large volumes of network events and generate alerts in near real time. Additional explanation calculations can increase processing requirements and response latency.

The computational cost varies by model, dataset, explanation method, and implementation. Some methods are suitable for individual alerts but may be expensive when applied to every event in a high-volume environment.

Possible approaches include generating explanations only for high-priority alerts, using efficient approximation techniques, and separating real-time detection from more detailed offline analysis. These strategies require empirical evaluation to ensure that important information is not lost.

6.4. Adversarial Manipulation

Cyber attackers may attempt to evade detection by changing observable behavior or manipulating the data used by security models. Explainability systems introduce additional considerations because attackers may seek to exploit information about the factors influencing a model's predictions.

For example, if a system consistently relies on a narrow set of features, an attacker may attempt to alter those features while preserving malicious behavior.

Researchers must consequently evaluate the robustness of both the predictive model and its explanation process. Explanation interfaces should also avoid disclosing unnecessary operational details that could weaken security.

6.5. Privacy and Sensitive Information

Cybersecurity datasets may contain sensitive information about users, devices, organizations, and communication patterns. Explanations that reveal detailed input features can unintentionally disclose confidential information.

This concern is particularly important when models are trained on data from multiple organizations or when explanation records are shared across teams.

Privacy-preserving learning, careful access controls, data minimization, and appropriate redaction can help reduce these risks. Explanation systems should communicate enough information for legitimate security analysis without exposing unnecessary personal or organizational details.

6.6. Human Factors and Overreliance

An explanation that appears clear may be accepted without sufficient scrutiny. Analysts may place excessive confidence in an AI-generated rationale, particularly when the system presents information in authoritative language or visually emphasizes selected features.

This creates a risk of automation bias, in which users rely too heavily on machine-generated outputs.

Effective explanation interfaces should communicate uncertainty, distinguish model outputs from verified evidence, and encourage independent assessment when the consequences of an error are significant. Training and organizational procedures are also important for ensuring appropriate human oversight.

7. Proposed Framework for Explainable Cybersecurity

A practical explainable cybersecurity framework can be organized into five interconnected components.

1. Data collection and preprocessing: The system collects relevant network, endpoint, application, or authentication data. Preprocessing checks data quality, handles missing values, and applies appropriate privacy safeguards.

2. Threat detection model: A machine learning model evaluates the processed data and produces a prediction, such as normal activity, suspicious behavior, or a potential security incident.

3. Explanation generation: An appropriate XAI method generates an explanation for the prediction. A decision tree may expose its decision path, while SHAP or LIME may provide feature-based explanations for more complex models.

4. Analyst interface: The system presents the prediction, supporting features, relevant context, and limitations in a format appropriate for security professionals. The interface should clearly distinguish observed evidence from inferred explanations.

5. Verification and feedback: Analysts review the alert and determine whether further investigation or action is required. Confirmed outcomes may be used to improve future models, subject to data quality, privacy, and governance requirements.

The framework should be evaluated through predictive metrics such as precision, recall, and false-positive rate, together with explanation-related measures such as fidelity, stability, computational cost, and analyst usefulness. Importantly, the framework is a conceptual proposal rather than an experimentally validated implementation. Its effectiveness would need to be tested using representative datasets, realistic operational conditions, and clearly defined evaluation procedures.

8. Future Research Directions

8.1. Real-Time Explainable Security Systems

Future research should investigate methods that generate useful explanations with minimal additional latency. This is particularly important for high-volume network monitoring and time-sensitive incident response.

Efficient explanation algorithms and selective explanation strategies may help reduce computational overhead. Researchers should measure whether these approaches maintain explanation quality under realistic workloads.

8.2. Robust and Trustworthy Explanations

Further research is needed to determine how explanations behave when input data changes, attacks evolve, or models encounter unfamiliar conditions.

Robust explanation frameworks should be evaluated for stability, fidelity, and resistance to deliberate manipulation. Researchers should also investigate how to communicate uncertainty when a model encounters data outside its training distribution.

8.3. Hybrid AI Approaches

Combining interpretable rules with machine learning may offer a useful balance between predictive flexibility and transparent decision-making. A hybrid system could use a learned model to identify suspicious behavior while applying understandable security policies to guide subsequent decisions. However, combining approaches does not automatically produce a trustworthy system. Future studies should examine how the components interact, whether explanations remain consistent, and how conflicts between model predictions and security rules are resolved.

8.4. Privacy-Preserving Explainability

Privacy-preserving approaches could help organizations benefit from explainable AI without unnecessarily sharing sensitive data. Potential directions include federated learning, privacy-aware feature attribution, and controlled access to explanation records.

Such approaches must be evaluated for their effects on predictive quality, explanation detail, computational requirements, and privacy protection.

8.5. Human-Centered Explanation Interfaces

Different users require different forms of explanation. A network engineer may need detailed technical features, whereas a manager may require a concise description of the potential business impact.

Future research should examine how explanation formats influence analyst accuracy, investigation time, workload, and confidence calibration. User studies can help determine whether an explanation genuinely improves decision-making rather than merely appearing understandable.

8.6. Standardized Evaluation Frameworks

Comparing XAI methods is difficult when studies use different datasets, models, evaluation metrics, and explanation requirements.

A standardized evaluation framework should include predictive performance, explanation fidelity, stability, latency, privacy implications, and human usefulness. Evaluation should also consider changing data distributions and realistic adversarial conditions.

Consistent evaluation procedures would make it easier to compare methods and identify which approaches are appropriate for specific cybersecurity tasks.

9. Discussion

The literature reviewed in this paper suggests that explainable AI can strengthen the practical use of machine learning in cybersecurity by helping users interpret predictions and investigate suspicious activities. Its value is especially relevant when decisions require technical scrutiny, documentation, or human approval.

However, explanation should not be treated as a substitute for model validation. A model may generate a persuasive explanation while relying on biased, incomplete, or misleading patterns. Likewise, a feature identified as important may be correlated with a prediction without representing the actual cause of malicious behavior.

The usefulness of XAI therefore depends on several interconnected factors: the quality of the underlying data, the suitability of the predictive model, the reliability of the explanation method, and the ability of the intended user to interpret the result correctly.

A well-designed system should combine explanation with established security practices, including access control, independent evidence collection, continuous monitoring, incident response procedures, and periodic model evaluation.

Another important consideration is that explanation requirements differ between applications. Malware analysis may benefit from detailed feature attribution, whereas a high-volume network monitoring system may require concise explanations that can be generated quickly. A single explanation technique is unlikely to satisfy every operational requirement.

Organizations should consequently select XAI methods according to their specific threat models, technical resources, privacy obligations, and decision-making needs. Further empirical work is required to determine how these methods perform across different security environments.

10. Conclusion

Explainable artificial intelligence represents an important research direction in the development of transparent and trustworthy cybersecurity systems. Although machine learning and deep learning can support the identification of malicious activities, their predictions may be difficult to interpret without additional explanation mechanisms.

This paper reviewed major XAI approaches, including decision trees, rule-based models, LIME, SHAP, and feature importance analysis. It discussed their potential applications in network intrusion detection, malware classification, phishing identification, anomaly detection, and security operations.

The review also identified significant challenges involving interpretability, explanation reliability, computational overhead, adversarial manipulation, privacy, and human overreliance. These challenges demonstrate that explainability must be designed and evaluated as an integral component of cybersecurity rather than added solely to improve the appearance of transparency.

Future research should prioritize robust explanation methods, efficient real-time implementations, privacy-preserving techniques, hybrid AI architectures, human-centered interfaces, and standardized evaluation procedures. Experimental validation under realistic conditions will be necessary to determine whether these approaches improve security outcomes.

In conclusion, explainable AI can help bridge the gap between automated threat detection and human understanding. When combined with appropriate validation, independent security evidence, and responsible human oversight, it can contribute to more interpretable and accountable cybersecurity systems.

References

1. Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10, 93104–93139. <https://doi.org/10.1109/ACCESS.2022.3204051>
2. Rjoub, G., Bentahar, J., Wahab, O. A., Mizouni, R., Song, A., Cohen, R., Otrok, H., & Mourad, A. (2023). A survey on explainable artificial intelligence for cybersecurity. *IEEE Transactions on Network and Service Management*, 20(4), 5115–5140. <https://doi.org/10.1109/TNSM.2023.3282740>
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
5. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30 (pp. 4765–4774).
6. Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>
7. Yan, H., et al. (2022). Explainable machine learning in cybersecurity: A survey. *International Journal of Intelligent Systems*. <https://doi.org/10.1002/int.23088>