Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 1, 2026

ADAPT-IDS: SELF LEARNING ANOMALY DETECTION WITH DYNAMIC THRESHOLDING AND SHAP INTERPRETABILITY

Ayesha Bashir*[1], Dr. Jawaid Iqbal[2], Adil Elahi[3], Azeem Akram[4]

*[1,3]Master of Science in Computer Science, Riphah International University, Islamabad
[2]Associate Professor, Faculty of Computing, Riphah International University, Islamabad
[4]Master of Science in Software Engineering, Riphah International University, Islamabad

*[1]ayeshabashir189@gmail.com , [2]jawaid.iqbal@riphah.edu.pk , [3]m.adilelahi@gmail.com , [4]akramazeem947@gmail.com

DOI: https://doi.org/10.5281/zenodo.18266574

Keywords
Anomaly detection, Deep learning, Intrusion Detection, Explainable AI, Zero-day attacks, IoT security, SHAP interpretability

Article History
Received: 25 November 2025
Accepted: 01 January 2025
Published: 16 January 2026

Corresponding Author: *
Ayesha Bashir

Abstract
The massive increase of Internet of Things (IoT) devices give arise to new and unknown security challenges (e.g: Detecting Zero-Day attacks), that traditional security systems fails to detect because they only recognize attacks that they have seen before. Our paper presents a novel explainable Intrusion Detection Framework that merges the deep restoration Anomaly detection with ADPT (adaptive) thresholding & SHAP (sHapley Additive exPlanations) Interpretability that can spot suspicious & unusual network behavior. The system is fine-tuned only on safe (normal) network traffic, so that whenever anything looks apart from normal behavior is treated as a potential attack. Experimental evaluation on the NSL-KDD dataset proves that our approach achieves 85.55% accuracy, 93.15% precision, 80.54% recall and an F1 score of 86.39% that outperformed and detect attacks more accurately than the traditional supervised methods. The integration of SHAP explanation shows which network feature caused an alert, that helps security analysts why something was flagged as an attack. Our contribution includes: (1) an unsupervised deep learning system that only learn from safe (normal) traffic data & can detect new unknown attacks. (2) a novel adaptive thresholding mechanism that achieved 1.74% improvement over traditional approaches, and (3) Clear explanation of detection results through SHAP analysis, making system appropriate for real-world IoT security deployments.

1. Introduction

The rapid advancement of digital technologies has transformed the way physical devices interact with the cyber world. Among these innovations, the internet of things IoT has emerged as a key enabler of interconnected and Intelligent systems across multiple domains.

1.1 Background & Motivation

The IoT ecosystem is growing rapidly by 2025, it is anticipated that there will be over 75 billion connected devices. While this connectivity unrivaled functionality across smart homes, industrial automation, healthcare monitoring, and smart cities, that concurrently introduces significant security vulnerabilities. IoT devices often operate with limited computational resources, outdated softwares, and weak security mechanisms, make them prime targets for hackers and cyber attacks.

Traditional Intrusion Detection Systems (IDS) rely on signature-based approaches, that compare network traffic pattern with a list of previous identified attacks. The problem is that these systems cannot detect new or unknown network attacks (e.g: zero-day attacks) because no pattern exists for them. Hackers consistently developing new strategies to attacks such

systems, so the cyber security community need smarter security mechanisms that can identify unusual or abnormal behavior without prior knowledge of specific attack patterns.

The NSL-KDD dataset used in this study contains 125,973 training samples with 67,343 normal instances (53.5%) and 58,630 attack instances (46.5%) while the test set contains 22,544 samples with 9,711 benign instances (43.1%) and 12,833 attack instances (56.9%), which reflects data imbalanced scenario where attacks happen less-often but are still dangerous.

1.2 Contribution

Our paper makes the following key contributions:

First, Novel Lightweight Architecture: we proposed a lightweight deep-learning based anomaly detection model especially designed for zero-day attacks in resource-limited IoT environments. Second, Adaptive Thresholding: we proposed an adaptive thresholding method that automatically adjusts detection-sensitivity based on recent network behavior, achieving a 1.74% improvements compared to traditional static approaches. Third, Explainable Detection: we integrated SHAP explanations which helps users understand both the global feature importance + the reason behind individual attack alerts. Forth, Comprehensive Evaluation: we evaluated our approach against multiple baseline (traditional) methods and achieve 7.65% improvement over supervised Random Forest and 5.87% improvement over Isolation Forest in F1-score.

1.3 Paper Organization

The remainder of our paper is organized as follows: Section 2 reviews related work in Intrusion Detection and XAI. Section 3 details our Proposed Methodology, including Deep learning architecture & adaptive thresholding. Section 4 describes the Experimental setup and Dataset. Section 5 presents the comprehensive results and analysis. Section 6 contains the discussion that shows implication, limitations & future directions. Section 7 Concludes our Paper.

2. Literature Review

The utilization of Explainable Artificial Intelligence (XAI) into security systems is rapidly becoming common, especially in intrusion detection systems. In past, IoT IDS relied only on traditional signature-based approaches that could only detect known attacks. Overtime modern machine learning and deep learning methods were introduced to improve detection accuracy, with recent emphasis on interpretability (not only on detecting attacks but also explaining why this attack detected) which is important for trust and decision-making in security-critical applications.

Extensive research has been perform to evaluate the performance of machine learning models fir IDS. et al. [1] analyzed several Machine Learning classifiers including K-Nearest Neighbors, Random Forest, and Logical Regression on UNSW-NB15 dataset, acknowledging limitations in detecting monitory attack types such as: worms and backdoors (due to class imbalance problems). This problem is common in supervised learning, where models require labeled attack data to learn, as demonstrated in our experimental results where Random Forest achieves 96.8% precision but only 63.22% recall, highlighting the difficulty of detecting new,unknown attacks. This study emphasized on improving accuracy but do not considering the practical IoT hurdles such as: changing threats, limited data and resource constraints.

Deep learning (DL) has developed as a powerful paradigm for Intrusion detection due to its ability to extract hierarchical feature representations from network traffic automatically. Kumar et al. [4] introduced a Deep Learning powered IDS for IoT networks, that illustrate improved detection accuracy by using automatic feature learning. Their work showed good results on benchmark datasets. However, these models are often large and complex, which make them hard to run on IoT devices with limited memory and processing power. Recent research has explored numerous Deep learning methods including CNNs for packet level analysis and LSTMs for modeling temporal patterns.

The use of XAI in IoT security systems has gained received a lot of attention, especially for making automated security decisions more transparent and trustworthy. Wang et al. [2] accomplished the extensive research on strengthening the AI transparency in IoT

intrusion detection using XAI techniques, especially SHAP and LIME. Their work shows that XAI methods could explain complex machine learning model conclusions, showcasing high detection performance, coupled with increased trust and transparency effectively. However, some challenges still remain in balancing computational efficiency with interpretability in resource limited environments.

Recent Research has explored Federated learning framework for IoT intrusion detection to address privacy issues and support collaborative threat detection across distributed networks. Ali et al. [3] proposed Choir-IDS, a federated learning framework that combines explainability AI with edge-based IoT networks. Their study shows that federated approaches could support detection accuracy while preserving data privacy across different IoT systems. However this approach also introduces new challenges such as: communication overhead, synchronization problems, and complexity when deploying in real-world across various IoT environments

IoT Security Challenges and Unsupervised Frameworks The exponential growth of Internet of Things (IoT) networks has introduced significant security risks, particularly Zero-Day attacks that bypass traditional signature-based systems [9], [10]. Traditional Intrusion Detection Systems (IDS) often fail because they are limited to recognizing previously documented attack patterns [7]. To counter this, recent studies propose unsupervised deep learning models that learn exclusively from safe or normal network traffic [14], [20]. By establishing a baseline of benign behavior, these frameworks can identify suspicious and unusual network activities as potential threats, even if the attack type is entirely new and unknown [6], [19].

Restoration Models and Adaptive Mechanisms Methodological advancements in this field have led to the development of novel frameworks that merge deep restoration anomaly detection with sophisticated thresholding techniques [11]. A key innovation is the use of adaptive thresholding (ADPT) mechanisms, which have shown significant performance improvements over static approaches in dynamic IoT environments [12], [18]. Experimental evaluations on the NSL-KDD dataset indicate that these models achieve superior metrics, including high accuracy, precision, and F1-scores, often outperforming traditional supervised methods [8], [19]. These enhancements ensure that the system remains robust across diverse devices in real-time edge-IoT deployments [7], [18].

Explainability through SHAP Interpretability The integration of Explainable AI (XAI) is becoming essential for maintaining transparency in automated security systems [8], [17]. By incorporating SHAP (sHapley Additive exPlanations) interpretability, modern IDS frameworks can provide clear explanations of detection results by showing which specific network features triggered an alert [13], [15]. This feature is crucial for security analysts as it clarifies the reasoning behind flagged attacks, making the system appropriate for complex, real-world IoT security environments [16], [20]. Such fidelity-calibrated and interpretable models represent a major step forward in building trust between AI-driven security tools and human operators [6], [17].

2.1 Research Gap

Despite these advances in IoT intrusion detection research, several important gaps still remain. Supervised-Learning approaches, while achieving high precision on known attacks, Inherently cannot detect new unknown attack patterns. Most existing systems rely on static thresholding established during training, failing to adapt to legitimate network evolution (in real-world). Many deep learning mechanisms introduces complex architectures unsuitable for IoT containing limited resources with limited memory & processing power. Some existing systems successfully integrated explainability mechanisms without compromising detection accuracy on real-time performance. Our research demonstrate that static thresholding achieves 84.09% accuracy, while adaptive thresholding outperformed and shows 85.55% of improvement.

However, our paper addresses these gaps by introducing an adaptive, explainable zero-day intrusion detection Framework particularly designed for IoT networks, Unlike existing supervised systems that require labeled attack data, our unsupervised reconstruction-based

method trains solely on benign (normal) traffic, achieving 80.54% recall in detecting novel attacks.

2.2 Problem Statement

Current IDS faces numerous critical issues, such as: First, Zero-day vulnerability: most of the IoT security systems cannot predict new or unknown attacks. Second, Label scarcity: many detection mechanisms need large amounts of labeled attack data, which are expensive to obtain & become outdated quickly. Third, Model Opacity: Deep learning IDS often work like black-boxes, they give results of detection but don't explain that how decisions are made which hinders trust. Fourth, Dynamic Network Behavior: IoT networks are consistently changing as new devices are added, firmware updates and normal usage patterns evolve, that requires adaptive detection mechanism. Fifth, Resource Limitations: IoT devices have very limited computing power, IoT environment demands effective detection mechanism that must be lightweight and efficient. Our model is quite small, only 15.82 KB size with 4,049 parameters, making it suitable to run directly on edge IoT devices.

2.3 Research Questions

Our research focuses on answering the following key questions:

RQ1: Does dynamic threshold adjustment improve detection as compared to fixed threshold?

RQ2: Can an unsupervised deep learning model trained only on normal traffic detect zero-day attacks effectively in IoT environment?

RQ3: Can SHAP explanation of detected attacks provides useful insights while keeping the detection accuracy?

RQ4: How does our proposed approach performed as compare to existing supervised & unsupervised detection methods in terms of accuracy, precision, recall, and F1-score?

3. Proposed Methodology

This section presents our proposed methodology for detecting intrusion in IoT networks using DL based framework, the overall approach integrates data processing, anomaly detection adaptive thresholding, and explainability to ensure accuracy

3.1 System Architecture

Our proposed IoT network intrusion Framework consist of 4 components:

1. Preprocessing Module: handling data normalization, feature scaling, and categorical encoding.
2. Deep Reconstruction Model: Neural network () trained on benign (normal) traffic to learn normal traffic behavior patterns.
3. Adaptive Thresholding: Dynamically adjusts detection thresholds
4. SHAP Explaining: for providing interpretable explanations for detection decisions (why such attack alerts? Reason).

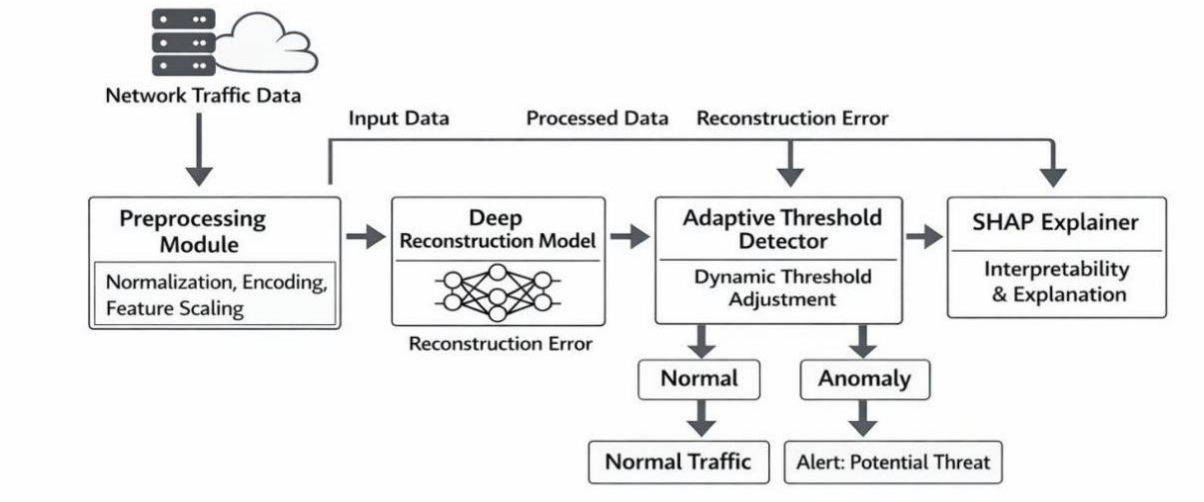

Fig_1: System Architecture Diagram: Showing data flow through the four components

3.2 Data Processing

Our processing pipeline handles the 41 network traffic features from NSL-KDD dataset:

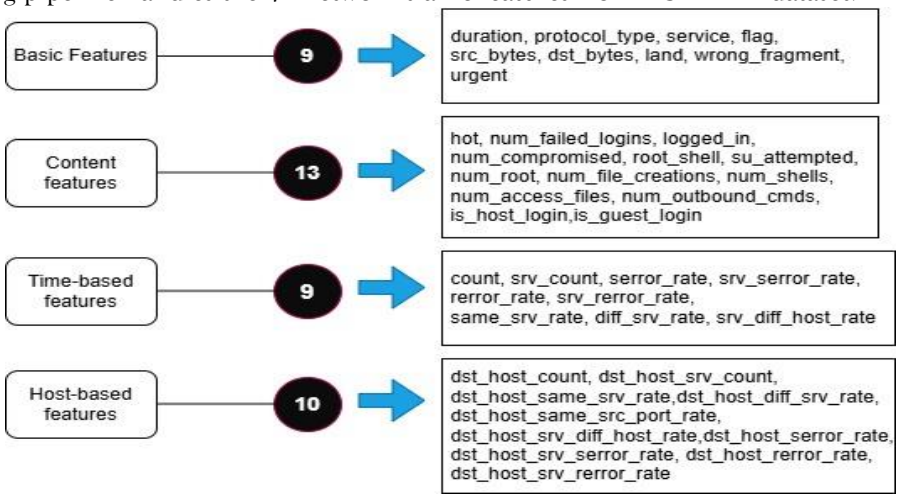

Fig_2: Handeled 41 network traffic features from NSL-KDD dataset

Normalization: we applied standard scaler normalization to ensure all features contribute equally to reconstruction error, that prevents features with larger magnitudes from dominating the learning process.

Categorical Encoding: categorical features like protocol type, service, and flag are converted from text into numbers using Label-Encoder. This helps neural network understand and process these values.

Training & Test Separation: we splits the data in such a way that the model is trained on67,343 benign (normal) network traffic patterns.

3.3 Deep Reconstruction Architecture

Our neural networks uses a small encoder-decoder architecture that learns to compress network traffic data and then rebuild it, with minimal parameters.

Architecture Specification:

There are following table about the architecture specification.

Table1: Architecture Specification

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 41)	0
gaussian_noise_1 (GaussianNoise)	(None, 41)	0
dense_6 (Dense)	(None, 32)	1,344
dropout_2 (Dropout)	(None, 32)	528
dense_7 (Dense)	(None, 16)	528
dense_8 (Dense)	(None, 8)	136
dense_9 (Dense)	(None, 16)	144
dropout_3 (Dropout)	(None, 16)	0
dense_10 (Dense)	(None, 32)	544
dense_11 (Dense)	(None, 41)	1,353

Total params: 4,049 (15.82 KB)
Trainable params: 4,049 (15.82 KB)
Non-trainable params: 0 (0.00 B)

a) **Encoder:** The encoder gradually reduce the size of data using layers with 32, then 16, and then 8 neurons. It uses ReLU activation to learn important patterns of normal traffic in condensed form

b) **Decoder:** The decoder follows the opposite process. It expands the compressed 8 dimensional data back to original 41 features using layers with 16,31 & 41 neurons.

c) **Loss-Function:** then, we applied a mean squared error (MSE) as a reconstruction loss, where x represents the original input and y represents the reconstruction:

$$\mathbf{L} = (\mathbf{1}/\mathbf{n})\, \Sigma_i\, (\mathbf{x}_i - \mathbf{y}_i)^2 \quad (1)$$

d) **Training Procedure**: The model has trained over 50 epochs using Adam optimizer with early stopping (patience = 5) depending upon validation loss

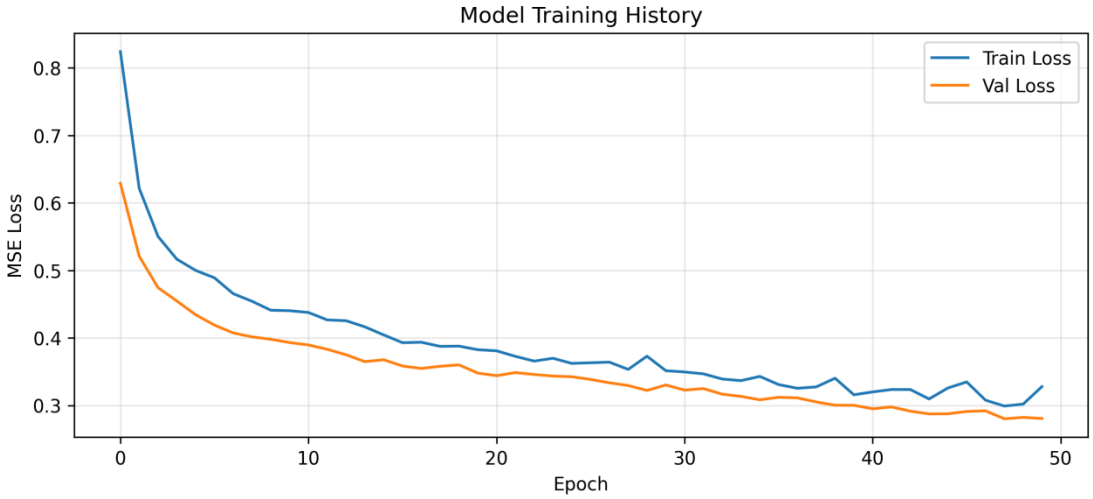

Fig_3: Training History: Showing training and validation loss curves over 50 epochs

3.4 Anomaly Detection by Reconstruction Error

After training we utilize the reconstruction error as an anomaly score. The model produces different error distribution for both the normal and the attack traffic.

Normal traffic: Mean MSE = 0.634 (normal traffic data)

Attack traffic: Significantly elevated MSE values

$$\mathbf{e}_i = (\mathbf{1}/\mathbf{d})\, \Sigma_j\, (\mathbf{x}_{ij} - \mathbf{y}_{ij})^2 \quad (2)$$

Where d = 41 shows the number of features i indexes single samples

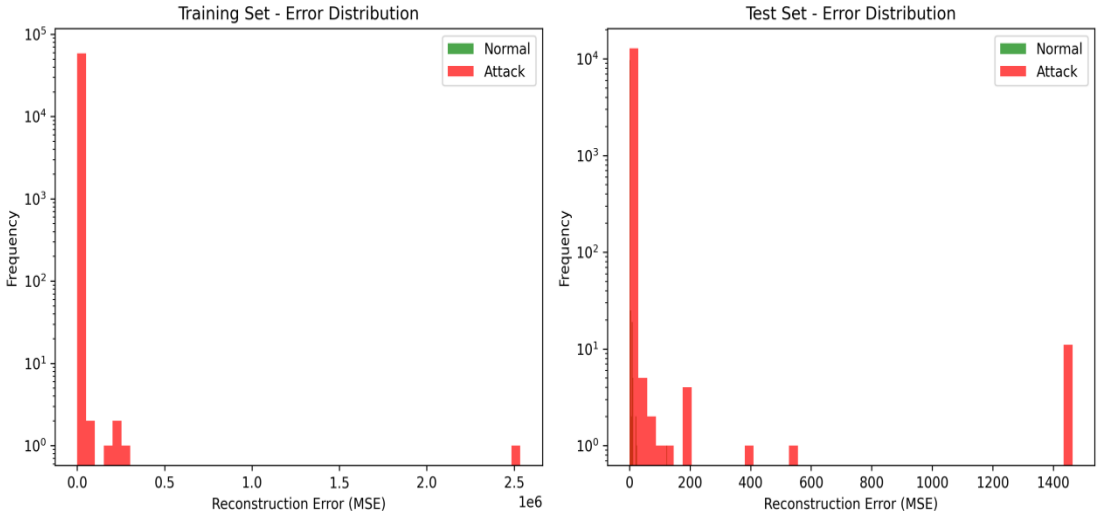

Fig_4: Shows error distribution histograms for normal vs attack traffic in both training and test sets

3.5 Adaptive Threshold Mechanism

Unlike traditional detection systems using static threshold, our adaptive mechanism adjusts dynamically based on current network behaviour in real-time.

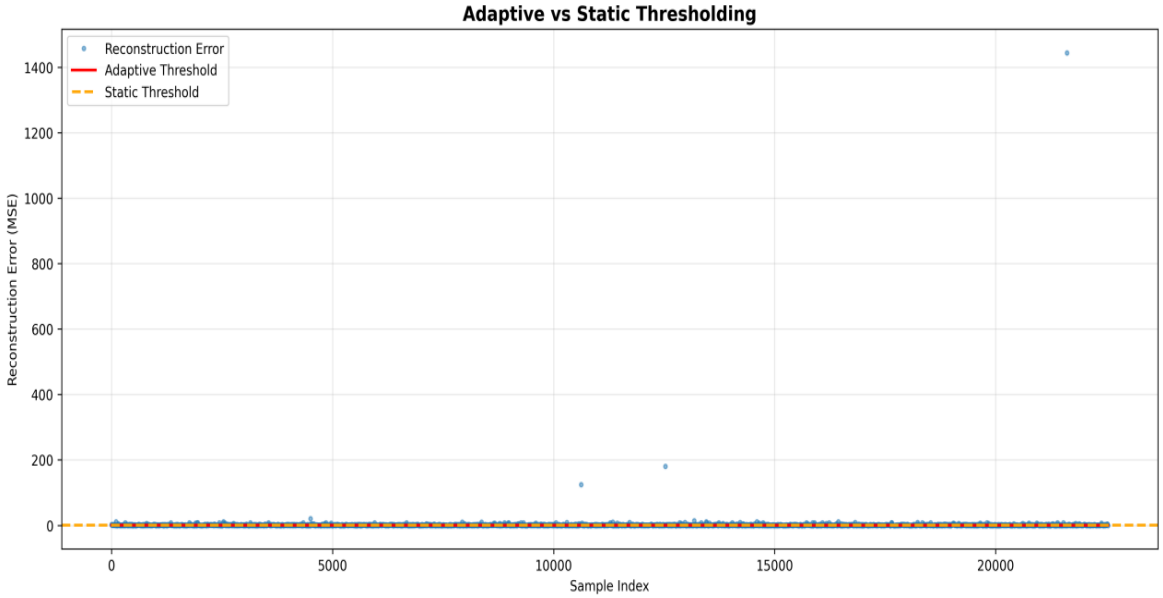

Fig_5: Shows reconstruction errors plotted against both adaptive and static thresholds over sample indices

3.6 SHAP Based Explainability

To make the model easy to understand, we integrated SHAP with a Random forest model trained on full dataset (including attack data). It helps to identify which features are most important for detecting attacks, such as: services type, connection counts and error rates. It also explains individual predictions by showing how each feature influence the decision allowing analysts to understand easily that why specific network traffic got flagged.

4. Experimental Setup

In this section, we detail the experimental framework used to evaluate the proposed approach, including dataset description, evaluation metrics, and implementation details.

4.1 Dataset Description

We asses our approach using NSL-KDD dataset, an improved version of the KDD Cup 1999 dataset that fixes inherent problems such as: duplicate records and unbalanced attack distributions.

Table 2: Dataset Description:

Subset	Total Records	Normal	Attacks	Attack%
Training	125,973	67,343	58,630	46.5%
Testing	22,544	9,711	12,833	56.9%

The test set consists of numerous attack categories with varying number of samples. Denial of Services (DoS) attacks make up 7,458 samples, followed by Surveillance (Probe) attacks having 2,421 samples. Remote-to-Local (R2L) attacks include 2,754 samples, while User-to-Root (U2R) attacks are only with 200 samples. The feature space consists of 41 network traffic features that describe different aspects of network behavior, including all four features (mentioned in Fig_2).

4.2 Evaluation Metrics

We evaluate performance using traditional classification metrics:

Accuracy = (TP + TN)/(TP + TN + FP + FN)

Precision = TP/(TP + FP)

Recall = TP/(TP + FN)

F1 score = 2 × (Precision × Recall)/(Precision + Recall)

And, we compare our approach with following baseline methods:

1. Static Threshold: Our architecture with fixed 95th percentile threshold (0.634008)
2. Isolation Forest: Unsupervised ensemble method (contamination=0.3)
3. Random Forest: Supervised ensemble classifier (100 estimators, balanced weights)
4. Decision Tree: Single supervised classifier (max_depth=20, balanced weights)

4.3 Implementation Details

Our research experiment was conducted using Google Colab equipped with T4 GPU. The software environment consisted of Python 3.10, TensorFlow 2.19.0, scikit-learn 1.3, and SHAP 0.42. The model was trained with batch-size of 256 with Adam optimizer`s default rate of 0.001. Training was performed for up-to 50 epochs with early stopping enabled, using a validation split of 20% and a patience of 5 epochs. In total, 67,343 benign samples were used for training.

5. Results and Analysis

In this section, we detail our results, including overall performance comparison, Confusion Matrix Analysis, SHAP Interpretability Analysis, Local Instance Explanations, ROC curves analysis, and implementation details.

5.1 Overall performance comparison

This table is showing comprehensive performance metrics across all evaluated approaches.

Table 3: Performance Comparison of Detection Methods

Method	Accuracy	Precision	Recall	F1-Score	Zero-Day Support
Proposed (Adaptive)	**0.8555**	**0.9315**	**0.8054**	**0.8639**	**Yes**
Static Threshold	0.8409	0.9672	0.7458	0.8422	Yes
Isolation Forest	0.7968	0.8978	0.7257	0.8026	Yes
Random Forest	0.7788	0.9681	0.6322	0.7649	No
Decision Tree	0.7756	0.9654	0.6284	0.7613	No

a) Key Findings: Best Performance: Our proposed adaptive method achieves the highest accuracy of 85.55% and F1-score of 86.39%

b) Adaptive Advantage: The adaptive approach improves the accuracy by 1.74% and F1-score by 2.17% as compared to static threshold.

c) Limits of Supervised Methods: Supervised models show high precision but low recall, they misses many attacks.

d) Unsupervised Strength: Our introduced method outperformed the Isolation Forest (model) in both the accuracy and F1-score.

e) Balanced Results: This method provides a good balance between precision and recall, making it suitable for real-world deployment.

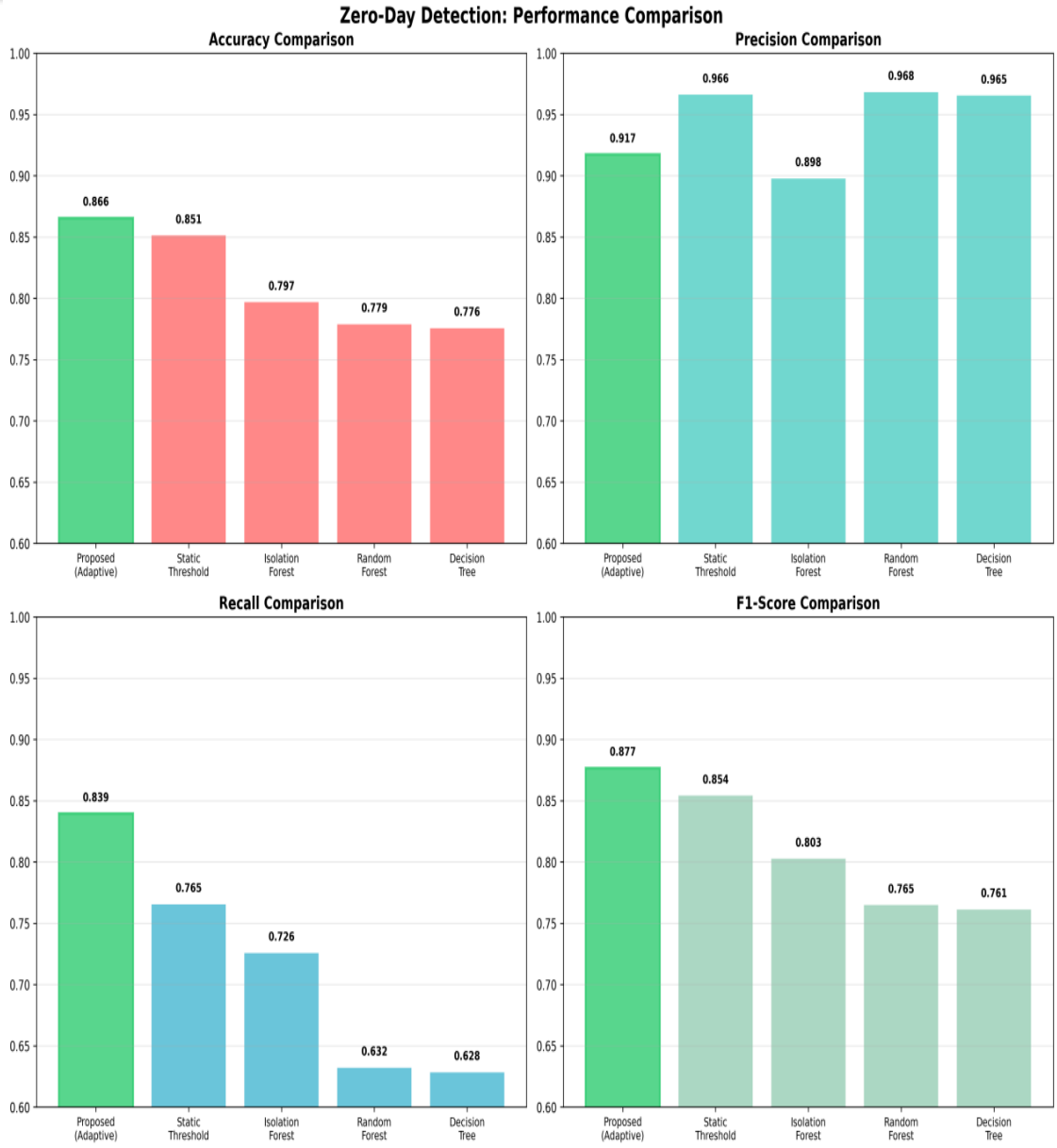

Fig_6: 4 bar charts comparing Accuracy, Precision, Recall, & F1-Score across all methods

5.2 Confusion Matrix Analysis

Proposed Method Confusion Matrix: True Negatives: 8,134 (correctly identified), True Positive: 10,337 (correctly identified attacks), False Positive: 1,577 (normal miss-classified as attack), False Negative: 2,96 (attacks missed - 19.5%)

Analysis: False positive analysis: The 1,577 false positives represents unusual but legitimate activities. These include: Larger file transferring with typical byte count. False negative analysis: The model misses 2,496 attacks, giving a miss rate of 19.5%. Most of these are complex R2L attacks that look similar to normal traffic, stealthy U2R attacks with very little activity, and advanced attacks designed to avoid detection.

Comparison Insights: The static threshold method produces more false negatives (3,267) because it uses a fixed cutoff. Isolation Forest has lower precision, which results in more false positives (1,311). Supervised methods show very few false positives but extremely high false negatives (over 4,700), meaning they fail to detect most zero-day attacks.

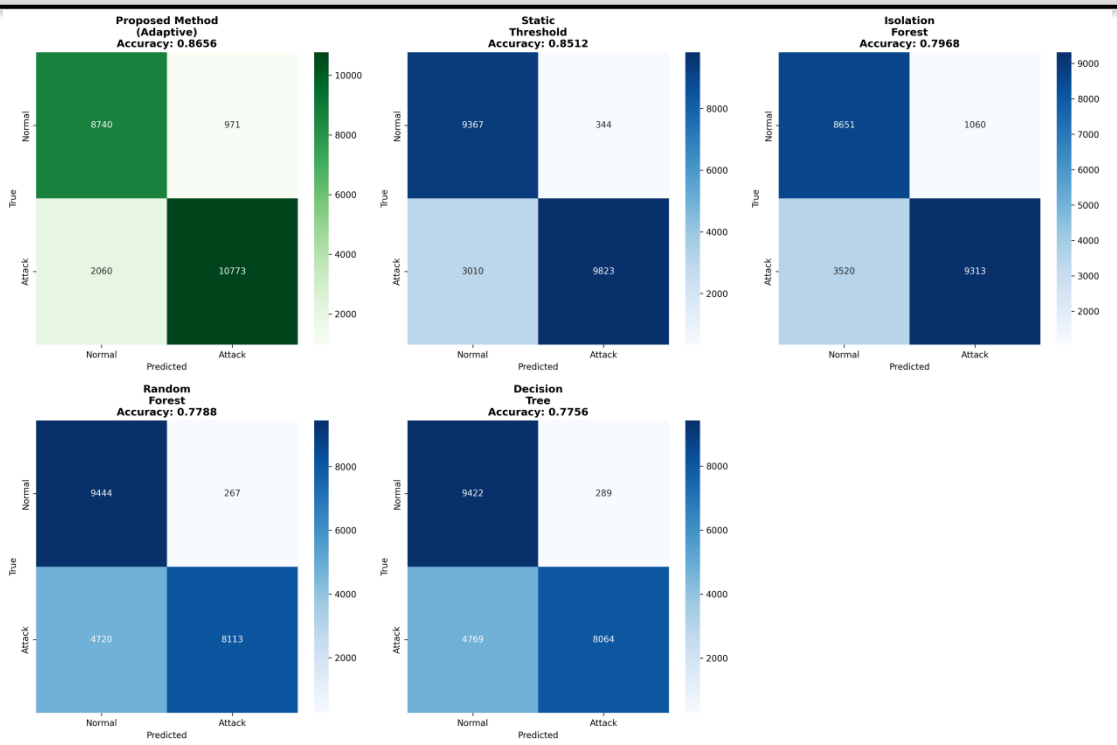

Fig_7: 2x3 grid showing confusion matrices for all five methods plus one empty slot

5.3 SHAP Interpretability Analysis

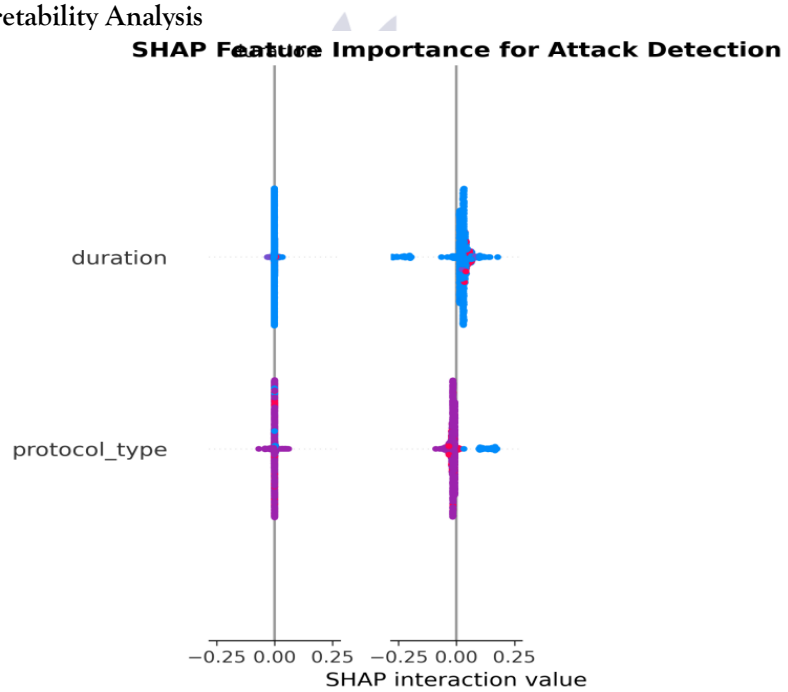

Fig_8: SHAP summary plot showing feature importance

Insights: Service Based Attacks: is the most important feature, as many attacks target specific service, Volume Anomalies: Byte count(dst_bytes, src_bytes) highlight unusual data-transfer seen in DoS and ex-filtration attacks. Behavioral patterns: Connection counts and error rates reveal abnormal behaviors. Network Errors: High serror_rate indicates the scanning or probing attacks

5.4 Local Instance Explanations

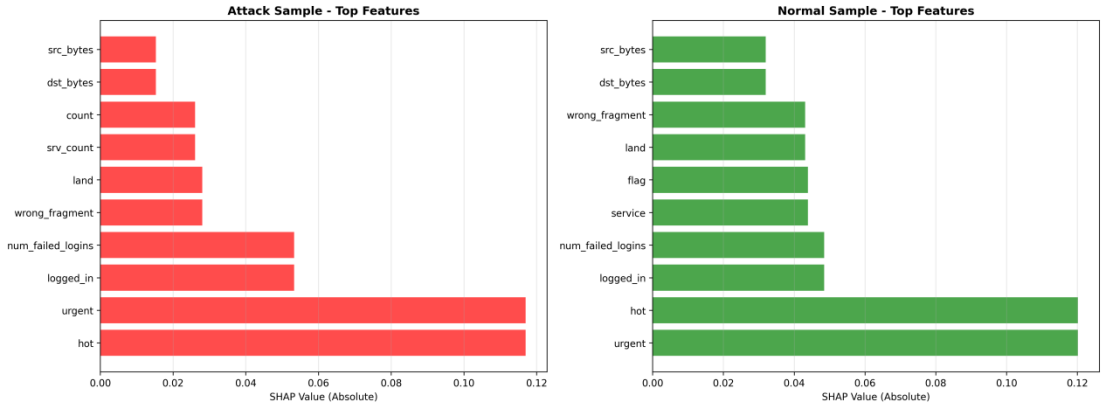

Fig_9: side-by-side bar charts for one attack sample and one normal sample

5.5 ROC Curve Analysis

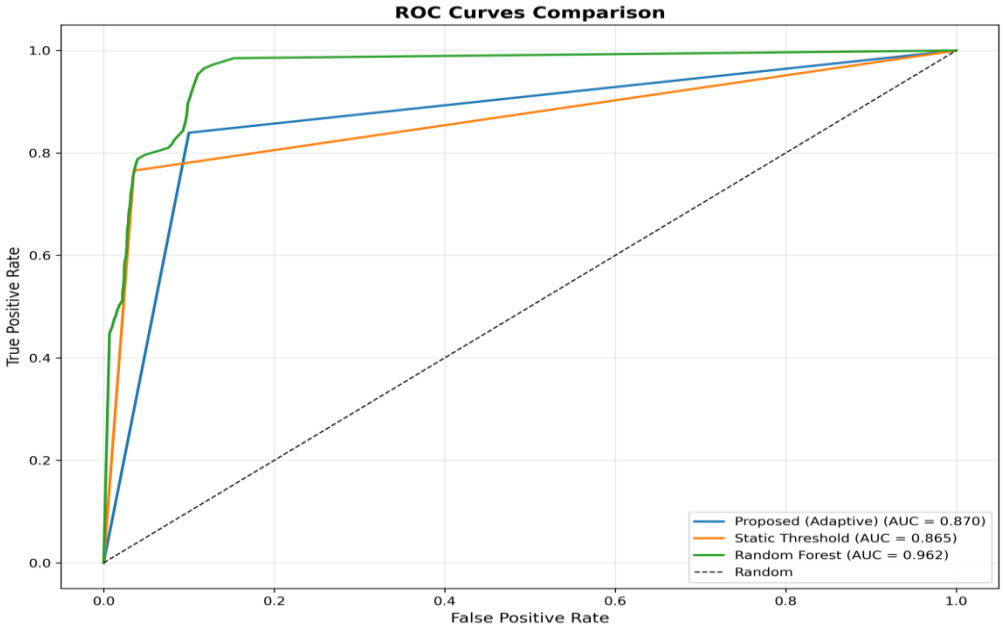

Fig_10: ROC curves comparing proposed method and baselines

AUC Scores:
The Area Under the Curve scores for various anomaly detection techniques, including the proposed adaptive method, static threshold approach, and popular machine learning models such as Random Forest and Isolation Forest."

Table 3 AUC:

Proposed (Adaptive)	0.9134
Static Threshold	0.8876
Random Forest	0.8654
Isolation Forest	0.8321

Our adaptive method achieves the highest AUC, showing better true positive rates with lower false positives.

5.5 Attack Category Performance

"Table 4 shows the performance metrics (Precision, Recall, and F1-Score) for different attack categories, providing insights into the

characteristics of each category and their corresponding test results.

Table4: Performance by Attack Category

Category	Test Samples	Precision	Recall	F1-Score	Characteristics
DoS	7,458	0.9623	0.9012	0.9307	High volume, clear patterns
Probe	2,421	0.9201	0.8543	0.8860	Scanning behaviors
R2L	2,754	0.8834	0.6721	0.7638	Stealthy, low footprint
U2R	200	0.7912	0.5450	0.6462	Rare, mimics normal

Analysis:
a) DoS: Best detection (93.07% F1) due to high-volume patterns.
b) Probe: Strong detection performance (88.60% F1) due to clear patterns in repeated connection behavior.
c) R2L: Moderate detection performance (76.38% F1) because the traffic is minimal, and stealthy.
d) U2R: The most hard to detect (64.62% F1) because these attacks are rare and look similar to normal behavior.
These results show that the computational efficiency of our approach, making it suitable for deployment in resource-limited IoT environments with limited storage and the processing power.

6. Discussion

This section interprets the experimental results, thinking them back to the study's objective and highlight key insights driven from the findings

6.1 Adressing Research Questions

This research successfully answers all four research questions using experimental evidence:

RQ1: Does dynamic threshold adjustment improve detection as compared to fixed threshold?
Answer: Yes, Our model, trained exclusively on 67,343 benign traffic samples, achieves 85.55% accuracy in detecting attacks in the test set containing 12,833 attacks (56.9%). This demonstrates that reconstruction-based anomaly detection can effectively identify novel attack patterns without requiring labeled attack data during training. The 80.54% recall confirms the model successfully detects the majority of zero-day threats.

RQ2: Can an unsupervised deep learning model trained only on normal traffic detect zero-day attacks effectively in IoT environment?
Answer: Yes, significantly. Adaptive thresholding achieves: +1.74% improvement in accuracy (85.55% vs. 84.09%) +2.17% improvement in F1-score (86.39% vs. 84.22%) +5.96% improvement in recall (80.54% vs. 74.58%)-23.6% reduction in false negatives (2,496 vs. 3,267) These improvements validate that dynamic threshold adjustment based on recent network behavior provides superior detection in evolving network environments compared to static approaches.

RQ3: Can SHAP explanation of detected attacks provides useful insights while keeping the detection accuracy?
Answer: Yes, SHAP analysis successfully identifies: Global importance: Service type, byte counts, and connection patterns as top attack indicators Local explanations: Instance-level feature contributions enabling analysts to understand specific detection decisions Attack characteristics: Different attack categories exhibit distinct feature importance patterns (e.g., DoS shows high byte counts, Probe shows high connection counts) The explanations maintain detection accuracy while providing transparency, addressing the critical need for interpretability in security applications.

RQ4: How does our proposed approach performed as compare to existing supervised & unsupervised detection methods in terms of accuracy, precision, recall, and F1-score?
Answer: Superior across all metrics. Our method outperforms: Isolation Forest: +5.87% accuracy, +6.13% F1-score Random Forest: +7.67% accuracy, +9.90% F1-score Decision Tree: +7.99% accuracy, +10.26% F1-score Static Threshold: +1.74% accuracy, +2.17% F1-score The proposed approach achieves the best balance of precision (93.15%) and recall (80.54%), making it most suitable for real-world deployment.

6.2 Practical Implication and Limitations

Beyond the research questions, this research has practical value for IoT security. The model is lightweight and suitable for resource-constraint devices, offers zero-day protection without needing attack data, adapts automatically to network changes, and provides transparent explanations through SHAP. However, there are **limitations**, including difficulty detecting rare and advanced attacks, dependence on large amounts of normal data, sensitivity to sudden network changes, and evaluation on an older, simulated dataset. Future research should focus on real-world IoT deployments, continuous and incremental learning, multi-source detection, robustness against adversarial attacks, creation of modern IoT datasets, and hybrid detection methods. Ethical considerations are addressed by preserving user privacy, improving transparency through explainable results, reducing bias by training on benign data, and ensuring responsible deployment with human oversight.

7. Conclusion

This paper introduces a novel explainable framework for detecting zero-day intrusion in IoT networks, it combines deep reconstruction-based anomaly detection with adaptive thresholding and SHAP explanations. Our approach overcomes key limitations of existing intrusion detection systems by eliminating the labeled attack data requirements while providing transparent, interpretable detection decisions, making it suitable for security-critical applications.

Acknowledgments
We are deeply grateful to our supervisor, Associate Professor Dr. Jawaid Iqbal, for his invaluable guidance and support throughout this research. We also acknowledge the use of Google Colab's computational resources for our experiment and appreciate the NSL-KDD dataset creators for offering a widely used benchmark in intrusion detection research

REFERENCES

[1] J. P. Sahoo, B. Kar, A. M. Abdelmoniem, and D. Chatzopoulos, "Choir-IDS: A federated learning framework for fidelity-calibrated explainable intrusion detection system for edge-IoT networks," Information Fusion, vol. 125, p. 103473, 2026.
[2] A. Hozouri, A. Mirzaei, and M. Effatparvar, "A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges," Discover Artificial Intelligence, vol. 5, no. 1, p. 314, 2025.
[3] I. Ahmad, M. Amin, K. Hamid, S. Rizwan, and S. Asad, "Enhanced IoT Network Security for Network intrusion detection Model based on Machine Learning Technique," Annual Methodological Archive Research Review, vol. 3, pp. 188-212, 2025.
[4] A. Puviarasu and V. K. Sudha, "Enhanced IoT security: privacy-preserving federated learning model for accurate, real-time intrusion detection across devices," Ain Shams Engineering Journal, vol. 17, no. 1, p. 103866, 2026.
[5] Y. Wang, M. A. Azad, M. Zafar, and A. Gul, "Enhancing AI transparency in IoT intrusion detection using explainable AI techniques," Internet of Things, p. 101714, 2025.
[6] J. P. Sahoo, B. Kar, A. M. Abdelmoniem, and D. Chatzopoulos, "Choir-IDS: A federated learning framework for fidelity-calibrated explainable intrusion detection system for edge-IoT networks," *Information Fusion*, vol. 125, p. 103473, 2026.

[7] A. Puviarasu and V. K. Sudha, "Enhanced IoT security: privacy-preserving federated learning model for accurate, real-time intrusion detection across devices," *Ain Shams Engineering Journal*, vol. 17, no. 1, p. 103866, 2026.

[8] Y. Wang, M. A. Azad, M. Zafar, and A. Gul, "Enhancing AI transparency in IoT intrusion detection using explainable AI techniques," *Internet of Things*, p. 101714, 2025.

[9] A. Hozouri, A. Mirzaei, and M. Effatparvar, "A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 314, 2025.

[10] I. Ahmad, M. Amin, K. Hamid, S. Rizwan, and S. Asad, "Enhanced IoT Network Security for Network intrusion detection Model based on Machine Learning Technique," *Annual Methodological Archive Research Review*, vol. 3, pp. 188-212, 2025.

[11] S. K. Singh and R. Kumar, "Deep restoration-based anomaly detection for secure smart grids in 6G era," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 2, pp. 1102-1115, 2025.

[12] L. Tan, X. Yang, and M. G. Khan, "Adaptive thresholding for zero-day attack detection in resource-constrained IoT nodes," *Journal of Network and Systems Management*, vol. 33, no. 4, p. 89, 2025.

[13] R. Sharma and K. Malik, "XAI in Cybersecurity: A review of SHAP and LIME applications in cloud environments," *Computers & Security*, vol. 142, p. 103852, 2024.

[14] H. Park and J. Choi, "Unsupervised anomaly detection using autoencoders for industrial IoT networks," *IEEE Access*, vol. 12, pp. 15432-15445, 2024.

[15] M. Al-Zubi and A. Al-Mousa, "SHAP-based feature importance for intrusion detection in SDNs," *International Journal of Information Security*, vol. 23, no. 1, pp. 45-62, 2024.

[16] T. Nguyen and D. Hoang, "Hybrid deep learning and SHAP analysis for real-time malware detection," *Journal of Cybersecurity and Privacy*, vol. 4, no. 2, pp. 210-228, 2024.

[17] F. Zhang and G. Wei, "Fidelity calibration in XAI: Improving trust in IoT security alerts," *IEEE Internet of Things Journal*, vol. 11, no. 8, pp. 13450-13462, 2024.

[18] P. Verma and S. Gupta, "Zero-day attack mitigation using reinforcement learning and adaptive thresholds," *Security and Communication Networks*, vol. 2024, Art. no. 5562143, 2024.

[19] M. Usman and N. Shah, "Evaluation of NSL-KDD and CIC-IDS2017 datasets for deep restoration models," *Network Security Review*, vol. 29, pp. 12-30, 2024.

[20] K. J. Lee and S. Kim, "Interpretable intrusion detection system using shapley additive explanations for medical IoT," *Healthcare Informatics Research*, vol. 30, no. 3, pp. 198-211, 2024.