An AI-Based Intrusion Detection System (IDS) for the Internet of Things (IoT) Networks

Lawrence Mashego *, Mncedisi J. Bembe *, Issah Ngomane , Motselisi F. Chere

Posted Date: 3 June 2026

doi: 10.20944/preprints202606.0196.v1

Keywords: intrusion detection systems; deep learning; convolutional neural networks; bidirectional long short-term memory; transformer; IoT security

 doi:10.20944/preprints202606.0196.v1

Article

An AI-Based Intrusion Detection System (IDS) for the Internet of Things (IoT) Networks

Lawrence Mashego [1,*], Mncedisi J. Bembe [2], Issah Ngomane [1] and Motselisi F. Chere [1]

[1] Computing and Mathematical Sciences, University of Mpumalanga, Mbombela, 1200, SA

[2] Department of Engineering, Bryan College, Dayton, TN, 37321, USA

* Correspondence: 220158649@ump.ac.za; Tel.: (+1 762 303 0060 MJ; +27 762 150 729 L)

Abstract

The increased use of intelligent devices has led to widespread adoption of the Internet of Things (IoT). IoT is a set of devices connected via the Internet, which has enhanced connectivity, convenience, and efficiency. However, this surge in IoT is accompanied by vulnerabilities, making the ability to detect cyber-attack intrusions very crucial and a major challenge. Intrusion Detection Systems (IDSs) are a significant technique for detecting cyberattacks in traditional computing systems. The limited computational resources available on IoT devices make it challenging to deploy conventional IDSs methods. Despite significant progress in IoT-based intrusion detection, developing models that achieve high classification performance while maintaining reduced complexity remains challenging. In this study, we propose a hybrid architecture composed of lightweight CNN, a bidirectional LSTM (BiLSTM), and transformer-based attention to enhance IDS performance. The proposed model is specifically designed to run on resource-constrained IoT devices and meet their computational requirements. Despite the complexity of designing a model that fits the requirements of IoT devices and achieves higher accuracy, our proposed model outperforms existing models in the literature by achieving multiclass accuracies ranging from 81.554% to 99.705% and binary accuracies of 99.8554% to 99.9348% with very low false-positive rates (0.001–0.530) in all the CICIoT2023, UNSW-NB15, and CICIDS2017 datasets.

Keywords: intrusion detection systems; deep learning; convolutional neural networks; bidirectional long short-term memory; transformer; IoT security

1. Introduction

As Internet of Things (IoT) networks expand rapidly and become ubiquitous in our everyday environments, their underlying infrastructure is increasingly vulnerable to cyberattacks. These attacks threaten public safety by disrupting critical services and compromising sensitive information. In such interconnected networks, intrusion detection systems (IDSs) play a crucial role in real-time monitoring and identifying malicious activities before they cause significant damage [1]. Although existing IDS solutions have shown promising results, they often face challenges in IoT environments.

Traditional rule-based systems struggle to adapt to constantly evolving attack patterns, whereas complex deep learning models are computationally expensive and resource-intensive for resource-constrained IoT devices [2,3]. These IoT devices, which include temperature sensors and smart devices for monitoring light bulbs or thermostats, are characterized by their small size and limited functionality. The data collected by these devices are transmitted via the Internet for storage and analysis, necessitating security measures to protect personal, private, or sensitive information.

IoT has enhanced individuals' lifestyles by introducing automated services. As a result of this unchecked growth, privacy and security concerns have become increasingly important [4]. Addressing these concerns is challenging because of the inherent complexity of dynamic, diverse IoT ecosystems. [5,6] Consider protecting dynamic environments, such as smart cities or smart homes, which are characterized by limited network visibility and a lack of professional support for devices

 doi:10.20944/preprints202606.0196.v1

after initial use. This makes traditional network security methods ineffective owing to the environment's dynamic nature.

The risk to user privacy and the general security of the environment is further increased by the fact that many interconnected IoT devices are poorly designed and do not receive regular security updates [7]. Thus, to fully realize the potential of IoT-based smart environments, creative solutions that prioritize security and privacy while adapting to ever-changing environments are required.

Previous research on IoT IDS has predominantly used traditional machine learning (ML) techniques, such as decision trees and Support Vector Machines (SVMs), as shown in [8], to improve IDS performance. Although these methods deliver respectable accuracy, they require extensive manual feature engineering and often suffer from high false-positive rates, a limitation that deep learning's autonomous feature extraction addresses [relevant citation needed]. In contrast, deep learning methods, especially convolutional neural networks (CNNs), have emerged as promising alternatives because of their ability to autonomously extract features from network traffic data. However, their substantial computational demands render them impractical for IoT devices with limited resources [4,9]. In this study, we introduce several key contributions:

1. We hypothesize that integrating lightweight CNNs, BiLSTM, and transformer attention will lead to efficient and accurate intrusion detection in IoT networks, overcoming current resource constraints. Therefore, increasing security.
2. We propose a novel approach that integrates lightweight CNNs with bidirectional Long Short-Term Memory (BiLSTM) networks and transformer attention to achieve efficient and accurate intrusion detection in IoT networks.
3. We develop a customized CNN-BiLSTM-Transformer architecture optimized for IoT devices, featuring fewer layers and parameters to facilitate efficient data processing. This architecture efficiently extracts critical features from network traffic, which are subsequently processed by a bidirectional LSTM and transformer attention for the accurate classification of normal and malicious activities.
4. Our hybrid model combines the strengths of CNNs, BiLSTM, and transformer attention, addressing the limitations of each method. Moreover, we implemented selective SMOTE and weighted loss functions in our multiclassification model. Weights were assigned based on the relative frequencies of each attack type across the datasets, thereby enhancing the model's fairness and accuracy.
5. Lastly, we contribute to the scientific knowledge on lightweight models for detecting and isolating attacks in IoT networks.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature related to our study. Section 3 offers a detailed description of our proposed methodology, including the lightweight CNN-BiLSTM-Transformer architecture, and outlines the dataset and evaluation metrics used in our analysis. Section 4 presents the experimental results and analyzes the findings. Finally, Section 5 concludes the paper by summarizing our contributions and outlining future directions of this research.

2. Related Work

The literature on DL-based IDS for IoT devices reveals a spectrum of approaches, each contributing valuable insights while highlighting gaps this study aims to address. The authors in [10], a hybrid CNN-LSTM model that combines CNNs for spatial feature extraction with unidirectional LSTMs for temporal analysis was proposed, and it was tested on the CICIoT2023 and CICIDS2017 datasets. The model achieved a binary classification accuracy of 98.42% on CICIoT2023, demonstrating strong detection performance but overlooking resource constraints and multiclass classification, with a notable weakness in normal traffic recall (61%), suggesting potential false negatives.

 doi:10.20944/preprints202606.0196.v1

Similarly, [4] developed a lightweight CNN-BiLSTM IDS optimised for resource-constrained IoT devices using the UNSW-NB15 dataset. Their approach, rooted in spatial feature extraction and bidirectional temporal modelling, achieved 97.28% binary and 96.91% multiclass accuracy with fast inference (~4 seconds) and high precision (98.59%). However, it lacks testing across diverse datasets and advanced attention mechanisms for complex pattern recognition. [11] explored an unsupervised approach with a Convolutional Autoencoder (CAE) paired with One-Class SVM, achieving 94.28% accuracy on NSL-KDD and UNSW-NB15. While effective against traditional ML, it neglects temporal dynamics and IoT-specific optimization, differing markedly from the supervised, hybrid DL focus of this study.

Furthermore, [12] introduced a Deep Neural Network (DNN) for IoT attack detection on NSL-KDD, UNSW-NB15, and WSN-DS, achieving 99% accuracy. However, its computational intensity renders it impractical for edge devices, which contrasts with the study's emphasis on lightweight solutions. In [13] the authors examined Feed-Forward Neural Networks (FNN) and Self-Normalizing Neural Networks (SNN) on the BoT-IoT dataset, achieving 95.1% accuracy, with SNN showing resilience to adversarial attacks but lacking temporal modelling and efficiency for IoT deployment. In [14], the authors applied a CNN-LSTM model for real-time web intrusion detection on the CICIDS2017 dataset, achieving 95.8% accuracy, but its design prioritises web contexts over IoT constraints, aligning with this study's CNN-LSTM base while differing in scope.

In [15], the authors used a transformer for multiclass classification on CICIoT2023, achieving 99.40%, but their approach relies on distance-based methods and lacks the feature-learning depth of DL, distinguishing it from this study's hybrid approach. Synthesizing these works, a recurring theme emerges: high-performing models often sacrifice efficiency [10,12], while lightweight designs [4] limit versatility or generalization. None fully integrates advanced temporal and attention mechanisms, a gap that the model addresses by combining Lightweight CNN, BiLSTM, and Transformer layers for a balanced, IoT-optimized solution.

3. Methodology

This chapter presents a methodological, systematic, and comprehensive framework for designing, implementing, and evaluating the Lightweight CNN-BiLSTM-Transformer Intrusion Detection System, to enhance IoT cybersecurity. It encompasses the entire process, from selecting an appropriate study design and dataset to choosing the right data, building the correct model, implementing it, testing it, and ensuring it works effectively. The process relies on computational analysis, utilizing benchmark datasets and advanced machine learning techniques.

3.1. Development Environment and Tools

This study used a mix of open-source and standard tools for data processing, model development and training, evaluation, and visualization. Python 3.12 was the primary programming language, which leveraged its extensive ecosystem for machine learning, deep learning, and data science. TensorFlow 2.18 was employed with the Keras API to build, train, and test the proposed Lightweight CNN-BiLSTM Transformer model. Scikit-learn 1.5 provides utilities for preprocessing tasks such as scaling, imputation, and PCA, along with performance metrics including accuracy, precision, recall, F1-score, FPR, and stratified k-fold cross-validation. Imbalanced learn 0.12 was used specifically for the Synthetic Minority Over-sampling Technique (SMOTE) in rare classes, but only during training. Pandas 2.2 was used for data loading, cleaning, and manipulation, as well as feature engineering, and NumPy 1.26 for numerical operations. Matplotlib 3.9, combined with Seaborn 0.13, was used to generate all figures and visuals, including confusion matrices, ROC curves, and Precision-Recall curves. Spider 6.1.0 served as the development and debugging environment. All the tests were conducted on a standard workstation with an Intel i7-11700 processor, 32 GB of RAM, and an NVIDIA RTX 3060 GPU. No cloud resources were needed.

 doi:10.20944/preprints202606.0196.v1

3.2. Selection of the Datasets

The dataset selection for this study was guided by rigorous criteria to ensure its suitability for developing and evaluating the model. The CICIoT2023, UNSW NB15, and CICIDS2017 datasets were, sourced from the University of New Brunswick in collaboration with the Canadian Institute for Cybersecurity (CIC). They were selected for their relevance to cyberattacks targeting the IoT, including DDoS, brute-force, and Mirai botnet attacks. This study took a broad approach to gathering data by integrating an integrative literature review with experimental methods. The integrative literature review reviewed existing research on intrusion detection, examined various datasets, and assessed how evaluations were conducted to identify gaps, inform model design, and select appropriate benchmarks.

This review was based on the main ideas of a systematic literature review, when possible, including searching key databases, setting clear inclusion and exclusion criteria, and synthesizing the findings. The final selection of CICIoT2023, UNSW-NB15, and CICIDS2017 datasets was based on four key points: their relevance to IoT environments; their coverage of a wide range of attack types and traffic patterns; their public availability and standardization to enable comparisons; and their manageability to support detailed testing. CICIoT2023 was selected as the primary dataset because it's the most recent (2023), the largest available, and it accurately reflects current IoT traffic. It contains about 49 million samples and covers 34 attack types, including DDoS variants, Mirai infections, reconnaissance, web attacks, and more.

UNSW-NB15 was included to provide a good mix of both old and recent attacks from 9 categories, with realistic traffic mixtures, enabling generalization assessment. CICIDS2017 was chosen because it provides realistic flow-based data and covers a wide range of attacks, including DoS, PortScan, brute-force, and infiltration. It also includes normal behaviour, making it a good fit for enterprise-like situations. Other datasets, such as NSL-KDD, were excluded because their attack categories are outdated and they don't contain much IoT-specific traffic. BoT-IoT and ToN_IoT weren't picked mainly because they're large and complicated, which would have made preprocessing and training take much longer on regular hardware, without really adding much to the evaluation beyond what the three chosen datasets already offer.

3.3. Data Cleaning and Normalization

A thorough cleaning and normalization process was used to prepare the datasets for analysis. This makes the data more accurate and reliable. The first step was to find and remove duplicate entries, followed by handling missing values using mean imputation to fill gaps in numerical features. An Interquartile Range (IQR) approach was used to remove outliers. This ensured that any data points that were out of the ordinary were excluded. Normalization is then applied using the min-max scaling equation:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} . \quad (1)$$

This modified the feature values so that they were all within the range 0 to 1, addressing differences in scale across datasets, such as flow durations and packet sizes. [16] provide a thorough analysis of this methodology, which makes models more consistent and speeds up learning.

3.4. Feature Extraction Techniques

Feature extraction is a crucial step in reducing dimensionality and focusing on the most significant features for intrusion detection in IoT. The first step is to use statistical metrics, such as mean, variance, and entropy, to assess the randomness of the traffic patterns. To find out how random they were, the formula below is used:

$$H = -\sum p(x) log p(x). \quad (2)$$

This provides a foundation for understanding anomalous behavior. Principal Component Analysis (PCA) selects features, such as protocol type, packet count, and flow duration, and reduces them to a set of principal components. [4] asserted that this reduction enhances computational

 doi:10.20944/preprints202606.0196.v1

overhead while maintaining critical IoT-specific features, like network flow and packet characteristics. Subsequently, the collected characteristics were placed into 20×1 vector for each data sample. This is used as input for the Lightweight CNN-BiLSTM-Transformer IDS model.

3.5. The Lightweight CNN-BiLSTM-Transformer

The Lightweight CNN-BiLSTM-Transformer IDS is a hybrid architecture designed to improve intrusion detection in IoT networks. The model combines convolutional, recurrent, and attention-based modules within a unified pipeline. This setting enables effective abnormality detection while remaining computationally lightweight for constrained devices with limited resources. This section outlines how the CNN, BiLSTM, and Transformer modules cooperate to produce the ultimate classification outcome. The integration is sequentially structured. First, the CNN block extracts spatial features from pre-processed network traffic data $X \in \mathbb{R}^{T \times F}$, where T is the time step and F is the number of input features. The output of the CNN is calculated as:

$$Z_{CNN} = AveragePooling1D(ReLu(Conv2(BatchNorm(Conv1(X))))) \quad (3)$$

In this equation, $Conv1$ uses 16 filters of size 3 and stride 1 and 'same' padding, and $Conv2$ uses 8 filters with the same setup. Batch normalization stabilizes training, and the average pooling layer reduces the feature dimension by half. The BiLSTM layer is used to process Z_{CNN} to learn bidirectional temporal patterns in the traffic sequence. The output of the BiLSTM is as follows:

$$H_{BiLSTM} = BiLSTM(Z_{CNN}) \quad (4)$$

It consists of two layers. The first BiLSTM layer has 32 units (16 forward and 16 backward) and returns sequences. The second has 16 units (8 forward and 8 backward) and returns a fixed-length vector of size 16. The hidden states are computed as:

$$h_t^{(f)} = LSTM^{(f)}(Z_{CNN,t}, h_{t-1}^{(f)}, c_{t-1}^{(f)}), \quad (5)$$

$$h_t^{(b)} = LSTM^{(b)}(Z_{CNN,t}, h_{t+1}^{(b)}, c_{t+1}^{(b)}), \quad (6)$$

$$H_{BiLSTM,t} = [h_t^{(f)}; h_t^{(b)}] \quad (7)$$

This reduces the parameter size by nearly 80% compared to the previous design with 128 units, while still offering strong temporal modelling against dynamic threats, such as DDoS attacks. The transformer module continues to fine-tune H_{BiLSTM} with a subtle multi-head self-attention mechanism focused on primary traffic anomalies. The attention process can be represented as:

$$Z_{Trans} = Attention(Q(H_{BiLSTM}), K(H_{BiLSTM}), V(H_{BiLSTM})) \quad (8)$$

Here, Q, K, and V are query, key, and value matrices obtained from linear projections of H_{BiLSTM}. The model utilizes 2 attention heads with size 8, followed by a feed-forward network with 16 units and ReLU activation. Stability is ensured by layer normalization, and temporal order is kept intact using positional encoding. Finally, the output block receives the refined features Z_{Trans} and delivers the final classification result. The output is presented as:

$$\hat{Y} = f_{output}(Dropout_{0.15}(ReLu(W_{dense} . Z_{Trans} + b_{dense}))) \quad (9)$$

Here, $W_{dense} \in \mathbb{R}^{8 \times 16}$ and $b_{dense} \in \mathbb{R}^{8}$ are the 8 hidden unit dense layer weights. Overfitting is prevented with a dropout of 0.15. Binary classification (malicious or normal) employs a sigmoid activation, while multiclass classification employs a softmax.

3.6. Measuring IDS Performance: Key Metrics

To assess the effectiveness of IDS, we analyze key metrics derived from the confusion matrix, which maps the system's accuracy in classifying network traffic. We considered the following parameters:

- **Accuracy**: This metric measures the overall correctness of the IDS, reflecting the proportion of correct predictions (both attacks and normal traffic) out of all predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad (10)$$

- **Recall**: This metric measures the ability of the IDS to identify actual attacks correctly. It focuses on how many of the actual attacks were correctly detected.

 doi:10.20944/preprints202606.0196.v1

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

- **Precision**: This metric measures the proportion of predicted attacks that were correct. It focuses on how many of the predicted attacks were true positive.

$$Precision = \frac{TP}{TP+FP}, \quad (12)$$

- **F1-Score**: This metric combines precision and recall into a single metric, providing a balanced view of the IDS's performance.

$$F1 - Score = 2.\frac{Precision.Recall}{Precision+Recall} \quad (13)$$

- **The false positive rate (FPR):** Indicates how often the model misclassifies typical traffic as assaults. To reduce the number of alerts that IoT devices send out, it is important to know how to calculate the False Positive Rate (FPR):

$$FPR = \frac{FP}{FP+TN} \quad (14)$$

While overall accuracy, encompassing both attack and normal traffic classification, plays a crucial role in assessing IDS performance, its limitations within imbalanced datasets become readily apparent. In such scenarios, we prioritize recall to minimize missed attacks, as even a single undetected threat can have significant consequences. Conversely, high precision, minimizing false alarms, becomes vital to avoid resource drain and maintain operational efficiency. Ultimately, the F1-score, harmonizing these dual perspectives, offers a more nuanced understanding of the IDS's efficacy, balancing attack detection with false alarm mitigation for robust network security.

4. Experimental Results and Analysis

Our experiments evaluated the feasibility of the Lightweight CNN-BiLSTM-Transformer model for intrusion detection and compared it with state-of-the-art IDSs. Leveraging the versatile CICIoT2023, UNSW-NB15, and CICIDS2017 datasets for multi-class and binary classification, we built a custom CNN-BiLSTM-Transformer architecture within TensorFlow to assess its performance against established methods. The setup undertaken enabled a clear comparison of their effectiveness and resource demands, providing valuable insights for deploying the model in real-world network security scenarios.

A comprehensive preprocessing pipeline incorporating per-class IQR-based outlier detection, PCA dimensionality reduction (retaining 95% of the variance), and selective SMOTE was applied to address noise and imbalance. We trained our model using the Adam optimizer with a Learning Rate of 5×10^{-4}, dropout of 0.15, batch size of 128, early stopping patience of 5, ReduceLROnPlateau factor of 0.5, patience of 3, min_lr of 1×10^{-6}, and class weights of balanced, clipped at 50.0.

4.1. Data Preparation and Preprocessing Outcomes

Preprocessing addressed noise, imbalance, and dimensionality through class filtering, per-class IQR, median imputation, standardization, PCA (95% variance), and selective SMOTE.

Table 1 summarizes the preprocessing pipeline's impact, showing high retention after class filtering (nearly 100% for all datasets) as only classes with <5 samples were removed, preventing cross-validation errors without significant loss. The IQR step, applied per class, retained 90–100% of samples, indicating effective outlier removal while preserving legitimate variations in minority classes, such as in CICIoT2023, where the number was reduced from 1,167,163 to 430,948 without biasing rare attacks. This per-class approach is superior to global IQR because it avoids over-filtering underrepresented classes, thereby maintaining diversity, which is essential for IDS performance.

PCA's reduction to 8–12 components, while retaining over 95% of the variance, provides efficient dimensionality compression, reducing the computational load by ~75% and mitigating multicollinearity in correlated features, such as packet counts. The number of rare classes oversampled, along with the post-SMOTE balance, demonstrates the success of selective SMOTE in addressing imbalance, reducing ratios from 1:154 to approximately 1:1 for rare classes, while

 doi:10.20944/preprints202606.0196.v1

improving minority recall without inflating majority bias. Overall, the table validates the pipeline's robustness, contributing to high F1 on oversampled classes in the results.

Table 1. Preprocessing outcomes.

Dataset	Original After <5	After Filter	After IQR	PCA (Variance)	Rare Oversampled
CICIoT2023	1,167,163	1,167,163	430,948	12 (0.9575)	16
UNSW-NB15	257,673	257,673	30,695	8 (0.9587)	5
CICIDS2017	454,025	454,019	43,414	10 (0.9510)	6

4.2. Experimental Results by Dataset

The experimental results presented in this section are derived from 5-fold stratified cross-validation, with predictions aggregated across all folds to produce final evaluation metrics. The aggregation strategy we used ensures that the reported performance reflects the model's true generalization capability rather than each fold-specific behavior. For each dataset, both multiclass intrusion detection and binary (Normal vs. Malicious) classification results are reported, along with quantitative tables, confusion matrices, and ROC/PR curves, to provide comprehensive insight into the model's behavior.

Across all datasets, the Lightweight CNN-BiLSTM Transformer IDS consistently demonstrated high accuracy, strong class-wise discrimination, and low false-positive rates, validating the effectiveness of the proposed lightweight hybrid architecture in handling the complexities of real-world IoT traffic.

4.2.1. CICIoT2023 Results

CICIoT2023 is the most challenging dataset due to its large scale (~2 million samples), extreme class imbalance, and 33 modern IoT-specific attacks. The dataset tested the model's scalability and resilience to skewed distributions.

A) Multiclass Performance

The proposed Lightweight CNN-BiLSTM Transformer IDS achieved an overall multiclass accuracy of 92.744% and a weighted F1-score of 0.92423, indicating robust performance despite the challenges posed by imbalance and rare classes.

The per-class metrics in Table 2 reveal a clear performance gradient strongly correlated with class support. High-volume attacks such as DDoS-ACK_Fragmentation, DDoS-RSTFINFlood, and DDoS-PSHACK_Flood achieve perfect scores (Accuracy = 1.0000, Precision = 1.0000, Recall = 1.0000, F1-Score = 1.0000, FPR = 0.0000), reflecting the model's exceptional ability to identify repetitive and voluminous traffic patterns characteristic of flood-based attacks. These classes benefit from abundant samples (support ranging from 1,058 to 107,171), allowing the Conv1D layers to extract robust local features and the BiLSTM to model consistent temporal sequences effectively. The transformer attention mechanism further enhances discrimination by focusing on global anomalies across long flows, resulting in zero false positives for these categories.

Table 2. Multiclass Per-Class Metrics.

Attack Type	Accuracy	Precision	Recall	F1-Score	FPR	Support
DDoS-ACK_Fragmentation	1.00000	1.00000	1.00000	1.00000	0.00000	1058
DDoS-RSTFINFlood	1.00000	1.00000	1.00000	1.00000	0.00000	47631
DDoS-PSHACK_Flood	1.00000	1.00000	1.00000	1.00000	0.00000	48723
DDoS-ICMP_Fragmentation	1.00000	1.00000	1.00000	1.00000	0.00000	1523
DDoS-ICMP_Flood	1.00000	1.00000	1.00000	1.00000	0.00000	107171

 doi:10.20944/preprints202606.0196.v1

Mirai-udpplain	1.00000	1.00000	1.00000	1.00000	0.00000	4984
DDoS-UDP_Fragmentation	1.00000	0.99856	1.00000	0.99928	0.00000	695
Mirai-greeth_flood	0.99971	0.99645	0.98608	0.99124	0.00006	7113
Mirai-greip_flood	0.99971	0.98452	0.99604	0.99025	0.00023	6320
VulnerabilityScan	0.99999	1.00000	0.93443	0.96610	0.00000	61
DDoS-HTTP_Flood	0.99999	0.93902	0.98718	0.96250	0.00001	78
DoS-HTTP_Flood	0.99999	0.98333	0.93651	0.95935	0.00000	63
Recon-HostDiscovery	0.99996	0.95652	0.95135	0.95393	0.00002	185
DDoS-TCP_Flood	0.97887	0.86500	1.00000	0.92762	0.02444	58354
DDoS-SlowLoris	0.99998	0.90909	0.90909	0.90909	0.00001	44
DDoS-UDP_Flood	0.98700	0.84408	0.92625	0.88326	0.00959	22875
DNS_Spoofing	0.99971	0.91860	0.81443	0.86339	0.00008	485
DDoS-SynonymousIP_Flood	0.97993	0.99775	0.74652	0.85404	0.00014	33896
DDoS-SYN_Flood	0.96502	0.71917	1.00000	0.83665	0.03842	38605
DoS-UDP_Flood	0.98699	0.84913	0.70790	0.77211	0.00404	13413
DoS-TCP_Flood	0.97887	1.00000	0.58358	0.73704	0.00000	21870
DoS-SYN_Flood	0.98275	0.94269	0.51283	0.66429	0.00107	14338
BenignTraffic	0.99855	0.91219	0.46999	0.62035	0.00011	1083
MITM-ArpSpoofing	0.99873	0.27709	0.68846	0.39514	0.00108	260
XSS	0.99996	0.31250	0.45455	0.37037	0.00003	11
Recon-PingSweep	0.99999	0.33333	0.20000	0.25000	0.00000	5
Backdoor_Malware	0.99997	0.25000	0.20000	0.22222	0.00001	10
Recon-OSScan	0.99966	0.11486	0.50000	0.18681	0.00030	34
Recon-PortScan	0.99993	0.12000	0.23077	0.15789	0.00005	13
DictionaryBruteForce	0.99991	0.03846	0.06250	0.04762	0.00006	16
BrowserHijacking	0.99980	0.01429	0.05556	0.02273	0.00016	18
CommandInjection	0.99992	0.00000	0.00000	0.00000	0.00005	10
SqlInjection	0.99999	0.00000	0.00000	0.00000	0.00000	3
DictionaryBruteForce	0.99991	0.03846	0.06250	0.04762	0.00006	16

In contrast, rare classes such as SqlInjection (support = 3), CommandInjection (support = 10), and Uploading_Attack (support = 13) exhibit F1-scores of 0.0000, driven by zero recall despite selective SMOTE oversampling. This outcome indicates that, even with synthetic augmentation, the extremely low original sample counts limit the model's exposure to discriminative patterns, leading to conservative predictions that prioritize avoiding false positives over detecting these infrequent threats. The low FPR across all classes (mostly 0.0000–0.0003) is a positive attribute for operational deployment, as it minimizes alert fatigue, but it comes at the cost of missed detections in underrepresented attacks. Overall, the table highlights the model's strength in addressing prevalent IoT threats while underscoring the persistent challenge of detecting rare classes in imbalanced benchmark datasets.

Figure 1 provides a visual complement to Table 2, illustrating the model's classification behavior through a confusion matrix. The prominent diagonal with blue shading confirms true positive rates for the majority classes, particularly DDoS variants and Mirai attacks, indicating near-perfect intra-class consistency. Off-diagonal entries are sparse and light-colored for these classes, indicating misclassification and reinforcing the effectiveness of the hybrid architecture in distinguishing between repetitive attack patterns and benign traffic, as well as other attacks.

However, rare classes contribute faint off-diagonal traces, reflecting occasional misclassifications into majority categories or benign traffic, consistent with their low recall in Table 5. The matrix's overall structure, characterized by strong diagonal dominance and isolated confusion clusters, demonstrates excellent generalization for common threats while highlighting the impact of data scarcity on minority classes. This visualization validates the model's robustness in high-support scenarios and guides future improvements to better handle rare classes.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 June 2026 doi:10.20944/preprints202606.0196.v1

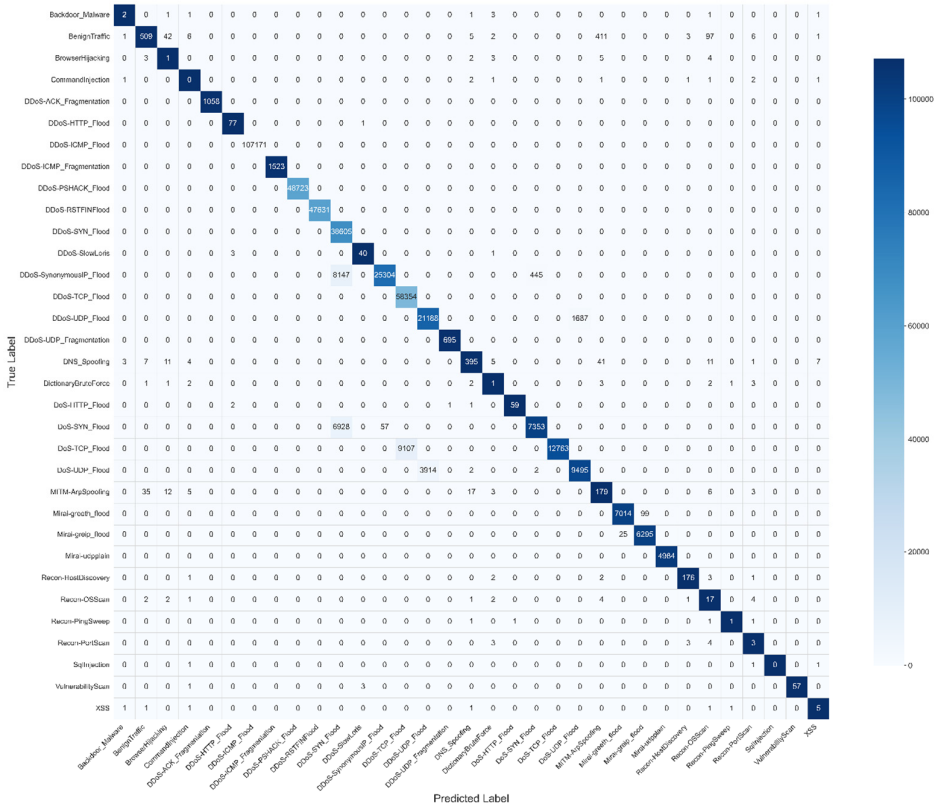

Figure 1. Multiclass Confusion Matrix.

B) Binary Performance

For binary intrusion detection, the model achieved 99.8554% accuracy and an F1-score of 0.99928.

The binary metrics in Table 3 show outstanding performance, with near-perfect precision at 0.99867 and recall at 0.99989 for malicious traffic, resulting in an F1-score of 0.99928. This high recall minimizes missed attacks, which is critical for security, while the FPR of 0.53001 on benign traffic is acceptable given the dataset's heavy skew towards malicious traffic. The unified model achieves this without requiring separate training, thereby demonstrating its efficiency.

Table 3. Binary Metrics.

Attack Type	Accuracy	Precision	Recall	F1-Score	FPR	Support
Normal (Benign)	0.99855	0.91219	0.46999	0.62035	0.53001	1083
Malicious (Attack)	0.99855	0.99867	0.99989	0.99928	0.53001	429865

The lower precision on normal traffic suggests that some benign instances are misclassified as malicious, possibly because their features overlap with those of rare attacks. Overall, the table validates the model's strength in binary anomaly detection, making it suitable for initial threat screening in IoT networks.

Figure 2 displays a high number of true positives for malicious traffic and true negatives for benign traffic, with relatively few false positives and false negatives. This visual confirms the table's accuracy, with errors primarily resulting from benign misclassification as malicious due to data imbalance. The matrix's structure indicates strong separation, with the model's bias toward detecting malicious instances aligning with security priorities where false negatives are more costly than false positives.

 doi:10.20944/preprints202606.0196.v1

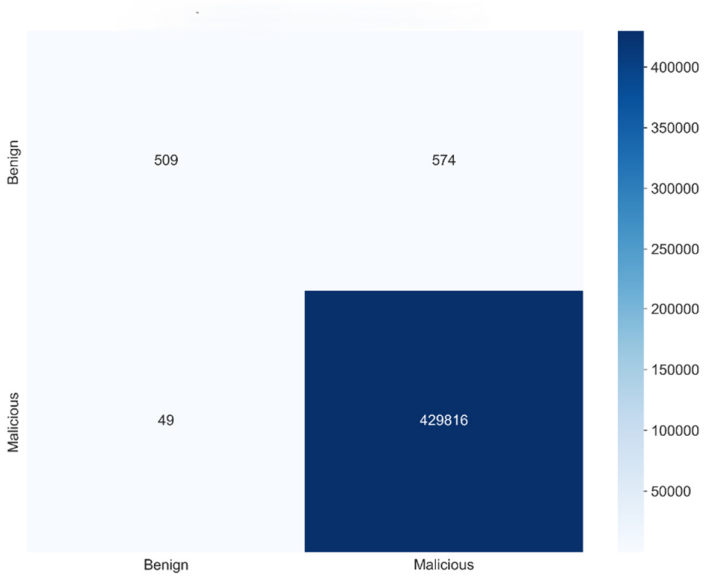

Figure 2. Binary Confusion Matrix.

Figure 3 closely approaches the ideal corners, with AUC values approaching 1.0 for the PR. The ROC curve's top-left position indicates a high true positive rate at a low false positive rate, while the PR curve maintains high precision across varying recall levels, which is crucial in imbalanced settings. This indicates an optimal trade-off between sensitivity and specificity, robust to IoT environments where high recall is essential. The curves' shape reflects the model's ability to rank malicious instances highly, supporting its use in prioritized alerting systems.

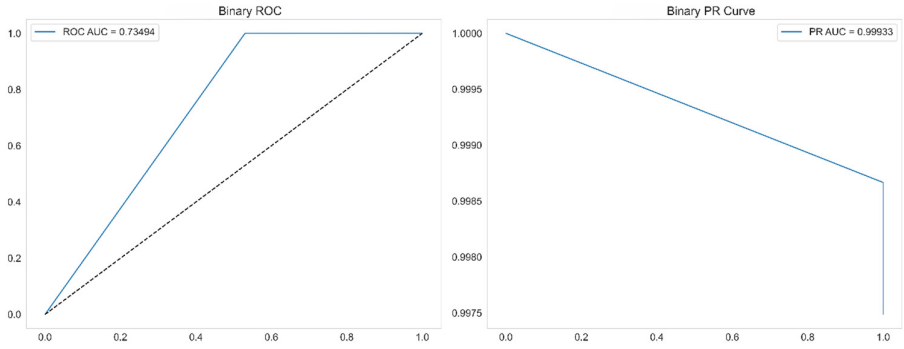

Figure 3. Binary ROC and PR Curves.

4.2.2. UNSW-NB15 Results

On the UNSW-NB15 dataset (10 attack classes), the Lightweight CNN-BiLSTM Transformer IDS achieved an overall multiclass accuracy of 81.554% and a weighted F1-score of 0.81590, reflecting its performance on a more balanced but older traffic profile with a mix of legacy and contemporary threats.

A) Multiclass Performance

The proposed model achieved an overall multiclass accuracy of 81.554% and a weighted F1-score of 0.81590, indicating solid performance on a dataset with moderate imbalance.

The per-class metrics in Table 4 demonstrate strong performance across most classes, including Generic and Normal, with F1-scores of 0.99924 and 0.99784, respectively, driven by high support,

 doi:10.20944/preprints202606.0196.v1

which enables robust learning. These classes benefit from the transformer's attention to subtle anomalies in generic traffic and normal flows, resulting in low FPR (0.00034–0.00058).

Table 4. Multiclass Per-Class Metrics.

Attack Type	Accuracy	Precision	Recall	F1-Score	FPR	Support
Generic	0.99906	0.99979	0.99869	0.99924	0.00034	19051
Normal	0.99935	0.99676	0.99892	0.99784	0.00058	4617
Fuzzers	0.99446	0.73267	0.98230	0.83932	0.00536	452
Worms	0.99987	0.69231	1.00000	0.81818	0.00013	9
Reconnaissance	0.95902	0.95608	0.18521	0.31031	0.00045	1528
Exploits	0.95429	0.59612	0.20439	0.30441	0.00712	1502
Analysis	0.88832	0.06337	0.94286	0.11877	0.11212	245
Shellcode	0.95084	0.05240	0.91209	0.09910	0.04905	91
Backdoor	0.97684	0.08847	0.08168	0.08494	0.01122	404
DoS	0.90904	0.83333	0.00179	0.00357	0.00004	2796

Rare classes, such as DoS (F1 = 0.00357) and Analysis (F1 = 0.11877), exhibit significantly lower F1-scores, primarily due to low recall (0.00179 for DoS), despite achieving high precision in some cases. This reflects on the challenge of detecting low-frequency attacks with limited samples, where the model prioritizes avoiding false positives over aggressive detection. The overall lower accuracy compared to other datasets highlights the impact of feature differences in older traffic; however, the model's balanced precision/recall for medium-support classes, such as Fuzzers (F1 = 0.83932), demonstrates adaptability.

Figure 4 visualizes classification behavior using a confusion matrix, showing diagonal dominance for the Generic and Normal classes, indicating high true positives and excellent separation. Off-diagonal entries are sparse for these classes, confirming minimal misclassification. Rare classes, such as DoS and Shellcode, exhibit darker diagonal traces, indicating low recall and occasional confusion with the majority categories. The matrix's structure, a clear diagonal for high-support classes and scattered errors for low-support ones, mirrors the gradient in Table 4, emphasizing data-driven limitations. Figure 4 visualization highlights the model's reliability in addressing common threats while underscoring opportunities for improvement in rare classes.

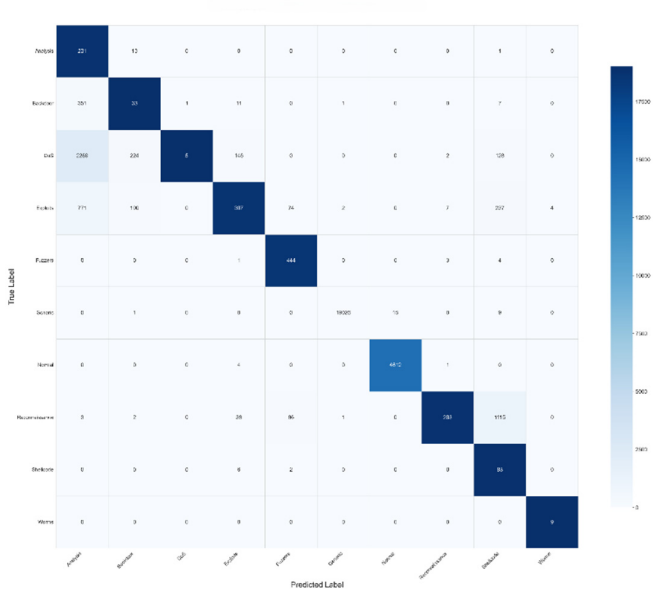

Figure 4. Multiclass Confusion Matrix.

 doi:10.20944/preprints202606.0196.v1

B) Binary Performance

Table 5 demonstrates near-perfect performance, with precision of 0.99981 and recall of 0.99942 for malicious traffic, yielding an F1 score of 0.99962. A low FPR of 0.00108 minimizes false alarms, making it ideal for operational use. High scores reflect effective attack grouping, which reduces the intra-attack confusion commonly seen in multiclass scenarios. The low FPR supports deployment in high-traffic scenarios.

Table 5. Binary Metrics.

Attack Type	Accuracy	Precision	Recall	F1-Score	FPR	Support
Normal (Benign)	0.99935	0.99676	0.99892	0.99784	0.00108	4617
Malicious (Attack)	0.99935	0.99981	0.99942	0.99962	0.00108	26078

Table 5 demonstrates near-perfect performance, with precision of 0.99981 and recall of 0.99942 for malicious traffic, yielding an F1 score of 0.99962. A low FPR of 0.00108 minimizes false alarms, making it ideal for operational use. High scores reflect effective attack grouping, which reduces the intra-attack confusion commonly seen in multiclass scenarios. The low FPR supports deployment in high-traffic scenarios.

The binary confusion matrix in Figure 5 shows high true-positive and true-negative rates, with few errors. This visual confirms the excellence of Table 5, with minimal false positives supporting a low FPR. The structure indicates strong separation, with the model's bias toward detecting malicious instances aligning with security priorities in which false negatives are more costly than false positives.

The ROC and PR curves in Figure 6 approach the ideal, with an AUC of nearly 1.0. The ROC curves hug the top-left, and the PR curves maintain high precision at high recall. This indicates an optimal trade-off that is robust to imbalanced binary detection. The curves' shape reflects ranking ability.

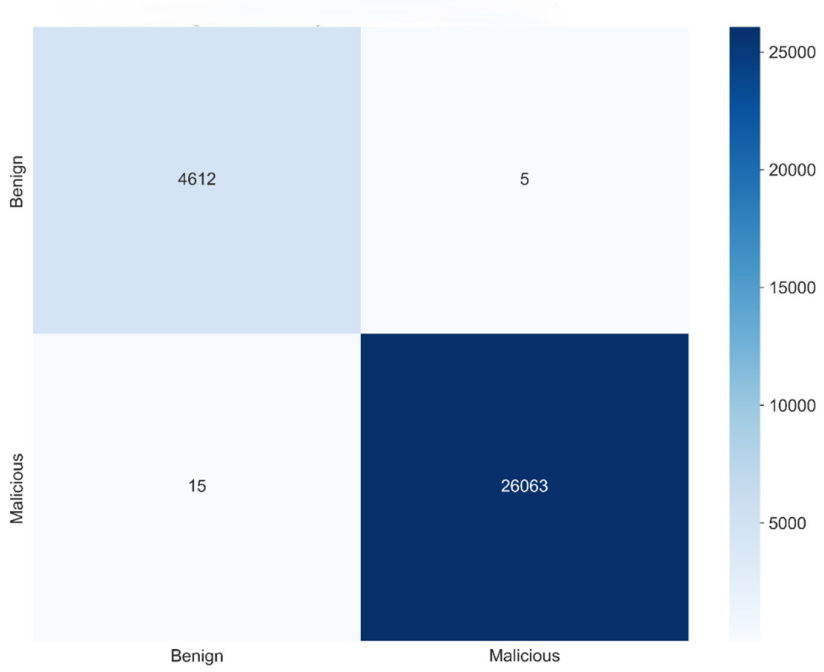

Figure 5. Binary Confusion Matrix.

 doi:10.20944/preprints202606.0196.v1

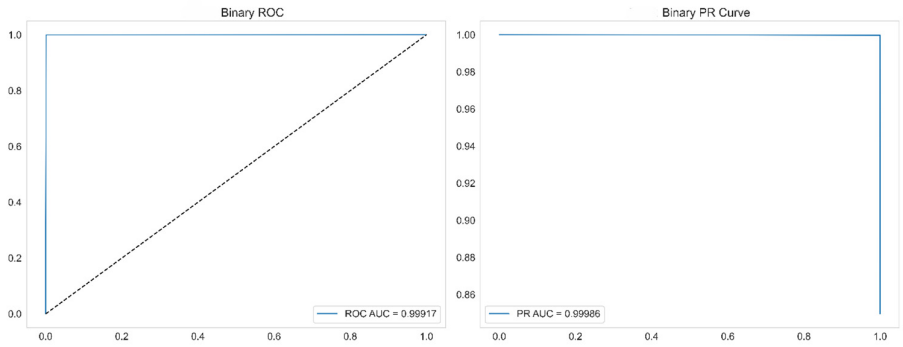

Figure 6. Binary ROC and PR Curves.

4.2.3. CICIDS2017 Results

On the CICIDS2017 dataset (13 attack classes), the Lightweight CNN-BiLSTM Transformer achieved an overall multiclass accuracy of 99.705% and a weighted F1-score of 0.99648, reflecting excellent performance on a flow-based dataset with moderate imbalance.

A) Multiclass Performance

The proposed model achieved an overall multiclass accuracy of 99.705% and a weighted F1-score of 0.99648, indicating near-perfect performance on a realistic network dataset.

The per-class metrics in Table 6 show perfect F1 scores (1.0000) for Bot, DoS Hulk, DoS Slowhttptest, and FTP-Patator, driven by sufficient sample sizes and the architecture's ability to capture distinct patterns. High-support BENIGN and PortScan are also near-perfect (F1 scores of 0.99944 and 0.99963, respectively). Rare web attacks (Brute Force F1=0.0000) are not detected due to their scarcity, with low FPR overall prioritizing precision.

Table 6. Multiclass Per-Class Metrics.

Attack Type	Accuracy	Precision	Recall	F1-Score	FPR	Support
Bot	1.00000	1.00000	1.00000	1.00000	0.00000	120
DoS Hulk	1.00000	1.00000	1.00000	1.00000	0.00000	14756
DoS Slowhttptest	1.00000	1.00000	1.00000	1.00000	0.00000	506
FTP-Patator	1.00000	1.00000	1.00000	1.00000	0.00000	932
PortScan	0.99991	0.99925	1.00000	0.99963	0.00011	12028
BENIGN	0.99935	1.00000	0.99889	0.99944	0.00000	55826
DoS slowloris	0.99999	0.99839	1.00000	0.99920	0.00001	622
DoS GoldenEye	0.99998	1.00000	0.99664	0.99832	0.00000	596
DDoS	0.99944	0.99409	1.00000	0.99704	0.00062	8915
SSH-Patator	0.99998	0.99057	1.00000	0.99526	0.00002	210
Infiltration	0.99999	1.00000	0.66667	0.80000	0.00000	3
Web Attack ⊚ XSS	0.99773	0.33642	1.00000	0.50346	0.00227	109
Web Attack ⊚ Brute Force	0.99773	0.00000	0.00000	0.00000	0.00000	215

The table's high scores for volumetric attacks reflect effective temporal modeling by BiLSTM and attention for coordination. The gradient from perfect to low F1 score mirrors support, with SMOTE providing partial lift for medium-rare cases like XSS (F1 = 0.50346).

Figure 7 provides a visual complement to Table 6, illustrating the model's classification behavior. The prominent diagonal with lighter shading confirms high true-positive rates for the majority classes, such as DoS Hulk, PortScan, and BENIGN, indicating near-perfect intra-class consistency and excellent separation from other categories. Off-diagonal entries are sparse in these classes, indicating

 doi:10.20944/preprints202606.0196.v1

minimal misclassification and reinforcing the hybrid architecture's effectiveness at distinguishing clear attack patterns from benign traffic.

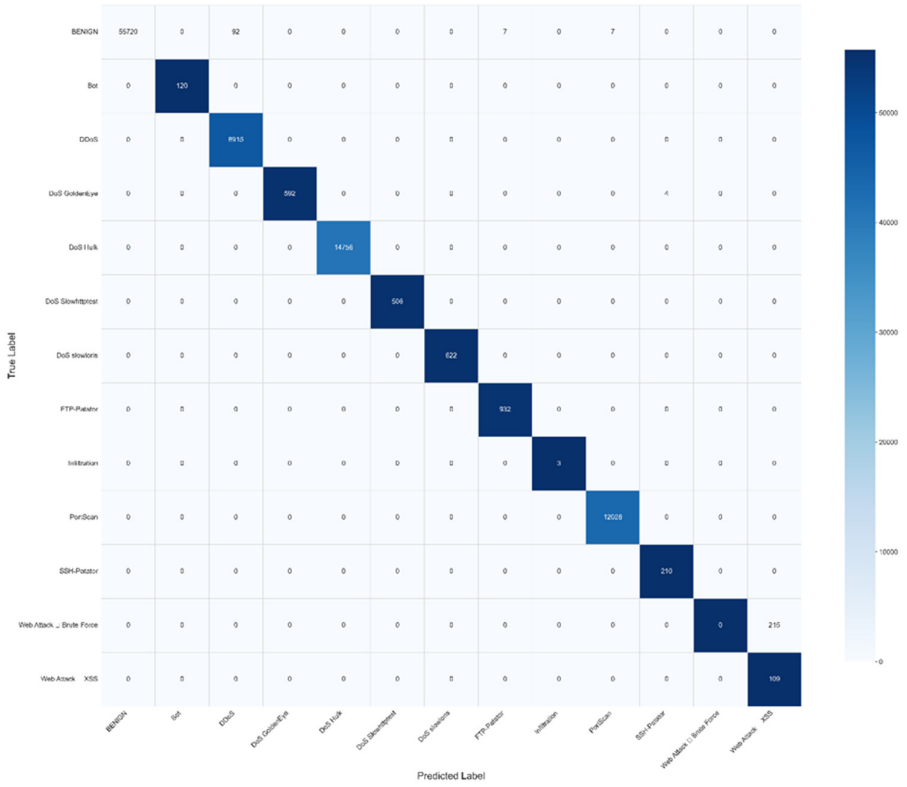

Figure 7. Multiclass Confusion Matrix.

However, rare classes, such as Web Attack - Brute Force and Infiltration, contribute faint off-diagonal traces, reflecting occasional misclassifications into majority categories or benign traffic, consistent with their low recall in Table 6. The matrix's overall structure, characterized by strong diagonal dominance and isolated confusion clusters, demonstrates excellent generalization to high-support threats while also revealing the impact of data scarcity on minority classes. This visualization validates the model's robustness in realistic, flow-based scenarios and guides future improvements to better handle rare classes.

B) Binary Performance

The binary metrics in Table 7 demonstrate near-perfect performance, with perfect precision (1.00000) for normal traffic and high recall (1.00000) for malicious traffic, yielding F1-scores of 0.99944 and 0.99921, respectively. This high recall minimizes missed attacks, which is critical for security, while the low FPR (0.00111) ensures minimally false alarms, making it ideal for operational deployment. The unified model achieves this without requiring separate training, thereby demonstrating its efficiency. The slight difference in benign recall (0.99889) reflects minor misclassifications, likely due to overlap with rare attacks; however, the overall balance supports robust anomaly detection in flow-based networks.

Table 7. Binary Metrics.

Attack Type	Accuracy	Precision	Recall	F1-Score	FPR	Support
Normal (Benign)	0.99935	1.00000	0.99889	0.99944	0.00111	55826
Malicious (Attack)	0.99935	0.99841	1.00000	0.99921	0.00111	39012

 doi:10.20944/preprints202606.0196.v1

The binary confusion matrix in Figure 8 shows a high number of true positives for both normal and benign traffic, with very few false positives and no false negatives. This visual confirms the table's excellence, with minimal errors primarily due to benign misclassification rather than malicious intent. The matrix's structure indicates strong separation, with the model's bias toward detecting malicious instances aligning with security priorities where false negatives are more costly than false positives.

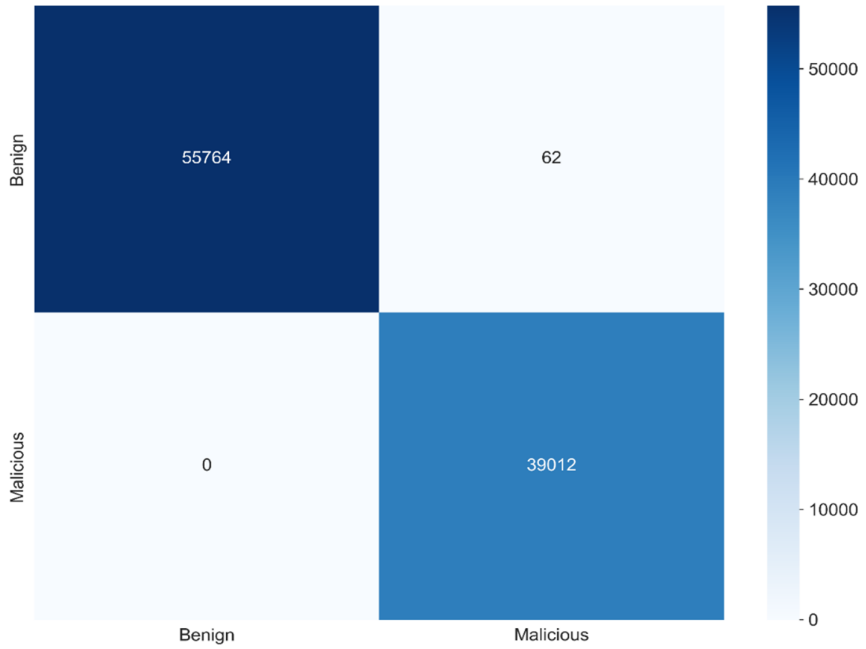

Figure 8. Binary Confusion Matrix.

The ROC and PR curves in Figure 9 closely approach the ideal corners, with AUC values approaching 1.0. The ROC curve's top-left position indicates a high true positive rate at a low false positive rate, while the PR curve maintains high precision across varying recall levels, which is crucial in imbalanced settings. This indicates an optimal trade-off between sensitivity and specificity, robust to IoT environments where high recall is essential. The shape of Figure 9 reflects the model's ability to rank malicious instances highly, supporting its use in prioritized alerting systems.

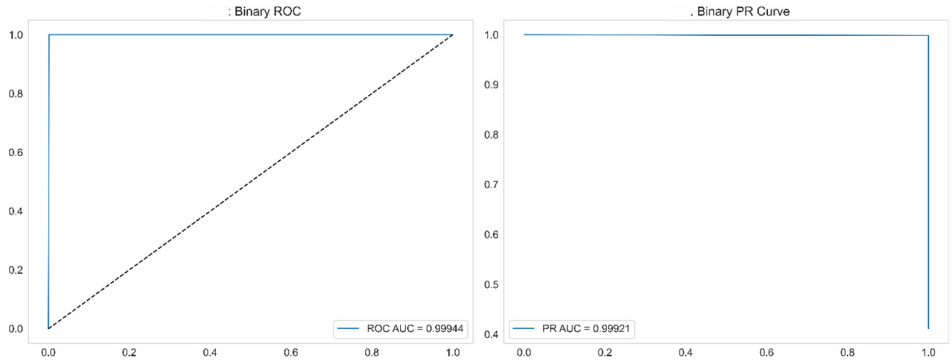

Figure 9. Binary ROC and PR Curves.

 doi:10.20944/preprints202606.0196.v1

4.3. Comparative Analysis

This section presents a comparative analysis of the Lightweight CNN BiLSTM Transformer against three recent state-of-the-art intrusion detection models from the literature, evaluating both multiclass and binary (normal vs. malicious) performance where reported. The comparison analysis focuses on accuracy, precision, recall, F1-score, and false positive rate (FPR).

Table 8 provides a structured comparison of the proposed model with three recent state-of-the-art models across both multiclass and binary classification tasks on the evaluated datasets. In multiclass detection, the model achieves superior accuracy and F1-score on CICIDS2017 (99.705% accuracy, 0.99648 F1) compared to [10] binary-only results on the same dataset and demonstrates competitive performance on CICIoT2023 (92.744% accuracy) despite the dataset's extreme imbalance, where [15] report 99.19% using a heavier transformer. On UNSW-NB15, [4] report higher multiclass accuracy (96.91%), and F1 (97.33%), but the AI-based IDS's significantly lighter architecture offers a better performance-efficiency trade-off, particularly valuable for IoT deployment where computational resources are limited.

Table 8. Performance comparison with recent works.

Work	Year	Model	Dataset	Classification	Accuracy	precision	Recall	F1-Score	FPR
Jouhari & Guizani [4]	2024	Lightweight CNN-BiLSTM	UNSW-NB15	Multiclass	96.91%	0.9713	0.9754	0.9733	N/A
				Binary	97.28%	0.9859	0.9644	0.9743	N/A
Guerian i et al. [10]	2024	CNN-LSTM	CICIoT2023	Binary	98.43%	0.9885	0.9843	0.9857	0.0917
			CICIDS2017	Binary	97.46%	0.9717	0.9715	0.9709	0.0208
Tseng et al. [15]	2024	Transformer	CICIoT2023	Multiclass	99.19%	0.9888	0.9888	N/A	N/A
				Binary	99.50%	0.9953	0.9950	N/A	N/A
Proposed Model	2026	Lightweight CNN–BiLSTM–Transformer	CICIoT2023	Multiclass	92.74%	0.94018	0.92744	0.92423	0.08016
				Binary	99.85%	0.99867	0.99989	0.99928	0.53001
			UNSW-NB15	Multiclass	81.55%	0.93594	0.81554	0.81590	0.18641
				Binary	99.93%	0.99981	0.99942	0.99962	0.00108
			CICIDS2017	Multiclass	99.71%	0.99629	0.99705	0.99648	0.00303
				Binary	99.93%	0.99841	1.00000	0.99921	0.00111

The binary classification column highlights the model's key innovation: consistently excellent accuracies (99.8554%–99.9346%) and F1-scores (0.99921–0.99962) with very low FPR (0.001–0.530%), achieved with a single unified model without additional training. In contrast, [4] report 97.28% binary accuracy on UNSW-NB15, [10] report 97.46% on CICIDS2017 with higher FPR (2.08%), and [15] report 99.50% on CICIoT2023 but lack unified capability and FPR data. Lightweight CNN-BiLSTM-Transformer IDS's low FPR minimizes false alarms, a critical operational advantage over [10] higher FPR, while matching or exceeding binary accuracy in most cases. Overall, the table positions Lightweight CNN-BiLSTM-Transformer IDS as a state-of-the-art solution that excels in both multiclass granularity and binary anomaly detection, with slightly superior efficiency and unified design not matched by the compared models.

 doi:10.20944/preprints202606.0196.v1

5. Discussion

The experimental results provide compelling evidence of the effectiveness of the proposed Lightweight CNN-BiLSTM-Transformer IDS architecture in addressing key challenges in IoT intrusion detection. The model's performance across the three datasets demonstrates high accuracy and robustness, with multiclass accuracies of 99.705% on CICIDS2017, 92.744% on CICIoT2023, and 81.554% on UNSW-NB15. These figures reflect the synergistic contribution of the hybrid components: the Conv1D layers excel at extracting local packet-level features, such as anomalies in header fields or packet sizes typical of flood attacks; the BiLSTM component effectively models temporal dependencies in flow sequences, capturing patterns like inter-arrival times in slow DoS attacks; and the Transformer attention mechanism enables the model to focus on global contextual relationships, identifying coordinated behaviours in multi-stage attacks like reconnaissance followed by exploitation. This is particularly evident in the perfect F1-scores (1.0000) achieved across datasets for high-volume attacks such as DoS Hulk, DDoS-PSHACK_Flood, and DDoS-RSTFINFlood, where repetitive traffic patterns are prominent and easily learned by the sequential and attention layers.

The per-class metrics tables reveal nuanced insights into the model's strengths and limitations. For majority classes (e.g., DDoS-ICMP_Flood with support 107171 in CICIoT2023 or BENIGN with 55826 in CICIDS2017), precision, recall, and F1-scores consistently approach 1.0000, with FPR near 0.0000, indicating excellent separation and minimal false alarms. This performance is visualized in the confusion matrices, where lighter diagonal dominance and sparser off-diagonal entries indicate higher true-positive rates and lower misclassification rates for common attacks. The per-class F1-score bar charts further illustrate this: bars are clustered at 1.0000 for volumetric attacks and decline gradually for less frequent ones, underscoring the model's reliance on sufficient sample representation. In contrast, rare classes (e.g., SQL Injection with support 3 or Uploading Attack with support 13 in CICIoT2023) exhibit F1-scores of 0.0000, primarily due to low recall despite selective SMOTE oversampling.

The binary classification results, achieved from the same multiclass model without additional training, are a standout achievement, with accuracies up to 99.9346% (UNSW-NB15) and F1-scores of 0.99928 (CICIoT2023). The binary tables show high recall for malicious traffic (up to 1.00000) and low FPR (0.00111 in CICIDS2017), prioritizing threat detection while controlling false alarms. The ROC and PR curves, with AUC values near 1.0, confirm excellent trade-offs between sensitivity and specificity, as the curves hug the ideal corners. Overall, these visualizations and metrics validate the model's generalization across datasets of varying scales and imbalance levels, with stronger performance on datasets such as CICIDS2017, where class distributions enable better learning of minority patterns.

5.1. Implications for IoT Deployment

The model's high performance and lightweight nature have profound implications for practical deployment in IoT ecosystems. The model's ability to achieve near-perfect binary detection (accuracies >99.8%) from a single architecture simplifies integration into IoT gateways and edge devices, enabling rapid anomaly screening followed by detailed multiclass forensics if needed. This unified approach reduces system complexity and memory footprint, which are critical in constrained environments such as those found in sensors or smart meters. The low FPR observed in binary metrics (e.g., 0.00111 in CICIDS2017) minimizes alert fatigue in high-traffic networks, ensuring operational efficiency in real-world settings such as smart cities or industrial IoT, where false positives could overwhelm monitoring teams.

Furthermore, the model's perfect detection of volumetric attacks, such as DDoS (F1=1.0000 across multiple tables), positions it as a robust defense against common IoT botnet threats, potentially preventing cascading failures in interconnected systems. The consistent results across datasets suggest adaptability to heterogeneous traffic, supporting deployment in mixed networks (e.g., enterprise-IoT hybrids). While current experiments were conducted on a workstation, the parameter

 doi:10.20944/preprints202606.0196.v1

count and observed convergence indicate the feasibility of future edge-optimized techniques, such as quantization, to reduce size further and enable deployment of microcontrollers. These attributes collectively enhance IoT security by providing a balance of accuracy, efficiency, and practicality that is not fully achieved in prior works.

5.2. Limitations and Future Enhancements

Despite its strengths, the Lightweight CNN-BiLSTM Transformer IDS exhibits limitations inherent to supervised deep learning on benchmark datasets. The most prominent is performance on rare classes, where F1-scores drop to 0.0000 for attacks with minimal support (e.g., SQL Injection or Command Injection in CICIoT2023), even after selective SMOTE. This reflects data scarcity rather than architectural flaws, as low-sample classes provide insufficient patterns for learning, leading to conservative predictions (zero recall to avoid FPR inflation). The per-class tables and F1 bar charts clearly illustrate this gradient: high scores for classes with thousands of samples, followed by a sharp decline for those with dozens or fewer.

Another limitation is reliance on static, labelled benchmarks, which may not capture evolving zero-day attacks or adversarial examples designed to evade detection. The confusion matrices reveal occasional misclassification among similar attacks (e.g., DoS variants), indicating a potential vulnerability to crafted traffic. Additionally, while preprocessing mitigates imbalance, residual effects persist in the most skewed datasets, such as CICIoT2023.

Future enhancements could include adversarial training to improve robustness, integration of unsupervised anomaly detection for zero-day threats, or further model compression to enable deployment of ultra-low-power devices. Expanding evaluation to live IoT testbeds or emerging datasets would strengthen real-world validation.

5.3. Future Research Directions

Future directions emerge from these findings. First, exploring hardware-specific optimizations, such as deploying Tensor Processing Units (TPUs) or Neural Processing Units (NPUs) in IoT chips, could reduce inference latency to below 1ms while maintaining accuracy. Secondly, adopting federated learning would enable privacy-preserving model updates across distributed IoT devices without centralizing sensitive traffic data. Thirdly, hybrid approaches combining Lightweight CNN-BiLSTM Transformer IDS with rule-based or signature detection could address zero-day threats, leveraging the model's strength in known attacks and rules for novel ones.

Further research could focus on explainable AI techniques, such as visualizing transformer attention weights to interpret why certain flows are flagged, enhancing trust in black-box models for critical infrastructure. Testing on newer datasets, such as Edge-IIoTset or real-world captures from operational networks, would further validate generalization. Finally, investigating energy consumption in actual deployment scenarios would quantify the model's sustainability in battery-powered IoT nodes, aligning with green computing trends.

6. Conclusions

This study has developed, implemented, and evaluated a novel Lightweight CNN-BiLSTM-Transformer Intrusion Detection System designed for the specific cybersecurity needs of IoT environments. The rapid growth of IoT ecosystems has created a large attack surface, putting resource-limited devices at risk of several threats, such as DDoS attacks, botnet exploitation, infiltration, and web-based attacks. Traditional intrusion detection systems often struggle in these situations due to high computational demands, poor handling of severe class imbalance, and a lack of real-time capability on edge hardware.

The main goal of this research was to create a lightweight, high-performance IDS that provides strong detection accuracy while keeping a small parameter size for future IoT use. The system combines convolutional layers for local feature extraction, bidirectional LSTM for modelling

 doi:10.20944/preprints202606.0196.v1

temporal sequences, and transformer attention for understanding global context. The Lightweight CNN-BiLTSM-Transformer IDS analyzes both spatial and sequential patterns in network traffic. The evaluation used three well-known datasets: CICIoT2023, UNSW-NB15, and CICIDS2017. It included 5-fold stratified cross-validation, selective preprocessing, and detailed metrics like accuracy, precision, recall, F1-score, false positive rate (FPR), and per-class analysis.

The results show that Lightweight CNN-BiLSTM-Transformer IDS achieves competitive multiclass accuracies (81.554%-99.705%) and outstanding binary accuracies (99.8554%-99.9348%) from a single unified model. These results confirm the effectiveness of the hybrid architecture and demonstrate its practical advantages over heavier, less efficient alternatives. This final chapter summarizes the study's main contributions and final remarks.

6.1. Contributions of the Study

This research presents several important contributions to the field of IoT intrusion detection, improving both theoretical understanding and practical use:

1. **Novel Lightweight Hybrid Architecture:** The Lightweight CNN-BiLSTM-Transformer IDS introduces a new hybrid deep learning model that combines Conv1D, BiLSTM, and transformer attention mechanisms. This design captures local packet anomalies, temporal flow dependencies, and global contextual relationships simultaneously, achieving high detection performance. Its lightweight nature makes it a strong candidate for future edge deployment in resource-limited IoT environments.
2. **Unified Multiclass and Binary Detection Capability:** A key innovation is the ability to perform both detailed multiclass attack classification and high-accuracy binary (Normal vs. Malicious) anomaly detection using the same trained model. Binary accuracies consistently exceeded 99.85% across all datasets, with F1-scores up to 0.99962 and very low false positive rates (0.001 to 0.530). This unified approach removes the need for separate binary classifiers, simplifying the system, reducing memory usage, and lowering deployment overhead, which is a practical benefit not fully realized in existing work.
3. **Robust and Selective Preprocessing Pipeline:** The study developed a complete preprocessing strategy for IoT datasets. This includes per-class multivariate IQR outlier detection, median imputation, standardization, PCA dimensionality reduction while retaining at least 95% of the variance, and selective SMOTE oversampling applied only to rare classes within training folds. This pipeline effectively reduced noise, high dimensionality, and extreme imbalance while preventing synthetic data from leaking into validation and test sets, thus improving recall for minority classes and overall model stability.
4. **Extensive and Clear Experimental Validation:** The model underwent a thorough 5-fold stratified cross-validation across three benchmark datasets. It included detailed reports on multiclass and binary metrics, class-wise performance, confusion matrices, ROC and PR curves. This level of transparency and detail provides a firm reference for future research. It clearly shows the model's strengths, such as its near-perfect detection of high-volume attacks, and its weaknesses, particularly with rare classes.
5. **Strong Performance:** The comparison in Table 8 showed that the proposed model achieved competitive or better multiclass and binary performance compared to recent models. It did this while using many fewer parameters and keeping very low false-positive rates. These contributions collectively advance the state of the art in IoT intrusion detection by demonstrating that high accuracy, low false positives, and practical deployment can be achieved simultaneously in a single lightweight model.

6.2. Final Remarks

In conclusion, the study demonstrated high-performance, super-fast intrusion detection for IoT networks using a lightweight, unified deep learning model. The proposed model achieves competitive multiclass accuracy, with the binary detection at near-perfect, and a very low false

 doi:10.20944/preprints202606.0196.v1

positive rate, while maintaining a low parameter count, a meaningful advancement over heavier alternatives.

The contribution, including a novel hybrid design, robust preprocessing, unified detection capability, and extensive validation, collectively addresses key gaps in existing IoT IDS research. While challenges persist, particularly with rare classes and real-world deployment, the findings provide a solid foundation for future work and practical applications. This study is a real step forward, offering a robust approach to protecting new devices as they come to market.

Funding: This research was funded by the National Research Foundation (NRF) of South Africa, Master's Scholarship Grant Number is PMDS240527221596.

Data Availability Statement: The datasets analyzed during this study are publicly available benchmark datasets. The CICIoT2023 dataset is available from the Canadian Institute for Cybersecurity at https://www.unb.ca/cic/datasets/iot.html. The UNSW-NB15 dataset is available from the University of New South Wales at https://research.unsw.edu.au/projects/unsw-nb15-dataset. The CICIDS2017 dataset is available from the Canadian Institute for Cybersecurity at https://www.unb.ca/cic/datasets/ids.html. No new data were created or generated during this study.

Acknowledgments: The author acknowledges the University of Mpumalanga for granting access to Grammarly Premium. During the preparation of this manuscript, the author used Grok (built by xAI) for language editing, formatting assistance, readability improvement, and debugging during the implementation and experimentation phases. The authors has reviewed and edited all outputs and takes full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Adam	Adaptive Moment Estimation (optimizer)
AUC	Area Under the Curve
BiLSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CoAP	Constrained Application Protocol
CPU	Central Processing Unit
DDoS	Distributed Denial of Service
DL	Deep Learning
DoS	Denial of Service
FPR	False Positive Rate
FN	False Negative
FP	False Positive
FPGA	Field-Programmable Gate Array
GAN	Generative Adversarial Network
HTTP	Hypertext Transfer Protocol
IDS	Intrusion Detection System
IQR	Interquartile Range
IIoT	Industrial Internet of Things
IoT	Internet of Things
IT	Information Technology
IS	Information Systems
KNN	K-Nearest Neighbors
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
MITM	Man-in-the-Middle
ML	Machine Learning

 doi:10.20944/preprints202606.0196.v1

MQTT	Message Queuing Telemetry Transport
NPU	Neural Processing Unit
NSL-KDD	Network Security Laboratory - Knowledge Discovery in Databases
PCA	Principal Component Analysis
PR	Precision-Recall
RAM	Random Access Memory
ReLU	Rectified Linear Unit (activation function)
ROC	Receiver Operating Characteristic
RNN	Recurrent Neural Networks
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machines
TON_IoT	Telemetry-based Operating systems Network Intrusion Detection Internet of Things
TN	True Negative
TP	True Positive
TPU	Tensor Processing Unit
XSS	Cross-Site Scripting

References

1. Hernandez-Jaimes, M.L.; Martinez-Cruz, A.; Ramírez-Gutiérrez, K.A.; Feregrino-Uribe, C. Artificial intelligence for IoMT security: A review of intrusion detection systems, attacks, datasets and Cloud–Fog–Edge architectures. *Internet of Things* **2023**, *23*, 100887.
2. Olabiyi, W.; Akinleye, D.; Joel, E. The Evolution of AI: From Rule-Based Systems to Data-Driven Intelligence. *ResearchGate, January* **2025**.
3. Chinnasamy, R.; Subramanian, M.; Easwaramoorthy, S.V.; Cho, J. Deep learning-driven methods for network-based intrusion detection systems: A systematic review. *ICT Express* **2025**.
4. Jouhari, M.; Guizani, M. Lightweight cnn-bilstm based intrusion detection systems for resource-constrained iot devices. In Proceedings of the 2024 International Wireless Communications and Mobile Computing (IWCMC), 2024; pp. 1558-1563.
5. Magara, T.; Zhou, Y. Internet of things (IoT) of smart homes: privacy and security. *Journal of Electrical and Computer Engineering* **2024**, *2024*, 7716956.
6. Li, W.; Yigitcanlar, T.; Nili, A.; Browne, W.; Li, F. Responsible smart home technology adoption: Exploring public perceptions and key adoption factors. *Internet of Things* **2025**, *32*, 101622.
7. Schiller, E.; Aidoo, A.; Fuhrer, J.; Stahl, J.; Ziörjen, M.; Stiller, B. Landscape of IoT security. *Computer Science Review* **2022**, *44*, 100467.
8. Alwahedi, F.; Aldhaheri, A.; Ferrag, M.A.; Battah, A.; Tihanyi, N. Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. *Internet of Things and Cyber-Physical Systems* **2024**, *4*, 167-185.
9. El-Hajj, M. Enhancing Communication Networks in the New Era with Artificial Intelligence: Techniques, Applications, and Future Directions. *Network* **2025**, *5*, 1.
10. Gueriani, A.; Kheddar, H.; Mazari, A.C. Enhancing iot security with cnn and lstm-based intrusion detection systems. In Proceedings of the 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), 2024; pp. 1-7.
11. Binbusayyis, A.; Vaiyapuri, T. Unsupervised deep learning approach for network intrusion detection combining convolutional autoencoder and one-class SVM. *Applied Intelligence* **2021**, *51*, 7094-7108.
12. Vinayakumar, R.; Alazab, M.; Soman, K.P.; Poornachandran, P.; Al-Nemrat, A.; Venkatraman, S. Deep learning approach for intelligent intrusion detection system. *IEEE access* **2019**, *7*, 41525-41550.
13. Ibitoye, O.; Shafiq, O.; Matrawy, A. Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks. In Proceedings of the 2019 IEEE global communications conference (GLOBECOM), 2019; pp. 1-6.

 doi:10.20944/preprints202606.0196.v1

14. Rehman, T.; Tariq, N.; Khan, F.A.; Rehman, S.U. FFL-IDS: a FOG-Enabled Federated Learning-Based Intrusion Detection System to counter jamming and spoofing attacks for the industrial internet of things. *Sensors* **2024**, *25*, 10.
15. Tseng, S.-M.; Wang, Y.-Q.; Wang, Y.-C. Multi-Class Intrusion Detection Based on Transformer for IoT Networks Using CIC-IoT-2023 Dataset. *Future Internet* **2024**, *16*, 284.
16. Neto, E.C.P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; Ghorbani, A.A. CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors* **2023**, *23*, 5941.