International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

Explainable AI (XAI) for Interpretable Cyber Threat Prediction

Mohammed Sadath P[1], R. Kaviyarasi[2]
[1]Research scholar - Yenepoya (Deemed to be University), Bangalore, Karnataka, India.
[2]Associate Professor - Yenepoya (Deemed to be University), Bangalore, Karnataka, India.
Email ID: *sadathmsc@gmail.com[1], r.kaviyarasi.blr@yenepoya.edu.in[2]*

Abstract
Due to the rapid evolution of cyber threats such as intrusions, botnets, DDoS assaults, and insider threats, malware, advanced persistent threats (APTs), detection and mitigation now require machine learning (ML) and deep learning (DL) models that are smart. Nonetheless, due to the "black-box" nature, they can hinder trust, interpretability, and it will uptake in high-stakes settings-an understanding of how the decision is made by the stakeholders is important. When transparency is introduced into ML/DL frameworks, Explainable Artificial Intelligence (XAI) helps to fill the gap and it allows human to understand and to provide any performance issues. The synthesized review incorporates 15 researches on XAI applications in cybersecurity to IDS, malware analysis, cyber risk assessment, threat prediction in IoT, finance & decentralized smart grid. Common examples of explainable AI (XAI) techniques include SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME), and by using decision trees or rule-based models, or using hybrid ensembles such as XGBoost, Random forest or Convolution Neural Networks (CNNs). Techniques applied on datasets like NSL-KDD, UNSW-NB15, and CICIDS2017; achieve high metrics like 99% accuracy, precision, and AUC. They also allow local and global interpretability, revealing feature importance (like network traffic patterns, behavioural logs) and causal reasoning. The results showed the advantages of XAI in terms of less false positive and false negatives, analyst accuracy improvement (up to 31%), and increased manage trust. Dataset imbalances, scalability, standardization, and interpretability versus accuracy trade-offs remain issues. Surveys stress the importance of benchmarking framework, real-time explainability, and privacy-preserving AI and hybrid models. The goal of this study is to make cyber security stronger and clearer. It plans to do this using a few different methods: federated learning, systems where people are involved in the decision-making, and by following ethical guidelines. So, XAI helps us understand what AI-powered defences are doing, which makes managing cyber threats more responsible and effective.

Keywords: *Cybersecurity; Explainable Artificial Intelligence (XAI); Intrusion Detection Systems (IDS); Malware Detection; Hybrid Ensemble Models; Deep Learning; Machine Learning; SHAP; LIME; Threat Prediction.*

1. Introduction
The rapid growth of internet digital devices and the environment has created that can expand attack surface for malicious actors. Cyber threats like ransom ware, malware, Advanced Persistent Threats (APTs) and network intrusions has grown in both frequencies by causing economic loss. In the year 2020, cyber damage cost the global economy approximately $945 billion, and the average cost of a data breach reached $4.88 million by 2024, Annual cybercrime has projected to exceed $15 trillion by

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

2030 [4, 9]. Old Traditional methods such as intrusion detection systems (IDS), firewalls, signature-based and antivirus is increasingly ineffective against sophisticated, evolving threats including malware and zero-day exploits **[1, 7]**. The cybersecurity community has increasingly resorted to machine learning (ML) and deep learning (DL) techniques to overcome these constraints. These data-driven techniques have proven to be remarkably effective in a range of tasks, such as malware detection and botnet identification, because they provide great precision and the ability to recognize previously uncovered patterns. network intrusion detection, and malicious classification [2, 3, 6]. ML and DL models develop in different environment incorporating millions of parameters and including many layers, they unavoidably become opaque "black-box" systems. Because of this opacity, even though these models can provide excellent predicted performance, human operators are unable to access their internal reasoning processes, and that makes it practically impossible to comprehend the reasoning behind a specific conclusion [8, 12, 14]. The black-box style of ML and DL models presents difficulties in cyber security settings. Security experts and network administrators must not only identify threats but also know from it in order to prioritize incidents, satisfy regulatory requirements, repair vulnerabilities, and discover attacks. In high-stakes fields like health care, banking, and other national infrastructure, binary predictions are adequate; analysts require clear, clear reasons that can bolster accountability and support their model decisions [13, 15]. Its restrictions, such as the Protection Regulation (GDPR) and EU General Data, which demand and will receive some explanations for this kind of impact choices, further emphasize the legal requirement for model integrity. [3]. A proposed paradigm to close the gap between model accuracy and human interpretability is Explainable Artificial Intelligence (XAI). A wide range of computational methods, such as SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), decision tree rule extraction, and feature importance ranking, are included in XAI. These methods are intended to make complex ML/DL model predictions transparent, intelligible, and useful [1, 2, 11]. XAI enables cybersecurity experts to trust AI-driven systems, uncover adversarial vulnerabilities, detect model biases, and make well-informed judgments by offering both local explanations (explaining individual predictions) and global explanations (revealing overall model behavior) [3, 8]. Additionally, XAI makes a crucial distinction that has frequently been missed in earlier research: interpretable models—those that are transparent by design, such decision trees—and post-hoc explainability techniques applied to black-box models [1, 2]. Even though there are a larger number of studies that use XAI for the cybersecurity domains, somewhat a bigger picture is missing. While there are existing survey articles on XAI in certain application domains, such as healthcare and finance, a comprehensive review of the landscape of XAI for cybersecurity is missing. For example, the application areas such as malware analysis, intrusion detection, detection of advanced persistent threat (APT), IoT botnet detection, distributed denial-of-service (DDoS) attack detection, detection of malicious domain, insider threat detection, etc., have been studied separately without being compiled together [3, 4, 15]. Most of these prior works focused on improving the accuracy of classification only, ignoring the interpretability and transparency of the underlying AI/ML models [2, 14]. Also, evaluation is mostly limited to benchmark datasets, such as KDD Cup, focusing on the Android malware only, without giving much attention to Windows, PDF, Linux, hardware-based malware, etc. [1]. This paper aims at XAI techniques and their applications in critical cyber security. Specifically, we explore and summarize the state-of-the-art research on XAI in cybersecurity, including intrusion detection systems (IDS) [8, 9, 14], malware analysis [1, 7], APT detection [4], IoT threat detection [6], malicious domain detection [10], DDoS attack detection [11], financial cybersecurity [13], smart grid security [5], and insider threat detection [15]. The review paper

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

contributes in the following way

- We present a structured classification of XAI techniques used in cyber security, categorizing methods based on their scope (global or local) model-dependency (model-agnostic vs. model-specific), and explanation type (feature-based, rule-based, attention-based, and example-based) [1, 3].
- A comprehensive synthesis of ML/DL approaches and XAI methods across key cybersecurity application domains, including IDS, malware detection, APT detection, and insider threat detection [2, 4, 12, 15].
- A critical analysis of existing research gaps, including dataset biases, limited platform coverage, and the underexplored adversarial robustness of XAI models themselves [1, 3, 11].
- Identification of future research directions, including the need for standardized evaluation benchmarks, cross-platform malware studies, real-time XAI integration, and regulatory-compliant explainability frameworks [3, 9, 13].

2. Literature Review

Explainable Artificial Intelligence (XAI) in cybersecurity focuses on intrusion detection systems (IDS), malware detection, and network security while continuously tracking the evolution from conventional detection methods to deep learning methods for model transparency.

2.1. Background: From Traditional Methods to AI-Based Security

Conventional cybersecurity relies on rule-based and signature-based approaches, but these are unable to withstand increasingly complex attacks [3, 15]. Machine learning (ML) and deep learning (DL), which have shown remarkable results in intrusion detection, spam filtering, fraud detection, and malware identification, it can be used as a result [3, 7, 15]. Intrusion detection systems (IDS) generally: network-based (NIDS) and host-based (HIDS) [14]. Both are beneficial, but they are ineffective against sophisticated threats like Advanced Persistent Threats (APTs), which are focused, long-term campaigns intended to avoid detection, and zero-day attacks. [4].

2.2. Why Explain ability Matters: The 'Black-Box' Problem

The literature often refers to the "black-box" issue of deep learning models that are too complex for analysts to grasp and often put accuracy over explainability [1, 3, 6, 11, 13]. This opacity poses practical challenges that includes analysts to verify alert assignment and submit trust [2, 6], and the challenge finding bias and rank dangers in high-stakes industries like banking [13]. The clarity of automated choices is now legally affects the EU's GDPR, making interpretability a requirement for compliance. [3, 12].

2.3. Overview of XAI Methods Used in Security Research

Several XAI models used in cyber security are used in the literature:

- **LIME** modifies input data to find the characteristics that have the biggest impact on a particular choice [4, 6, 8].
- **SHAP** assigns feature importance scores based on game theory, which explains why a particular activity was flagged [4, 6, 14].
- **LRP** follows relevance scores backward through neural network layers [4].
- **Attention mechanisms** draw attention to aspects that the model prioritized when making decisions [4].
- **Rule-based models** (e.g., Decision Trees, LENs) offer transparent 'If-Then' logic that practitioners find intuitive and trustworthy [1, 2].

A significant different technique between post-hoc explain ability by applying XAI after training (e.g., SHAP, LIME) and intrinsic explain ability, where it is built into the model design. While predominantly employs post-hoc methods, many authors argue that that the embedded XAI from the outset yields more depended explanations [11].

2.4. XAI Applied to Malware Analysis

In Various studies we found that, XAI for enhanced

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

malware detection. Static analysis had the features like API calls and other opcodes from PE files, dynamic analysis uses system calls from controlled running environments, and hybrid approaches combine both [1].Among the notable systems are XRan, which uses CNNs with LIME and SHAP for ransomware detection [1], TabLSTMNet, which combines dual architectures with LIME and SHAP for Android malware classification [1], PAIRED, which is a lightweight, highly accurate explainable Android detection system [1], and LEMNA, which is intended for high-fidelity explanations of PDF malware [1]. PREEMPT used Random Forest and Decision Trees with hardware performance counters to address hardware-level threats [1]. Post-hoc XAI integration is necessary since models such as Random Forest, SVM, and LSTM showed excellent precision in harmful domain detection but remained mostly "black-box" [10].

2.5. XAI for Intrusion Detection and Network Security

Although it is becoming more popular, XAI for IDS is still less frequent than in industries like computer vision and healthcare [8]. Wang et al.'s SHAP-based framework for local and global explanations, Patil et al.'s combination of ensemble methods and LIME for increased reliability, and Marino et al.'s use of adversarial strategies to explain misclassifications are important studies [8]. Deep CNNs for traffic classification, Fish Swarm Optimization for feature selection, and evolutionary algorithms for optimization are methods used for IoT and industrial security [7]. In DDoS detection, while CNN and RNN models accurately identify traffic patterns, XAI-specific frameworks for DDoS remain sparse – a major gap in the literature [11]. LSTM models capture long-term temporal patterns, while auto-encoders identify behavioral abnormalities for APTs [4]. On benchmark datasets, ensemble trees and SHAP have demonstrated up to 100% accuracy; however, real-time deployment is hampered by lengthy training cycles [4].

2.6. Smart Grid and Domain-Specific Applications

XAI has been used for smart grid stability prediction in addition to cybersecurity. Arzamasov et al. and Breviglieri et al. used decision trees and optimized deep learning to obtain excellent accuracy, while Ucar and Allal et al. used SHAP, PDP, and LIME, pointing out that PDP lacks granular detail and LIME yields inconsistent results between runs [5]. This demonstrates that the fundamental issues facing XAI go beyond cybersecurity.

2.7. Identified Gaps and Future Directions

The evaluated literature reveals a number of common gaps. With little research done on Windows, Linux, and macOS, malware detection research is overly dependent on the Android platform [1]. A lot of research relies on out-of-date or unbalanced databases, including DREBIN, which is skewed toward benign samples [1]. The computational expense of XAI in real-time situations is still mostly neglected [7, 12], and XAI is primarily applied post-hoc rather than integrated into model construction [11]. Furthermore, as LIME and SHAP have been demonstrated to be vulnerable to adversarial manipulation [3], XAI tools itself contain security flaws. Lastly, the lack of a uniform framework for assessing and contrasting XAI techniques has been repeatedly noted in the literature [8].

3. Methodology

The research approaches for XAI in cybersecurity are examined in this synopsis, which includes datasets, tools, algorithms, workflows, and mathematical models.

3.1. Description of the Dataset

The assessed publications used a range of benchmark datasets. The extent and nature of the attack or threat under study were significantly influenced by the dataset selection. Shown in Table 1.

Table 1 Comparative Overview of Datasets for IDS, APT Detection, and Cybersecurity Analytics.

Paper(s)	Dataset	Size / Key Features	Purpose

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

[2]	KDD Benchmark	Standard network traffic records	IDS rule extraction with Decision Trees
[4]	UNSW-NB15	49 features, 9 attack types	APT / general intrusion detection
[4]	CICIDS2017	78 features, 3M+ samples	Large-scale APT detection
[4]	ZYELL	Real-world traffic (Probing, DoS)	Real-world APT scenario testing
[5]	UCI Smart Grid	10,000 instances, 12 variables	Smart grid stability prediction
[6, 11, 14]	UNSW-NB15	175,341 train / 82,332 test	IDS hybrid and DDoS classification
[8]	ADFA-LD	System call sequences (30 per input)	Anomaly detection (MLP-based IDS)
[9]	NSL-KDD	22,544 records, 41 features → 10	Ensemble ANN+SVM IDS
[10]	Custom Domain Dataset	90,000 samples, 32 features	Malicious domain detection
[11]	NSL-KDD (DDoS subset)	Neptune & Smurf attack records	DDoS-specific CNN detection
[13]	Financial Cybersecurity	1.2M fraud + 4.8M intrusion + 930K auth logs	Cyber risk in financial services
[14]	UNSW-NB15	2.5M+ records, 9 attack types	Comprehensive IDS evaluation
[15]	CERT r6.2, NSL-KDD, CICIDS2017	Multiple large-scale datasets	Insider threat detection (ITD) review

Because of its variety of assault methods, the UNSW-NB15 dataset was the most popular [4, 6, 11, 14]. Instead, papers [1] and [3] carried out thorough literature searches from 2011 to 2024 using IEEE Xplore, Google Scholar, ScienceDirect, ACM, Springer, ResearchGate, and arXiv.

3.2. Tools, hardware and libraries used

XAI Libraries: Almost all experimental papers used SHAP for both local and global explanations [4, 5, 6, 8, 9, 10, 11, 13, 14], LIME for instance-level explanations via input perturbation [2, 6, 7, 8, 10, 11, 14], ELI5 for model weight visualization [6, 14], LRP for tracing relevance through neural layers [4], Grad-CAM for CNN-based explanations [2], and surrogate-rule explanations for financial cybersecurity models [13].

ML/DL Frameworks: TensorFlow/Keras or PyTorch for deep learning architectures [4, 7, 8, 9, 11], XGBoost, CatBoost, and AdaBoost for ensemble methods [5, 6, 10, 13, 14], Flask for real-time pipeline deployment [9], Cuckoo Sandbox for dynamic malware analysis [2], and Scikit-learn for standard models (Random Forest, SVM, Decision Trees).

3.3. Algorithms and Model Architectures

Ensemble Methods: XGBoost and GBoost used gradient boosting [5, 6, 10, 13, 14], AdaBoost and CatBoost provided accuracy-interpretability trade-offs [5, 10, 14], Random Forest was a popular meta-learner for stacking ensembles in papers [9] [1, 6, 9, 10, 14], and M3L employed Laplace smoothing for IDS classification [6].

Deep Learning: Auto encoders modeled normal behavior to flag deviations [4], MLPs classified anomalies from system call sequences [8], a stacked ANN-SVM ensemble used Random Forest as a meta-learner [9], CNNs extracted spatial features from

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

network traffic and malware data, including a 1D CNN for DDoS detection [2, 4, 7, 11], RNNs and LSTMs captured temporal patterns, especially for APT detection [4, 9], and Sparse Denoising Auto encoders handled noisy or incomplete input data [7].

3.4. General Research Workflow

From data collection to XAI-based interpretation, a consistent pipeline appears in all experimental studies. This generalized workflow is shown in the table below. Instead of using an experimental pipeline, survey-based papers [1, 3, 12, 15] used a structured literature review procedure that included keyword searching, abstract screening, full-text review, taxonomy construction, and gap identification. Shown in Table 2.

Table 2 Workflow Pipeline for Cybersecurity Analytics: From Data Collection to XAI-Driven Deployment

Step	Stage	Description
1	**Data Collection**	Gather benchmark datasets (e.g., UNSW-NB15, NSL-KDD, KDD) or conduct systematic literature search across IEEE Xplore, Google Scholar, ACM, etc. [1, 3, 4, 6, 9, 11, 14]
2	**Preprocessing**	Clean data, apply Min-Max normalisation to scale features to [0,1], encode categorical labels, and remove duplicates or irrelevant entries [7, 11, 14]
3	**Feature Selection**	Reduce dimensionality using methods such as Recursive Feature Elimination (RFE) [9], Sequential Forward Selection [10], or Mayfly Optimization Algorithm (MOA) [7]
4	**Model Training**	Train ML/DL models (e.g., Random Forest, CNN, LSTM, MLP, XGBoost). Tune hyper parameters using algorithms like HOA [7] or cross-validation
5	**XAI Integration**	Apply post-hoc techniques — SHAP for global/local feature importance, LIME for instance-level explanations, LRP for layer-level neural network analysis [4, 6, 8, 9, 11]
6	**Validation**	Validate explanations via perturbation analysis (modifying 'important' features to confirm model response changes) [8], user surveys [8], or comparison against ground-truth labels
7	**Evaluation**	Measure performance using accuracy, precision, recall, F1-score, AUC, False Positive Rate (FPR), plus XAI-specific metrics: fidelity, stability, and explanation complexity [12, 13]
8	**Deployment (optional)**	Deploy as real-time pipeline using Flask on cloud infrastructure (Amazon EC2) with live packet capture [9]

3.5. Mathematical Models

SHAP uses Shapley values from cooperative game theory, applied both globally and locally, to calculate the average marginal contribution of each feature across all feasible feature subsets [4, 6, 8, 9, 11]. Smart Grid (DSGC): A second-order differential equation governing price-frequency relationships and physical oscillator dynamics is used in Paper [5] to simulate grid stability. Here, τ stands for reaction time, γ for price elasticity, p for nominal power, and Δf for frequency deviation. Recursive Feature Elimination (RFE): Paper [9] reduces 41 features to

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

the 10 most predictive for IDS by iteratively eliminating the least significant features using Gini-based significance scores. In order to prevent high-magnitude features from controlling model training, Min-Max Normalization is used in publications [7] and [11] to scale all feature values to [0, 1]. Evaluation Metrics: Papers [12] and [13] include XAI-specific metrics, such as fidelity (the degree to which explanations accurately reflect model behavior), stability (consistency across similar inputs), and complexity (interpretability for human analysts), in addition to standard metrics (Precision, Recall, F1, AUC-ROC, FPR) [12, 13].

4. Results and Discussion

Malware analysis, intrusion detection, APT detection, smart grid security, and financial cyber risk are among the cybersecurity domains where XAI applications are presented in this part. These applications are assessed through performance outcomes, comparative analysis, significance, and critical discussion.

4.1. Results: Model Performance

4.1.1. Overall Performance Summary

Accuracy was consistently higher than 93% in all experimental publications. Key performance metrics from the analyzed studies are combined in the table 3 below.

Table 3 Consolidated model performance across Reviewed papers [1–15].

Paper	Model / Method	Dataset	Accuracy / Key Metric
[1]	Ensemble + Image-based	Various malware datasets	99.87% accuracy
[1]	PAIRED (CNN + XAI)	Android DREBIN	97.98% accuracy (84% fewer features)
[2]	SVM (Watson et al.)	Performance counter data	90% accuracy
[2]	3D CNN (Abdelsalam)	Cloud network traffic	90% accuracy
[2]	CNN + LIME	DREBIN (5,560 apps)	~98% accuracy
[3]	XAI Malware Classifier	>8,000 samples	97% test accuracy
[3]	BiLSTM-XAI	Honeypot dataset	97.2% detection rate
[4]	Ensemble Trees + SHAP	NF-BoT-IoT v2 / NF-ToN-IoT-v2	100% accuracy
[4]	Tree-CNN Algorithm	Multiple attack datasets	98% avg. accuracy, 36% faster
[4]	RFC + SHAP	CICIDS / Hop Skip Jump	98.5–100% accuracy
[5]	ANN + SHAP/ICE	UCI Smart Grid (10,000 inst.)	AUC 99.4%, CA 96.2%
[6]	M3L + ELI5	UNSW-NB15	97.38% accuracy (post-optimization)
[6]	Random Forest + ELI5	UNSW-NB15	95.54% accuracy
[7]	LXAIDM-CTLSN (SDAE)	NSLKDD2015 / CICIDS2017	99.09% accuracy
[8]	MLP + LIME/SHAP	ADFA-LD	93.71% accuracy, 98.06% precision
[9]	ANN+SVM+RF Ensemble	NSL-KDD	99.40% accuracy
[10]	XGBoost + SHAP/LIME	Custom 90,000-sample dataset	AUC 0.9991, Acc. 0.9856

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

[11]	1D CNN + SHAP/LIME	NSL-KDD (DDoS subset)	Acc. 94%, F1 94%
[13]	SHAP + Surrogate rules	Financial (1.2M+ records)	Fidelity r=0.69–0.76 with F1
[14]	XGBoost / CatBoost	UNSW-NB15	87% accuracy (top models)
[15]	KNN (SHAP)	LYCOS-IDS2017	AUC 0.99
[15]	Auto-Encoders	CICIDS2017	Accuracy up to 0.92

4.1.2. Top-Performing Models and Key Figures

Ensemble-based systems achieved the highest accuracy. Decision Tree and Random Forest with SHAP reached 100% on IoT datasets [4]; the ANN+SVM+Random Forest ensemble achieved 99.40% on NSL-KDD [9]; and the Sparse Denoising Auto encoder model reached 99.09% on NSLKDD2015 and CICIDS2017 [7]. In malware detection, an image-based ensemble achieved 99.87% accuracy, while PAIRED reduced features by 84% — from 214 to 35 — while maintaining 97.98% accuracy [1]. For smart grid prediction, SHAP identified reaction time (τ) as the most influential predictor, with the ANN achieving 99.4% AUC and 96.2% accuracy [5]. The 1D CNN for DDoS detection achieved 94% accuracy and F1-score [11]. In financial cybersecurity, SHAP-based interpretability reduced ambiguous analyst alerts by 18–27% and improved decision accuracy by 22–31% [13].

4.1.3. XAI-Specific Findings

With an average change in model output of 65.73% when the top-10 system calls were substituted, SHAP proved to be the most successful XAI tool for both global and local explanations [8]. Technical analysts liked SHAP for mathematical consistency, whereas non-technical consumers preferred LIME for visual clarity [8]. In addition, when LIME and ensemble models were integrated, accuracy was improved by up to 15%, showing that XAI can improve accuracy, not just explain it [12]. Although it was established that sttl and ct_srv_dst were the most significant indicators of malicious activities in UNSW-NB15 [14], SHAP plots showed that reaction time (τ) and price elasticity (γ) were the most significant smart grid attributes [5].

4.2. Comparison with Existing Studies

A lot of models are compared in the reviewed studies. These results are categorized in the table below. The major compromise is based on interpretability and deep learning accuracy, as Decision Trees are interpretable but not accurate, while CNNs, LSTMs, and DNNs achieve maximum accuracy but remain opaque [2, 3, 11]. This is evident in Paper [6] when a standalone Decision Tree was trained at 88.66%, but ELI5-based feature selection improved the accuracy of the M3L classifier to 97.38%, narrowing the gap but at the cost of interpretability. Similarly, XGBoost and CatBoost surpassed Gaussian Naive Bayes, which was trained at 87%, compared to CatBoost at 73%, even though it was faster and more suitable for real-time use [14]. The Tree-CNN hybrid model, as described in Paper [4], performed better than both approaches individually, resulting in a reduction of execution time of 36%, but also ensuring accuracy of 98%. According to a research paper [15], SOMs were able to achieve accuracy of 0.91, and for KNN, AUC of 0.99 was achieved, which proves that under proper circumstances, even simpler models, coupled with proper XAI, can compete with deep learning. It is also important to note that, as described in research [3], both LIME and SHAP are vulnerable to adversarial attacks, which is a new threat. [3]. Shown in Table 4.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

Table 4 A comparative overview of interpretability, constraints, and models in various publications [1–15].

Approach	Representative Papers	Accuracy Range	Interpretability	Limitation
Ensemble ML + XAI	[1],[4],[6],[9],[10]	95–100%	High (SHAP/LIME)	Slow training; limited real-time use
Deep Learning (CNN/LSTM)	[2],[3],[7],[11]	93–99%	Medium (post-hoc)	Black-box; computationally heavy
White-box Models (DT/Rules)	[2],[12],[14]	87–89% (DT baseline)	Very High (intrinsic)	Lower raw accuracy vs. DL
Hybrid Ensemble (ANN+SVM+RF)	[9]	99.40%	High (SHAP global)	Complex pipeline design
Anomaly Detection (AE/SOM)	[15]	0.91–0.92 AUC	Moderate	High false-positive risk
XAI in Finance/Smart Grid	[5],[13]	AUC 0.99 / r=0.76	High (SHAP/ICE)	Domain-specific, limited transfer

5. Interpretation: Why the Results Matter

Though it is evident from the above discussion that ML's maturity as a cybersecurity tool is established from the accuracy of all studies, it is also evident that there is deeper significance when considering the practical and legal scenario.The need for explainable automated judgments is now a requirement under GDPR, making XAI a compliance requirement rather than a "nice-to-have" improvement [3]. The financial cybersecurity data also indicates that interpretability can improve human performance, increasing accuracy by 22-31% and reducing unclear alarms by 18-27% [13]. Furthermore, using the PAIRED system, it is evident that XAI can improve models so that they are simpler, faster, and more auditable, as it is capable of reducing features by 84% while achieving 97.98% accuracy. Last but not least, SHAP's identification of setsockopt as a common deceptive system call also indicates that XAI can reveal failure modes of models, which might otherwise be missed using accuracy metrics alone—knowing why a model is wrong is as important as knowing how often it is right [8].

6. Discussion: Strengths, Limitations, and Patterns

Strengths: SHAP and LIME showed model-agnostic adaptability in financial services, smart grids, malware analysis, IDS, and APT detection [1, 2, 4, 5, 6, 9, 11, 13]. When coupled, they are most effective, with SHAP offering global significance and LIME offering visual local explanations [8, 12]. The notion that interpretability compromises performance was refuted by XAI-guided feature selection, which increased ensemble accuracy by up to 15% [12]. Real-time deployment of lightweight XAI-enhanced models is also becoming feasible [9, 14]. Limitations: LIME's explanations differ between runs, which makes high-stakes decisions less reliable [1, 5]. Comprehensive APT datasets are still hard to get by [4], and many research rely on out-of-date datasets like DREBIN and KDD [1, 3]. Windows, Linux, and macOS are understudied in

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

malware XAI research, which is primarily focused on Android [1]. The high computational costs of ensemble models with SHAP restrict their use in real-time [4, 12]. There is currently no widely recognized criterion for assessing explanation quality [3, 13], and LIME and SHAP are susceptible to adversary manipulation [3]. Emerging Patterns: Three patterns stand out: the field is converging on SHAP, LIME, and ELI5 as standard tools, but there is still disagreement over how to assess explanation quality, which is the most important open research question [3, 12, 13]. Ensemble models consistently outperform single architectures and integrate naturally with SHAP [4, 6, 9].

Conclusion

Since high-accuracy ML and DL models are still too difficult to implement in the real world without transparent decision-making, Explainable Artificial Intelligence (XAI) has become crucial in cybersecurity. In the areas of financial security, smart grids, malware analysis, intrusion detection, and APT detection, tools like SHAP and LIME successfully close this transparency gap by turning black-box models into reliable systems analysts. When combined with post-hoc XAI, ensemble models like XGBoost and Random Forest provide the best accuracy-to-explainability ratio, with many models surpassing 97–99% accuracy. Transparency improves human efficacy in addition to model performance, as demonstrated by the reduction of ambiguous analyst warnings by 18–27% and the improvement of decision correctness by 22–31% with XAI integration. XAI is further reinforced as a compliance requirement by regulatory frameworks such as the EU's GDPR. Significant obstacles still exist, though, such as LIME instability, explanation tools' adversarial vulnerability, a substantial reliance on out-of-date datasets like DREBIN and KDD, Android platform bias, and the lack of established measures for assessing explanation quality. Real-time deployment is further hampered by computationally demanding ensemble techniques. To close the gap between experimental findings and practical cybersecurity deployment, future research must give priority to multi-platform dataset production, adversarial robustness, real-time XAI capabilities, and human-in-the-loop feedback mechanisms.

Future Scope

In the future, the field must deal with a number of significant issues. The community must first provide more recent, balanced datasets that accurately represent how malware operates in the modern world in order to move past outdated benchmarks like DREBIN and KDD. Additionally, research has focused too much on Android; if results are to hold up in actual enterprise settings, future work must pay similar attention to Windows, Linux, macOS, and hardware-level threats.

Attackers will unavoidably attempt to game XAI tools as they become more popular. It will be crucial to create systems that are resistant to manipulation, both at the model and explanation levels. Alongside this, the discipline urgently requires established metrics to assess what constitutes a solid explanation, including consistency, accuracy, and whether or not a human analyst can truly grasp it. Practically speaking, the majority of existing XAI techniques are too computationally demanding for real-world implementation. Transitioning from research prototypes to production systems requires the development of real-time, lightweight approaches. Most significantly, future designs should regard analysts as active players rather than passive recipients; systems that allow security experts to debate, question, and improve model explanations will promote much higher levels of efficacy and trust. Lastly, investigating hardware-level inspection tools and developing modular architectures that combine post-hoc explanation techniques with built-in interpretability may lead to fascinating new developments in the field.

References

[1]. Manthena, H., Shajarian, S., Kimmell, J. C., Abdelsalam, M., Khorsandroo, S., & Gupta,

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

M. (2025). Explainable artificial intelligence (XAI) for malware analysis: A survey of techniques, applications, and open challenges. *IEEE Access, 13*, 61611–61640. https://doi.org/10.1109/ACCESS.2025.3555926

[2]. Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, *2021*, Article 6634811. https://doi.org/10.1155/2021/6634811

[3]. 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

[4]. Mutalib, N. H. A., Sabri, A. Q. M., Wahab, A. W. A., Abdullah, E. R. M. F., & AlDahoul, N. (2024). Explainable deep learning approach for advanced persistent threats (APTs) detection in cybersecurity: A review. *Artificial Intelligence Review*, *57*, 297. https://doi.org/10.1007/s10462-024-10890-4

[5]. Cifci, A. (2025). Interpretable prediction of a decentralized smart grid based on machine learning and explainable artificial intelligence. *IEEE Access*, *13*, 36285-36306. https://doi.org/10.1109/ACCESS.2025.3543759

[6]. Mohammed, S. J., & Nema, B. M. (2025). Threat detection based on explainable AI (XAI) and hybrid learning. *Mesopotamian Journal of Cybersecurity*, *5*(2), 477–490. https://doi.org/10.58496/MJCS/2025/029

[7]. Nalinipriya, G., Sree, S. R., Radhika, K., Lydia, E. L., Karim, F. K., Ishak, M. K., & Mostafa, S. M. (2025). Leveraging explainable artificial intelligence for early detection and mitigation of cyber threat in large-scale network environments. *Scientific Reports*, *15*, Article 24662. https://doi.org/10.1038/s41598-025-08597-9

[8]. Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on Multi-Layer Perceptron. *IEEE Access*, *12*, 30164-30175. https://doi.org/10.1109/ACCESS.2024.3368377

[9]. Alabdulatif, A. (2025). A novel ensemble of deep learning approach for cybersecurity intrusion detection with explainable artificial intelligence. *Applied Sciences*, *15*(14), 7984. https://doi.org/10.3390/app15147984

[10]. Aslam, N., Khan, I. U., Mirza, S., AlOwayed, A., Anis, F. M., Aljuaid, R. M., & Baageel, R. (2022). Interpretable machine learning models for malicious domains detection using explainable artificial intelligence (XAI). Sustainability, 14(12), 7375. https://doi.org/10.3390/su14127375

[11]. Salloum, S., & Norozpour, S. (2025). XAI-IDS: A transparent and interpretable framework for robust cybersecurity using explainable artificial intelligence. Shifra, 1, 69-80. https://doi.org/10.70470/SHIFRA/2025/004

[12]. Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, Article 1526221. https://doi.org/10.3389/frai.2025.1526221

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0209
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1385 - 1396

[13]. Ashfaq, S., & Chowdhury, T. K. (2023). Explainable Artificial Intelligence (XAI) approaches for cyber risk assessment in financial services. American Journal of Interdisciplinary Studies, 4(3), 96–135. https://doi.org/10.63125/3gjcb322

[14]. Mohale, V. Z., & Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. Frontiers in Computer Science, 7, Article 1520741. https://doi.org/10.3389/fcomp.2025.1520741

[15]. Alketbi, K. S., & Mehmood, A. (2025). A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. IEEE Access, 13, 121772-121800. https://doi.org/10.1109/ACCESS.2025.3587114