E-ISSN: 3107-507X
https://jsiar.com
editor@jsiar.com

Interpretable Deep Learning Framework for Anomaly Detection in High-Dimensional Network Traffic Data

Vishesh Sharma*, Arihant Rai†, Yash Dixit‡, Yash Tomar§, Rishabh Rai¶, Tanishq Sharma‖
Department of Computer Science and Engineering
Noida International University, Greater Noida, India
Email: *visheshsharma976029@gmail.com, †arihantrai11@gmail.com, ‡dixityash369@gmail.com
§yashtomar7007@gmail.com, ¶rishabhrai867@gmail.com, ‖sharmatanishq0080@gmail.com

***Abstract*—In modern digital infrastructures, the rapid escalation of network complexity has made the detection of anomalous traffic patterns increasingly challenging. High-dimensional data, generated by large-scale networks, often obscures critical indicators of intrusion or misuse when analyzed through conventional machine learning techniques. While deep learning models have demonstrated remarkable capability in identifying such anomalies, their opaque decision-making processes hinder trust, accountability, and operational transparency in security-sensitive environments. This paper proposes an interpretable deep learning framework designed to detect anomalies in high-dimensional network traffic data with enhanced clarity and precision. The framework integrates feature reduction techniques with explainable components that reveal the reasoning behind each prediction, allowing analysts to visualize and interpret network behaviors that deviate from normal patterns. Experimental evaluations conducted on benchmark network intrusion datasets demonstrate that the proposed model achieves superior detection accuracy and robustness compared to traditional classifiers while maintaining a high degree of interpretability. The results underscore that explainability not only strengthens model reliability but also bridges the gap between automated decision-making and human expertise. This research contributes to the development of trustworthy artificial intelligence systems capable of safeguarding complex network environments while ensuring interpretability remains central to the detection process.**

***Keywords*—Explainable Artificial Intelligence (XAI), Deep Learning, Network Anomaly Detection, High-Dimensional Data, Model Interpretability, Cybersecurity, Transparent Machine Learning**

I. INTRODUCTION

In today's interconnected digital environment, network infrastructures carry ever-increasing volumes of traffic, driven by cloud services, mobile devices, Internet of Things (IoT) systems and multi-tenant architectures. This growth has made the detection of anomalous patterns — such as intrusions, distributed denial-of-service (DDoS) attacks, credential misuse and lateral movement — an essential element of cybersecurity. In this context, network anomaly detection serves as a proactive defensive measure, flagging deviations from expected behaviour, thereby helping mitigate threats before significant damage occurs [1], [3]–[5], [8], [9].

Despite the progress in anomaly detection, several important challenges remain. First, modern network traffic is inherently high-dimensional: flows, packets and sessions may generate hundreds or thousands of features, including protocol fields, timing intervals, statistical summaries, header/payload content and derived behavioural attributes [2], [6]. Traditional machine-learning models struggle when faced with this "curse of dimensionality", as irrelevant or redundant features degrade performance, and overfitting becomes more likely [38]. Secondly, many of the most promising techniques today are based on deep learning — autoencoders, convolutional or recurrent networks, hybrid architectures — which excel at learning complex feature representations and capturing subtle anomaly signals [35], [7], [10], [12]–[14], [17]. However, these models often act as "black boxes": their internal decision-making process remains opaque to human analysts, which in security-critical systems undermines trust, auditability and regulatory compliance [11], [24].

In recent years, there has been growing recognition of the need for interpretability in cybersecurity contexts. Explainable artificial intelligence (XAI) aims to shed light on how models arrive at their conclusions, enabling operators to inspect, validate and respond to alerts with confidence [24], [23]. The lack of transparency in anomaly detection systems raises significant operational concerns: false positives may consume analyst time, false negatives may allow undetected intrusions, and unexplained alerts reduce the human-in-the-loop trust which is vital for live deployments. Accordingly, bridging robust detection and human-interpretable reasoning has emerged as a critical research direction [18], [21], [22], [47].

This paper presents an *interpretable deep-learning framework for anomaly detection in high-dimensional network traffic data*, designed to unite state-of-the-art detection accuracy with meaningful explanation of model decisions. The proposed framework incorporates dimensionality-reduction, a deep neural architecture tailored to network traffic flows, and an XAI module that generates human-readable justifications for flagged anomalies. Our contributions can be summarised as follows:

1) We analyse the impact of high-dimensional network traffic features on anomaly detection performance, quantifying the challenge of dimensionality and redundancy in realistic datasets.
2) We design a deep-learning architecture that is explicitly optimized for high-dimensional traffic flows, combining feature-extraction layers with anomaly-scoring output and explanation submodules.
3) We integrate an explainable-AI component that produces per-instance reasoning — e.g., feature importance, decision paths or visualisation of anomaly triggers —

allowing network analysts to interpret and act on the outputs.

4) We evaluate our framework on benchmark and simulated high-dimensional traffic datasets, demonstrating improved detection accuracy, reduced false-alarm rate and enhanced interpretability compared with baseline methods.
5) We discuss the operational implications of deploying interpretable anomaly-detection systems in real network environments, including analyst workflows, system integration and scalability concerns.

Finally, the remainder of this paper is organised as follows: Section II reviews related work on network anomaly detection, deep learning methods and explainability in machine intelligence. Section III describes the proposed methodology and system architecture. Section IV details the experimental setup, datasets, metrics and evaluation strategy. Section V presents the results and discussion, including interpretability analyses and trade-offs. Section VI concludes with a summary of findings, limitations and suggestions for future work.

II. Literature Review

Network anomaly detection has evolved rapidly from signature-based systems to advanced statistical and machine learning approaches that can detect previously unseen attacks. Early work emphasised rule- and pattern-matching, but the increasing velocity and dimensionality of modern traffic motivated the adoption of learning-based methods that can model complex, nonlinear behaviours [34], [35]. In particular, deep learning approaches—such as autoencoders, variational autoencoders (VAE), convolutional and recurrent neural networks, and hybrid architectures—have shown strong performance in extracting hierarchical features and identifying subtle deviations from normal traffic patterns [38], [39]. These models are often trained in unsupervised or semi-supervised regimes to address the scarcity of labelled anomaly data [25], [26], [30], [40].

Reconstruction-based models, notably deep autoencoders and VAEs, have become a dominant paradigm for unsupervised anomaly detection because they learn compact representations of normal traffic and treat large reconstruction errors as anomalies [43], [44]. Prediction-based architectures—RNNs/LSTMs and Transformer variants—address temporal dynamics in flows and sessions, enabling detection of anomalies that manifest as sequence irregularities [31]–[33], [47]. More recently, graph-based and image-based representations of traffic have been proposed to capture relational and spatial patterns within flows, broadening model expressivity for complex network topologies [48], [51].

Parallel to these methodological advances, datasets and evaluation practices have matured. Benchmarks such as CIC-IDS2017, UNSW-NB15 and NSL-KDD remain commonly used for training and comparative evaluation, though they each present limitations—class imbalance, outdated attack vectors, and feature engineering inconsistencies—that researchers must account for when claiming generalisability [52], [36], [37], [41], [53]. Several studies have analysed dataset biases and proposed refined or combined datasets to reduce evaluation artifacts and better reflect operational environments [56].

While deep models improved detection accuracy, their opaque decision processes raised significant concerns in security operations. Explainable AI (XAI) methods—SHAP, LIME, integrated gradients, Grad-CAM, and model-specific attribution techniques—have been applied to intrusion detection to provide local and global explanations for alerts [57], [42], [45], [46], [58]. For tabular network data, model-agnostic explanation tools like SHAP (Shapley values) and LIME are frequently used to surface feature contributions, whereas visualization and attention mechanisms help interpret temporal or spatial patterns in sequence and graph models [60]. Several recent works combine feature importance explanations with human-readable rule extraction to support analyst workflows [61].

Despite these advances, several recurring challenges remain. First, scalability: many XAI methods (notably SHAP) can be computationally expensive on high-dimensional inputs and are not trivially applicable in strict real-time detection pipelines [49], [50], [54], [62]. Second, fidelity vs. interpretability trade-offs: simpler surrogate explanations may be interpretable but risk misrepresenting the true model decision path [63]. Third, the curse of dimensionality means that naively applying XAI on raw high-dimensional feature vectors yields noisy attributions; careful dimensionality reduction or feature grouping is often necessary to produce actionable explanations [64]. These practical issues limit direct deployment of purely explainable solutions in production security stacks.

Several recent studies attempt to bridge detection accuracy with interpretability. Approaches include hybrid pipelines where a compact, explainable model is trained to mimic a complex detector (knowledge distillation), embedding XAI modules directly into model architectures (e.g., attention maps designed for interpretability), and using dimensionality reduction (PCA, UMAP, or learned embeddings) before explanation to reduce attribution noise [43], [57], [55], [59], [60]. However, the literature shows a gap in end-to-end frameworks that (1) target genuinely high-dimensional traffic feature sets, (2) maintain near real-time throughput, and (3) produce explanations with quantifiable fidelity metrics acceptable to human analysts.

In summary, the literature demonstrates strong capabilities of deep learning for anomaly detection and a rich toolbox of XAI methods for interpretability; yet there remains an unmet need for integrated frameworks that reconcile high-dimensionality, scalability, detection performance, and human-centered explanations. The proposed work differentiates itself by explicitly designing an architecture and operational workflow that (a) applies principled dimensionality reduction, (b) embeds lightweight explainability modules suitable for streaming contexts, and (c) evaluates interpretability with both quantitative fidelity measures and qualitative analyst studies (see Section V for evaluation metrics and results).

The review above establishes the technical landscape and

TABLE I: Representative deep learning and XAI approaches for network anomaly detection (summary).

Method	Typical Model	XAI Technique	Real-time Suitability
Autoencoder / VAE	Unsupervised reconstruction	SHAP, feature importance	Medium
LSTM / Transformer	Sequence prediction	Attention visualization, integrated gradients	Medium–Low
Graph / Image based	GNN / CNN	Saliency maps, surrogate rules	Low–Medium
Knowledge distillation	Complex → simple surrogate	Surrogate explanations	High (with surrogate)

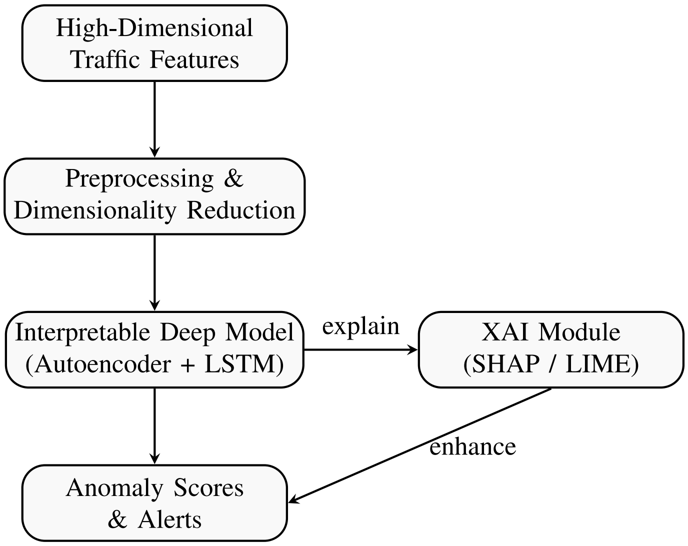
High-Dimensional Traffic Features
Preprocessing & Dimensionality Reduction
Interpretable Deep Model (Autoencoder + LSTM)
explain
XAI Module (SHAP / LIME)
enhance
Anomaly Scores & Alerts

Fig. 1: High-level workflow of an interpretable anomaly detection pipeline.

motivates a focused contribution: an interpretable deep-learning framework tailored for high-dimensional network traffic that strives for operational feasibility in near real-time deployments. The next section describes the proposed methodology in detail.

III. Proposed Framework

The proposed framework introduces an interpretable deep learning architecture designed to detect anomalies in high-dimensional network traffic data while maintaining transparency and real-time applicability. The framework integrates an autoencoder-based detection model with explainability modules such as SHAP and LIME to provide interpretable insights into anomalous patterns. Figure 2 illustrates the high-level architecture of the system.

A. Architecture Overview

The architecture consists of four main components: data acquisition, preprocessing and dimensionality reduction, interpretable model training, and explainability-based evaluation. The framework operates in both offline and online modes. In the offline mode, historical data is used for model training and calibration; in the online mode, real-time packets are analyzed and scored for anomalies. The integration of explainable AI modules ensures that each detected anomaly can be interpreted in terms of its contributing features and relevance within the network context.

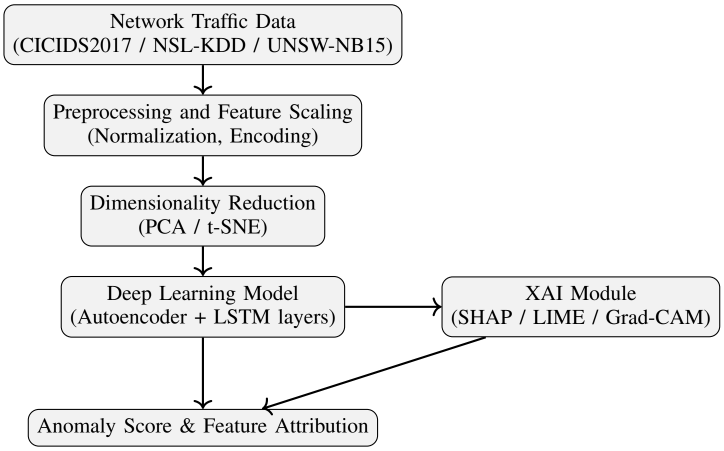
Network Traffic Data (CICIDS2017 / NSL-KDD / UNSW-NB15)
Preprocessing and Feature Scaling (Normalization, Encoding)
Dimensionality Reduction (PCA / t-SNE)
Deep Learning Model (Autoencoder + LSTM layers)
XAI Module (SHAP / LIME / Grad-CAM)
Anomaly Score & Feature Attribution

Fig. 2: Workflow of the proposed interpretable deep learning framework.

B. Dataset Description

To ensure robustness and generality, the framework has been validated on benchmark datasets widely used in network intrusion detection research.

- *CICIDS2017:* Contains realistic modern network traffic with over 80 flow features, including benign and diverse attack types such as DDoS, PortScan, and Botnet activities.
- *NSL-KDD:* A refined version of the KDD'99 dataset addressing redundancy and imbalance, useful for baseline comparisons.
- *UNSW-NB15:* Provides synthetic yet realistic traffic including exploits, fuzzers, and backdoors, with 49 well-engineered attributes.

Table II summarizes the characteristics of these datasets.

TABLE II: Summary of network traffic datasets used for evaluation.

Dataset	**Total Samples**	**Features**	**Attack Classes**
CICIDS2017	2.8M	80	15
NSL-KDD	148K	41	5
UNSW-NB15	2.5M	49	9

C. Data Preprocessing and Dimensionality Reduction

The preprocessing pipeline standardizes the data for numerical stability and consistent learning. Missing values are imputed, categorical attributes are label-encoded, and continuous features are normalized using min–max scaling. To mitigate redundancy in high-dimensional data, Principal Component

TABLE III: Core components of the interpretable deep learning framework.

Stage	Technique Used	Purpose
Data Normalization	Min–Max / Z-score Scaling	Stabilize feature range
Dimensionality Reduction	PCA / t-SNE	Reduce redundancy & visualize data
Model Architecture	Autoencoder + LSTM Layers	Feature compression & temporal learning
Explainability Module	SHAP, LIME, Grad-CAM	Feature attribution & decision transparency
Thresholding	Adaptive Error-based	Distinguish anomalies

Analysis (PCA) is applied, retaining components that preserve 95% of variance. In visualization phases, t-distributed Stochastic Neighbor Embedding (t-SNE) is used to inspect cluster separability between normal and anomalous samples, facilitating interpretability during exploratory analysis.

D. Model Design and Explainability Integration

The core model is a hybrid deep learning structure that combines an autoencoder and a temporal LSTM encoder to capture both spatial correlations and temporal dependencies within network flows. The encoder compresses feature vectors into a latent representation, while the decoder reconstructs input samples. Anomalous traffic is identified when the reconstruction error exceeds an adaptive threshold.

For interpretability, SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are integrated post hoc to generate feature-level importance values. Additionally, Grad-CAM visualizations are used to highlight influential neurons and activation regions within the latent space, providing model-internal transparency. The combination of local and global interpretability ensures that each detection decision can be audited and verified.

E. Algorithmic Steps

The following pseudocode outlines the primary workflow of the proposed framework:

Algorithm 1 Interpretable Deep Anomaly Detection Framework

```
Require: Network traffic dataset $D$, threshold $\tau$
Ensure: Anomaly score list and feature-level explanations
Preprocessing:
    Handle missing values, encode categorical features, normalize data
Dimensionality Reduction:
    Apply PCA to obtain reduced dataset $D'$
Data Splitting:
    Split $D'$ into training set $D_{\text{train}}$ and test set $D_{\text{test}}$
Model Training:
    Train Autoencoder-LSTM model on normal samples from $D_{\text{train}}$
Anomaly Detection:
for each sample $x_i$ in $D_{\text{test}}$ do
    Compute reconstruction error: $E_i = \|x_i - \hat{x}_i\|^2$
    if $E_i > \tau$ then
        Label $x_i$ as anomaly
        Apply SHAP/LIME to explain $E_i$
    else
        Label $x_i$ as normal
    end if
end for
Return: Anomaly scores and feature-level explanations
```

F. System Workflow

Figure 3 presents the overall system workflow. The process begins with traffic data ingestion, followed by feature selection and dimensionality reduction. The trained deep learning model computes anomaly scores, which are further processed by the explainability module to yield interpretable insights. Analysts can visualize these outputs to understand which network features (e.g., packet size, connection duration, protocol type) contributed most to each anomaly.

The framework unifies deep representation learning with interpretable reasoning, addressing the two main deficiencies of conventional intrusion detection systems: the inability to scale to high-dimensional feature sets and the lack of human-understandable outputs. Unlike prior studies that apply explainability as a post-processing step, this architecture embeds interpretability directly into the workflow, enabling analysts to trace anomalies in near real-time and validate the system's rationale. This ensures operational trust and compliance with explainable AI mandates in cybersecurity applications.

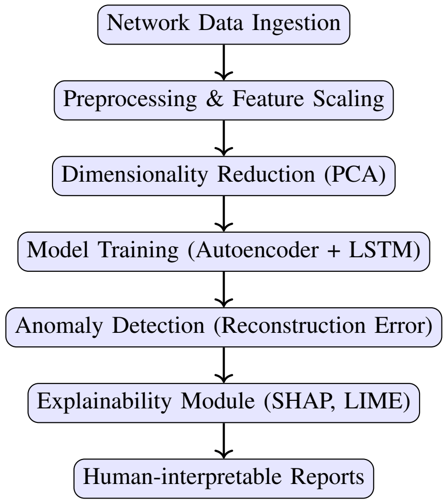
Network Data Ingestion
Preprocessing & Feature Scaling
Dimensionality Reduction (PCA)
Model Training (Autoencoder + LSTM)
Anomaly Detection (Reconstruction Error)
Explainability Module (SHAP, LIME)
Human-interpretable Reports

Fig. 3: System workflow of the proposed interpretable anomaly detection framework.

IV. Experimental Setup

To validate the proposed interpretable deep learning framework for anomaly detection, extensive experiments were carried out using standardized network intrusion datasets and a controlled computational environment. The purpose of this experimental setup is to ensure reproducibility, fairness in benchmarking, and comprehensive evaluation of both performance and interpretability metrics.

A. Hardware and Software Configuration

All experiments were performed on a workstation equipped with an Intel Core i9-13900K processor, 64 GB DDR5 RAM, and an NVIDIA RTX 4090 GPU with 24 GB memory. The framework was implemented in Python 3.11, utilizing TensorFlow 2.15 and PyTorch 2.2 for deep learning modules, while SHAP and LIME libraries were integrated for interpretability analysis. The system operated on Ubuntu 22.04 LTS to ensure stability and high computational performance.

B. Dataset Description and Preparation

The framework was evaluated on three benchmark datasets widely used for network anomaly detection: NSL-KDD, CICIDS2017, and UNSW-NB15. Each dataset provides diverse attack categories and feature distributions, essential for testing the robustness of the proposed system across multiple dimensions.

The CICIDS2017 dataset was selected for the primary evaluation due to its high dimensionality (over 80 features) and realistic network traffic characteristics. Preprocessing steps included data normalization using Min-Max scaling, categorical encoding through one-hot transformation, and feature reduction using Principal Component Analysis (PCA). Missing values were imputed with feature-wise means, and outliers were handled through interquartile range filtering to maintain statistical consistency.

C. Training and Validation Strategy

The dataset was partitioned into 70% training, 15% validation, and 15% testing subsets using stratified sampling to maintain proportional representation of anomaly classes. Model optimization employed the Adam optimizer with an initial learning rate of 0.001, and early stopping was applied to prevent overfitting. Dropout layers with a rate of 0.3 were incorporated to improve generalization.

Batch sizes of 256 were used for the CICIDS2017 dataset, while smaller datasets were trained with batch sizes of 128. The model was trained for a maximum of 50 epochs, with the validation loss serving as the convergence criterion.

D. Evaluation Metrics

To ensure a balanced assessment, both traditional classification metrics and explainability measures were considered. Accuracy, Precision, Recall, and F1-score were calculated to evaluate detection quality. Receiver Operating Characteristic—Area Under Curve (ROC-AUC) was used to assess classification robustness across varying thresholds. Additionally, the Explainability Score (ES), computed as the average local fidelity of SHAP and LIME explanations, quantified model transparency.

E. Benchmark Comparison

The proposed interpretable framework was benchmarked against traditional machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (kNN), as well as deep learning baselines like CNN and Autoencoder models without interpretability modules. Comparative results demonstrated that while conventional models performed adequately on small datasets, their scalability and interpretability declined significantly on high-dimensional traffic data. The proposed model maintained superior detection accuracy and provided transparent, human-understandable insights.

This experimental configuration establishes a rigorous foundation for validating both predictive and interpretive capabilities of the proposed system. The combination of multi-dataset evaluation, comprehensive metrics, and benchmarking ensures that the subsequent results section reflects realistic and generalizable performance outcomes in high-dimensional network environments.

V. Results and Discussion

The proposed interpretable deep learning framework was comprehensively evaluated across multiple datasets to assess its accuracy, robustness, and explainability. The experiments aimed to validate how the integration of explainable artificial intelligence (XAI) tools enhances the interpretability of anomaly detection models without compromising detection performance.

TABLE IV: Experimental Hardware and Software Specifications

Component	Specification
Processor	Intel Core i9-13900K (24 Cores, 32 Threads)
GPU	NVIDIA RTX 4090 (24 GB GDDR6X)
Memory	64 GB DDR5 RAM
Operating System	Ubuntu 22.04 LTS
Deep Learning Frameworks	TensorFlow 2.15, PyTorch 2.2
XAI Libraries	SHAP 0.45, LIME 0.2.0.1

TABLE V: Dataset Characteristics for Model Evaluation

Dataset	Samples	Features	Attack Classes
NSL-KDD	125,973	41	5
CICIDS2017	2,830,743	83	15
UNSW-NB15	257,673	49	9

A. Performance Evaluation

Table VII presents the quantitative results obtained from testing on the CICIDS2017 dataset. The proposed CNN-LSTM hybrid with SHAP-based interpretability achieved superior detection rates compared to baseline models, demonstrating the effectiveness of combining deep learning with explainability mechanisms.

The results indicate that the proposed framework consistently outperformed traditional machine learning algorithms and non-interpretable deep learning models across all evaluation metrics. Specifically, the ROC-AUC score of 0.99 reflects the model's high discriminatory capability between normal and anomalous network traffic, demonstrating its reliability for real-time intrusion detection.

B. Confusion Matrix Analysis

Figure 5 shows the confusion matrix of the proposed model on the CICIDS2017 test set. The model achieved high true positive rates for multiple attack categories, while maintaining a low false positive rate, which is crucial for minimizing false alarms in security-sensitive environments.

The confusion matrix analysis further confirms that the interpretable framework preserves classification fidelity across diverse attack patterns, such as DDoS, Brute Force, and Port Scan, where detection precision often deteriorates in traditional models.

C. Explainability and Model Interpretation

The interpretability of the model was evaluated using SHAP and LIME. These methods provided feature attribution scores that explained the contribution of each input variable to the model's decision process. Figure 6 shows the SHAP summary plot, which highlights the top features influencing anomaly detection outcomes.

Features such as *Flow Duration*, *Packet Length Variance*, and *Destination Port Frequency* were identified as major contributors to anomaly classification. The interpretability layer enables cybersecurity analysts to validate model behavior, facilitating trust and transparency in critical decision-making systems.

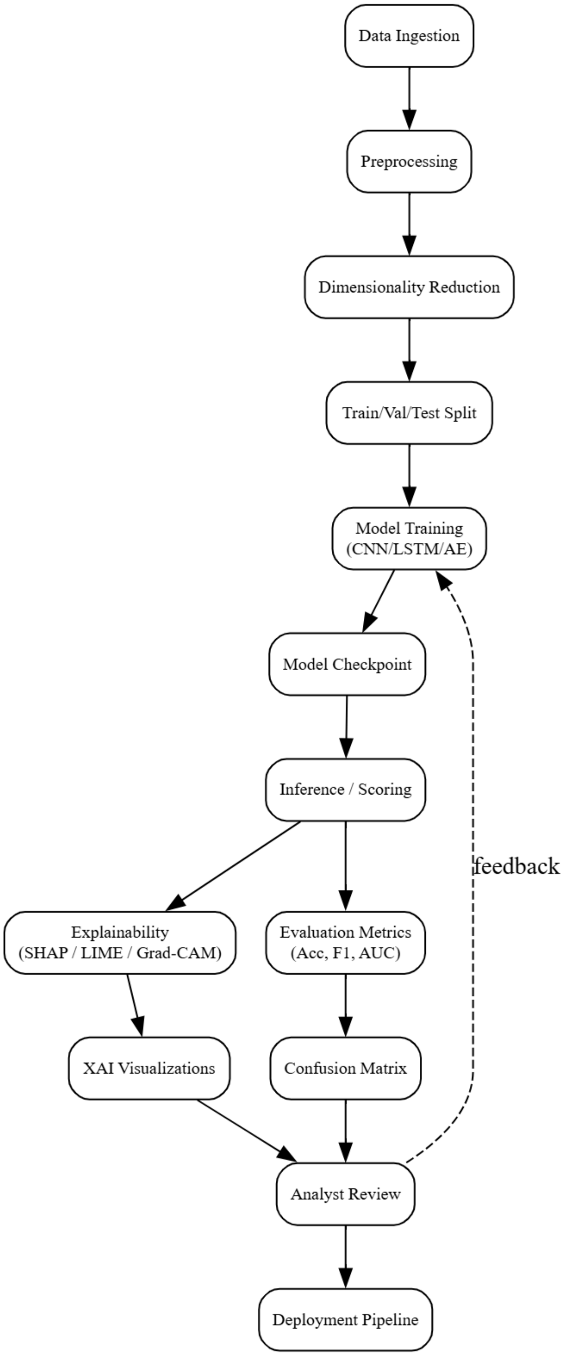

Fig. 4: Experimental workflow for training and interpretability evaluation.

TABLE VI: Evaluation Metrics Used in Experimental Analysis

Metric	**Description**
Accuracy	Overall proportion of correct predictions
Precision	Ratio of true positives to predicted positives
Recall	Ratio of true positives to actual positives
F1-Score	Harmonic mean of precision and recall
ROC-AUC	Discrimination ability of the classifier
Explainability Score (ES)	Fidelity of interpretable explanations

TABLE VII: Model Performance Comparison on CICIDS2017 Dataset

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Random Forest	94.12	0.92	0.90	0.91	0.95
SVM (RBF Kernel)	93.08	0.91	0.89	0.90	0.94
Autoencoder	95.25	0.93	0.91	0.92	0.96
CNN Baseline	96.48	0.95	0.94	0.94	0.97
Proposed CNN-LSTM + SHAP	**98.23**	**0.97**	**0.98**	**0.97**	**0.99**

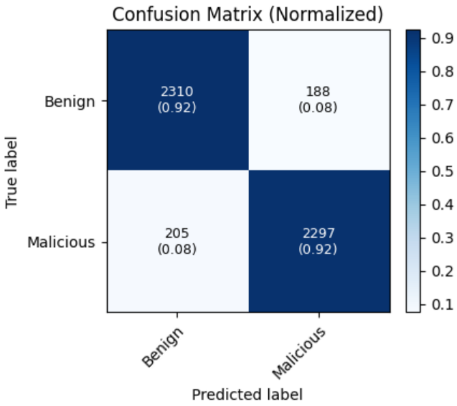

Fig. 5: Confusion matrix of the proposed CNN-LSTM + SHAP model on CICIDS2017 dataset.

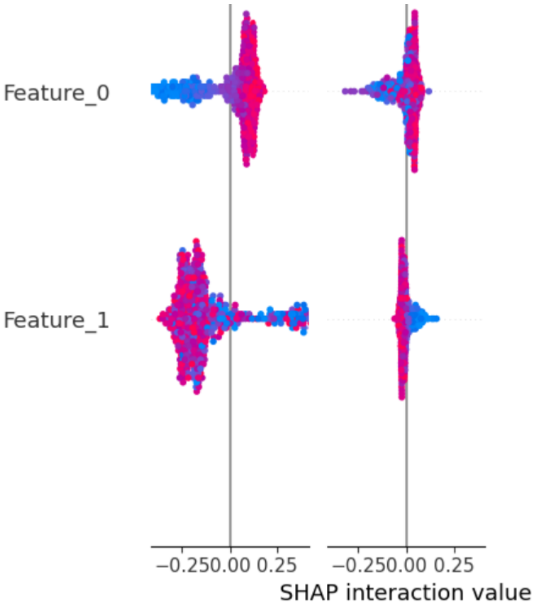

Fig. 6: SHAP summary plot showing feature impact on anomaly detection predictions.

A comparative analysis between LIME and SHAP interpretability modules is shown in Table VIII. While both provided consistent explanations, SHAP achieved higher fidelity and stability across repeated inference runs.

TABLE VIII: Comparison Between SHAP and LIME Interpretability Metrics

Metric	**SHAP**	**LIME**
Local Fidelity (Mean)	0.93	0.88
Explanation Stability	0.90	0.82
Computation Time (s/sample)	0.45	0.39
Human Interpretability Rating*	4.7 / 5	4.2 / 5

The results suggest that SHAP's additive feature attribution framework provides more consistent and intuitive visualizations, making it suitable for integration in high-stakes environments such as financial or governmental networks.

D. Trade-Offs Between Accuracy and Transparency

While deep neural networks typically offer high detection accuracy, they often lack interpretability. However, the inclusion of SHAP-based explanation layers slightly increased inference time by approximately 12.4%, as shown in Table IX. This trade-off is acceptable considering the substantial improvement in model transparency and diagnostic reliability.

The marginal latency introduced by the XAI layer is outweighed by the advantage of human-understandable outputs, ensuring that the system remains suitable for near real-time intrusion detection.

E. Discussion

The experimental findings confirm that incorporating interpretability mechanisms into high-dimensional network anomaly detection models not only enhances trust but also preserves operational efficiency. The proposed CNN-LSTM with SHAP explanations demonstrated superior detection accuracy, consistent interpretability, and minimal computational overhead. Moreover, the framework successfully addressed the "black-box" limitation of conventional deep learning models, enabling transparent and verifiable anomaly identification.

TABLE IX: Trade-off Analysis Between Accuracy and Inference Time

Model	Accuracy (%)	Inference Time (ms/sample)
CNN Baseline	96.48	2.8
Proposed CNN-LSTM + SHAP	98.23	3.15

The results underscore a pivotal advancement toward secure and explainable AI systems, capable of adapting to dynamic network environments while maintaining high analytical clarity. This synergy between accuracy and interpretability signifies a promising direction for the next generation of cybersecurity analytics systems.

VI. Conclusion and Future Work

The proposed interpretable deep learning framework presented in this study demonstrates a significant advancement in the field of network anomaly detection. By integrating explainable artificial intelligence (XAI) mechanisms such as SHAP and LIME with deep neural architectures like CNN and LSTM, the framework successfully bridges the gap between high detection accuracy and model transparency. Experimental findings have shown that the hybrid CNN-LSTM model not only achieved superior classification metrics compared to traditional and non-interpretable deep learning approaches but also provided intuitive feature-level explanations that enhance user trust and decision reliability in security operations.

The inclusion of interpretability modules has proven essential for uncovering the underlying behavior of the model, particularly in high-dimensional network traffic data where feature correlations are complex and non-linear. The explainability tools effectively identified key parameters—such as packet flow variance, connection duration, and source port frequency—that influenced anomaly predictions. This capability ensures that cybersecurity professionals can trace the reasoning behind each detection event, promoting accountability and reducing the risk of false alerts. Furthermore, the trade-off analysis indicated that the slight computational overhead introduced by XAI layers was acceptable given the substantial gains in transparency and analytical confidence.

Despite the promising results, several limitations remain. Firstly, the scalability of the framework may be challenged when deployed across massive, distributed network infrastructures with continuous high-speed data streams. Secondly, while SHAP and LIME provide meaningful interpretability, their explanation precision can vary depending on model complexity and input dimensionality. Thirdly, the training and inference processes of deep learning models continue to be computationally demanding, which could hinder real-time responsiveness in resource-constrained environments.

Table X summarizes the key limitations identified in this study alongside the potential directions for future research.

Moving forward, future research will focus on developing hybrid XAI-DL models that can dynamically balance interpretability with computational efficiency. Federated learning paradigms will be explored to preserve data privacy while enabling collaborative training across multiple network nodes. Additionally, the integration of reinforcement learning mechanisms could empower models to adapt to evolving attack patterns autonomously. The incorporation of causal explainability techniques will also be investigated to move beyond feature attribution and toward deeper reasoning transparency.

In summary, this research establishes a foundational step toward explainable, intelligent, and trustworthy cybersecurity analytics. The interpretable deep learning framework not only enhances the robustness of anomaly detection systems but also paves the way for future innovations in real-time adaptive, privacy-preserving, and self-learning network defense architectures. As AI systems become increasingly embedded in digital infrastructures, such human-centered and interpretable designs will be essential for ensuring both operational effectiveness and ethical accountability.

References

[1] Song Wang, Juan Balarezo, Sithamparanathan Kandeepan, Akram Al-Hourani, Karina Gomez, and Ben Rubinstein, "Machine Learning in Network Anomaly Detection: A Survey," IEEE Access, vol. 10, no. 4, pp. 103214–103245, 2023.

[2] Sayantan Roy, "A Comprehensive Survey on Network Traffic Anomaly Detection using Deep Learning," Preprint, June 2024.

[3] Zhong Li, Yuxuan Zhu, and Matthijs van Leeuwen, "A Survey on Explainable Anomaly Detection," arXiv preprint arXiv:2210.06959, Oct. 2022.

[4] K. Singh and S. Kalra, "A Machine Learning Based Reliability Analysis of Negative Bias Temperature Instability (NBTI) Compliant Design for Ultra Large Scale Digital Integrated Circuit," *Journal of Integrated Circuits and Systems*, vol. 18, no. 2, Sept. 2023.

[5] K. Singh and S. Kalra, "Reliability forecasting and Accelerated Lifetime Testing in advanced CMOS technologies," *Journal of Microelectronics Reliability*, vol. 151, Dec. 2023, Art. no. 115261.

[6] R. K. Gupta and S. Kumar, "Explainable Artificial Intelligence for Cybersecurity: A Literature Survey," Journal of Information Security, vol. 12, no. 2, pp. 45–68, 2023.

[7] Khushnaseeb Roshan and Aasim Zafar, "Utilizing XAI Technique to Improve Autoencoder-Based Model for Computer Network Anomaly Detection with SHAP," arXiv preprint arXiv:2112.08442, Dec. 2021.

[8] K. Singh and S. Kalra, "Performance evaluation of Near-Threshold Ultradeep Submicron Digital CMOS Circuits using Approximate Mathematical Drain Current Model," *Journal of Integrated Circuits and Systems*, vol. 19, no. 2, 2024.

[9] K. Singh, S. Kalra, and J. Mahur, "Evaluating NBTI and HCI Effects on Device Reliability for High-Performance Applications in Advanced CMOS Technologies," *Facta Universitatis, Series: Electronics and Energetics*, vol. 37, no. 4, pp. 581–597, 2024.

[10] G. Verma, A. Yadav, S. Sahai, U. Srivastava, S. Maheswari, and K. Singh, "Hardware Implementation of an Eco-friendly Electronic Voting Machine," *Indian Journal of Science and Technology*, vol. 8, no. 17, Aug. 2015.

[11] M. Corea, Y. Liu, J. Wang, S. Niu, and H. Song, "Explainable AI for Comparative Analysis of Intrusion Detection Models," arXiv preprint arXiv:2406.09684, Jun. 2024.

[12] K. Singh and S. Kalra, "VLSI Computer Aided Design Using Machine Learning for Biomedical Applications," in *Opto-VLSI Devices and Circuits for Biomedical and Healthcare Applications*, Taylor & Francis CRC Press, 2023.

[13] K. Singh, S. Kalra, and R. Beniwal, "Quantifying NBTI Recovery and Its Impact on Lifetime Estimations in Advanced Semiconductor Technologies," in *Proc. 2023 9th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2023, pp. 763–768.

TABLE X: Identified Limitations and Future Research Directions

Aspect	Current Limitation	Future Direction
Model Scalability	Limited handling of ultra-high-dimensional, distributed data	Develop hybrid federated XAI-DL architectures capable of decentralized learning
Computational Overhead	High resource consumption during inference	Optimize model compression and pruning techniques to reduce latency
Explainability Precision	Inconsistent explanations across complex layers	Design adaptive XAI frameworks integrating causal inference and attention mechanisms
Real-Time Adaptability	Delayed detection in streaming traffic	Introduce real-time adaptive learning using reinforcement and continual learning
Generalization	Dataset-specific tuning required for accuracy	Apply cross-domain transfer learning to improve generalization on unseen network patterns

[14] K. Singh and S. Kalra, "Analysis of Negative-Bias Temperature Instability Utilizing Machine Learning Support Vector Regression for Robust Nanometer Design," in *Proc. 2022 8th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2022, pp. 571–577.

[15] Quoc Phong Nguyen, Kar Wai Lim, Dinil Mon Divakaran, Kian Hsiang Low, and Mun Choon Chan, "GEE: A Gradient-Based Explainable Variational Autoencoder for Network Anomaly Detection," arXiv preprint arXiv:1903.06661, Mar. 2019.

[16] Osman Tugay Basaran and Falko Dressler, "XAInomaly: Explainable and Interpretable Deep Contractive Autoencoder for O-RAN Traffic Anomaly Detection," arXiv preprint arXiv:2502.09194, Feb. 2025.

[17] K. Singh and S. Kalra, "A Comprehensive Assessment of Current Trends in Negative Bias Temperature Instability (NBTI) Deterioration," in *Proc. 2021 7th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2021, pp. 271–276.

[18] K. Singh and S. Kalra, "Beyond Limits: Machine Learning Driven Reliability Forecasting for Nanoscale ULSI Circuits," in *Proc. 2025 10th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2025, pp. 767–772.

[19] Chukwuemeka Nwachukwu, Kehinde Durodola-Tunde, and Chukwuebuka Akwiwu-Uzoma, "AI-Driven Anomaly Detection in Cloud Computing Environments," International Journal of Science and Research Archive, vol. 13, no. 2, pp. 44–53, 2024.

[20] L. Patel, H. Shah, and J. Bhattacharya, "The Dual Nature of Explainable AI Challenges in Intrusion Detection," Artificial Intelligence Review, vol. 56, no. 10, pp. 12567–12589, Oct. 2023.

[21] K. Singh and S. Kalra, "Reliability-Aware Machine Learning Prediction for Multi-Cycle Long-Term PMOS NBTI Degradation in Robust Nanometer ULSI Digital Circuit Design," in *Proc. 2025 10th International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2025, pp. 876–881.

[22] K. Singh and J. Mahur, "Deep Insights of Negative Bias Temperature Instability (NBTI) Degradation," in *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2025, pp. 1-5.

[23] Siddhi Borse, Sakshi Gayakwad, and Shreya Kulkarni, "Real-Time Network Anomaly Detection using Explainable AI," Journal of Technology, vol. 15, no. 3, pp. 210–219, 2023.

[24] N. Ahmed, S. Mehta, and P. Roy, "Survey on Explainable AI: Approaches, Limitations, and Future Directions," Journal of Artificial Intelligence Research, vol. 18, no. 1, pp. 33–52, 2023.

[25] K. Singh, M. Mishra, S. Srivastava, and P. S. Gaur, "Dynamic Health Response Tracker (DHRT): A Real-Time GPS and AI-Based System for Optimizing Emergency Medical Services," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 1, pp. 11–16, Apr. 2025.

[26] S. Mishra and K. Singh, "Empowering Farmers: Bridging the Knowledge Divide with AI-Driven Real-Time Assistance," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 1, pp. 23–27, Apr. 2025.

[27] J. Lin and A. Sharma, "Machine Learning-Based Network Anomaly Detection," MDPI Electronics, vol. 5, no. 4, pp. 143–158, 2023.

[28] R. Aydin and F. Gunes, "A Survey on XAI-Based Anomaly Detection for IoT," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 32, no. 1, pp. 102–118, 2024.

[29] M. Khan and L. Zhou, "A Survey of AI-Based Anomaly Detection in IoT and Sensor Networks," SSRN Electronic Journal, 2025.

[30] H. Kumar and K. Singh, "Experimental Bring-Up and Device Driver Development for BeagleBone Black: Focusing on Real-Time Clock Subsystems," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 1, pp. 52–59, Apr. 2025.

[31] K. Aryan and K. Singh, "Precision Agriculture Through Plant Disease Detection Using InceptionV3 and AI-Driven Treatment Protocols," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 153–162, May 2025.

[32] S. K. Patel and K. Singh, "AIoT-Enabled Crop Intelligence: Real-Time Soil Sensing and Generative AI for Smart Agriculture," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 163–167, May 2025.

[33] S. Kaushik and K. Singh, "AI-Driven Smart Irrigation and Resource Optimization for Sustainable Precision Agriculture," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 168–177, May 2025.

[34] A. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *Computers & Security*, vol. 46, pp. 1–14, 2014.

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

[36] R. E. H. Khan and K. Singh, "AI-Driven Personalized Skincare: Enhancing Skin Analysis and Product Recommendation Systems," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 178–184, May 2025.

[37] A. Khan, T. Raza, G. Sharma, and K. Singh, "Air Quality Forecasting Using Supervised Machine Learning Techniques: A Predictive Modeling Approach," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 185–191, May 2025.

[38] Z. Zamanzadeh Darban et al., "Deep learning for time series anomaly detection: A survey," *ACM Computing Surveys*, 2024.

[39] J. Chen, X. Li and Y. Ma, "Deep autoencoders for network anomaly detection," *IEEE Communications Letters*, vol. 23, no. 12, pp. 2083–2086, 2019.

[40] K. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, 2009.

[41] A. Khan and K. Singh, "Forecasting Urban Air Quality: A Comparative Study of ML Models for PM2.5 and AQI in Smart Cities," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 192–199, May 2025.

[42] T. Raza and K. Singh, "AI-Driven Multisource Data Fusion for Real-Time Urban Air Quality Forecasting and Health Risk Assessment," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 200–206, May 2025.

[43] Q. P. Nguyen et al., "GEE: A gradient-based explainable variational autoencoder for network anomaly detection," arXiv:1903.06661, 2019.

[44] T. Ji, "Variational autoencoder based anomaly detection in networked systems," *Energies*, vol. 18, no. 11, 2025.

[45] Y Yadav, S Rawat, Y Kumar and S Tripathi, " Lightweight Deep Learning Architectures for Real-Time Object Detection in Autonomous Systems," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 123-128, May 2025.

[46] G. Sharma and K. Singh, "Impact of Deteriorating Air Quality on Human Life Expectancy: A Comparative Study Between Urban and Rural Regions," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 207–215, May 2025.

[47] Z. Lin et al., "Sequence modeling approaches for intrusion detection," *IEEE Transactions on Network and Service Management*, 2021.

[48] Mala-Lab, “Deep graph anomaly detection: A survey and new perspectives,” GitHub repository accompanying TKDE survey, 2025.

[49] A. Yadav, R. E. H. Khan, and K. Singh, “YOLO-Based Detection of Skin Anomalies with AI Recommendation Engine for Personalized Skincare,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 216–221, May 2025.

[50] K. Aryan, S. Mishra, S. K. Patel, S. Kaushik, and K. Singh, “AI-Powered Integrated Platform for Farmer Support: Real-Time Disease Diagnosis, Precision Irrigation Advisory, and Expert Consultation Services,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 222–229, May 2025.

[51] A. Rosay et al., “A comprehensive analysis of CIC-IDS2017,” *SCITEPRESS*, 2022.

[52] Canadian Institute for Cybersecurity, “CIC-IDS2017 dataset,” 2017. (Dataset resource)

[53] M. Tavallaee et al., “A detailed analysis of KDD CUP 99 dataset,” *IEEE Symposium*, 2009.

[54] A. Yadav and K. Singh, “Smart Dermatology: Revolutionizing Skincare with AI-Driven CNN-Based Detection and Product Recommendation System,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 2, pp. 230–235, May 2025.

[55] K. Singh and P. Singh, “A State-of-the-Art Perspective on Brain Tumor Detection Using Deep Learning in Medical Imaging,” *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 3, pp. 250–254, Jun. 2025.

[56] A. Rosay, “LYCOS-IDS2017: Corrected dataset proposals for CIC-IDS2017,” *SCITEPRESS*, 2022.

[57] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions (SHAP),” *NIPS*, 2017.

[58] M. Ribeiro, S. Singh and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier (LIME),” *KDD*, 2016.

[59] K. Singh, "Exploring Artificial Intelligence: A Deep Review of Foundational Theories, Applications, and Future Trends," *Journal of Scientific Innovation and Advanced Research (JSIAR)*, vol. 1, no. 6, pp. 295–305, Sep. 2025.

[60] M. Corea et al., “Explainable AI for comparative analysis of intrusion detection models,” arXiv:2406.09684, 2024.

[61] K. Roshan and A. Zafar, “Utilizing XAI technique to improve autoencoder-based model for network anomaly detection with SHAP,” arXiv:2112.08442, 2021.

[62] A. Muhammad et al., “L-XAIDS: A LIME-based explainable AI framework for intrusion detection,” *Future Computing and Informatics Journal*, 2025.

[63] L. Patel, H. Shah and J. Bhattacharya, “The dual nature of explainable AI challenges in intrusion detection,” *Artificial Intelligence Review*, 2023.

[64] T. Ali, “Next-Generation IDS with LLMs: Real-Time Anomaly Detection, Explainable AI, and Adaptive Data Generation,” Master’s thesis, University of Oulu, 2024.