EXPLAINABLE AI–DRIVEN CYBERSECURITY THREAT DETECTION: A MODERN FRAMEWORK FOR TRANSPARENT AND TRUSTWORTHY NETWORK DEFENSE

Dr. Jayesh Gamar[1]
Himani Gajjar[2]

1 Adhyapak Sahayak, Computer Science, V.P. & R.P.T.P. Science College, Vallabh Vidyanaagar, Anand, Gujarat, India
2 Assistant Professor, BCA, Bholabhai Patel College of Computer Studies, Gandhinagar Gujarat, India

ABSTRACT

The growing popularity of cloud computing, Internet of Things (IoT), Industrial Internet of Things (IIoT), Internet of Medical Things (IoMT), and next-generation (NextG) communication networks has introduced new complexities to the attack surface of modern computing networks. The traditional approaches to intrusion detection, relying on rule-based and signature matching strategies, have failed to handle the intensity, pace, and complexity of ever-evolving cyber-enabled attacks. Meanwhile, artificial intelligence (AI)-generated strategies have emerged successful in terms of detection efficacy, but their opaque black-box behavior has sparked deep concerns related to transparency, trust, regulatory requirements, and acceptance in the cyber community. In this regard, this paper describes an innovative explainable artificial intelligence (XAI)-based cybersecurity threat detection framework that integrates high-performance machine learning (ML) and deep learning (DL) engines with human-understandable explanation reasoning tools. The proposed XAI solution integrates feature optimization strategies, two-component ML/DL-based attack detection models, post-hoc explanation tools such as SHAP, LIME, ELI5, along with analyst-friendly visualization dashboards that could be associated with the MITRE ATT&CK knowledge base. Contrary to traditional studies on explainable intrusion detection, which revolve around model interpretability, the proposed solution is an operational, analyst-driven framework pertaining to XAI, which encompasses intrusion detection, explanation, visualization, as well as human feedback, all combined as an efficient workflow process following SOC, as well as the knowledge domain of MITRE ATT&CK. The framework is intended to enable various operational settings, such as IoT, IoMT, IIoT, Enterprise, and Government networks, while preserving near-state-of-the-art detection performance. With the incorporation of human-in-the-loop decision-making capabilities, the framework helps in building analyst trust, auditability, and ultimately achieves faster and more intelligent response actions for detected incidents. To practically and effectively embody the idea of Explainable AI in a cybersecurity defense solution, a strong architectural design, process flow, and application-centric analysis have been operationalized.

Keywords: **Explainable Artificial Intelligence, Intrusion Detection Systems, Cybersecurity, Threat Detection, SHAP, LIME**

1. INTRODUCTION

Shifts in the use of interconnected digital resources have generated a new operational environment within cyberspace, and new and developing technologies such as IoT, IIoT, and NextG wireless communications offer scalability and fully automated solutions, but also expose a larger attack profile to sophisticated cyber-attacks [1], [7]. Conventional intrusion detectors with rule or signature-based methodologies do not guarantee protections against novel attack patterns such as zero-day attacks or advanced persistent threats (APTs) [4].

As alternatives, machine learning algorithms, in conjunction with deep learning algorithms, prove to be effective in achieving a certain level of precision in the detection process in various benchmarking data sets [1], [7], [11]. Yet, the interpretability of the results effectively stands in the way of real-world application of these algorithms in the detection process, as security professionals need clear explanations for the alerts received and the manner in which the attacks act in order for them to take relevant measures [2], [3].

It also appears in contemporary research the significance of artificial intelligence in improving national and enterprise-level cybersecurity infrastructure, especially through automation and threat intelligence tools [15].

Although the existing XAI-driven intrusion detection approaches appear promising in the sense of interpretability, the relevant state-of-the-art work generally targets only a specific aspect of the problem, for example, explanation of the model, accuracy of the intrusion detection mechanism, among others.

1.1 Research Motivation and Key Contributions

Despite increasing attention in XAI research, much of the existing work concentrates on each isolated aspect and lacks end-to-end systems' integration. This paper fills in this research gap by introducing a holistic and layered approach for the integration of optimized data processing tasks and ML/DL models for detection along with visualization tools for security analysts. The main point of contribution is in showing how each can be effectively assembled together as a complete architecture in the area of cybersecurity.

The major novel contribution of the work may be listed as follows:

1. End-to-end operational architecture implementing the use of the XAI-IDS system in its pre-processing, hybrid machine-learning and deep-learning mechanisms, explanatory, visual, and analysis phases.
2. A two-level explanation technique, which integrates global model-level explanation and local instance-level explanation w.r.t. the MITRE ATT&CK techniques and tactics.
3. Analyst-in-the-loop feedback system that facilitates adaptive learning, trust calibration, and system improvement.
4. A framework that is not dependent on deployment scenarios in IoT, IoMT, IIoT, Next
5. A trust-aware explanation, focusing on dependability, action ability, and auditability in cybersecurity operations.

2. EXPLAINABLE AI–DRIVEN CYBERSECURITY FRAMEWORK

As seen in Chart 1 below, our framework has a layered architecture, displaying the interaction between data processing, detection models, explanation modules, and analyst interfaces in a layered manner.

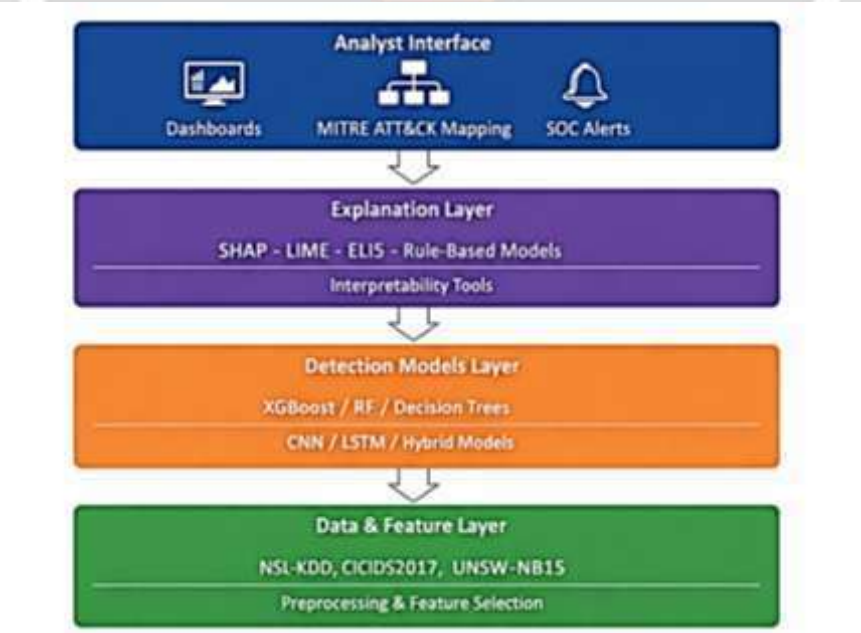

Chart-1: Layered Architecture of the Explainable AI Cybersecurity Framework

In Chart-1, there is a module-based structure that comprises four main layers, which include data/feature processing, detection models, mechanisms for explainability, as well as mechanisms for interfaces that are Analyst-facing.

2.1 Trust-Aware Explainability Layer

Even with improved transparency, not all explanations may be credible or of use to analysts. Towards a solution, a trust-aware explainability component is introduced in the proposed solution to analyze the credibility and consistency of explanations developed. The component is based on confidence levels of explanations, temporal consistency, as well as consistency with domain-specific knowledge such as MITRE ATT&CK. Low confidence explanations can be removed, and the model emphasizes stable features.

Trust Dimension	Description
Confidence	Reliability of explanation output
Stability	Consistency across similar attack instances
Action ability	Relevance to mitigation decisions

Table -1: Trust Dimensions Employed in the Proposed Trust-Aware Explainability Layer

2.2 Data and Feature Layer

The data and feature engineering level constitutes the building block for the framework. It obtains the unprocessed network traffic data from the sensors, logs, and network-monitoring tools and prepares them for model learning as processed data representations. Benchmark data sets like NSL-KDD, CICIDS2017, UNSW-NB15, and NF-CSE-CIC-IDS2018 are employed to handle various attack behaviors and network scenarios [4]. Preprocessing methods adopted for improving the data quality include min–max normalization, standardization, encoding categorical variables, and noise filtering. Metaheuristic-based feature selection methods, such as Mayfly optimization and Hiking algorithms, were employed for the problems of high dimensionality and associated computational costs. These techniques will help in identifying the most informative features that further enhance the detection accuracy and help reduce model complexity. Feature selection is guided not only by detection performance but also by explainability and relevance, ensuring that selected features remain interpretable to analysts.

2.3 Detection Models

The detection module applies both the classical machine learning techniques and deep models and their variations. Supervised machine learning algorithms: Decision Trees, Random Forest, and XGB, provide very good baseline models with inbuilt interpretability properties [2, 3]. Deep models, such as CNN, LSTMs, and SDAE, learn the temporal and spatial correlations in the vast majority of the reported works, which have all shown high detection accuracy above 97% in the detection task and IoMT as one of its most critical application areas [5, 8].

Layer	Techniques Used	Purpose	Representative Studies
Data Collection	NSL-KDD, CICIDS User, Min-Max normalization encoding Noise	Capture various patterns of normal and attack traffic on heterogeneous networks	Nalinipriya et al. [1], Arreche et al. [9]
Preprocessing	Metaheuristics (Mayfly, Hiking	Ensure clean and uniform network traffic to boost model robustness	Mohale & Obagbuwa [2], Arreche et al. [9]

Feature Selection	XGBoost, Random Forest, Decision Trees	Dimensionality reduction, feature selection	Nalinipriya et al. [1], Hafid et al. [8]
Detection Models	CNN, LSTM, SDAE, CNN-L	Precise and interpretable intrusion detection models	Gupta & Misra [3], Reddy et al. [11]
Deep Learning Models	CNN, LSTM, SDAE, hybrid CNN–LSTM	Learn complex temporal and spatial attack patterns	Hossain et al. [7], Nalinipriya et al. [1]
Explainability Module	SHAP, LIME, ELI5, rule-based models	Offer global and local explanations for alerts	Mohale & Obagbuwa [2], Dasari et al. [13]
Analyst Interface	Dashboards, Visual Analytics, MITRE ATT&CK Mapping	Facilitate an understanding, validation, and response to security threats by SOC analysts	Kütük et al. [6], Kalutharage et al. [10]
Feedback Loop	Analyst feedback, model training	Improvement and adaptation in the face of new threats	Gupta & Misra [3], Malik et al. [5]

Table -2: Example Layered Pipeline for Explainable AI–Based Threat Detection

The key layers, techniques, and purposes are summarized in Table 2 of the proposed framework, illustrating how detection performance and explainability are jointly achieved. The requirement for model choice in this proposed approach emphasizes explanations—performance trade-offs over meritorious correctness, achieving compatibility with explanation tools and interpretation by an analyst.

3. EXPLAINABILITY AND ANALYST INTERFACE

Current frameworks in XAI-IDS state that transparent models and strong attack detection systems ought to be combined to develop effective systems that anticipate transparency and evolving cyberattacks [17]. Explainability is achieved by using post-hoc XAI tools like SHAP, LIME, or ELI5 that offer global importance values for features as well as local interpretations for specific alerts [2], [12], [13].

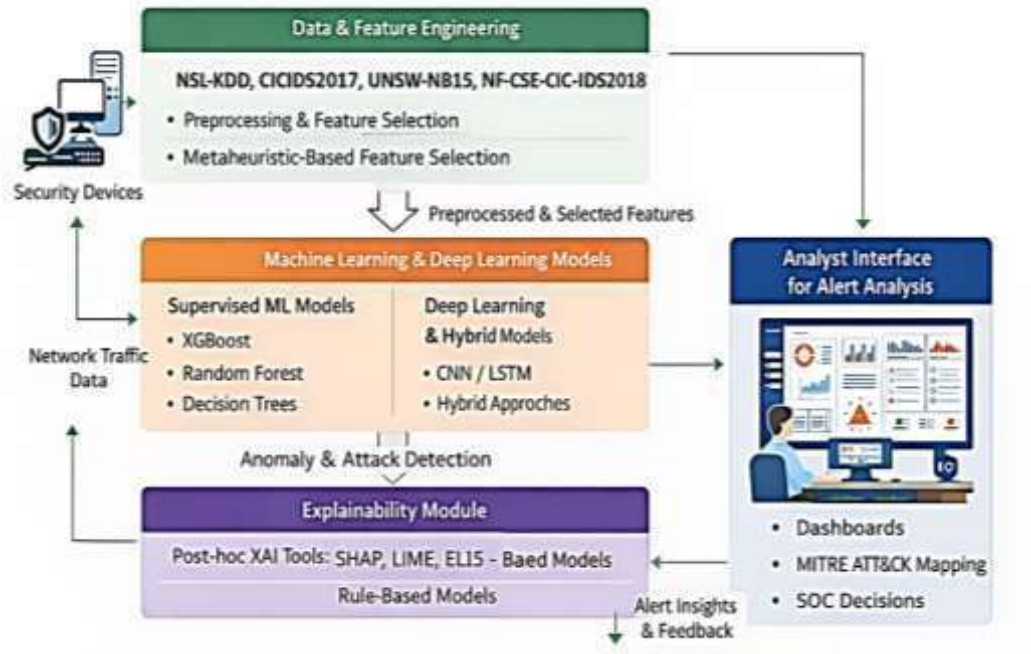

Fig-1: End-to-End Workflow of the Proposed Explainable Threat Detection System

Fig 1 shows the entire process chain, right from the ingestion of the data and detection process all the way to explanation generation and feedback for analysts.

3.1 Analyst-Facing Visualization

The role of the interface to the analyst cannot be understated when it comes to operational adoption. The interface presents the analyst with notifications, importance scores, and attack trends over a period of time. The integration with the MITRE ATT&CK framework helps an analyst understand attack tactics used by attackers [6], [10]. Current hybrid approaches for intrusion detection have recently confirmed the value of coupling real-time visualizations with machine learning algorithms in achieving a positive impact in Security Operation Centers (SOC) in terms of analyst response time and awareness of the situation [16]. In contrast to the dominant alert-oriented designs of dashboards developed thus far, the designed visualization layer targets explanation clarity, trusted information, and attack awareness. This is the kind of design that best fits natural processes within a SOC.

Fig-2: Analyst-Facing Explainability and Visualization Interface

Fig. 2 illustrates an instance of an alerting dashboard in the Security Operation Center where the alerts regarding explainability, feature importance, and MITRE ATT\&CK techniques can be visualized together to respond to an incident in an efficient and confident manner.

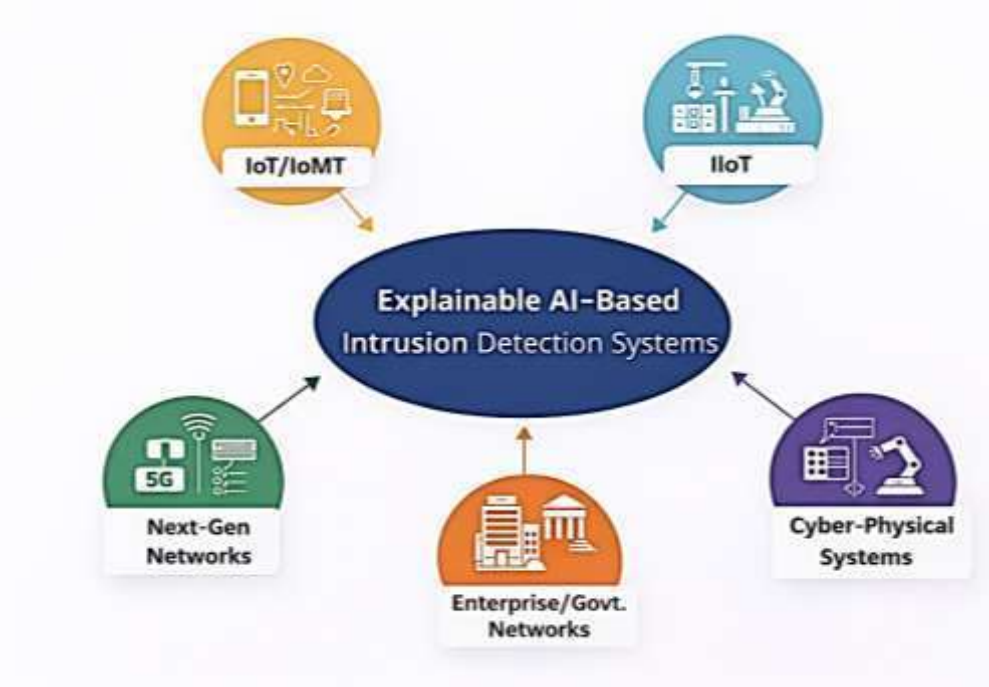

Chart-2: Application Domains of Explainable AI–Based Threat Detection

Chart-2 showcases prominent application sectors, such as IoT, IoMT, IIoT, NextG communications, and an enterprise network. From the chart, it is evident that Explainable AI-Intrusion Detection is universally applicable at various sites

4. CONCLUSION

The novelty of this research lies in applying the principles of Explainable AI in a cybersecurity domain using trust-aware Explainability, analyst workflow methods, and Explainability-driven response strategies in an integrated manner. This work provided a broad framework for AI-driven cybersecurity threat detection that is explainable and makes use of optimized machine-learning or deep-learning models, post-hoc explainability, and visualization support for analysts. Using a combination of accuracy, explainability, and human interpretability, the framework provided in the work fills very essential gaps in traditional black-box approaches used in AI. The inclusion of domain knowledge in the framework, based on MITRE ATT&CK mapping, and its support for current deployment environments, such as federated learning, adds to its applicability. Future work will be focused on real-time, adaptive, and large-scale testing.

5. REFERENCE

[1] Nalinipriya, G., Sree, R., Radhika, K., Lydia, L., Karim, F., Ishak, M., & Mostafa, S. (2025). Leveraging explainable artificial intelligence for early detection and mitigation of cyber threat in large-scale network environments. *Scientific Reports*, 15. https://doi.org/10.1038/s41598-025-08597-9

[2] Mohale, V., & Obagbuwa, I. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: enhancing transparency and interpretability. *Frontiers in Computer Science*. https://doi.org/10.3389/fcomp.2025.1520741

[3] Gupta, L., & Misra, D. (2025). Cybersecurity Threat Detection Through Explainable Artificial Intelligence (XAI): A Data-Driven Framework. *International Research Journal of MMC*. https://doi.org/10.3126/irjmmc.v6i2.80687

[4] Mutalib, N., Sabri, A., Wahab, A., Abdullah, E., & Aldahoul, N. (2024). Explainable deep learning approach for advanced persistent threats (APTs) detection in cybersecurity: a review. *Artificial Intelligence Review*, 57. https://doi.org/10.1007/s10462-024-10890-4

[5] Malik, A., Arshid, K., Noonari, N., & Munir, R. (2025). Artificial Intelligence-Driven Cybersecurity Framework Using Machine Learning for Advanced Threat Detection and Prevention. *Scholars Journal of Engineering and Technology*. https://doi.org/10.36347/sjet.2025.v13i06.005

[6] Kütük, A., Çoşkun, Ö., Kütük, H., & Kök, İ. (2025). Machine Learning Models and Explainable Artificial Intelligence Approaches for Intrusion Detection in IoT Networks. *The European Journal of Research and Development*. https://doi.org/10.56038/ejrnd.v5i1.630

[7] Hossain, M., Alam, K., Monir, M., Hoque, M., & Ahmed, T. (2025). Explainable AI Meets Synthetic Data: A Deep Learning Framework for Detecting Network Intrusion in NextG Network Infrastructure. *IEEE Access*, 13, 114979-115001. https://doi.org/10.1109/access.2025.3585783

[8] Hafid, A., Rahouti, M., & Aledhari, M. (2025). Optimizing Intrusion Detection in IoMT Networks Through Interpretable and Cost-Aware Machine Learning. *Mathematics*. https://doi.org/10.3390/math13101574

[9] Arreche, O., Guntur, T., & Abdallah, M. (2024). XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems. *Applied Sciences*. https://doi.org/10.3390/app14104170

[10] Kalutharage, C., Liu, X., & Chrysoulas, C. (2025). Neurosymbolic learning and domain knowledge-driven explainable AI for enhanced IoT network attack detection and response. *Comput. Secur.*, 151, 104318. https://doi.org/10.1016/j.cose.2025.104318

[11] Reddy, D., Mrudula, B., Goud, S., Saida, B., & Ramesh, M. (2025). Network Shield: Machine Learning Based Threat Detection. *International Research Journal on Advanced Engineering and Management (IRJAEM)*. https://doi.org/10.47392/irjaem.2025.0286

[12] Arreche, O., Guntur, T., Roberts, J., & Abdallah, M. (2024). E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection. *IEEE Access*, 12, 23954-23988. https://doi.org/10.1109/access.2024.3365140

[13] Dasari, A., Bisawas, S., & Purkayastha, B. (2025). Enhanced Network Intrusion Detection Systems With Explainable Artificial Intelligence for Network Security. *International Journal of Communication Systems*, 38. https://doi.org/10.1002/dac.70209

[14] Shoukat, S., Gao, T., Javeed, D., Saeed, M., & Adil, M. (2024). Trust my IDS: An explainable AI integrated deep learning-based transparent threat detection system for industrial networks. *Comput. Secur.*, 149, 104191. https://doi.org/10.1016/j.cose.2024.104191

[15] Kumar, K. (2025). Artificial Intelligence for Improving Cybersecurity Framework. *INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*. https://doi.org/10.55041/ijsrem48949

[16] G, M., Lutimath, K., N, M., & R, P. (2025). A Hybrid Machine Learning Approach for Network Intrusion Detection with Real-Time Visualization. *Indian Journal of Computer Science and Technology*. https://doi.org/10.59256/indjcst.20250402009

[17] Salloum, S., & Norozpour, S. (2025). XAI-IDS: A Transparent and Interpretable Framework for Robust Cybersecurity Using Explainable Artificial Intelligence. *SHIFRA*. https://doi.org/10.70470/shifra/2025/004