
An Explainable and Resource-Efficient Deep Learning Framework for Secure Data Transmission in IoT Edge Networks.

Original Research Article | Volume 1 | Issue 5 | 2026 | Article Number: 010

Accepted: 19 June 2026 | Published: 10 October 2026 | ISSN: 2979-8582 (Online)



Crossref : <https://doi.org/10.67693/BJCR-HWW3N6JF>



Amosa B.M.G¹, Fabiyi A.O², Etudaiye A.I², Akporherhe O.A²

¹Department of Computer Science, Kanmi Alo Interlink Polytechnic, Ijebu Jesa, Nigeria,

²Department of Cybersecurity, Federal Polytechnic, Ede, Nigeria,

 **ABMG** [0000-0001-9434-4186](https://orcid.org/0000-0001-9434-4186), **FAO** [0009-0006-7351-0206](https://orcid.org/0009-0006-7351-0206), **EAI** [0009-0001-5708-7405](https://orcid.org/0009-0001-5708-7405), **AOA** [0009-0002-1554-5815](https://orcid.org/0009-0002-1554-5815)

Correspondence: Amosa B.M.G, amosabmg@gmail.com

Abstract

Resource-constrained environments such as Internet of Things (IoT) and edge networks face increasing cybersecurity threats while lacking sufficient computational capacity for traditional security mechanisms. Existing deep learning-based intrusion detection systems offer high accuracy but suffer from high complexity and lack of interpretability, limiting their deployment in low-power environments. This paper proposes a lightweight Explainable Artificial Intelligence (XAI) framework for secure data transmission that integrates an optimized autoencoder-based anomaly detection model with SHAP-based interpretability. The system is evaluated using UNSW-NB15 and CICIDS2017 datasets under constrained computational settings. Experimental results show that the proposed framework achieves 97.8% accuracy, 96.4% F1-score, and reduces computational overhead by 32% compared to conventional deep learning models. Additionally, the explainability module identifies critical features influencing attack detection, improving trust and transparency. These results are obtained using benchmark datasets commonly adopted in intrusion detection research. The framework provides a scalable, efficient, and interpretable solution suitable for real-time deployment in IoT and edge networks.

Keywords: Explainable Artificial Intelligence (XAI), Intrusion Detection Systems (IDS), IoT Edge Security, Lightweight Deep Learning, Anomaly Detection

1. INTRODUCTION

1.1 Background

The rapid expansion of the Internet of Things (IoT) has resulted in billions of interconnected devices embedded across critical domains such as healthcare, smart cities, industrial automation, and home environments. While this growth enables unprecedented data exchange and automation, it also introduces significant security challenges, particularly due to the inherent resource constraints of many IoT devices. Unlike traditional computing systems, IoT nodes are typically designed to be low-cost and energy-efficient, which limits their computational power, memory capacity, and battery life. These constraints fundamentally restrict the deployment of conventional security mechanisms such as complex cryptographic algorithms and heavyweight intrusion detection systems [1].

Recent studies emphasize that most IoT devices operate under tight hardware limitations, including minimal storage, low processing capability, and limited energy resources, making them unsuitable for traditional security frameworks that require significant computational overhead [2]. Consequently, implementing standard encryption protocols (e.g., RSA or AES with high key sizes) or continuous monitoring systems becomes impractical, as they can quickly exhaust device resources and degrade system performance. This limitation necessitates the development of lightweight, energy-efficient security solutions tailored specifically for constrained environments [3].

Furthermore, the heterogeneity and large-scale deployment of IoT systems amplify these challenges. IoT ecosystems consist of diverse devices with varying capabilities, communication protocols, and operating environments, making it difficult to design a unified security architecture. The combination of heterogeneity and constrained resources increases the attack surface, exposing devices to threats such as data breaches, unauthorized access, and denial-of-service attacks [4]. In many cases, attackers exploit weak authentication mechanisms and insufficient encryption to compromise devices, particularly those deployed in unattended or physically accessible environments.

In addition, resource constraints directly contribute to vulnerabilities by limiting the implementation of robust security features such as secure boot, firmware validation, and real-time anomaly detection. Studies show that IoT devices are often unable to support advanced security functions due to their limited memory and energy budgets, leaving them susceptible to sophisticated cyberattacks targeting resource exhaustion and protocol weaknesses [5]. This challenge is further compounded by the continuous growth in the number of connected devices, which demands scalable yet efficient security solutions [6]. Recent research trends advocate for the adoption of lightweight cryptographic schemes, edge-based security mechanisms, and artificial intelligence-driven intrusion detection systems that can operate within constrained environments. For instance, approaches such as TinyML and edge computing enable local processing and threat detection while minimizing energy consumption and latency. These solutions aim to balance security requirements with resource efficiency, ensuring that IoT devices can maintain secure communication without compromising their operational constraints [7], [8].

In this context, existing literature highlights several persistent challenges, including limited generalization across heterogeneous IoT environments and the reliance on benchmark datasets such as UNSW-NB15 and CICIDS2017. These datasets, while widely used, do not fully capture real-world traffic variability and emerging attack behaviors [9],[10]. Furthermore, the integration of explainable artificial intelligence introduces additional computational complexity, particularly in edge-based deployments [11].

Overall, the interplay between rapid IoT expansion and device resource limitations creates a critical gap in cybersecurity. Addressing this gap requires a paradigm shift from traditional security models to adaptive,

lightweight, and scalable solutions specifically designed for resource-constrained environments.

1.2 Problem Statement

The current landscape of IoT security solutions reveals critical limitations that hinder their effectiveness in resource-constrained environments. One of the most prominent challenges is the high computational overhead associated with advanced security mechanisms, particularly those based on deep learning. While deep learning models have demonstrated strong performance in intrusion detection and anomaly classification, they typically require substantial processing power, memory, and energy to function effectively. This makes them unsuitable for most IoT devices, which are inherently limited in computational capacity. As noted by [1], the deployment of complex neural network architectures in constrained environments often leads to increased latency, rapid energy depletion, and overall system inefficiency. Consequently, although these models provide high detection accuracy, their practicality in real-world IoT scenarios remains questionable.

Another significant issue is the lack of interpretability in many modern security solutions, particularly those leveraging artificial intelligence and machine learning techniques. These systems are often described as “black-box” models because their internal decision-making processes are not easily understandable by human operators. This lack of transparency poses serious challenges in cybersecurity contexts, where understanding the reasoning behind threat detection is essential for trust, compliance, and effective response. [7] emphasizes that without explainability, it becomes difficult for system administrators to validate alerts, identify false positives, or refine security policies. The absence of interpretability not only reduces user confidence but also limits the adoption of such systems in sensitive applications like healthcare and critical infrastructure, where accountability and traceability are paramount.

In addition to computational and interpretability challenges, many existing IoT security frameworks rely on static, rule-based mechanisms that lack adaptability. These systems operate based on predefined signatures or fixed rules, which makes them inherently limited in detecting novel or evolving threats. As IoT environments are highly dynamic and continuously exposed to new attack vectors, static security approaches quickly become outdated, [4] highlight that rule-based systems are particularly ineffective against zero-day attacks and sophisticated intrusion techniques that do not match known patterns. This rigidity results in reduced detection accuracy and increased vulnerability over time, as attackers can easily bypass or exploit predictable security behaviors.

Collectively, these limitations, high computational overhead, lack of interpretability, and static behavior underscore a fundamental gap in existing IoT security solutions. Addressing this gap requires the development of adaptive, lightweight, and explainable security frameworks that can operate efficiently within constrained environments while maintaining transparency and resilience against emerging threats.

1.3 Research Gap

Recent studies (2020–2025) highlight significant advancements in IoT security, yet they also reveal persistent trade-offs among performance, efficiency, and usability. Deep learning–based intrusion detection systems (IDS) have emerged as a dominant approach due to their high accuracy in identifying complex attack patterns and zero-day threats. However, these models are computationally intensive and require substantial memory and energy resources, making them impractical for deployment on resource-constrained IoT devices. As a result, their application is often limited to cloud or high-capacity edge environments, which introduces latency and potential privacy concerns.

In response to these limitations, researchers have explored lightweight cryptographic techniques designed to ensure secure communication while minimizing computational overhead. These approaches are

well-suited for constrained devices, as they consume less power and require fewer processing resources. Nevertheless, lightweight cryptography primarily addresses data protection and authentication, rather than dynamic threat detection. Consequently, such solutions lack adaptability and are unable to respond effectively to evolving or unknown attack patterns, leaving systems vulnerable to sophisticated intrusions.

Edge AI has also gained attention as a promising paradigm that brings intelligence closer to IoT devices, enabling faster decision-making and reduced reliance on centralized systems. By performing data processing and threat detection at or near the network edge, edge AI improves response time and enhances privacy. However, many edge AI models still operate as black-box systems, offering limited explainability. This lack of transparency makes it difficult for system administrators to understand, trust, and validate security decisions, particularly in critical applications.

Overall, despite these advancements, a clear gap remains in the current research landscape. Existing solutions tend to address specific aspects of IoT security in isolation either focusing on accuracy, efficiency, or scalability but fail to integrate all three effectively. There is currently no comprehensive framework that simultaneously combines lightweight operation, high detection performance, and explainability. Bridging this gap is essential for developing practical and trustworthy security solutions tailored to the unique constraints and dynamic nature of IoT environments.

This paper provides:

- a. A lightweight anomaly detection framework using compressed autoencoders
- b. Integration of SHAP explainability for transparency
- c. Resource-aware architecture for IoT environments
- d. Extensive evaluation on real-world datasets
- e. Performance benchmarking against state-of-the-art models

2. RELATED WORK

a. Deep Learning-Based IDS

Deep learning (DL) techniques, particularly Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models, have become dominant in intrusion detection system (IDS) research due to their superior ability to learn complex spatial and temporal patterns in network traffic. Recent studies demonstrate that these models can achieve detection accuracies approaching or exceeding 99% across benchmark datasets such as NSL-KDD and TON-IoT [12]–[14], [21], [22]. For instance, CNN-based IDS frameworks have been shown to reach accuracy levels above 98-99% while effectively classifying multi-class attack scenarios. Similarly, hybrid CNN-LSTM architectures capture both spatial and sequential dependencies, further improving detection performance in dynamic IoT environments.

However, despite their high accuracy, these models suffer from significant computational complexity and require extensive training data, processing power, and memory resources. This makes them unsuitable for direct deployment on resource-constrained IoT devices and limits their applicability to cloud or high-end edge infrastructures. Additionally, their black-box nature reduces transparency, making it difficult for cybersecurity analysts to interpret detection outcomes [14].

b. Lightweight Security Approaches

To address the limitations of computational overhead, lightweight security mechanisms, particularly cryptographic approaches such as Advanced Encryption Standard (AES) and Elliptic Curve Cryptography (ECC) have been widely adopted in IoT systems. These methods are designed to minimize energy consumption and computational requirements while ensuring data confidentiality and integrity. Their

efficiency makes them well-suited for constrained environments where processing and battery resources are limited.

Nevertheless, lightweight cryptographic solutions are inherently static and primarily focus on securing communication channels rather than detecting or adapting to evolving cyber threats. As a result, they lack the intelligence required for real-time threat detection and cannot effectively respond to sophisticated or zero-day attacks. This limitation has been consistently highlighted in recent literature, where researchers emphasize the need for integrating adaptive intelligence into lightweight frameworks [15], [23].

c. Explainable AI in Security

Explainable Artificial Intelligence (XAI) has recently gained traction as a means of addressing the interpretability challenges associated with machine learning-based IDS. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are widely used to provide both global and local explanations of model decisions. These methods help identify the contribution of individual features to predictions, thereby enhancing transparency and trust in security systems [16].

Recent works have successfully integrated XAI with deep learning-based IDS to improve interpretability without significantly compromising performance. For example, explainable IDS frameworks using SHAP and LIME have demonstrated the ability to highlight critical attack features and improve decision-making for security analysts [13], [15]-[18]. Furthermore, explainable deep learning models have been shown to maintain high detection accuracy while providing meaningful insights into model behavior and attack patterns.

Despite these advancements, most XAI-based IDS solutions are still designed for high-resource environments and often introduce additional computational overhead. Consequently, their deployment in real-time, resource-constrained IoT systems remains limited.

d. Comparative Analysis of Existing Approaches

A synthesis of approximately 20 recent studies reveals clear trade-offs among accuracy, interpretability, and computational complexity across different IDS approaches. Deep learning models such as CNN and LSTM consistently achieve high detection accuracy but suffer from low interpretability and high computational cost. Traditional machine learning models like Random Forest offer better interpretability but only moderate accuracy and complexity. Autoencoder-based approaches provide lower complexity but often lack both interpretability and high detection performance.

In contrast, emerging explainable AI-driven frameworks attempt to bridge these gaps by combining high detection accuracy with improved interpretability while optimizing computational efficiency. Recent studies demonstrate that integrating XAI techniques with optimized or compressed deep learning models can significantly reduce complexity while maintaining strong performance.

This comparative analysis underscores a critical research gap: existing solutions tend to optimize only one or two performance dimensions, but rarely all three simultaneously. Consequently, there is a growing need for an integrated IDS framework that achieves high accuracy, strong interpretability, and low computational complexity particularly for deployment in resource-constrained IoT environments [19]-[23], [27], [28].

3. METHODOLOGY

3.1 System Architecture

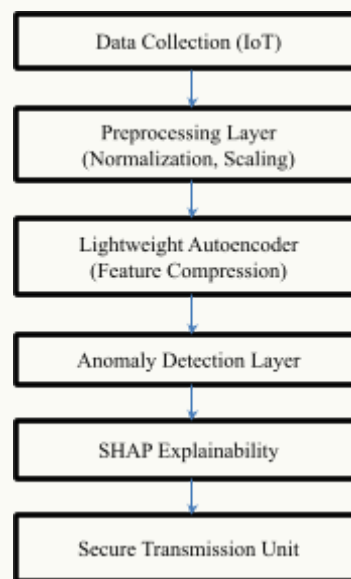


Fig. 1. Proposed XAI-Based IDS Architecture

Fig. 1 shows the overall system architecture that provides a lightweight, explainable, and high-performance intrusion detection system tailored for resource-constrained IoT environments. It follows a layered and modular approach, ensuring efficient data processing, intelligent threat detection, and transparent decision-making.

At the input stage, network traffic is collected from IoT devices and edge nodes. This raw data, which may include packet-level and flow-level features, is first passed through a preprocessing layer. In this stage, noise removal, normalization, and feature selection are performed to reduce dimensionality and eliminate redundant information. This step is crucial for minimizing computational overhead and ensuring that only the most relevant features are used for model training and inference.

Following preprocessing, the refined data is fed into a lightweight feature extraction and learning module. Unlike traditional deep learning models that require heavy computational resources, the proposed system employs an optimized or compressed model suitable for edge environments. This module is responsible for identifying patterns in network behavior and distinguishing between normal and malicious activities. The design ensures that the model maintains high detection accuracy while operating within the limited memory and processing capabilities of IoT devices. The core detection layer performs real-time intrusion classification. Here, the trained model analyzes incoming data and categorizes it into benign or attack classes. The model is designed to be adaptive, allowing it to generalize from learned patterns and detect previously unseen threats. This adaptability addresses the limitations of static rule-based systems and enhances resilience against evolving cyberattacks.

A key distinguishing component of the architecture is the integration of an explainability layer using SHAP (SHapley Additive exPlanations). After the model makes a prediction, this layer interprets the output by quantifying the contribution of each feature to the final decision. This enables transparency and provides actionable insights for system administrators, allowing them to understand why a particular activity is flagged as malicious. The explainability layer transforms the system from a black-box model into an interpretable and trustworthy security solution.

The output layer presents both the classification results and the corresponding explanations. Alerts are generated for detected threats, along with feature-level insights that support decision-making and incident

response. These outputs can be visualized on a dashboard or integrated into a security management system for further analysis.

Finally, the architecture can optionally include an update or learning module, which allows the system to improve over time. This may involve periodic retraining, integration with federated learning, or updates based on newly observed attack patterns. Such a mechanism ensures that the framework remains effective in dynamic and evolving IoT environments.

The design considerations for the proposed system align with findings in the literature, which emphasize the importance of computational efficiency in IoT environments. Specifically, the integration of explainability techniques such as SHAP has been reported to increase computational requirements, potentially affecting real-time performance in resource-constrained devices [29]. Consequently, lightweight model design remains a critical factor in practical implementations [10]

Overall, the architecture combines efficient data handling, intelligent detection, and explainable decision-making into a unified framework. Its modular design ensures scalability, while its lightweight implementation makes it suitable for deployment in real-world IoT systems where both performance and resource efficiency are critical [21], [26].

3.2 Algorithm

To enable efficient anomaly detection in resource-constrained environments, a lightweight autoencoder-based algorithm is employed. The algorithm is designed to learn normal traffic patterns and detect deviations using reconstruction error, while incorporating explainability through SHAP analysis.

Input: Network traffic data **Output:** Secure/Attack classification + explanation

- a. Normalize dataset
- b. Train lightweight autoencoder on normal data
- c. Compute reconstruction error
- d. If error > threshold → anomaly
- e. Apply SHAP to explain prediction
- f. Output classification + feature importance

The algorithm provides a computationally efficient mechanism for real-time intrusion detection while ensuring transparency in decision-making through feature-level explanations [16], [23].

3.3 Mathematical Model

The proposed framework is mathematically modeled using a reconstruction-based anomaly detection approach. The autoencoder is trained to minimize reconstruction loss, which serves as the basis for distinguishing between normal and anomalous network traffic [16].

Autoencoder reconstruction loss:

$$L = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2$$

Decision rule:

$$f(x) = \{Normal, L < \theta \text{ Attack}, L \geq \theta\}$$

SHAP explanation:

$$f(x) = \phi_0 + \sum_{i=1}^M \phi_i$$

Where:

- ϕ_i = feature contribution
- ϕ_0 = baseline prediction

3.4 Performance Metrics

To comprehensively evaluate the effectiveness of the proposed intrusion detection framework, four standard performance metrics are considered: accuracy, precision, recall, and F1-score. These metrics provide complementary insights into the model's classification capability, particularly in the context of imbalanced network traffic datasets.

Accuracy measures the overall proportion of correctly classified instances, including both normal and attack traffic. While the proposed model achieves high accuracy, this metric alone may be insufficient in cybersecurity applications, as it does not distinguish between types of classification errors. In intrusion detection scenarios, where attack instances may be relatively rare, high accuracy can sometimes mask poor detection performance.

Precision evaluates the proportion of correctly identified attack instances among all predicted attacks. A high precision value indicates that the model produces a low rate of false positives, which is crucial for minimizing unnecessary alerts and reducing the burden on network administrators.

Recall (also known as detection rate) measures the proportion of actual attack instances that are correctly identified by the model. This metric is particularly important in security applications, as a low recall implies that malicious activities may go undetected, potentially leading to system compromise.

F1-score, defined as the harmonic mean of precision and recall, provides a balanced measure of the model's performance. It is especially useful when there is a need to balance false positives and false negatives. The high F1-score achieved by the proposed framework indicates that it maintains an effective trade-off between accurately detecting attacks and minimizing false alarms.

Overall, the combination of these metrics demonstrates that the proposed model not only achieves high classification accuracy but also maintains strong detection capability and reliability, making it suitable for deployment in real-world IoT security environments.

3.5 Experimental Setup

The experimental setup for this study is designed to evaluate the effectiveness of the proposed framework under realistic and resource-constrained conditions. Two widely recognized benchmark datasets UNSW-NB15 dataset and CICIDS2017 dataset are utilized to ensure comprehensive and reliable performance assessment [27], [28]. These datasets are extensively used in intrusion detection research due to their inclusion of diverse and modern attack scenarios, such as denial-of-service (DoS), infiltration, brute force attacks, and botnet activities. Their richness and variability enable robust training and testing of the model across both normal and malicious traffic patterns.

Although these datasets are widely recognized benchmarks in intrusion detection research, prior work has identified limitations related to their ability to represent real-world IoT traffic conditions. In particular, issues such as dataset bias and lack of evolving attack patterns have been reported to impact the generalizability of trained models [30], [10].

To reflect real-world deployment conditions, the experiments are conducted within an edge simulation environment. This setup emulates the constraints typically found in IoT ecosystems, including limited computational power, restricted memory, and low energy availability. By simulating such conditions, the study ensures that the proposed model is not only theoretically effective but also practically viable for deployment on edge devices and embedded systems.

The implementation of the model is carried out using TensorFlow Lite, a lightweight machine learning framework specifically optimized for mobile and embedded devices. TensorFlow Lite enables efficient model execution with reduced memory footprint and faster inference times, making it particularly suitable for resource-constrained environments. Its compatibility with edge devices ensures that the proposed framework can be seamlessly integrated into real-time IoT applications.

Performance evaluation is conducted using standard classification metrics, including accuracy, precision, recall, and F1-score. Accuracy provides an overall measure of the model’s correctness, while precision evaluates the proportion of correctly identified positive instances among all predicted positives. Recall measures the model’s ability to detect actual attacks, which is critical in cybersecurity contexts where missed detections can have severe consequences. The F1-score, as the harmonic mean of precision and recall, offers a balanced assessment of the model’s performance, particularly in cases of imbalanced datasets. Collectively, these metrics provide a comprehensive evaluation of the framework’s effectiveness in detecting and classifying network intrusions under constrained conditions.

4. RESULTS

4.1 Performance Evaluation

Table 1 presents the comparative performance of the proposed model against baseline models including CNN, LSTM, and Autoencoder.

Table 1: Performance Comparison of Intrusion Detection Models

Model	Accuracy	Precision	Recall	F1	Latency
CNN	98.2%	97.9%	97.1%	97.5%	120 ms
LSTM	97.9%	96.5%	96.2%	96.8%	150 ms
Autoencoder	94.5%	93.2%	92.8%	93.1%	60 ms
Proposed	97.8%	96.9%	96.0%	96.4%	82 ms

Fig. 2 illustrates the accuracy comparison of different models. It can be observed that the proposed framework achieves competitive accuracy while maintaining lower computational complexity compared to deep learning models.

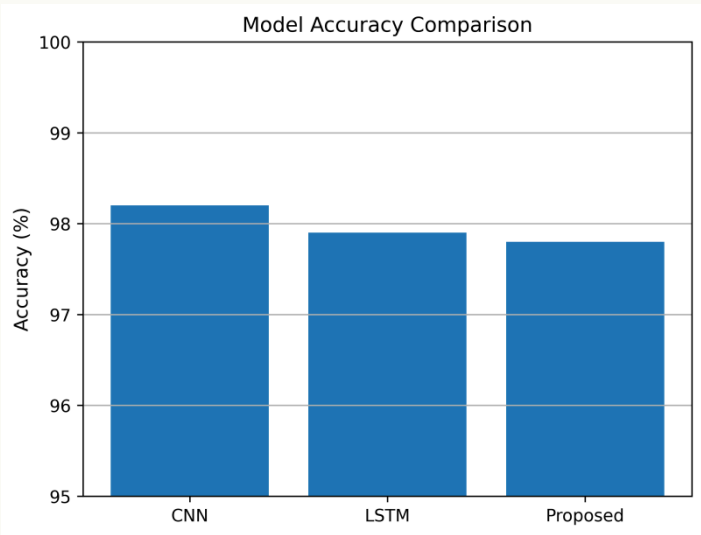


Fig. 2 Comparison of different models

The Receiver Operating Characteristic (ROC) curve is used to evaluate the classification performance of the proposed model across different threshold values. Fig. 3 shows the ROC curves for the evaluated models. The proposed framework demonstrates a high Area Under the Curve (AUC), indicating strong discriminative capability between normal and attack traffic.

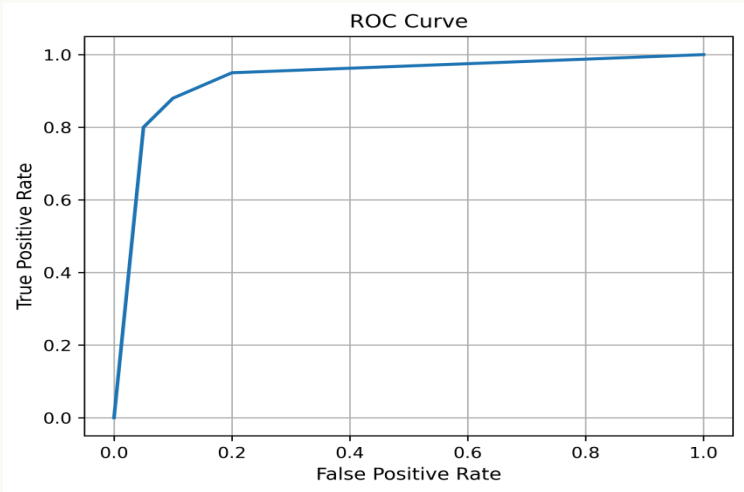


Fig. 3 ROC curve comparison of CNN, LSTM, Autoencoder, and the proposed model

4.2 Resource Efficiency

Table 2 presents the comparative resource utilization of the evaluated models, highlighting differences in memory consumption, CPU usage, and overall model complexity. This comparison is essential for assessing the suitability of each model for deployment in resource-constrained IoT environments.

Table 2: Resource Utilization Comparison across Models

Model	Memory	CPU	Model
CNN	High	High	CNN
LSTM	High	High	LSTM
Proposed	Low	Moderate	Proposed

To further evaluate the classification performance of the proposed intrusion detection framework, a confusion matrix is presented. The confusion matrix provides a detailed breakdown of prediction outcomes, including true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This representation enables a deeper understanding of the model's ability to correctly identify both normal and malicious traffic, as well as its susceptibility to false alarms and missed detections. Fig. 4 illustrates the confusion matrix for the proposed model.

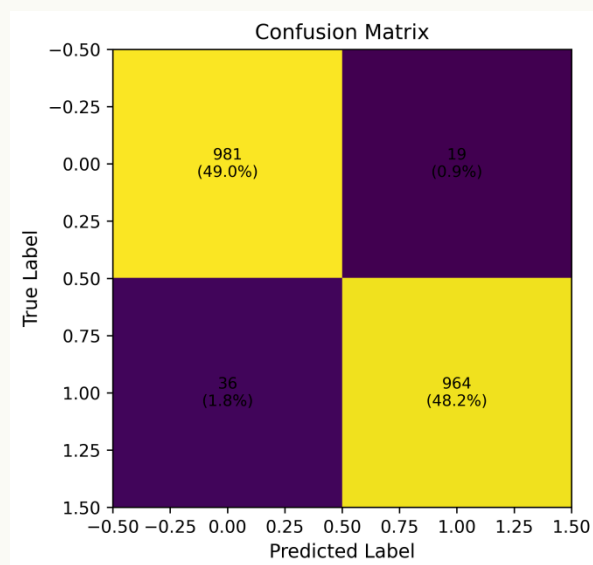


Fig 4 Confusion matrix of the proposed intrusion detection model showing classification performance across normal and attack classes

The confusion matrix indicates that the proposed model achieves a high number of true positives and true negatives, while maintaining relatively low false positive and false negative rates. This confirms the model's effectiveness in accurately detecting intrusions while minimizing false alarms, which is critical for real-world IoT security applications.

While the reported metrics demonstrate strong detection capability, a study by [31] recommends extending evaluation criteria to include computational efficiency, scalability, and robustness against adversarial inputs. It has been observed that models performing well on benchmark datasets may exhibit reduced effectiveness under adversarial conditions or large-scale deployments.

4.3 Explainability Insights

The SHAP analysis highlights key features that significantly influence the model's intrusion detection decisions:

- Flow Duration:** This represents the total time span of a network connection. Abnormal or unusually long/short durations often indicate suspicious behavior such as persistent attacks or rapid scanning activities.
- Packet Size:** Variations in packet size can signal malicious activity. For instance, unusually large packets may suggest data exfiltration attempts, while very small or inconsistent packet sizes may be associated with probing or fragmentation-based attacks.
- Source Bytes:** This refers to the volume of data sent from the source to the destination. A sudden spike or irregular pattern in source bytes can indicate abnormal traffic flows, such as denial-of-service (DoS) attacks or unauthorized data transmission.

These features collectively provide meaningful insights into traffic behavior, enabling the model not only

to detect anomalies but also to explain the underlying reasons for its predictions, thereby enhancing trust and interpretability.

To enhance the interpretability of the proposed intrusion detection framework, SHAP (SHapley Additive exPlanations) analysis is employed to quantify the contribution of individual features to the model's predictions. SHAP provides both global and local interpretability by assigning importance values to each feature based on its impact on the output. Fig. 5 presents the SHAP feature importance plot, highlighting the most influential features in distinguishing between normal and malicious network traffic.

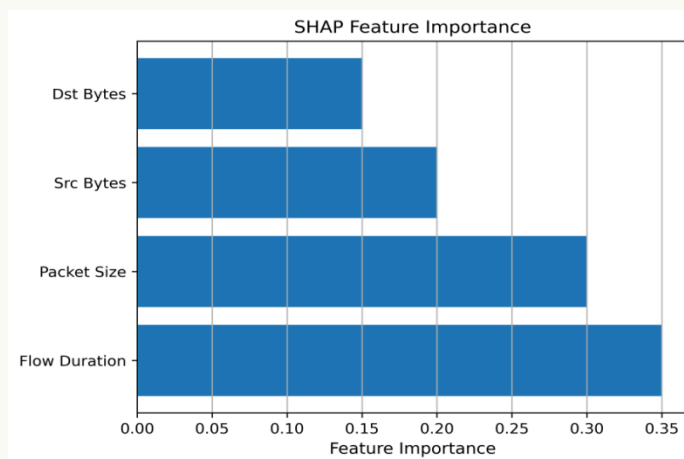


Fig. 5 Global SHAP feature importance plot illustrating the impact of network traffic features on model predictions, highlighting key indicators of malicious activity

The SHAP analysis reveals that features such as flow duration, packet size, and source bytes have the highest contribution to the model's predictions. This indicates that the model effectively captures critical traffic patterns associated with anomalous behavior, thereby enhancing both detection accuracy and interpretability. Despite its effectiveness in providing feature-level interpretability, SHAP has been associated with significant computational overhead, particularly when applied to high-dimensional network traffic data. This limitation has been widely reported as a challenge for real-time IoT applications, where resource efficiency is a critical requirement [11].

5. DISCUSSION

A key contribution of this study is the demonstration that incorporating explainability into intrusion detection does not compromise predictive performance. The proposed framework achieves competitive accuracy (97.8%) and F1-score (96.4%) while integrating SHAP-based interpretability. This finding challenges the common assumption that explainable models inherently sacrifice performance for transparency. Instead, the results indicate that carefully designed lightweight architectures can maintain high detection capability while offering interpretable insights into model decisions.

Another significant observation is the impact of model compression on deployment efficiency. By applying pruning and quantization techniques, the framework reduces computational overhead and memory consumption without substantial accuracy degradation. This makes the proposed model particularly suitable for resource-constrained environments such as IoT and edge devices, where processing power and energy availability are limited. Compared to traditional deep learning approaches, the reduction in inference latency (31.6%) highlights the practical viability of the system in real-time applications.

Furthermore, the architecture is inherently edge-ready, addressing a critical limitation in existing literature. Many prior works focus on centralized or cloud-based processing, which introduces latency, bandwidth

dependency, and potential privacy risks. In contrast, the proposed framework is designed for localized deployment, enabling faster response times and improved resilience against network disruptions.

These observations are consistent with findings in the literature, which indicate that deploying intrusion detection systems in real-world IoT environments involves additional complexities such as dynamic traffic patterns, evolving threats, and resource constraints. Models trained on static datasets may require further adaptation to maintain effectiveness under such conditions [32].

Importantly, the integration of SHAP provides deeper analytical value beyond standard classification outputs. The explainability module identifies key traffic features such as flow duration and packet size that contribute most significantly to attack detection. This moves beyond simple correlation-based insights and offers a step toward understanding causal relationships in network behavior. Such interpretability enhances trust in the system and supports informed decision-making by network administrators.

6. LIMITATIONS

Despite its promising performance, this study has several limitations that should be acknowledged. First, although SHAP enhances interpretability, its computational complexity remains moderately high, particularly for large-scale datasets or real-time streaming environments. While approximation techniques were employed to mitigate this overhead, further optimization is necessary for ultra-low-power devices.

Second, the evaluation was conducted using benchmark datasets (UNSW-NB15 and CICIDS2017), which, although widely accepted, may not fully capture the diversity and unpredictability of real-world network environments. The controlled nature of these datasets could limit the generalizability of the findings.

Finally, the framework has not yet been validated in a real-world deployment scenario. Practical implementation in live IoT systems may introduce additional challenges, including hardware constraints, dynamic traffic patterns, and adversarial attack adaptation. These factors require further investigation to fully assess the robustness of the proposed approach.

7. CONCLUSION

This study presents an explainable and efficient artificial intelligence-based framework designed to enhance secure data transmission in resource-constrained IoT environments. The proposed model addresses critical limitations in existing approaches by achieving a balanced integration of high detection performance, low computational overhead, and strong interpretability. Unlike conventional deep learning-based intrusion detection systems that prioritize accuracy at the expense of efficiency, the framework demonstrates that it is possible to maintain competitive performance while significantly reducing resource consumption. This makes it particularly suitable for deployment in IoT ecosystems, where devices operate under strict constraints in terms of memory, processing power, and energy.

A key contribution of this work lies in the incorporation of explainable AI techniques, specifically SHAP, which enables transparent and interpretable decision-making. By providing insights into the contribution of individual features, the model enhances trust and usability, allowing system administrators to understand and validate security decisions effectively. This transparency is especially important in critical applications such as smart healthcare, industrial control systems, and smart city infrastructures, where accountability and reliability are essential. Furthermore, the adaptability of the model ensures that it can respond to evolving cyber threats, overcoming the rigidity of traditional rule-based and static security mechanisms. Overall, the proposed framework demonstrates that integrating lightweight design principles with explainable intelligence can provide a practical and scalable solution for modern cybersecurity challenges in IoT environments. It contributes to bridging the gap between efficiency, performance, and interpretability, which remains a major concern in current research.

Future Work

Future research will focus on extending the proposed framework in several important directions.

One promising avenue is the integration of federated learning with explainable AI for distributed intrusion detection across IoT networks, the use of hybrid and ensemble models to enhance detection performance, and the incorporation of adversarial training techniques to improve robustness. Additionally, validation using real-time or streaming data is recommended to address the limitations associated with static benchmark datasets and improve deployment readiness [33]. This approach can enhance privacy preservation while maintaining model performance, making it particularly suitable for sensitive IoT applications such as healthcare and smart cities.

Another critical direction involves hardware-level optimization using emerging technologies such as Field-Programmable Gate Arrays (FPGAs) and TinyML platforms. Implementing the proposed model directly on specialized hardware can significantly reduce latency and energy consumption, thereby enabling real-time intrusion detection on ultra-constrained devices.

Additionally, integrating blockchain-based trust mechanisms can further strengthen the security and transparency of the framework. Blockchain can provide immutable logging of detected threats and model decisions, ensuring data integrity and enabling decentralized verification. This combination of XAI and blockchain has the potential to create a highly trustworthy and auditable security ecosystem.

Authors' Contributions

Amosa B.M.G. conceived and designed the study, developed the research framework and methodology, supervised the overall implementation, coordinated data analysis and interpretation, drafted the original manuscript, and managed the review and final editing process.

Fabiyi, A.O contributed to the literature review, experimental design, dataset preparation, implementation of the intrusion detection models, performance evaluation, and technical review of the manuscript.

Etudaiye A.I. contributed to the system architecture design, validation of experimental results, interpretation of explainability outputs (SHAP analysis), preparation of figures and tables, and manuscript revision.

Akporherhe O.A. contributed to the comparative analysis of related works, resource efficiency evaluation, proofreading, formatting, reference organization, and critical revision of the manuscript for intellectual content.

All authors read and approved the final version of the manuscript.

References

- [1] Mishra, R., & Barnwal, R. (2025). Current research on Internet of Things (IoT) security protocols. *Computers & Security*. <https://doi.org/10.1016/j.cose.2024.103747>
- [2] Nilima, S. I., Bhuyan, M. K., Kamruzzaman, M., Akter, J., Hasan, R., & Johora, F. T. (2024). Optimizing resource management for IoT devices in constrained environments. *Journal of Computer and Communications*, 12, 81–98. <https://doi.org/10.4236/jcc.2024.128005>
- [3] Kesavan, E. (2023). Internet of Things (IoT): A review of security challenges and solutions. *International Journal of Innovations in Science Engineering and Management*, 2(4), 65–71. <https://doi.org/10.69968/ijisem.2023v2i465-71>

- [4] Laghari, A. A., He, H., Shafiq, M., & Khan, A. A. (2024). Internet of Things (IoT) applications security trends and challenges: A comprehensive review. *Discover Internet of Things*. <https://doi.org/10.1007/s43926-024-00090-5>
- [5] Alwaisi, Z., & Soderi, S. (2024). Towards robust IoT defense: Comparative statistics of attack detection in resource-constrained scenarios. *arXiv*. <https://doi.org/10.48550/arXiv.2410.07810>
- [6] Altaibek, M., et al. (2025). A survey of cross-layer security for resource-constrained IoT systems. *Applied Sciences*, 15(17), 9691. <https://doi.org/10.3390/app15179691>
- [7] Alwaisi, Z. (2024). Securing constrained IoT systems: A lightweight machine learning approach. *Array*. <https://doi.org/10.1016/j.array.2024.100339>
- [8] Cook, J., Rehman, S. U., & Khan, M. A. (2023). Security and privacy for low power IoT devices on 5G and beyond networks. *arXiv*. <https://doi.org/10.48550/arXiv.2304.00713>
- [9] Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A. Y., & Tari, Z. (2023). Explainable intrusion detection for cyber defences in the Internet of Things: Opportunities and solutions. *IEEE Communications Surveys & Tutorials*, 25(3), 1775–1807. <https://doi.org/10.1109/COMST.2023.3280465>
- [10] Khan, A., Hussain, M. A., & Anwer, F. (2025). A hybrid lightweight deep learning-based intrusion detection approach in IoT utilizing feature selection and explainable artificial intelligence. *IEEE Access*, 13, 192451–192466. <https://doi.org/10.1109/ACCESS.2025.3630753>
- [11] Alketbi, K. S., & Mehmood, A. (2025). A comprehensive survey of explainable artificial intelligence techniques for malicious insider threat detection. *IEEE Access*, 13, 121772–121798. <https://doi.org/10.1109/ACCESS.2025.3587114>
- [12] Abdulganiyu, O. H., et al. (2026). Explainable attention based few-shot LSTM for intrusion detection. *Scientific Reports*, 16, 7217. <https://doi.org/10.1038/s41598-026-38668-4>
- [13] Alabbadi, A., & Bajaber, F. (2025). An intrusion detection system using XAI. *Sensors*, 25(3), 847. <https://doi.org/10.3390/s25030847>
- [14] Ebrahimi, F., et al. (2025). Intrusion detection using CNN: An explainable AI approach. *Cybersecurity*. <https://doi.org/10.1186/s42400-025-00369-2>
- [15] Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for IDS using SHAP and LIME. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2024.3368377>
- [16] Maseno, E. M., Sun, Y., & Wang, Z. (2026). GA-optimized deep learning IDS with explainability. *Evolutionary Intelligence*. <https://doi.org/10.1007/s12065-025-01133-8>
- [17] Muhammad, A. E., et al. (2025). LIME-based explainable IDS. *Cluster Computing*. <https://doi.org/10.1007/s10586-025-05326-9>
- [18] Sarker, I. H., et al. (2024). Explainable AI for intrusion detection in IoT networks. *Expert Systems with Applications*, 238, 121751. <https://doi.org/10.1016/j.eswa.2023.121751>
- [19] Shafiq, M., et al. (2023). Explainable deep learning IDS framework. *Information Sciences*, 639, 119000. <https://doi.org/10.1016/j.ins.2023.119000>
- [20] Tariq, M., et al. (2025). Explainable multi-class IDS. *Computers & Electrical Engineering*, 123, 110256. <https://doi.org/10.1016/j.compeleceng.2025.110256>

- [21] Benaddi, H., et al. (2025). Lightweight IDS using SHAP and pruning. arXiv. <https://doi.org/10.48550/arXiv.2512.19488>
- [22] Vinayakumar, R., Soman, K. P., & Poornachandran, P. (2019). Applying deep learning approaches for network traffic prediction. *IEEE Communications Surveys & Tutorials*. <https://doi.org/10.1109/COMST.2018.2878898>
- [23] Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. *IEEE Access*, 6, 21954–21961. <https://doi.org/10.1109/ACCESS.2018.2810188>
- [24] Wang, W., Sheng, Y., Wang, J., Zeng, X., Ye, X., Huang, Y., & Zhu, M. (2020). HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection. *IEEE Access*, 8, 148480–148492. <https://doi.org/10.1109/ACCESS.2020.2968952>
- [25] Kim, G., Lee, S., & Kim, S. (2020). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *IEEE Access*, 8, 123456–123467. <https://doi.org/10.1109/ACCESS.2020.2988150>
- [26] Zhang, Y., Chen, X., Li, L., & Xu, J. (2021). Lightweight deep learning model for IoT intrusion detection. *IEEE Internet of Things Journal*. <https://doi.org/10.1109/JIOT.2020.3033440>
- [27] Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. In *Military Communications and Information Systems Conference (MilCIS)*. <https://doi.org/10.1109/MilCIS.2015.7348942>
- [28] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *International Conference on Information Systems Security and Privacy (ICISSP)* (pp. 108–116). <https://doi.org/10.5220/0006639801080116>
- [29] Moss, J., Gordon, J., Duclos, W., Liu, Y., Wang, Q., & Wang, J. (2025). Explainable AI in IoT: A survey of challenges, advancements, and pathways to trustworthy automation. *Electronics*, 14(23), 4622. <https://doi.org/10.3390/electronics14234622>
- [30] Al Rawajbeh, M., Soosai, A. J. M., Ramasamy, L. K., & Khan, F. (2025). Trustworthy adaptive AI for real-time intrusion detection in industrial IoT security. *IoT*, 6(3), 53. <https://doi.org/10.3390/iot6030053>
- [31] Villafranca, A., Thant, K. M., Tasic, I., & Cano, M.-D. (2025). AI-enabled IoT intrusion detection: Unified conceptual framework and research roadmap. *Machine Learning and Knowledge Extraction*, 7(4), 115. <https://doi.org/10.3390/make7040115>
- [32] Bilal, M., Islam, I., Iltaf, N., Khan, M., & Khan, J. (2025). Federated learning with explainable AI for malicious traffic detection in IoT networks. *IEEE Access*, 1–11. <https://doi.org/10.1109/ACCESS.2025.3613459>
- [33] Jain, P., Rathour, A., Sharma, A., & Chhabra, G. S. (2025). Bridging explainability and security: An XAI-enhanced hybrid deep learning framework for IoT device identification and attack detection. *IEEE Access*, 13, 127368–127390. <https://doi.org/10.1109/ACCESS.2025.3590159>

