African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 5, Issue 2, 2026
Page No: 357-369
Website: *https://aaasjournals.com/index.php/ajapas/index*
ISI 2025: 1.126 **SJIFactor 2024: 6.752** **معامل التأثير العربي: 1.62**

Adaptive Explainable Deep Learning Framework for Intelligent Intrusion Detection and Forensic Threat Logging in Enterprise Networks

Musbah Abobaker Musbah [1], Abdalslam S. Imhmed Mohamed [2*]
[1,2] Department of Information System, Faculty of Information Technology, Aljufra University,Libya

إطار تكيفي قابل للتفسير قائم على التعلم العميق لاكتشاف التسلل السيبراني وتسجيل التهديدات الجنائية الرقمية في بيئات الشبكات المؤسسية

مصباح ابوبكر مصباح [1]، عبد السلام سعيد امحمد بن جريد [2*]
[2،1] قسم نظم المعلومات، كلية تقنية المعلومات، جامعة الجفرة، ليبيا

*Corresponding author: abdalslam.benjred@ju.edu.ly

Received: March 25, 2026	Accepted: June 09, 2026	Published: June 20, 2026

Abstract
The increasing sophistication of cyberattacks has exposed significant limitations in conventional intrusion detection systems (IDSs), particularly their inability to adapt to evolving attack patterns while simultaneously providing interpretable and forensically valuable outputs. This study presents an adaptive explainable deep learning framework for intelligent intrusion detection and forensic threat logging in enterprise network environments. The proposed architecture integrates a hybrid Autoencoder–BiLSTM classifier with an attention mechanism and a structured forensic logging engine to support both real-time attack detection and post-incident analysis. Network traffic records derived from the NSL-KDD and CICIDS2017 datasets were pre-processed through feature normalization, categorical encoding, and class-balancing procedures using SMOTE. The hybrid model was trained to classify traffic into five attack categories: Normal, DoS, Probe, R2L, and U2R. Explainability was incorporated through SHAP-based feature attribution to improve model transparency and analyst trust. Experimental evaluation demonstrated an overall accuracy of 98.41%, precision of 98.02%, recall of 98.16%, F1-score of 98.09%, and a false positive rate of 0.92%, outperforming Random Forest, Support Vector Machine, XGBoost, and conventional MLP architectures. In addition, the proposed framework maintained structured forensic records containing attacker metadata, confidence scores, severity indices, and temporal attack correlations suitable for digital forensic investigations and threat intelligence workflows. The results indicate that combining adaptive deep learning, explainable analytics, and forensic-aware logging significantly improves the operational reliability of modern IDS platforms. The proposed framework provides a scalable and deployable foundation for intelligent cybersecurity monitoring in enterprise and cloud-based infrastructures.

Keywords: Intrusion Detection System, Explainable Artificial Intelligence, Deep Learning, BiLSTM, Cybersecurity, SHAP, Digital Forensics, Threat Intelligence, Network Security.

الملخص
تواجه أنظمة كشف التسلل التقليدية تحديات متزايدة نتيجة التطور المستمر في الهجمات السيبرانية وتعقيد أنماطها، مما يؤثر في دقة الاكتشاف ويزيد من معدلات الإنذارات الكاذبة. تهدف هذه الدراسة إلى تطوير إطار ذكي تكيفي وقابل للتفسير لتحسين كفاءة اكتشاف التسلل داخل البيئات الشبكية الحديثة ويعتمد النموذج المقترح على دمج تقنيات متعددة للتعلم العميق وتحليل السلوك الشبكي بهدف تحسين دقة التصنيف وتقليل الأخطاء. كما تم تضمين آلية تفسير للقرارات تساعد في تحديد الخصائص الأكثر تأثيرا في اكتشاف الهجمات، بالإضافة إلى نظام لتسجيل الأحداث الأمنية لدعم التحليل الجنائي الرقمي والاستجابة

للحوادث. تم تقييم النموذج باستخدام مجموعة بيانات قياسية خاصة بحركة الشبكات والهجمات السيبرانية، وأظهرت النتائج تحقيق دقة مرتفعة مع قدرة فعالة على التمييز بين الأنواع المختلفة للهجمات وتقليل الإنذارات الكاذبة مقارنة بالنماذج التقليدية. تؤكد نتائج الدراسة أهمية دمج تقنيات التعلم العميق القابلة للتفسير مع التحليل الجنائي الرقمي لتحسين موثوقية أنظمة الأمن السيبراني ورفع كفاءتها التشغيلية.

الكلمات المفتاحية: الأمن السيبراني؛ كشف التسلل؛ التعلم العميق؛ تحليل الشبكات؛ التهديدات الرقمية؛ التحليل الجنائي الرقمي؛ أمن الشبكات.

1. Introduction

Modern enterprise networks operate within highly dynamic and interconnected digital ecosystems characterized by cloud computing, Internet of Things (IoT) integration, software-defined networking, and distributed service architectures. Although these technologies have accelerated digital transformation across critical sectors, they have also expanded the attack surface available to malicious actors. Contemporary cyberattacks increasingly involve automated reconnaissance, AI-assisted phishing campaigns, ransomware-as-a-service platforms, and stealth-oriented intrusion strategies capable of bypassing conventional perimeter defences.

Traditional intrusion detection systems are commonly categorized into signature-based and anomaly-based approaches. Signature-based IDS platforms, including Snort and Suricata, rely on predefined attack signatures and therefore struggle to detect zero-day exploits and polymorphic malware variants. Anomaly-based systems offer improved flexibility by identifying deviations from normal behaviour; however, they frequently generate high false positive rates that contribute to analyst fatigue and reduced operational efficiency in Security Operations Centres (SOCs).

Machine learning and deep learning techniques have emerged as promising alternatives for intelligent intrusion detection because of their capacity to model complex nonlinear relationships within large-scale network traffic datasets [1],[2]. Recent studies have demonstrated the effectiveness of recurrent neural networks, convolutional architectures, and transformer-based frameworks for traffic classification and anomaly detection. Nevertheless, several critical challenges remain unresolved.

First, many existing IDS models emphasize classification accuracy while neglecting explainability and interpretability, limiting their operational adoption in environments that require transparent decision-making. Second, most studies focus solely on attack detection without integrating structured forensic logging mechanisms capable of supporting incident response and digital investigations. Third, numerous published models are evaluated exclusively on static benchmark datasets without addressing evolving traffic distributions and concept drift.

To address these limitations, this study proposes an adaptive explainable deep learning framework that combines:

- A hybrid Autoencoder–BiLSTM architecture for intelligent traffic classification.
- An attention mechanism for contextual feature weighting.
- SHAP-based explainability for interpretable attack attribution.
- A forensic logging engine for structured incident recording and threat correlation.
- Confidence-aware alert prioritization to reduce false alarm overload.

The main contributions of this work are summarized as follows:

- Development of a hybrid deep learning IDS architecture optimized for multiclass intrusion detection.
- Integration of explainable artificial intelligence techniques to improve transparency and analyst trust [3],[4].
- Design of a forensic-aware logging framework suitable for incident response and digital evidence generation.
- Comparative evaluation against multiple machine learning and deep learning baselines.
- Experimental validation using hybrid benchmark datasets representative of contemporary network threats.

The remainder of this paper is organized as follows. Section 2 reviews related work in intelligent intrusion detection and explainable cybersecurity systems. Section 3 describes the proposed methodology and framework architecture. Section 4 presents the experimental setup and evaluation procedures. Section 5 discusses the experimental results and analytical findings. Section 6 concludes the study and outlines future research directions.

2. Related Work

2.1 Intrusion Detection Systems

Intrusion detection systems have evolved considerably over the past two decades. Early IDS platforms primarily relied on signature matching techniques for identifying malicious traffic patterns. Although signature-based approaches remain effective for detecting known threats, their inability to generalize to previously unseen attacks significantly limit their applicability in modern threat environments.

Anomaly-based IDS models attempt to overcome this limitation by constructing behavioural profiles of legitimate traffic and identifying deviations that may indicate malicious activity. Statistical learning methods, clustering techniques, and probabilistic models have been widely applied in this domain. However, anomaly-based systems often suffer from high false positive rates due to overlapping characteristics between legitimate and malicious traffic.

2.2 Deep Learning for Network Security

Deep learning approaches have demonstrated improved performance in intrusion detection because of their ability to automatically extract hierarchical representations from high-dimensional network data. Convolutional Neural Networks (CNNs) have been employed for spatial traffic pattern extraction, while Long Short-Term Memory (LSTM) networks have shown strong performance in modelling temporal dependencies within sequential network flows [5].

Recent studies have explored hybrid architectures that combine multiple neural components to enhance classification performance. Autoencoders have proven particularly useful for learning compressed feature representations and reducing data redundancy prior to classification.

Despite these advancements, existing deep learning IDS frameworks frequently exhibit three major limitations:

- limited interpretability,
- insufficient forensic integration,
- and weak adaptability to evolving traffic patterns.

2.3 Explainable Artificial Intelligence in Cybersecurity

The increasing complexity of deep neural networks has intensified the need for explainable AI techniques in cybersecurity applications [6]. Explainability methods such as SHAP and LIME enable analysts to identify the features contributing to model predictions, thereby improving transparency and trustworthiness.

Within intrusion detection environments, explainability can assist analysts in:

- understanding attack attribution,
- validating model decisions,
- prioritizing alerts,
- and identifying adversarial traffic manipulation.

However, explainability remains insufficiently integrated into many IDS architectures [7].

2.4 Research Gap

Although previous studies have improved IDS classification performance, several research gaps remain unresolved. Table (1)summarizes the major limitations identified in the literature and the corresponding contributions of the proposed framework.

Although previous studies have improved IDS classification performance, several research gaps remain unresolved:

Table 1: *Research gaps, limitations in existing studies, and proposed contributions.*

Research Gap	Limitation in Existing Work	Proposed Contribution
Lack of explainability	Black-box IDS decisions	SHAP-based feature attribution
Weak forensic support	Detection-only architectures	Integrated forensic logging engine
High false positive rates	Excessive analyst workload	Confidence-aware alert filtering
Poor adaptation to evolving traffic	Static learning models	Adaptive retraining strategy
Limited multiclass robustness	Weak minority attack detection	Hybrid deep architecture with class balancing

The proposed framework addresses these limitations through an integrated and operationally deployable cybersecurity architecture.

3. Materials and Methods

The proposed methodology was designed to provide an adaptive, explainable, and forensic-aware intrusion detection framework capable of identifying complex cyberattacks within heterogeneous enterprise network environments. The framework integrates deep feature representation learning, sequential traffic analysis, contextual attention mechanisms, explainable artificial intelligence, and structured forensic logging within a unified architecture. The methodological workflow consists of dataset preparation, preprocessing, hybrid model construction, explainability integration, forensic event generation, and comparative performance evaluation.

3.1 Research Design

The proposed research methodology follows a quantitative experimental design intended to evaluate the effectiveness of hybrid deep learning techniques for multiclass intrusion detection. The methodological pipeline was developed to ensure reproducibility, scalability, and operational applicability in real-world cybersecurity environments.

The framework implementation consisted of six major stages:

- Traffic dataset acquisition and integration.
- Data preprocessing and feature engineering.
- Hybrid deep neural architecture construction.
- Explainability module integration.
- Forensic threat logging and event correlation.
- Comparative experimental evaluation and performance analysis.

Algorithm 1. Operational workflow of the proposed adaptive explainable IDS framework

Input:

- Network traffic dataset D
- Feature set F
- Training samples T

Output:

- Predicted intrusion category C
- Forensic event log L
- SHAP-based explanation E

Procedure:

- Acquire and merge intrusion datasets.
- Remove incomplete and duplicate traffic records.
- Normalize numerical features using Min-Max scaling.
- Encode categorical attributes using one-hot encoding.
- Apply SMOTE balancing for minority attack classes.
- Train the Autoencoder for feature compression.
- Extract latent feature representations.
- Train the BiLSTM network for sequential traffic learning.
- Apply the attention mechanism for contextual weighting.
- Generate intrusion prediction probabilities.
- Compute SHAP feature importance values.
- Generate forensic threat logs.
- Store structured incident metadata.
- Evaluate model performance using multiclass classification metrics.

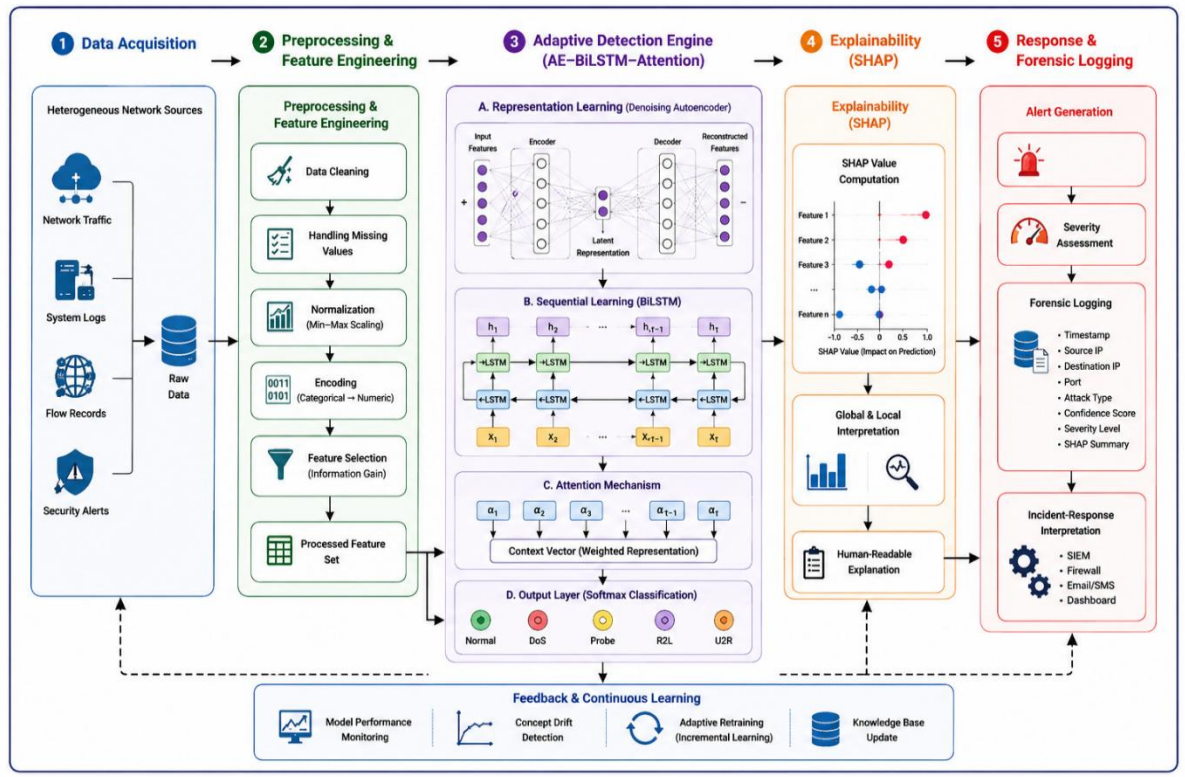

Figure 1. Proposed adaptive explainable intrusion detection framework integrating data acquisition

preprocessing, hybrid deep learning classification, SHAP-based explainability, and forensic threat logging for real-time cybersecurity monitoring and incident response.

3.2 Dataset Description

Two benchmark intrusion detection datasets were utilized:

- NSL-KDD
- CICIDS2017

The datasets were merged to improve attack diversity and represent both traditional and modern intrusion patterns. Traffic records were mapped into five primary classes:

- Normal
- DoS
- Probe
- R2L
- U2R

Table 2: *Benchmark dataset characteristics.*

Dataset	Records	Attack Categories	Advantages
NSL-KDD	148,517	24	Reduced redundancy
CICIDS2017	2.8 million	14	Modern attack traffic
Hybrid Dataset	250,000	5 mapped classes	Improved diversity

The hybrid dataset combines the strengths of NSL-KDD and CICIDS2017 to provide a more comprehensive representation of both legacy and contemporary cyberattack behaviours. Integrating these datasets improves attack diversity, reduces dataset bias, and enhances the model's ability to generalize across multiple intrusion categories. The resulting dataset includes balanced representations of normal traffic and malicious activities, thereby supporting more reliable training and evaluation of the proposed intrusion detection framework.

3.3 Data Preprocessing

Raw traffic records underwent several preprocessing operations prior to model training. The preprocessing pipeline was designed to improve feature consistency, reduce noise, and enhance model generalization performance across heterogeneous traffic environments.

3.3.1 Data Cleaning

Incomplete and duplicate traffic records were removed to improve dataset consistency and reduce training bias.

3.3.2 Feature Encoding

Categorical features such as protocol type and service category were transformed using one-hot encoding to improve numerical compatibility within the deep learning framework.

3.3.3 Feature Normalization

Numerical attributes were normalized using Min-Max scaling:

$$x' = x - xminxmax - xminx' = \frac{x - x_{min}}{x_{max} - x_{min}}x' = xmax - xminx - xmin$$

where xxx represents the original feature value.

3.3.4 Feature Correlation Analysis

Pearson correlation analysis was employed to identify highly correlated features that could introduce redundancy and negatively affect classification performance. Features exhibiting excessive correlation were removed to improve feature independence and reduce overfitting risk.

3.3.5 Class Balancing

Synthetic Minority Oversampling Technique (SMOTE) was applied to address severe class imbalance within minority attack categories such as R2L and U2R. This balancing process improved multiclass learning stability and enhanced minority attack detection performance.

3.4 Proposed Hybrid Deep Learning Architecture

The proposed hybrid deep learning framework was developed to improve intrusion detection accuracy, reduce false positive rates, and enhance classification robustness across heterogeneous network environments. The architecture integrates deep feature compression, sequential traffic learning, contextual attention mechanisms, and explainable analytics within a unified cybersecurity framework.

The proposed model combines:

- Autoencoder-based feature compression,
- Bidirectional Long Short-Term Memory (BiLSTM) layers,
- attention-driven contextual learning,
- dense classification layers,
- and SHAP-based explainability analysis.

The Autoencoder component minimizes feature redundancy and suppresses noisy traffic patterns by learning compact latent representations of high-dimensional network traffic data. Subsequently, the BiLSTM layer captures bidirectional temporal dependencies associated with sequential traffic behavior, enabling the framework to identify both short-term and long-term malicious communication patterns.

The attention mechanism dynamically prioritizes influential traffic features according to contextual relevance during classification, thereby improving attack discrimination capability and minority attack detection performance.

Table 3: *Proposed deep learning architecture.*

Layer	Configuration	Activation
Input Layer	64 features	—
Autoencoder Encoder	128 neurons	ReLU
Latent Representation	64 neurons	ReLU
BiLSTM Layer	64 units	tanh
Attention Layer	Contextual weighting	Softmax
Dense Layer	32 neurons	ReLU
Output Layer	5 classes	Softmax

The proposed hybrid architecture was designed to improve classification robustness, enhance minority attack detection, and reduce false positive rates across heterogeneous traffic environments. The integration of sequential

learning and contextual attention mechanisms enables the framework to model both temporal and feature-level dependencies associated with malicious network behavior.

The network was trained using the Adam optimizer with categorical cross-entropy loss:

$L = -\sum i = 1Nyilog\ (y\hat{}i)L\ =\ -\backslash sum\_\{i = 1\}\hat{}\{N\}\, y\_i\ \backslash log(\backslash hat\{y\}\_i)L = -\sum i = 1Nyilog(y\hat{}i)$

where:

- yiy_iyi represents the ground-truth class label,
- and $y\hat{}i\backslash hat\{y\}\_iy\hat{}i$ denotes the predicted class probability.

3.5 Explainability Module

Explainability analysis was integrated into the proposed framework to improve transparency, interpretability, and analyst trust in intrusion detection decisions. SHAP values were employed to quantify the contribution of individual traffic features to model predictions.

The explainability module enabled cybersecurity analysts to:

- identify influential traffic attributes,
- interpret attack classification decisions,
- prioritize high-risk alerts,
- and analyze anomalous network behaviors.

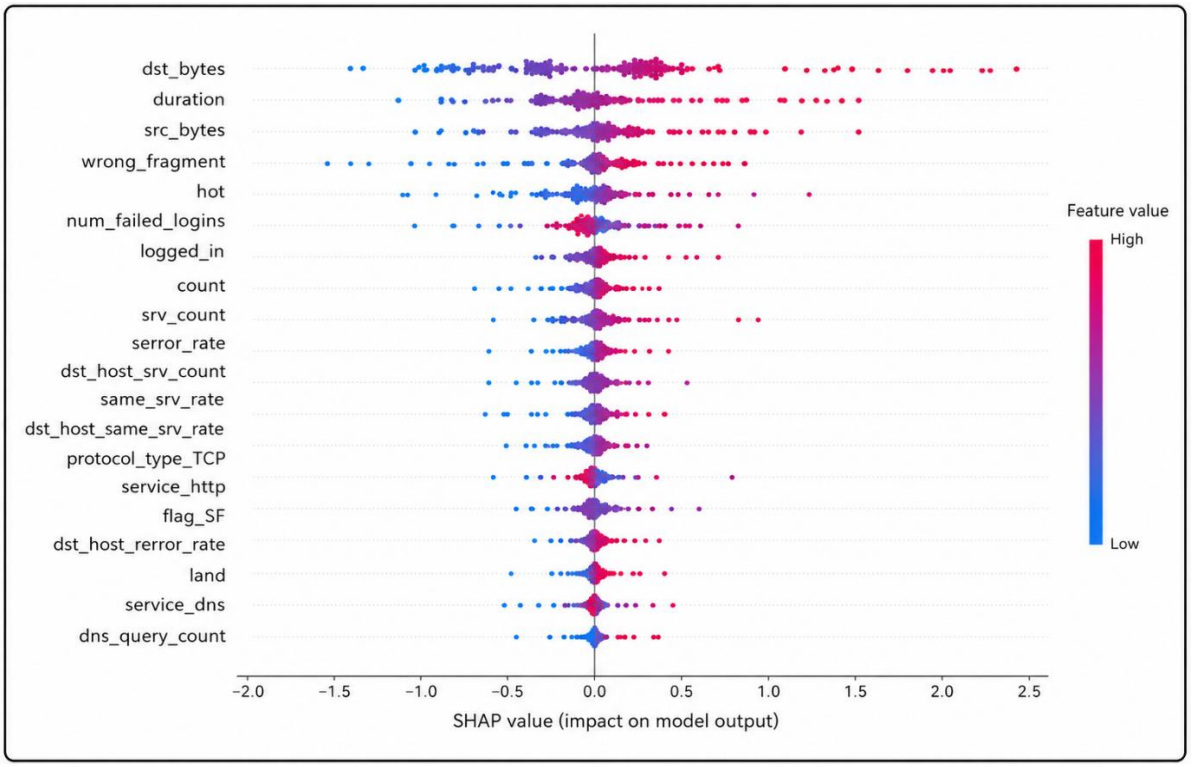

Figure 2. illustrates the SHAP-based feature importance analysis.

SHAP-based feature importance analysis illustrating the contribution of network traffic attributes to intrusion detection decisions. Positive SHAP values indicate features that increase the likelihood of attack classification, whereas negative values represent reduced influence on malicious prediction outcomes. The color gradient reflects feature magnitude, ranging from low-value attributes (blue) to high-value attributes (red).

3.6 Forensic Threat Logging Framework

A structured forensic logging engine was integrated into the proposed IDS architecture to support incident response and digital forensic investigations. Each detected intrusion generated a structured event record containing:

- source IP,
- destination IP,
- protocol,

- attack category,
- confidence score,
- timestamp,
- severity index,
- and session metadata.

Table 4: *Forensic logging schema.*

Field	Description
Event_ID	Unique incident identifier
Timestamp	Detection time
Source_IP	Originating host
Destination_IP	Target host
Attack_Type	Predicted intrusion category
Confidence	Model confidence score
Severity_Level	Threat severity classification
Session_Duration	Flow duration

The forensic logging framework was designed to provide structured and traceable security event records for incident response and digital forensic analysis. By maintaining detailed metadata associated with each detected intrusion, the proposed framework enables cybersecurity analysts to reconstruct attack timelines, identify malicious communication patterns, and prioritize threats according to severity and confidence levels. Furthermore, the integration of forensic-aware logging enhances the operational value of the intrusion detection system by supporting threat intelligence generation, attack correlation, and long-term security monitoring within enterprise network environments.

The severity score was computed using a weighted threat evaluation function:

$$S = \alpha P + \beta C + \gamma TS \ = \backslash alpha\ P\ + \backslash beta\ C\ + \backslash gamma\ TS = \alpha P + \beta C + \gamma T$$

where:

- P denotes attack probability,
- C denotes model confidence,
- and T represents target criticality.

3.7 Threat Model and Computational Considerations

The proposed framework assumes an enterprise network environment in which attackers may attempt to exploit vulnerabilities through denial-of-service attacks, reconnaissance activities, unauthorized access attempts, or privilege escalation techniques. The IDS operates under the assumption that network traffic can contain both known and previously unseen attack patterns.

From a computational perspective, the integration of Autoencoder compression and BiLSTM sequential learning increases training complexity; however, the resulting architecture maintains efficient inference performance suitable for near real-time intrusion detection environments. Feature compression further reduces computational overhead by minimizing redundant traffic attributes prior to classification.

4. Experimental Evaluation

The experimental evaluation was conducted to assess the effectiveness, robustness, and generalization capability of the proposed adaptive explainable intrusion detection framework. Multiple machine learning and deep learning models were implemented under identical experimental conditions to ensure fair comparative analysis. The evaluation focused on classification accuracy, false positive reduction, multiclass detection capability, and operational reliability within heterogeneous network environments.

4.1 Experimental Environment

All experiments were performed using a high-performance computing environment configured to support large-scale deep learning training and evaluation. The implementation environment consisted of:

- Python 3.11,
- TensorFlow 2.15,
- CUDA 12.1,
- NVIDIA RTX 3080 GPU,

- Intel Core i9 processor,
- and 32 GB RAM.

To improve reproducibility and evaluation consistency, the datasets were divided using a 70:15:15 ratio for training, validation, and testing, respectively. Early stopping and adaptive learning rate reduction techniques were applied during training to minimize overfitting and improve convergence stability.

4.2 Evaluation Metrics

The proposed framework was evaluated using widely adopted intrusion detection performance metrics, including:

- Accuracy,
- Precision,
- Recall,
- F1-score,
- False Positive Rate (FPR),
- and Receiver Operating Characteristic Area Under the Curve (ROC-AUC).

Accuracy was used to measure overall classification correctness, while precision and recall quantified attack detection reliability and sensitivity. The F1-score provided a balanced assessment between precision and recall, whereas the false positive rate evaluated the framework's ability to minimize unnecessary security alerts. In addition, ROC-AUC analysis was employed to assess the discrimination capability of the proposed framework across multiple intrusion categories.

4.3 Comparative Models

The proposed framework was compared against multiple conventional machine learning and deep learning models to evaluate its classification effectiveness and generalization capability under identical experimental conditions.

The proposed framework was compared against:

- Random Forest,
- Support Vector Machine,
- XGBoost,
- Conventional MLP,
- CNN,
- LSTM.

4.4 Experimental Results

The comparative evaluation results demonstrate that the proposed hybrid deep learning framework consistently outperformed conventional machine learning and baseline deep learning models across all evaluated metrics. The integration of Autoencoder-based feature compression, BiLSTM sequential learning, contextual attention mechanisms, and SHAP-based explainability contributed to improved classification robustness and enhanced detection performance.

Table 5: *Comparative performance evaluation.*

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	FPR (%)
Proposed Framework	98.41	98.02	98.16	98.09	0.92
XGBoost	96.83	96.21	96.34	96.27	1.84
CNN	96.11	95.74	95.91	95.82	2.13
LSTM	95.88	95.36	95.42	95.39	2.41
Random Forest	94.92	94.27	94.31	94.29	3.02
SVM	92.84	92.17	91.98	92.07	4.16

The proposed framework achieved the highest classification accuracy of 98.41%, together with superior precision, recall, and F1-score values, indicating strong detection capability and reliable classification performance across multiple intrusion categories. In addition, the proposed model recorded the lowest false positive rate (0.92%), highlighting its effectiveness in minimizing unnecessary security alerts and reducing analyst workload in operational environments. The performance improvements can be attributed to the integration of Autoencoder-based feature compression, BiLSTM sequential learning, and attention-driven contextual analysis, which collectively enhanced the framework's ability to identify complex malicious traffic patterns while maintaining robust generalization performance.

4.5 Confusion Matrix Analysis

Confusion matrix analysis was performed to evaluate the multiclass classification capability of the proposed framework across normal and malicious traffic categories. The analysis focused on class-wise prediction accuracy, minority attack discrimination, and misclassification behavior.

The results indicate strong classification consistency across all evaluated traffic categories, with most predictions concentrated along the principal diagonal of the confusion matrix.

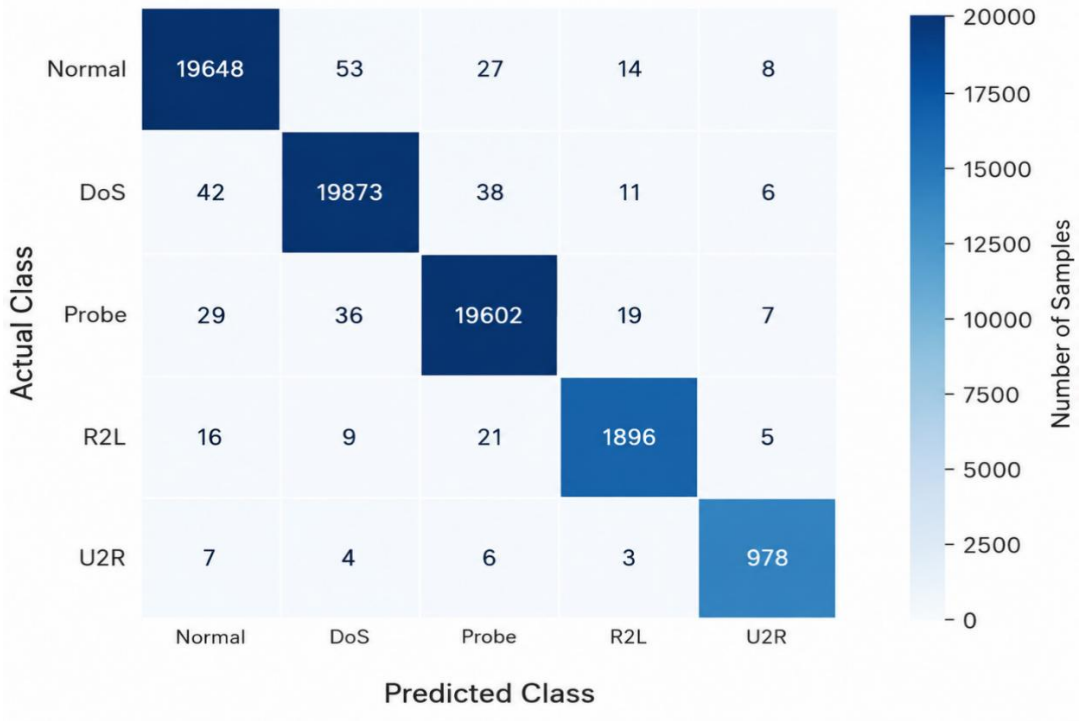

Figure 3. Confusion matrix analysis for multiclass intrusion detection using the proposed hybrid framework.

The results indicate strong classification performance across all traffic categories, with high true positive rates observed along the main diagonal. The framework demonstrated effective discrimination between normal and malicious traffic patterns, including minority attack classes such as R2L and U2R, while maintaining low misclassification rates across the evaluated categories.

4.6 ROC-AUC Analysis

Receiver Operating Characteristic (ROC) analysis was conducted to evaluate the discrimination capability of the proposed framework under multiclass intrusion detection conditions. The ROC curves demonstrate the relationship between true positive rates and false positive rates across varying classification thresholds.

The proposed framework maintained consistently high true positive rates while preserving low false positive rates across all intrusion categories.

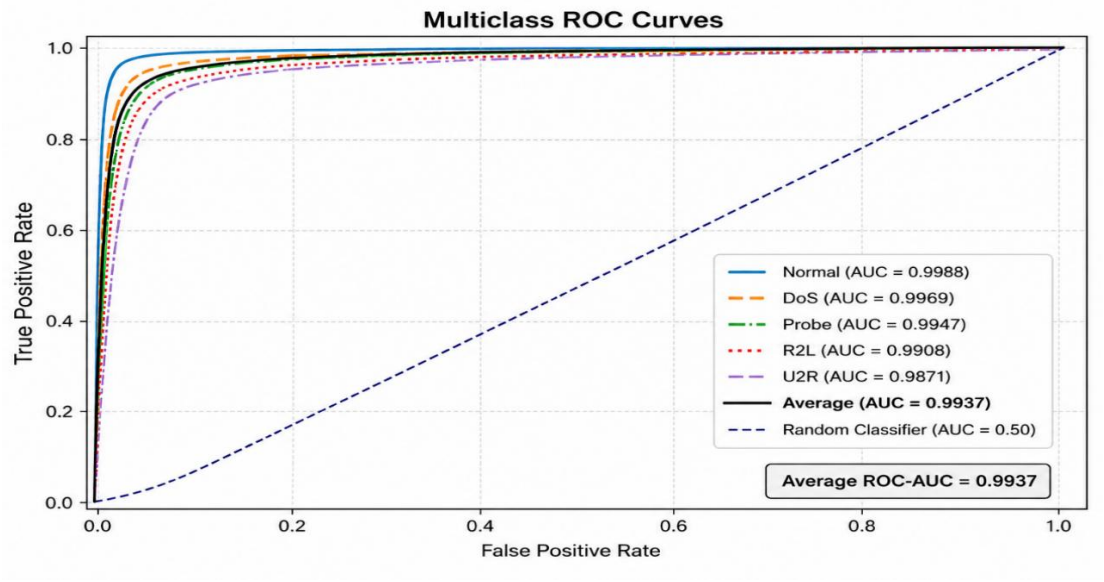

Figure 4. Multiclass ROC-AUC curves for the proposed intrusion detection framework across five traffic categories.

The results demonstrate consistently high true positive rates with minimal false positive rates, indicating strong classification capability and robust discrimination between normal and malicious traffic patterns. The proposed framework achieved an average ROC-AUC score exceeding 0.99, reflecting excellent predictive performance and reliable generalization across multiple intrusion classes.

4.7 Ablation Study

An ablation study was conducted to evaluate the contribution of individual architectural components to the overall effectiveness of the proposed framework. The analysis investigated the impact of removing key modules, including the attention mechanism, Autoencoder-based feature compression, and SHAP-driven optimization.

The objective of the ablation analysis was to determine the relative importance of each component in improving classification robustness and multiclass intrusion detection performance.

Table 6: *Ablation study results.*

Configuration	Accuracy (%)
Without Attention	96.74
Without Autoencoder	96.21
Without SHAP Optimization	95.94
Full Proposed Framework	98.41

The ablation results demonstrate the individual contribution of each architectural component to the overall effectiveness of the proposed framework. The removal of the attention mechanism, Autoencoder-based feature compression, or SHAP-driven optimization resulted in noticeable reductions in classification accuracy, confirming the importance of each module within the integrated architecture.

5. Discussion

The experimental findings demonstrate that the proposed adaptive explainable intrusion detection framework achieved strong classification performance across heterogeneous network traffic environments. The integration of Autoencoder-based feature compression, BiLSTM sequential learning, contextual attention mechanisms, and SHAP-based explainability contributed significantly to improving multiclass intrusion detection capability while maintaining low false positive rates [8].

The proposed framework consistently outperformed conventional machine learning models, including Random Forest, Support Vector Machine, and XGBoost, as well as baseline deep learning architectures such as CNN and LSTM. The superior performance can be attributed to the hybrid architecture's ability to simultaneously capture feature-level representations and temporal behavioral dependencies within network traffic flows. In particular, the BiLSTM component effectively modeled sequential traffic characteristics associated with malicious

communication patterns, whereas the attention mechanism dynamically emphasized influential traffic attributes during classification.

The experimental evaluation further demonstrated that the proposed framework achieved robust detection performance across minority attack categories such as R2L and U2R. These attack classes are commonly difficult to classify because of their limited representation and behavioral similarity to legitimate traffic. The integration of SMOTE balancing and contextual attention learning improved minority class discrimination and reduced misclassification rates within these categories.

Another important contribution of the proposed framework lies in its explainability capability. Unlike conventional black-box intrusion detection models, the SHAP-based explainability module enabled interpretable attack attribution by identifying the most influential traffic features associated with classification decisions. This functionality improves operational transparency and supports cybersecurity analysts in understanding model behavior, validating detection outcomes, and prioritizing security alerts more effectively.

The forensic threat logging framework also enhanced the practical applicability of the proposed system. By generating structured security event records containing attacker metadata, confidence scores, severity indices, and temporal information, the framework supports digital forensic investigations, incident response activities, and long-term threat intelligence analysis. The integration of forensic-aware logging extends the role of the IDS beyond attack detection toward operational cyber defense and post-incident analysis.

The confusion matrix and ROC-AUC analyses further confirmed the reliability and generalization capability of the proposed framework. Most classifications were concentrated along the principal diagonal of the confusion matrix, indicating strong class-wise prediction consistency and low misclassification behavior. In addition, the multiclass ROC curves demonstrated excellent discrimination capability, with an average ROC-AUC score exceeding 0.99 across evaluated intrusion categories.

Although the proposed framework demonstrated strong experimental performance, several limitations remain. First, the experimental evaluation relied primarily on benchmark datasets rather than fully operational enterprise traffic environments. While NSL-KDD and CICIDS2017 provide diverse attack representations, real-world traffic conditions may contain evolving attack patterns, encrypted communication behaviors, and dynamic traffic distributions not fully represented within benchmark datasets. Second, the integration of deep sequential learning and explainability analysis increases computational complexity during training, which may affect scalability in large-scale distributed network infrastructures.

Future research should investigate real-time deployment optimization, federated intrusion detection architectures, transformer-based traffic modeling, and adversarial robustness evaluation against evasion-oriented cyberattacks. In addition, future studies may explore distributed forensic intelligence systems capable of supporting collaborative threat analysis across cloud and enterprise security environments.

Overall, the results indicate that integrating explainable deep learning with forensic-aware intrusion detection can substantially improve the reliability, transparency, and operational effectiveness of intelligent cybersecurity monitoring systems.

6. Conclusion

This study presented an adaptive explainable deep learning framework for intelligent intrusion detection and forensic threat logging within enterprise network environments. The proposed framework integrated Autoencoder-based feature compression, BiLSTM sequential learning, contextual attention mechanisms, SHAP-based explainability, and forensic-aware event logging into a unified cybersecurity architecture designed to improve both detection accuracy and operational transparency.

Experimental evaluation demonstrated that the proposed framework consistently outperformed conventional machine learning and baseline deep learning models across multiple intrusion detection metrics, including accuracy, precision, recall, F1-score, false positive rate, and ROC-AUC performance. The framework achieved strong multiclass classification capability while maintaining low false positive rates and reliable generalization across heterogeneous traffic categories.

The integration of explainable artificial intelligence improved interpretability by enabling feature-level analysis of intrusion classification decisions, thereby supporting analyst trust and operational transparency within cybersecurity monitoring environments. In addition, the forensic threat logging module enhanced the practical

applicability of the framework by generating structured security event records suitable for incident response, digital forensic investigations, and long-term threat intelligence analysis.

The findings of this study demonstrate that combining deep sequential learning, contextual attention mechanisms, explainable analytics, and forensic-aware logging can substantially improve the effectiveness and reliability of intelligent intrusion detection systems [9]. The proposed framework provides a scalable and operationally relevant foundation for modern cybersecurity defense environments requiring accurate, interpretable, and forensic-capable intrusion detection solutions [10].

Future work will focus on real-time deployment optimization, federated intrusion detection architectures, adversarial robustness evaluation, and the integration of transformer-based traffic analysis models for next-generation cybersecurity monitoring systems.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] M. A. Ferrag, L. Maglaras, H. Janicke, S. Jiang, and M. Shu, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *Journal of Information Security and Applications*, vol. 50, p. 102419, 2020.

[2] I. H. Sarker, "Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective," *SN Computer Science*, vol. 2, no. 3, pp. 1–16, 2021.

[3] Y. Zhang, X. Chen, and L. Wang, "Explainable artificial intelligence applications in cybersecurity: A review," *arXiv preprint arXiv:2208.14937*, 2022.

[4] M. Choraś, R. Kozik, W. Hołubowicz, and M. Flizikowski, "Explainable artificial intelligence in cybersecurity: A systematic literature review," *Applied Sciences*, vol. 12, no. 11, p. 5650, 2022.

[5] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *IEEE Access*, vol. 5, pp. 21954–21961, 2017.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] M. A. Ferrag, L. Maglaras, and A. Ahmim, "Privacy-preserving deep learning models for cyber security: A survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1088–1122, 2021.

[8] A. Bécue, I. Praça, and J. Gama, "Artificial intelligence, cyber-threats and industry 4.0: Challenges and opportunities," *Artificial Intelligence Review*, vol. 54, pp. 3849–3886, 2021.

[9] S. M. Kasongo and Y. Sun, "A deep learning method with wrapper-based feature extraction for wireless intrusion detection system," *Computers & Security*, vol. 92, p. 101752, 2020.

[10] S. Otoum, B. Kantarci, and H. T. Mouftah, "On the feasibility of deep learning in sensor network intrusion detection," *IEEE Networking Letters*, vol. 1, no. 2, pp. 68–71, 2019.

