Innovative Journal of Computing, Security and Digital Technologies
Volume 1, Issue 1, Jan – Apr 2026
Page no : 01-09

Explainable Artificial Intelligence for Cybersecurity Threat Detection: Methods, Challenges, and Future Directions

Pawlicka[1], R. Kozik [2]

[1,2]Department of Computer Science,University of Westford, London, United Kingdom.

Abstract

Artificial intelligence (AI) is increasingly used to detect cybersecurity threats such as malicious network traffic, malware, phishing, and abnormal user activity. Although machine-learning and deep-learning models can identify complex patterns, their decisions may be difficult for security analysts to interpret. This limitation is important in cybersecurity, where an alert may lead to account restrictions, incident escalation, or defensive actions. Explainable artificial intelligence (XAI) aims to make model predictions more understandable by identifying the factors that influenced a decision. This paper reviews major XAI methods and presents a practical framework for integrating explanations into cybersecurity threat detection. The discussion covers model-specific and model-agnostic techniques, local and global explanations, evaluation criteria, operational challenges, and future research directions. A proposed workflow combines data preparation, threat classification, explanation generation, analyst review, and feedback-based monitoring. The paper emphasizes that explanations should be evaluated for fidelity, stability, usefulness, and resistance to manipulation—not merely visual clarity. It also distinguishes published evidence from the proposed framework: no new experimental results are claimed. The article concludes that explainability can support threat investigation and model governance, but must be integrated with security controls, privacy safeguards, and human oversight.

Index Terms— Explainable artificial intelligence, cybersecurity, threat detection, intrusion detection, machine learning, SHAP, LIME, interpretable AI.

Publication Date : 10.01.2026

I. INTRODUCTION

Cybersecurity systems must detect threats across networks, endpoints, cloud services, and digital applications. Modern threat detection increasingly uses machine learning (ML) to identify patterns that may be difficult to express as fixed rules. These systems can support intrusion detection, malware classification, phishing detection, and user-behavior analysis. However, a model's prediction alone may not provide enough information for an analyst to determine why an event was flagged or what action should follow.

This problem is especially relevant when complex models are used. A high-performing classifier may identify a suspicious connection but provide limited insight into the features that influenced its prediction. An analyst may need to know whether the alert was driven by

How to Cite
Explainable Artificial Intelligence for Cybersecurity Threat Detection: Methods, Challenges, and Future Directions. (2026). Innovative Journal of Computing, Security and Digital Technologies, 1(1), 01-09.
https://ijcdst.in/index.php/about/article/view/4

unusual connection duration, a high number of failed requests, unexpected protocol use, or a combination of signals. Without a meaningful explanation, it can be difficult to validate alerts, investigate false positives, and communicate decisions to other teams.

Explainable artificial intelligence (XAI) refers to methods that make the behavior or outputs of AI systems more understandable to people. In cybersecurity, explanations may help analysts inspect model decisions, identify suspicious patterns, and assess whether a prediction is consistent with other evidence. A survey of XAI for cybersecurity identifies the need to consider both the threat context and the limitations of explanation methods.

This paper examines how XAI methods can be integrated into cybersecurity threat detection. It focuses on four questions:

1. Which XAI methods are commonly used to interpret cybersecurity models?
2. How can explanations support threat detection and investigation?
3. What technical and operational risks limit the reliability of explanations?
4. What framework and research directions can improve practical use?

The main contribution is a conceptual workflow that links threat detection with explanation generation, analyst review, and continuous monitoring. The framework is intended as a design proposal, not as a tested system. Its performance must be established through implementation and controlled evaluation.

II. BACKGROUND AND RELATED WORK

A. AI-Based Cybersecurity Threat Detection

AI-based threat detection learns patterns from security data and uses them to classify or rank events. Depending on the application, input data may include network-flow features, endpoint events, system calls, authentication records, or email characteristics. Models may be trained to distinguish known attack categories or to identify deviations from a learned baseline.

Threat detection is not a single prediction task. A useful system must also consider alert volume, false positives, response time, and the consequences of missed detections. A model that classifies traffic accurately on a benchmark may still be difficult to use in a live environment if it generates too many alerts or performs poorly when traffic patterns change.

B. Explainability in Cybersecurity

XAI methods are often grouped into two broad categories:

- Intrinsic interpretability: The model is designed to be understandable, such as a shallow decision tree or a rule-based classifier.
- Post-hoc explanation: An explanation method is applied to an already trained model to describe its predictions.

Explanations may also be local, explaining an individual prediction, or global, describing general model behavior. Local explanations can support investigation of a particular alert. Global explanations can help model developers understand which features influence predictions across a dataset.

Research has applied XAI to network intrusion detection and evaluated methods such as SHAP and LIME across models and datasets. One evaluation framework examined explanation properties including descriptive accuracy, sparsity, stability, efficiency, robustness, and completeness.

C. Research Gap

Existing work demonstrates that XAI methods can be applied to cybersecurity models, but several practical questions remain. An explanation may be understandable yet fail to reflect the model's actual behavior. Different explanation methods may produce different feature rankings. Explanations may also be computationally expensive or vulnerable to manipulated inputs.

Recent research on intrusion detection highlights the need to evaluate explanation quality alongside detection performance and to consider robustness, deployment context, and human-in-the-loop workflows.

The research gap addressed in this paper is therefore not simply the absence of explanations. It is the need for a structured threat-detection workflow in which explanations are generated, evaluated, presented to analysts, and monitored for reliability.

III. XAI METHODS FOR CYBERSECURITY

A. SHAP

SHAP (SHapley Additive exPlanations) assigns contribution values to input features for a model prediction. In a cybersecurity setting, a local SHAP explanation may indicate which network-flow attributes contributed most to an alert. Aggregated explanations may help identify features that commonly influence a model across many events.

SHAP is useful for local and global analysis, but its output must be interpreted carefully. A feature contribution indicates how the model used that feature; it does not establish that the feature caused an attack. Correlated features and differences in background data can also affect interpretation.

B. LIME

LIME (Local Interpretable Model-Agnostic Explanations) approximates a model's behavior near a specific input using a simpler interpretable model. For an intrusion alert, LIME may provide a compact explanation of which input features were influential in the local prediction.

Its local approximation can be useful for individual cases, but results may vary with the perturbation strategy, sampling, and selected neighborhood. A security analyst should not treat a single LIME explanation as a complete description of the model.

C. Feature Importance and Rule-Based Explanations

Feature-importance methods estimate which inputs are influential across a model or dataset. Rule-based explanations express decisions through conditions that analysts can inspect, such as a combination of unusual traffic volume and an unexpected protocol state.

These methods may be easier to communicate than complex visualizations. However, feature importance alone may not explain a specific alert, and rules extracted from a complex model may only approximate its behavior.

D. Model-Specific Explanations

Some explanation methods are designed for particular model families. For example, gradient-based methods may be used with neural networks, while tree-based models may expose feature contributions or decision paths. Model-specific methods can be efficient, but their applicability depends on the model architecture and input representation.

E. Method Selection

No single explanation method is best for every cybersecurity task. Selection should consider the model type, data format, analyst's question, runtime constraints, and required explanation scope.

Table I. Comparison of XAI Methods

Method	Explanation scope	Potential cybersecurity use	Main limitation
SHAP	Local and global	Feature contributions for alerts and model review	Computational cost; sensitive to feature dependence
LIME	Local	Explaining an individual alert	Local approximation may be unstable
Feature importance	Mainly global	Model inspection and feature ranking	May not explain individual decisions
Rule-based explanation	Local or global	Analyst-readable decision logic	Rules may simplify model behavior
Gradient-based methods	Usually local	Neural-network interpretation	Depends on model architecture and gradients

IV. PROPOSED XAI-ENABLED THREAT DETECTION FRAMEWORK

The proposed framework places explanation generation after a threat-detection model and before analyst review. The explanation is treated as supporting evidence, not as an independent security verdict.

A. Framework Components

1. Data collection: Collect relevant network, endpoint, or application events.
2. Data preparation: Validate fields, handle missing values, normalize data where required, and apply privacy controls.
3. Threat detection model: Produce a threat score or predicted class.
4. Explanation engine: Generate a local explanation for the event and, where needed, a global explanation for model monitoring.
5. Explanation validation: Check whether the explanation is sufficiently faithful, stable, and understandable for its intended use.

6. Analyst review: Present the alert, explanation, confidence information, and supporting evidence to a security analyst.
7. Feedback and monitoring: Record analyst feedback and monitor changes in model performance, explanation behavior, and input data.

B. Flowchart

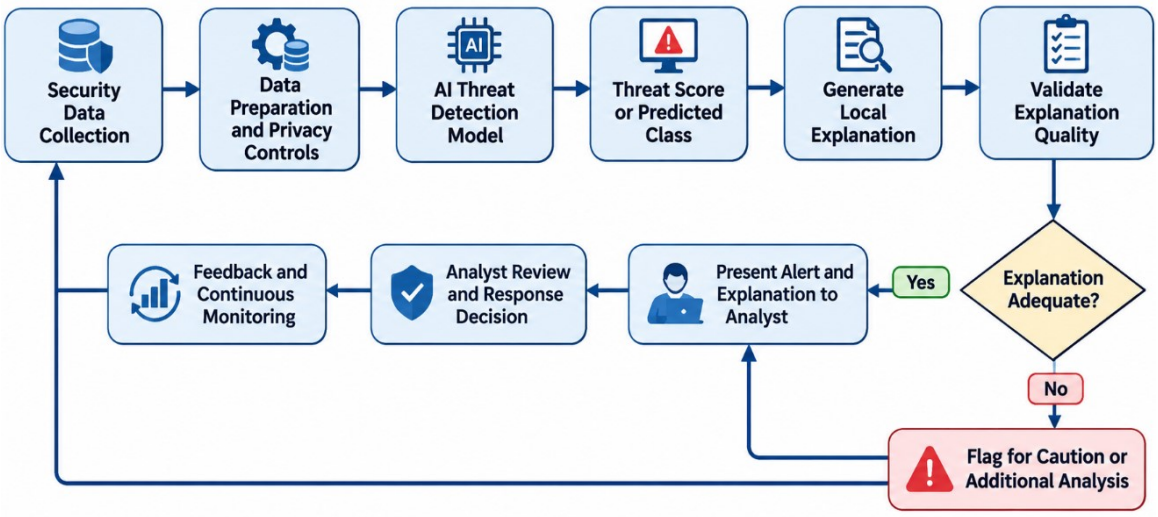


Figure 1. Proposed XAI-Enabled Cybersecurity Threat Detection Workflow

As shown in Figure 1, the proposed workflow generates an explanation after the threat model produces a prediction. The explanation is checked before it is presented to an analyst, and analyst feedback supports subsequent monitoring.

C. Explanation Presentation

A useful analyst-facing explanation should contain:

- the predicted threat category and model score;
- the most influential input features;
- the direction of each feature's contribution, where supported;
- relevant event context and timestamp;
- a clear statement that the explanation is model-derived and may be incomplete;
- links to underlying security evidence, where available.

An explanation should not replace the original event data or incident-response procedure. Analysts should be able to inspect the evidence that produced the prediction.

V. EVALUATION FRAMEWORK

A. Detection Performance

The detection model should be evaluated separately from the explanation method. Common classification metrics include precision, recall, F1-score, and false-positive rate. In cybersecurity, recall is important when missed attacks have high cost, while precision is important when excessive false alerts consume analyst time.

For a binary classification task:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Here, (TP), (FP), and (FN) represent true positives, false positives, and false negatives.

B. Explanation Quality

Explanation evaluation should consider more than whether a visualization looks clear. Useful dimensions include:

- Fidelity: Does the explanation accurately represent the model's behavior?
- Stability: Do similar inputs receive reasonably consistent explanations?
- Sparsity: Is the explanation concise enough to be inspected?
- Efficiency: Can explanations be generated within operational time limits?
- Robustness: Does the explanation remain meaningful under small, reasonable input changes?
- Completeness: Does the explanation capture the relevant aspects of the model decision?

These dimensions are consistent with prior work evaluating XAI methods for network intrusion detection.

C. Analyst-Centered Evaluation

Technical explanation metrics do not fully measure whether analysts can use an explanation. A human evaluation could ask participants to review alerts with and without explanations and measure decision accuracy, time to decision, confidence calibration, and agreement between analysts.

Such an evaluation should use a documented protocol and avoid assuming that faster decisions are necessarily better. Analysts should also be tested on whether they can identify misleading or insufficient explanations.

D. Reproducible Experimental Design

A future implementation should specify the dataset, preprocessing steps, model architecture, train-test split, explanation method, software versions, and hardware environment. The comparison should include the same detection model with and without the explanation interface, where appropriate.

The paper should report actual measured results only after the framework has been implemented and tested. No numerical performance improvement is claimed in this conceptual study.

VI. CHALLENGES AND LIMITATIONS

A. Explanation Fidelity

An explanation may provide a plausible story without accurately describing the model's internal behavior. This risk is significant when analysts use explanations to justify blocking, escalation, or containment actions. Explanation methods should therefore be tested against the model they explain.

B. Adversarial Manipulation

Attackers may attempt to evade detection by changing input features. In some settings, information exposed by an explanation could also help an attacker infer which features influence a model. This creates a tension between transparency and information disclosure.

C. Data Quality and Dataset Shift

Cybersecurity data can contain missing fields, noisy labels, duplicated events, and changing attack patterns. A model trained on historical data may behave differently when deployed in a new environment. Explanations can also shift when the input distribution changes.

D. Privacy and Sensitive Information

Security records may contain personal identifiers, device details, or confidential operational information. Explanation interfaces should avoid exposing unnecessary sensitive data. Access control, data minimization, and appropriate retention policies are important.

E. Computational Overhead

Some explanation methods require additional model evaluations or complex calculations. This may create delays in high-volume or real-time systems. Practical deployments should measure explanation latency and define when explanations are generated—for every event, only for high-risk alerts, or on analyst request.

F. Human Factors

Analysts may over-trust explanations, especially when they are presented as precise numerical values or polished visualizations. Interfaces should communicate uncertainty and make it easy to inspect source evidence. Training and clear operational guidance are also necessary.

VII. FUTURE RESEARCH DIRECTIONS

A. Robust and Faithful Explanations

Future work should investigate explanation methods that remain faithful under input perturbations and adversarial conditions. Robustness should be tested rather than assumed.

B. Real-Time Explanation Generation

Research should examine methods that reduce explanation latency while preserving useful information. Adaptive explanation strategies may generate detailed explanations for high-risk events and lightweight summaries for routine alerts.

C. Human-in-the-Loop Security Operations

Future systems should be evaluated with security analysts in realistic workflows. Research should determine whether explanations improve triage quality, reduce avoidable investigation effort, and support appropriate trust.

D. Cross-Dataset and Cross-Environment Generalization

An explanation method that works on one dataset may not transfer to another network or organization. Evaluation should include multiple environments and document differences in data, model configuration, and threat categories.

E. Standardized Evaluation

The field would benefit from consistent reporting of detection performance, explanation quality, computational cost, and analyst outcomes. Standardized evaluation would make comparisons more meaningful and reduce the risk of treating visually appealing explanations as proof of reliability.

VIII. CONCLUSION

Explainable artificial intelligence can make cybersecurity threat detection more transparent by showing which input features influenced a model's prediction. Methods such as SHAP, LIME, feature importance, and rule-based explanations offer different ways to interpret model behavior, but their suitability depends on the detection task and operational setting. This paper proposed a conceptual workflow that connects threat detection, explanation generation, explanation validation, analyst review, and continuous monitoring. The framework emphasizes that explanations should support—not replace—security evidence and human judgment. It also identifies fidelity, stability, robustness, privacy, computational cost, and analyst usability as important challenges. Published research provides a foundation for applying and evaluating XAI in intrusion detection, but the framework presented here has not been independently implemented or experimentally validated. Future work should conduct controlled evaluations using documented datasets, models, and explanation metrics, alongside analyst-centered studies. Such evaluation is necessary to determine whether XAI improves cybersecurity decisions in practice rather than merely making model outputs easier to visualize.

REFERENCES

1. A. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
2. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.

3. “A Survey on Explainable Artificial Intelligence for Cybersecurity,” IEEE Transactions on Network and Service Management, vol. 20, no. 4, pp. 5115–5140, 2023, doi: 10.1109/TNSM.2023.3282740.
4. “E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection,” IEEE Access, vol. 12, pp. 23954–23988, 2024, doi: 10.1109/ACCESS.2024.3365140.
5. “Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron,” IEEE Access, vol. 12, pp. 30164–30175, 2024, doi: 10.1109/ACCESS.2024.3368377.
6. M. Pawlicki, A. Pawlicka, R. Kozik, et al., “The Survey on the Dual Nature of xAI Challenges in Intrusion Detection and Their Potential for AI Innovation,” Artificial Intelligence Review, vol. 57, article 330, 2024.
7. “A Comprehensive Survey of Explainable Artificial Intelligence Techniques for Malicious Insider Threat Detection,” IEEE Access, vol. 13, pp. 121772–121798, 2025, doi: 10.1109/ACCESS.2025.3587114.
8. “Explainable Deep Learning Approach for Advanced Persistent Threats (APTs) Detection in Cybersecurity: A Review,” Artificial Intelligence Review, 2024.
9. “Explainable AI-Based Intrusion Detection Systems for IoT Environments: A Systematic Literature Review,” 2026.
10. “Explainable AI-Based Intrusion Detection Systems for Industry 5.0 and Adversarial XAI: A Systematic Review,” Information, vol. 16, no. 12, article 1036, 2025