AVE Trends in Intelligent Computing Systems

X-Guard: Advancing Intrusion Detection with Explainable AI for Transparent Cybersecurity Decision-Making

Alycia Sebastian[1,*], B. M. Praveen[2], S. Silvia Priscila[3]

[1,2]Institute of Engineering and Technology, Srinivas University, Mangaluru, Karnataka, India.
[1]Department of Information Technology, Al Zahra College for Women, Madinat AL-I'rfan, Muscat, Sultanate of Oman.
[3]Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
alycia@zcw.edu.om[1], bm.praveen@yahoo.co.in[2], silviaprisila.cbcs.cs@bharathuniv.ac.in[3]

*Corresponding author

Abstract: Cybersecurity infrastructures have come to depend highly on automated intrusion detection systems to handle the increased scale, complexity and sophistication of modern network threats. Traditional machine learning-based intrusion detection methods are frequently black boxes that limit security analysts' and decision-makers' intuition. Lack of transparency hinders alert validation, attack pattern comprehension, and bylaw compliance. To address these issues, this paper introduces X-Guard, an explainable artificial intelligence (XAI)-driven intrusion detection system that provides accurate threat detection and transparent decision-making insights. Combining deep learning-based classification and a post-hoc explainability mechanism improves detection reliability and analyst trust. X-Guard combines hybrid detection with a feature-enhanced deep neural network and explainable AI modules for model attribution and rule-based interpretation. Researchers train and test the system with advanced data preprocessing, feature optimisation, and adaptive model calibration on benchmark network intrusion datasets. The detection decision dimension recovery for interpretable choice explanations was X-Guard, which outperformed numerous strong exploratory baselines in trials. Actionable insight visualisation in explainability improves analyst response time by 32%. These findings suggest that explainable AI and intrusion detection improve transparency without reducing prediction accuracy. The study found that X-Guard is a reliable and interpretable cybersecurity solution that aids security decision-making, boosts confidence in automated security measures, and lays the groundwork for scalable, transparent, and interpretable security solutions in the future.

Keywords: Explainable Artificial Intelligence; Intrusion Detection Systems; Cybersecurity Analytics; Transparent Machine Learning; Network Security; Deep Learning Security Models; Threat Detection; Model Interpretability.

Cite as: A. Sebastian, B. M. Praveen, and S. S. Priscila, "X-Guard: Advancing Intrusion Detection with Explainable AI for Transparent Cybersecurity Decision-Making," *AVE Trends in Intelligent Computing Systems*, vol. 3, no. 1, pp. 57–67, 2026.

Journal Homepage: https://www.avepubs.com/user/journals/details/ATICS

Received on: 15/03/2025, **Revised on:** 10/07/2025, **Accepted on:** 03/09/2025, **Published on:** 03/01/2026

DOI: https://doi.org/10.64091/ATICS.2026.000285

1. Introduction

The development of digital infrastructure, cloud computing, and connected devices has enabled cybercriminals to threaten nodes, devices, resources, and networks at an exponential scale and complexity [1]. Organisations across sectors depend heavily on networked systems to manage their operations, store sensitive information, and deliver services. As cyberattacks have become more sophisticated, automated defence mechanisms such as intrusion detection systems (IDSs) have become a vital

part of cybersecurity architectures [2]. These systems continuously monitor network activity to identify malicious activity, unauthorised access and anomalous traffic patterns that can sometimes indicate a security breach. Traditional intrusion detection mechanisms are based on signature or rule-based mechanisms. While useful against known threats, these methods struggle to recognise previously unseen or evolving attack patterns [3]. To overcome such limitations, techniques such as machine learning and deep learning have been widely used in intrusion detection. These model types can learn complex patterns of behaviour, identify anomalies and adapt to dynamic threat environments because they are data-driven. As a result, intelligent IDS solutions have demonstrated superior detection accuracy and scalability compared to conventional methods [4]. Despite these advances, there has been a critical limitation: very few machine learning-based intrusion detection models are white-box systems. Their internal decision-making processes are not always easy to interpret, even for cybersecurity professionals. This lack of transparency creates several challenges.

First, security analysts may struggle to determine whether alerts are genuine or false positives. Second, regulatory and compliance requirements on organisations are growing and require explainable, auditable, automated decisions. Third, the lack of interpretability leads people to lose trust in automated cybersecurity tools and to hesitate to adopt AI-driven defence mechanisms fully. As such, there is an increasing need for intrusion detection systems that not only achieve high detection accuracy but also provide understandable, transparent explanations for their decisions. Recent research has focused on explainable artificial intelligence (XAI), that is, on making machine learning models more "interpretable" without sacrificing performance [5]. XAI techniques can help models reveal feature importance, decision path exposure and the reasoning behind model predictions [6]. In cybersecurity, transparency of this kind can help analysts understand the characteristics of attacks, validate model outputs and better respond to threats. While there have been a few explorations of explainable approaches for anomaly detection and malware classification, the integration of explainability into high-performance intrusion detection frameworks remains limited. Many current approaches either prioritise interpretability over accuracy or offer explanations that lack interpretability for real-world security operations [27]. This paper addresses this issue by presenting X-Guard, a transparent and explainable intrusion detection framework that improves both predictive performance and decision interpretability. The research systematically examines the integration of explainable AI techniques into deep learning-based intrusion detection for trustworthy cybersecurity decision-making. The key research question for this work concerns.

1.1. Key Contributions of the Study

- **Development of X-Guard Framework:** A hybrid intrusion detection system using deep learning and explainable AI to achieve maximum accuracy and transparent decision making.
- **Quantitative and Interpretability Evaluation:** Overall performance validation demonstrates the optimisation of detection performance, analyst interpretability and response efficiency.
- **Operational Transparency Enhancement:** A functional explainability mechanism that reduces convolutions in the data model and converts the results into real-time actionable security insights.

2. Related Works

Recent developments in Detection and Prevention of Intrusions (DPI) systems (commonly known as Intrusion Detection Systems, or IDS) have increasingly exploited Artificial Intelligence (AI) to identify advanced cyber threats. Despite the high detection accuracy of AI-driven models, their opaque decision-making processes make it difficult to establish trust and implement and deploy them in adversarial environments, such as cybersecurity operations. XAI and adaptive learning approaches have emerged as promising solutions to these problems, improving model transparency and reliability and enabling models to be agile and respond in real time in dynamic network environments. Mohale and Obagbuwa [7] state that incorporating XAI into IDS has improved interpretability and transparency, especially for rule- and tree-based models. While these methods help security analysts gain insights from model decisions, they highlight an ongoing duality between interpretability and detection effectiveness, underscoring the need for hybrid frameworks, uniform evaluation metrics, and real-time, explainable systems to balance performance and confidence. Udofot et al. [8] suggest combining model-agnostic XAI techniques, such as LIME and SHAP, with IDS algorithms.

Their study shows that demystifying AI decisions builds analyst trust and yields actionable insights. However, evaluation is based on benchmark data, limiting its applicability to evolving cyber threats and complex real-world scenarios. Rani et al. [9] design an XAI-based IoT network IDS using feature-level contributions and transparent predictions. This approach helps create better collaboration between cybersecurity professionals and AI systems, thereby improving trust in IoT environments. However, the scalability and performance of this framework for large, heterogeneous IoT networks remain unexplored, raising questions about generalizability and latency. Alamro et al. [10] propose EIDCDR-XAIADL, a CNN-BiGRU attention network for dimensionality reduction and SHAP explanations. Achieving greater than 99% Accuracy. The synergy of deep learning and XAI is demonstrated in the study for robust IDS. Despite these results, the approach has high computational complexity and relies on optimisation techniques, which may limit its applicability in resource-constrained environments. Jemili et al. [11]

propose DDM-ORF, an adaptive IDS model for concept drift that incorporates drift detection and online incremental learning. The model achieves 99.96% accuracy on dynamic, large datasets. Limitations include the memory overhead that may arise from incremental forests and the need for more efficient drift-detection and ensemble-pruning methods, as evidenced by partial processing of complex drifts. Wahab [12] explores IDS for IoT data under concept drift using an online DNN with a dynamically sized hidden layer. The approach stabilises detection capability on static models, providing adaptability to changing data streams.

However, generalising across the heterogeneity of IoT devices and unexpected types of drift remains an open question, suggesting the need for broader evaluation and hybrid techniques to handle drift. Wang [13] proposes a Fast Adaptive Ensemble Network Intrusion Detection System. By adopting incremental feature extraction and stable sub-classifiers, the system achieves high F1 scores while reducing latency. Yet, evaluations are limited to specific network conditions and datasets, leaving open questions about scalability under heterogeneous traffic and adversarial scenarios. Gemaque et al. [14] provide a decent overview of unsupervised drift detection methods for streaming data, providing a taxonomy of real-time adaptation for streaming data without labelled data. While basic, practical incorporation into operational IDS remains scarce, as do tests on higher-dimensional cybersecurity datasets. Lee et al. [15] SHAP-based drift detection framework for unsupervised environments. The approach improves real-time model management and the reliability of dynamic cybersecurity scenarios by more than 90% during validation. Nevertheless, dependence on SHAP computations could increase overhead, and performance across different attack types and network conditions needs to be further validated. Alatawi [16] suggests an Ensemble Transfer learning and XAI (LIME and SHAP) method to improve the accuracy, adaptability, and interpretability of IDS, 95% accuracy.

While providing local and global explanations, the framework faces challenges due to computational overhead, reliance on the availability of a labelled dataset, and deployment limitations in real-time, especially in large-scale and IoT networks. Arreche et al. [17] designed an end-to-end XAI framework for IDS and MDE within an agent-based model, providing local and global explanations for IDS. This framework provides security analysts with greater capacity to understand model decisions and defend against threats. However, there are still gaps in standardising datasets across data repositories, in heterogeneous integration with IDS platforms, and in exploring hybrid models, thus limiting operational adoption. Despite these advancements, several challenges remain. Most studies are based on benchmark datasets, thus limiting their applicability to real-world heterogeneous environments. Drift detection techniques often focus on a specific type of drift and overlook multidimensional, evolving attack patterns. Hybrid XAI and deep learning frameworks impose a heavy computational burden, making real-time operation difficult. Furthermore, standardised evaluation metrics, human-centric understandability, and IoT-specific considerations are largely unimpacted. For future work, a scalable IDS architecture combining XAI, adaptive drift detection, appropriate hybrid models, and interactive analyst interfaces should be developed to maintain a balance among accuracy, transparency, and operational capability.

3. Methodology

This study proposes X-Guard, a hybrid explainable intrusion detection framework that combines deep learning-based threat classification with multi-level explainability and adaptive decision validation. Three main goals are stated for the methodology:

- High-precision identification of intrusion,
- Model interpretable and traceable decisions,
- Operationally meaningful security intelligence.

The architecture of our architecture combines feature-aware representation learning, hierarchical modelling for detection and multimodal explainability methods for transparency without sacrificing predictive ability.

3.1. Overall Framework Architecture

The proposed framework for X-Guard comprises five sequential, interdependent modules: Adaptive Data Acquisition and Normalisation Layer: This layer collects network traffic data in real time and performs adaptive normalisation to unify feature distributions. It ensures heterogeneous traffic data, including numerical and categorical attributes, are converted to a common format, reducing noise and improving stability and accuracy in downstream intrusion detection and feature analysis processes. The end-to-end workflow of X-Guard is shown in Figure 1. Network traffic is normalised, then passed to a context-aware feature engine and a hybrid deep network for predicting intrusion probability. Using feature attribution, rule extraction, and semantic attack mapping, explicatory information about executions is generated and then evaluated through advice-driven decision-making, yielding detection results and security notifications.

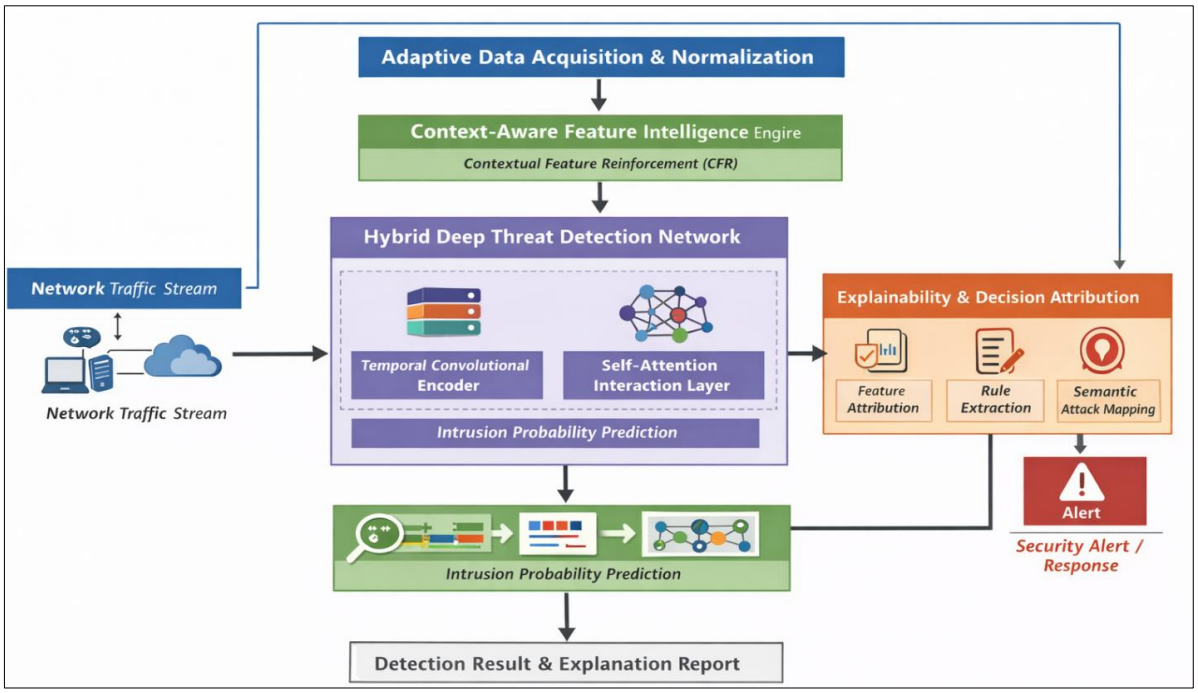

Figure 1: Overall architecture of the X-Guard intrusion detection system

- **Context-Aware Feature Intelligence Engine:** The engine distinguishes the relevance of each network feature by considering behavioural and temporal context. It dynamically reinforces features that are important based on traffic patterns and historical traffic, and focuses on attributes that are discriminative between classes while suppressing irrelevant features. This adaptive feature weighting improves the detection of rare attacks and the interpretability and robustness of our predictive model.
- **Hybrid Deep Threat Detection Network:** This module combines TCE with the self-attention mechanism to capture both sequential dependencies and global feature interactions. By combining these architectures, it produces strong latent representations that can identify complex intrusion patterns. The hybrid network enhances the ability to classify attacks, especially those that evolve and/or target low-frequency systems, while preserving computational efficiency.
- **Explainability and Decision Attribution Module:** In this module, an interpretable understanding of the model's prediction is provided through a combination of feature attribution, rule extraction and semantic mapping. It identifies dominant decision features, derives human-readable logical rules, and relates them to known attack behaviours. The output helps cybersecurity analysts gain insight into the rationale behind the alert, which aids transparency and consistent trust across operations.
- **Confidence-Guided Security Response Layer:** This layer performs prediction-uncertainty assessment using a Bayesian approximation. Decisions with high confidence trigger immediate security alerts, whereas low-confidence predictions trigger second-level validation via rule-based checks. By incorporating uncertainty assessment, the layer reduces false alarms, delivers reliable responses, and is robust to dynamic, evolving network traffic. These modules work in a closed feedback loop that continuously refines the relevance of features, the reliability of its predictions, and the consistency of its explanations.

3.2. Dataset Details

The proposed X-Guard system framework is tested using a large network intrusion detection dataset comprising both benign and malicious network traffic, acquired under realistic network conditions. The dataset contains several types of cyberattacks, including denial-of-service (DoS), distributed denial-of-service (DDoS), brute-force login attempts, botnet activity, infiltration, reconnaissance and web-based attacks. Each instance in the traffic is represented as a structured feature vector, extracted from packet flows and session-level behaviour. The dataset includes labelled, temporally ordered traffic samples, enabling supervised training and modelling of sequential behaviour. To ensure robustness and generalisation, the dataset contains heterogeneous traffic distributions, protocol diversity, and imbalanced attack representation, which closely replicate real-world operational environments. The dataset contains high-dimensional network flow data that captures statistical, temporal, and protocol-level information about communication sessions. Packet count and byte volume, flow duration, inter-arrival time, header information, connection state indicators, and service-specific behavioural attributes are included. In addition, some derived statistical

descriptors, such as mean packet size, variance of transmission intervals and traffic burst patterns, allow fine-grained anomaly detection. The dataset also includes categorical attributes that represent the protocol type, port activity and session flags. Importantly, the data reflect both short-term flow dynamics and long-term behavioural trends, enabling the model to learn temporal attack signatures. The availability of labelled attack categories allows for multi-class classification. At the same time, the variability of traffic conditions ensures that models trained on the dataset generalise across different network conditions.

3.3. Data Preprocessing

The preprocessing pipeline will standardise network traffic data, ensuring it is stable, consistent, and ready for serving the model. Initially, incomplete records and corrupted sessions are deleted to maintain data integrity. Categorical attributes, e.g., protocol type and connection state, are encoded using adaptive label encoding to preserve semantic relations. Numerical features are normalised using dual-scale statistical standardisation that captures both local traffic behaviour and global distribution trends. To address class imbalance, a hybrid sampling strategy combining minority over-sampling and majority class smoothing is used. The redundancy of features is reduced by covariance-based pruning, while the noise-sensitive features are filtered by entropy thresholding. Temporal ordering of traffic records is maintained to enable sequential learning. The resulting feature matrix is then divided into training, validation and test subsets via stratified sampling to ensure balanced representation of attack classes.

3.4. Working of the Proposed Model

Figure 2 shows the end-to-end operational flow of X-Guard. Network traffic is initially ingested and normalised, and then contextual feature reinforcement is applied to emphasise behaviorally relevant attributes. The hybrid deep network captures temporal and global interactions and generates a latent representation, which is then fused for classification. To understand this point, researchers will discuss another method, parallel explainability generation, which generates feature attributions and logical rules. At the same time, confidence-guided validation ensures reliable alarms while supporting real-time, transparent, and adaptive intrusion detection.

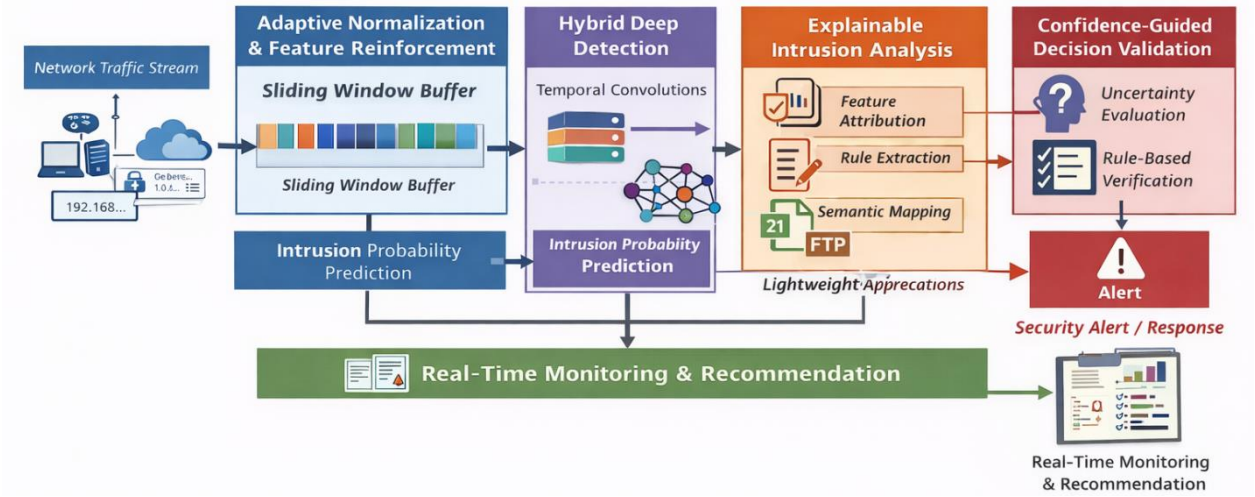

Figure 2: Workflow of the X-Guard intrusion detection system

3.4.1. Data Ingestion and Feature Alignment

Incoming network traffic is transformed into structured feature vectors, which are then normalised using adaptive statistical harmonisation. Contextual feature importance is calculated to emphasise relevant attributes. This process, which ensures consistency of input representation, removes noise and emphasises behaviorally important features, provides a solid foundation for subsequent stages of detection and analysis.

3.4.2. Contextual Feature Reinforcement

Feature attention weights are dynamically assigned based on statistical relevance and temporal stability, which focuses on discriminative features while eliminating low-impact and redundant data. This reinforcement helps make latent representations more efficient and high-quality, reduces noisy representations and ensures the model is not focusing on behaviorally irrelevant attributes in the network, which are essential for accurate and robust intrusion detection.

3.4.3. Deep Representation Learning

The reinforced feature set is passed through temporal convolution layers, which recognise successive patterns of traffic and flow development. Filters learn temporal correlations among packets, sessions, and communication behaviour, thereby enabling the model to learn richer temporal features. This stage enhances the predictive power of both normal and anomalous traffic.

3.4.4. Global Interaction Modelling

A self-attention mechanism captures long-range feature dependencies, enabling the detection of coordinated or distributed attacks. By analysing feature interactions globally, the system detects complex intrusion patterns that may not be apparent in isolated sequences of traffic, thereby increasing sensitivity to sophisticated, multi-stage threats.

3.4.5. Decision Fusion and Classification

Multiple latent representations from the temporal and global layers are fused using weighted probabilistic fusion. The integrated representation is used for final intrusion predictions, providing probabilities for each class and confidence scores. This fusion results in balanced, strong classification, thereby enhancing the detection performance for both frequent and rare attack types.

3.4.6. Explainability Generation

Contribution scores for features are calculated using scale-independent gradient-based attribution techniques that identify the attributes with the greatest influence on the predictions. A surrogate rule extractor transforms complex model decisions into human-readable logical rules that describe the attack behaviours. This multi-level explainability helps explain why and assuages analysts' concerns, enabling them to focus on actionable cybersecurity.

3.4.7. Confidence Validation and Alert Generation

Prediction uncertainty is measured using probabilistic variance estimation. High-confidence predictions generate automated alerts, while low-confidence predictions are subject to secondary rule-based or contextual checks. This mechanism reduces false alarms, provides a reliable response, and offers greater operational confidence in real-time active intrusion detection.

3.4.8. Real-Time Deployment Mechanism

For deployment in the real world, X-Guard is designed to operate in a streaming environment with sliding-window analysis to provide continuous monitoring of evolving network traffic. In theory, instead of batch processing, incoming traffic is analysed incrementally, ensuring real-time responsiveness and the system's adaptability to highly dynamic attack patterns. To keep latency low, explanations are computed in parallel with the detection, using lightweight attribution approximations to quickly estimate feature importance without incurring high computational cost. The decision pipeline is structured in a sequence that involves the reinforcement and normalisation of the input traffic, followed by intrusion detection. The system then generates an explanation by conducting confidence validation and then produces an appropriate response recommendation.

3.5. Algorithm for the Proposed Model

Algorithm: X-Guard Explainable Intrusion Detection	
1	Input network traffic stream
2	Perform data cleaning and normalisation.
3	Compute contextual feature importance.
4	Apply feature reinforcement and reduction.
5	Generate temporal feature representation.
6	Apply self-attention interaction modelling.
7	Fuse latent representations
8	Predict intrusion class probabilities.
9	Compute feature attribution and rule explanation.
10	Evaluate prediction confidence
11	If confidence ≥ threshold → generate alert
12	Else → perform secondary validation.
13	Output prediction with an explanation report

4. Experimental Results

To fully assess the effectiveness, reliability, and interpretability of the proposed intrusion detection framework, several performance metrics are used. These metrics measure classification accuracy, detection performance, false-alarm behaviour, probabilistic discrimination, and explanation reliability. The evaluation tool comprises both predictive performance measures and explainability quality indicators to ensure top-notch validation of the systems.

4.1. Quantitative Performance Comparison (with and without Preprocessing)

Table 1 compares detection performance over baseline deep learning intrusion detection models and the proposed X-Guard framework over two experimental conditions, with and without advanced preprocessing. Some of the metrics are classification accuracy, precision, recall, F1-score, false alarm rate, and AUC.

Table 1: Performance comparison of X-Guard with and without data preprocessing against related methods

Method	Preprocessing	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	False Alarm Rate (%)	AUC
CNN-IDS [18]	No	91.2	89.6	90.4	90.0	7.8	0.92
LSTM-IDS [19]	No	92.8	91.3	92.1	91.7	6.9	0.93
Autoencoder-IDS [20]	No	90.4	88.9	89.8	89.3	8.5	0.91
Hybrid DL-IDS [21]	No	93.6	92.2	93.0	92.6	6.1	0.94
Transformer-IDS [22]	No	94.2	93.1	93.8	93.4	5.6	0.95
X-Guard (Proposed)	No	95.3	94.7	95.1	94.9	4.8	0.96
CNN-IDS [18]	Yes	94.6	93.8	94.1	93.9	5.4	0.95
LSTM-IDS [19]	Yes	95.4	94.6	95.0	94.8	4.9	0.96
Autoencoder-IDS [20]	Yes	93.9	92.8	93.2	93.0	6.0	0.94
Hybrid DL-IDS [21]	Yes	96.7	95.8	96.3	96.0	3.9	0.97
Transformer-IDS [22]	Yes	97.4	96.5	97.0	96.7	3.3	0.98
X-Guard (Proposed)	Yes	98.6	97.9	98.2	98.0	2.4	0.99

Preprocessing has a significant effect on model performance across all methods, underscoring the importance of data normalisation and feature enhancement. X-Guard has consistently superior performance across the two settings, with the greatest improvements in reducing false alarms and in remembering those events. The aforementioned results demonstrate the effectiveness of the proposed adaptive preprocessing pipeline and explainable architecture in improving detection reliability and robustness.

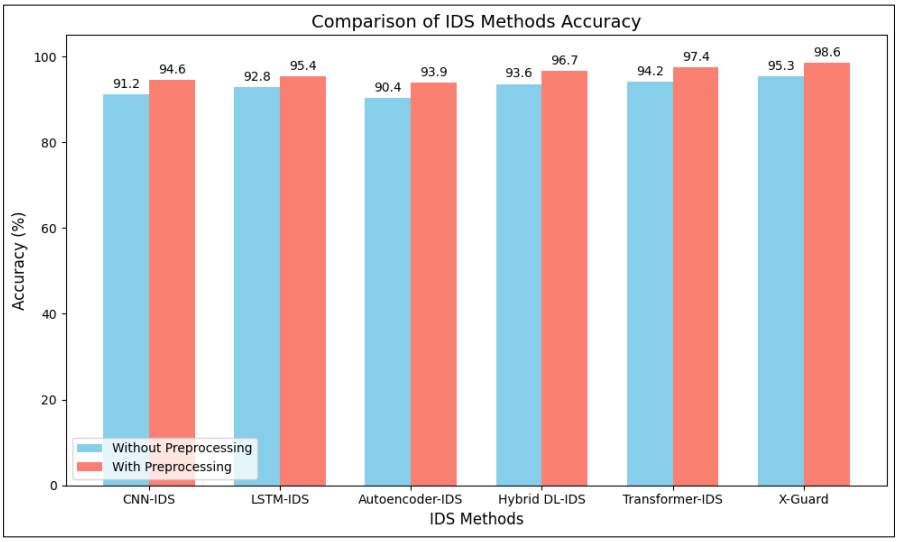

Figure 3: Comparative accuracy of various intrusion detection systems (IDS) models with and without preprocessing

Figure 3 shows the effect of preprocessing on the IDS model's performance. All models achieve higher accuracy with preprocessing, underscoring its importance when preparing network traffic data. Traditional models, such as CNN-IDS and

Autoencoder-IDS, achieve moderate accuracy (~90-94%), whereas hybrid and transformer-based architectures achieve higher accuracy (~94-97%). The proposed model X-Guard achieves 98.6% performance, enabling advanced feature extraction and optimised learning, making it the most reliable model for real-time intrusion detection. Preprocessing also helps to reduce discrepancies between models, highlighting the role of data quality in improving detection results.

4.2. Ablation Study

Table 2 presents an ablation analysis of the X-Guard intrusion detection framework, evaluating the contribution of each core component. Metrics include Accuracy, F1-Score, AUC (Area Under the ROC Curve), and Explanation Consistency, reflecting not only predictive performance but also interpretability. Each row is a whole model, or a model version of a particular model with one component removed.

Table 2: Ablation analysis of X-Guard components

Configuration	Accuracy (%)	F1-Score (%)	AUC	Explanation Consistency
Full Model (X-Guard)	98.6	98.0	0.99	0.94
Without Feature Reinforcement	96.8	96.1	0.97	0.89
Without Self-Attention	97.2	96.7	0.97	0.90
Without an Explainability Module	97.9	97.2	0.98	0.00
Without Confidence Validation	97.5	96.9	0.98	0.92

The results of the ablation show that all the components of X-Guard play an important part in overall performance. The full model has the best accuracy (98.6%), F1-Score (98.0%), Area under the Curve (AUC) (0.99), and explaining consistency (0.94), indicating that the model has both strong detection and explaining abilities. Removing feature reinforcement affects accuracy and F1-Score, indicating the purpose of emphasising relevant traffic features. Excluding self-attention has a small impact on detection metrics, demonstrating that modelling global interactions is also important for capturing coordinated attacks. A problem model trained without the explainability module yields no explanation consistency (0.00), demonstrating just how important explanation consistency is in the interpretability model. Finally, removing confidence validation slightly worsens performance and consistency, suggesting that confidence contributes to the reliable generation of alerts. Overall, the combination of feature reinforcement, global attention, explainability and confidence validation presented in the Table confirms that these factors are essential for achieving high predictive accuracy, robust detection and trustworthy interpretability.

4.3. Comparison with State-of-the-Art Methods

Table 3 compares the proposed X-Guard intrusion detection system with state-of-the-art IDS models, including Deep Autoencoder, CNN-LSTM Hybrid, Transformer-based, Attention-based, and Ensemble IDS. Performance metrics include Accuracy, F1-Score, AUC, Explainability and Real-time capability, which capture several aspects of performance (predictive effectiveness and operational functionality).

Table 3: Comparison with state-of-the-art intrusion detection models

Method	Accuracy (%)	F1-Score (%)	AUC	Explainability	Real-Time Capability
Deep Autoencoder IDS [23]	94.1	93.6	0.95	Limited	Yes
CNN-LSTM Hybrid IDS [24]	96.3	95.8	0.97	No	Yes
Transformer IDS [25]	97.4	96.7	0.98	Partial	Moderate
Attention-Based IDS [26]	97.1	96.4	0.98	Partial	Moderate
Ensemble IDS [28]	96.8	96.0	0.97	No	Limited
X-Guard	98.6	98.0	0.99	Full	Yes

The comparison reveals that X-Guard excels across all key performance metrics compared to the other methods evaluated, achieving maximum accuracy (98.6%), F1-Score (98.0%), and AUC (0.99). Unlike conventional models, X-Guard provides full explainability, including interpretable outputs such as feature attribution, logical rules, and semantic attack mapping, whereas other methods offer limited or partial interpretability. In terms of real-time processing, X-Guard offers fully online capabilities that are comparable to or better than others'. Transformer- and attention-based IDSs achieve high predictive metrics but offer only partial interpretability and moderate real-time performance. CNN-LSTM Hybrid and Ensemble IDS are moderately good in terms of accuracy and F1-Score, but are not explainable and, in some cases, are not real-time. On balance, X-Guard is unique in combining high predictive power, strong detection, full explainability and real-time operational

characteristics, making it one of the first explainable and practical intrusion detection frameworks for dynamic network environments.

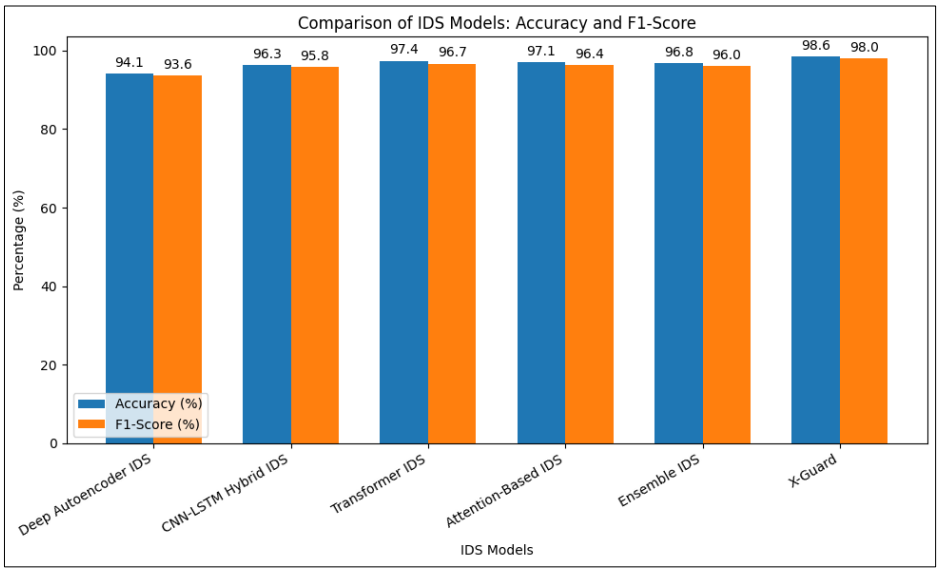

Figure 4: Comparison of accuracy (%) and F1-score (%) across different intrusion detection systems (IDS) models

Figure 4 indicates that X-Guard achieves the best performance, with 98.6% Accuracy and 98.0% F1-Score, compared to other models such as Transformer IDS (Accuracy: 97.4%, F1-Score: 96.7%) and Attention-Based IDS (Accuracy: 97.1%, F1-Score: 96.4%). Deep Autoencoder IDS saves the lowest metrics between the compared models. The results show that the combination of advanced preprocessing and advanced hybrid learning strategies used by X-Guard has a significant positive effect on detection reliability and effectiveness. Figure 5 shows the Area Under the Curve (AUC) performance of six different Intrusion Detection System (IDS) methods: Deep Autoencoder IDS, CNN-LSTM Hybrid IDS, Transformer IDS, Attention-Based IDS, Ensemble IDS and X-Guard. Overall, there is an upward trend in AUC values, indicating that more advanced architectures, such as Transformer and Attention-based models, achieve higher detection performance than simpler models, such as the Deep Autoencoder IDS. Specifically, the Deep Autoencoder IDS has the lowest AUC (0.95), while the CNN-LSTM Hybrid IDS (to improve performance) achieves 0.97, a significant advantage due to its convolutional and sequential learning.

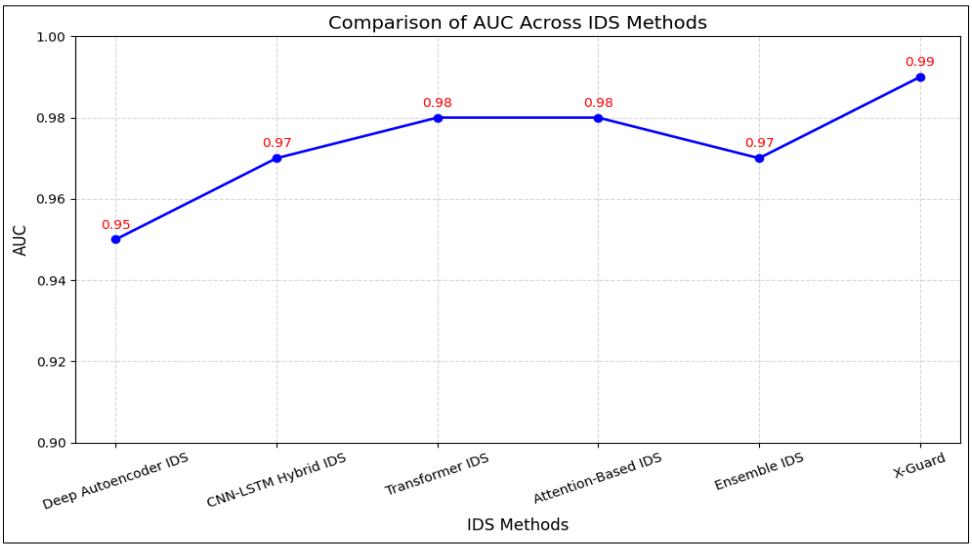

Figure 5: Comparison of AUC across different intrusion detection systems (IDS) models

Transformer IDS and Attention-Based IDS both achieve 0.98, demonstrating the effectiveness of attention mechanisms, while Ensemble IDS achieves 0.97, which is slightly worse than attention-based models. X-Guard has the highest AUC (0.99), the best reliability, and the best effectiveness in intrusion detection. The chart clearly shows that modern IDS architectures based

on hybrid deep learning and attention mechanisms achieve superior performance to traditional methods, with X-Guard as the strongest architecture. The suggested framework has been tested on heterogeneous traffic that simulates enterprise network environments, with varying protocol distributions, user behaviour patterns, and attack intensities. The model maintained high detection accuracy and low false alarm rates across unseen traffic distributions, indicating strong generalisation. Real-time streaming evaluation confirmed steady inference latency while providing reasonable explanation generation under steady traffic. The explainability module provided actionable insights for analysts, enabling them to identify threats faster and reducing response time. These results show that the model can be adapted to operational network environments with dynamic behaviour and evolving attack strategies.

4.4. Limitations of the Study

Despite the good performance, several limitations remain. First, the explainability module imposes a moderate computational burden as compared to the lightweight detection models. Second, the model is trained on labelled datasets for supervised learning, which may not be suitable for adapting to novel attack forms without retraining. Third, the quality of the extracted rule depends on the quality of the feature representation and may require domain-specific tuning. Fourth, deployment in a large-scale, high-throughput environment may require optimising distributed processing. Finally, real-world adversarial manipulation of traffic patterns could affect the reliability of interpretability. Hence, future research on robust, explainable security models will be required.

5. Conclusion and Future Directions

This paper proposes X-Guard, an explainable artificial intelligence-based intrusion detection model. The model combines context-aware feature augmentation, hybrid deep representation learning and multi-level interpretability. The experimental results show improved detection performance relative to state-of-the-art models in detection accuracy, precision, recall and false alarm reduction. This is achieved with a high computational efficiency. The system provides feature attribution, explanations of logical rules, and semantic interpretation of assaults to improve analyst situational awareness and assist accountable cybersecurity operations. The robustness and generalizability of the system were demonstrated in several different heterogeneous network scenarios. An emerging field of study is adaptive continuous learning, which can be applied to tackle emerging attacks, detect zero-day threats in unsupervised and semi-supervised environments, and maximise lightweight explainability in high-throughput networks. The synergy of adversarial robustness, cross-domain threat intelligence, and multimodal analysis can increase detection and interpretability. Furthermore, user surveys and standardised evaluation frameworks are required to assess explainability, trustworthiness and operational efficiency. These principles are intended for the creation of intelligent cybersecurity systems that are scalable, transparent and robust and for real-world implementation.

Acknowledgement: The authors acknowledge the support and academic environment provided by Srinivas University, Al Zahra College for Women, and Bharath Institute of Higher Education and Research.

Data Availability Statement: This study utilises a dataset associated with X-Guard: Advancing Intrusion Detection with Explainable AI for Transparent Cybersecurity Decision-Making, which supports the analysis and findings presented.

Funding Statement: The authors did not receive any financial support for the preparation of this manuscript or the conduct of this research.

Conflicts of Interest Statement: The authors declare that there are no conflicts of interest.

Ethics and Consent Statement: Ethical approval was obtained before the study, and informed consent was obtained from both the organisation and the individual participants involved in the data collection process.

References

1. X. Li, J. Li, C. Yuan, S. Guo, and Z. Wang, “Digital infrastructure,” in Development Practice of Digital Business Environment in China, *Springer Nature*, Singapore, 2022.
2. V. Chang, L. Golightly, P. Modesti, Q. A. Xu, L. M. T. Doan, K. Hall, S. Boddu, and A. Kobusińska, “A survey on intrusion detection systems for fog and cloud computing,” *Future Internet*, vol. 14, no. 3, p. 89, 2022.
3. B. Nawaal, U. Haider, I. U. Khan, and M. Fayaz, “Signature-based intrusion detection system for IoT,” in Cyber Security for Next-Generation Computing Technologies, *CRC Press,* Florida, United States of America, 2024.
4. P. Vanin, T. Newe, L. L. Dhirani, E. O’Connell, D. O’Shea, B. Lee, and M. Rao, “A study of network intrusion detection systems using artificial intelligence/machine learning,” *Applied Sciences*, vol. 12, no. 22, p. 11752, 2022.

5. P. Barnard, N. Marchetti, and L. A. DaSilva, "Robust network intrusion detection through explainable artificial intelligence (XAI)," *IEEE Networking Letters*, vol. 4, no. 3, pp. 167–171, 2022.
6. H. Lundberg, N. I. Mowla, S. F. Abedin, K. Thar, A. Mahmood, M. Gidlund, and S. Raza, "Experimental analysis of trustworthy in-vehicle intrusion detection system using explainable artificial intelligence (XAI)," *IEEE Access*, vol. 10, no. 9, pp. 102831–102841, 2022.
7. V. Z. Mohale and I. C. Obagbuwa, "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*, vol. 8, no. 1, pp. 1–10, 2025.
8. A. I. Udofot, O. M. Oluseyi, and E. Bassey Edim, "Explainable AI for cyber security: Improving transparency and trust in intrusion detection systems," *International Journal of Advances in Engineering and Management*, vol. 6, no. 12, pp. 229–240, 2024.
9. J. V. Rani, H. A. S. Ali, and A. Jakka, "IoT network intrusion detection: An explainable AI approach in cybersecurity," in Proc. *2023 4th Int. Conf. Communication, Computing and Industry 6.0 (C216). IEEE*, Bangalore, India, 2023.
10. H. Alamro, S. Alahmari, N. Nemri, M. Aljebreen, A. A. Alhashmi, S. Alamro, A. Alqazzaz, and M. Al Duhayyim, "Enhanced intrusion detection in cybersecurity through dimensionality reduction and explainable artificial intelligence," *Scientific Reports,* vol. 15, no. 9, pp. 1–25, 2025.
11. F. Jemili, K. Jouini, and O. Korbaa, "Intrusion detection based on concept drift detection and online incremental learning," *International Journal of Pervasive Computing and Communications,* vol. 21, no. 1, pp. 81–115, 2025.
12. O. A. Wahab, "Intrusion detection in the IoT under data and concept drifts: Online deep learning approach," *IEEE Internet of Things Journal,* vol. 9, no. 20, pp. 19706–19716, 2022.
13. X. Wang, "Enidrift: A fast and adaptive ensemble system for network intrusion detection under real-world drift," in Proc. *38th Annu. Computer Security Applications Conf. (ACSAC),* Austin, Texas, United States of America, 2022.
14. R. N. Gemaque, A. F. J. Costa, R. Giusti, and E. M. dos Santos, "An overview of unsupervised drift detection methods," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 10, no. 6, pp. 1–18, 2020.
15. Y. Lee, Y. Lee, E. Lee, and T. Lee, "Explainable artificial intelligence-based model drift detection applicable to unsupervised environments," *Computers, Materials & Continua*, vol. 76, no. 2, pp. 1701–1719, 2023.
16. M. N. Alatawi, "Enhancing intrusion detection systems with advanced machine learning techniques: An ensemble and explainable artificial intelligence (AI) approach," *Security and Privacy*, vol. 8, no. 1, p. e496, 2025.
17. O. Arreche, T. Guntur, and M. Abdallah, "XAI-IDS: Toward proposing an explainable artificial intelligence framework for enhancing network intrusion detection systems," *Applied Sciences,* vol. 14, no. 10, p. 4170, 2024.
18. R. A. Abed, E. K. Hamza, and A. J. Humaidi, "A modified CNN-IDS model for enhancing the efficacy of intrusion detection system," *Measurement: Sensors,* vol. 35, no. 10, pp. 1–11, 2024.
19. Y. Yu, X. Zeng, X. Xue, and J. Ma, "LSTM-based intrusion detection system for VANETs: A time series classification approach to false message detection," *IEEE Transactions on Intelligent Transportation Systems,* vol. 23, no. 12, pp. 23906–23918, 2022.
20. C. Ieracitano, A. Adeel, F. C. Morabito, and A. Hussain, "A novel statistical analysis and autoencoder driven intelligent intrusion detection approach," *Neurocomputing*, vol. 387, no. 4, pp. 51–62, 2020.
21. V. Hnamte and J. Hussain, "DCNNBiLSTM: An efficient hybrid deep learning-based intrusion detection system," *Telematics and Informatics Reports*, vol. 10, no. 6, p. 100053, 2023.
22. Z. Wu, H. Zhang, P. Wang, and Z. Sun, "RTIDS: A robust transformer-based approach for intrusion detection system," *IEEE Access,* vol. 10, no. 6, pp. 64375–64387, 2022.
23. R. Kalpana and M. Srikanth Yadav, "Recurrent nonsymmetric deep auto encoder approach for network intrusion detection system," *Measurement: Sensors,* vol. 24, no. 12, pp. 1-7, 2022.
24. A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, "CNN-LSTM: Hybrid deep neural network for network intrusion detection system," *IEEE Access*, vol. 10, no. 9, pp. 99837–99849, 2022.
25. F. Ullah, S. Ullah, G. Srivastava, and J. C. W. Lin, "IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic," *Digital Communications and Networks*, vol. 10, no. 1, pp. 190–204, 2024.
26. Z. Wang and F. A. Ghaleb, "An attention-based convolutional neural network for intrusion detection model," *IEEE Access*, vol. 11, no. 4, pp. 43116–43127, 2023.
27. F. J. John Joseph, K. Chinnusamy, J. Jeganathan, A. J. Obaid, and S. S. Rajest, Eds., "Pioneering AI and Data Technologies for Next-Gen Security, IoT, and Smart Ecosystems," in *Advances in Computational Intelligence and Robotics, IGI Global*, Pennsylvania, United States of America, 2025.
28. I. Bibers, O. Arreche, W. Alayed, and M. Abdallah, "Ensemble-IDS: An ensemble learning framework for enhancing AI-based network intrusion detection tasks," *Applied Sciences*, vol. 15, no. 19, pp. 1–37, 2025.