JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 04, Issue No. 01, 2026

ARTICLE TYPE: Research Article

Article ID: JEPCS-2026-0401-0008

A Federated Learning Framework with CNN-BiLSTM-Attention and Explainable AI for Intrusion Detection in IoMT Healthcare Networks

Aleeha Azher[1,*] and Muhammad Usman[1]

[1] Department of Computer Science, University of Central Punjab, Lahore, Pakistan

* Corresponding author: aleehaazher12@gmail.com

Received: March 21, 2026 **Accepted:** June 25, 2026

Abstract- Internet of Medical Things (IoMT) networks generate heterogeneous and privacy-sensitive traffic, while severe class imbalance can further complicate intrusion detection across healthcare institutions. This paper evaluates a federated intrusion detection framework that combines a CNN-BiLSTM-Attention classifier with FedAvg aggregation, node-level SMOTE, and SHAP-based explainability. Three simulated hospital nodes train local models and exchange only model weights, while raw training samples remain local. The framework is evaluated on the full ECU-IoHT dataset and a computationally constrained subset comprising 10 of 169 CICIoT2023 CSV files. On ECU-IoHT, the proposed model achieved 94.02% accuracy, a weighted F1-score of 0.9540, ROC-AUC of 0.9975, MCC of 0.8771, and a false alarm rate of 0.0043. For the rare DoS class, recall reached 0.91, but precision was 0.11, indicating a substantial false-positive trade-off. On the evaluated CICIoT2023 subset, the model achieved 62.13% accuracy, weighted precision of 0.7691, ROC-AUC of 0.9410, and a false alarm rate of 0.0301. In the reported SHAP analysis, Length and Protocol produced the largest displayed effects for the No Attack class. These results show that federated training, local class balancing, and post-hoc explanation can be combined in the evaluated IoMT intrusion detection process while keeping raw samples local. Broader validation requires non-IID clients, evaluation on all 169 CICIoT2023 CSV files, repeated runs, formal privacy mechanisms, and component-level ablations.

Keywords- Explainable artificial intelligence; federated learning; intrusion detection; Internet of Medical Things; SMOTE

1. Introduction

Interconnected medical devices have expanded the amount of clinical and operational data exchanged inside healthcare networks. The same connectivity also increases the attack surface [1]. During the early COVID-19 response, the World Health Organization reported that cyberattacks directed at the organization were more than five times higher than during the same period of the previous year [2]. Such incidents motivate intrusion detection that can operate on heterogeneous medical-device traffic without requiring unrestricted pooling of institutional data. IoMT environments include devices such as heart-rate sensors, blood-pressure monitors, and temperature sensors, often using different protocols and hardware capabilities. Signature-based intrusion detection is limited when attack behavior is previously unseen or changes over time. Machine learning (ML)-based intrusion detection systems (IDSs) can learn traffic patterns directly from data, but their deployment in healthcare introduces additional requirements. In particular, training data may be distributed across institutions, attack classes can be strongly imbalanced, and security teams may need an explanation for high-impact alerts.

Recent papers have addressed this problem partially. Ghourabi [1] combined LightGBM and BERT-Transformer and reported approximately 99% accuracy, with results approaching 100% in some evaluated settings across ECU-IoHT, ToN-IoT, Edge-IIoTset, and EMBER. Mosaiyebzadeh et al. [3] proposed SECIoHT-FL, which combines federated

learning, differential privacy, and SHAP explainability in an IoHT setting and reported 95.48% accuracy. These studies provide strong baselines, but the reviewed literature leaves a narrower question unresolved: how can federated IoMT intrusion detection address severe class imbalance locally while also providing feature-level explanations?

This paper evaluates that question using a CNN-BiLSTM-Attention model trained with FedAvg. The main contributions are as follows:

- A federated IDS is evaluated with three simulated hospital nodes so that raw training samples remain local and only model weights are aggregated.
- SMOTE is applied independently to the training partition at each node before local optimization. This design targets the severe imbalance in ECU-IoHT, where the DoS Attack class contains 639 samples compared with 77,920 Smurf Attack samples.
- SHAP KernelExplainer is applied to the final global model, using a k-means summarized background derived from 500 test instances and explanations for 50 selected test samples.
- The framework is evaluated on ECU-IoHT and on 10 of 169 CICIoT2023 partitions. On ECU-IoHT it achieves 94.02% accuracy, 0.9975 ROC-AUC, and 0.0043 FAR; on the evaluated CICIoT2023 subset it achieves 62.13% accuracy and 0.0301 FAR.

The remainder of the paper is organized as follows. Section 2 reviews ML IDS methods, IoMT security datasets, federated IDS approaches, and work on class imbalance and explainability. Section 3 summarizes the research gaps that motivate the framework. Section 4 introduces the technical components used by the method. Section 5 describes the datasets, preprocessing, federated setup, model architecture, SMOTE procedure, SHAP integration, and evaluation metrics. Section 6 presents the experimental results and limitations. Section 7 concludes the paper.

Table 1 compares representative recent IDS approaches with the proposed framework.

Table 1. *Comparison of related work.*

Ref	Year	Method	Dataset	Acc. (%)	F1	Attacks	Imbalance	FL	XAI
[1]	2022	LightGBM + BERT-Transformer	ECU-IoHT, ToN-IoT, Edge-IIoTset, EMBER	99-100	0.99-1.00	4-14	✗	✗	✗
[5]	2020	LightGBM (SwiftIDS)	KDD99, NSL-KDD, CICIDS2017	99.70	0.997	5	✗	✗	✗
[6]	2021	LightGBM + ADASYN	NSL-KDD, UNSW-NB15, CICIDS2017	99.91	N/R	N/R	✗	✗	✗
[8]	2022	CNN-BiLSTM + Attention + C5.0	KDD CUP 99	95.50	N/R	4	✗	✗	✗
[9]	2022	CNN-LSTM	CIC-IDS2017, UNSW-NB15, WSN-DS	99.64	0.996	6	✗	✗	✗
[13]	2020	RF, KNN, SVM, ANN	Custom EHMS testbed	92.45	N/R	1	✗	✗	✗
[3]	2025	FL + CNN/DNN + Diff. Privacy + SHAP	wustl-ehms-2020, ECU-IoHT	93.20-95.48	0.96	2-4	✓	✓	✓
[20]	2024	FL + FedAvg + DNN/LSTM/GRU	ToN-IoT	91.68	0.878	9	✗	✓	✗
[10]	2026	BiLSTM-CNN	UNSW-NB15	91.14	0.931	9	✗	✗	✗
[11]	2025	CNN-BiLSTM-DNN	IoT-23, Edge-IIoTset	99	0.99	10+	✗	✗	✗
[21]	2024	Attention + BiGRU + Inception-CNN	Edge-IIoTset, CIC-IDS2017, CICIoT2023	94.7-99.5	0.946-0.995	5-14	✓	✗	✗
[25]	2025	K-Means + PCA + SHAP + PDP + ALE	ICU-IoT testbed	N/R	N/R	3	✗	✗	✓
[24]	2025	XGBoost + LR Late Fusion + SHAP	CIC IoMT 2024	97	0.98	18	✗	✗	✓
[22]	2024	LSTM + SMOTE + RFE	CICIDS2017, NSL-KDD, UNSW-NB15	98.31-99.75	0.983-0.998	Binary	✓	✗	✗
Proposed	**2026**	**FL + CNN-BiLSTM-Attention + SHAP**	**ECU-IoHT, CICIoT2023**	**94.02**	**0.954**	**4 / 33**	✓	✓	✓

N/R = not reported. For the proposed framework, "4 / 33" denotes four ECU-IoHT attack types and 33 CICIoT2023 attack types.

2. Related Work

2.1. Machine Learning Approaches for IDS

Intrusion detection has progressed from classical machine-learning methods to hybrid deep models. LightGBM uses Gradient-Based One-Sided Sampling and Exclusive Feature Bundling to reduce training cost; Ke et al. [4] reported training speedups of more than 20 times over conventional GBDT implementations in some experiments while maintaining comparable accuracy. Jin et al. [5] used LightGBM in SwiftIDS, a parallel framework designed for real-time intrusion detection. Liu et al. [6] combined LightGBM with ADASYN oversampling and reported up to 99.91% accuracy on NSL-KDD, UNSW-NB15, and CICIDS2017.

Deep models are often used when temporal or local feature dependencies are important. Devlin et al. [7] introduced BERT as a bidirectional Transformer architecture. For network intrusion detection, Gao [8] and Halbouni et al. [9] reported hybrid CNN-LSTM designs that combine local feature extraction with sequence modeling. More recent studies extend this pattern with bidirectional recurrent layers and attention. Sadhwani et al. [10] studied a BiLSTM-CNN architecture for IoT intrusion detection, Agbor et al. [11] reported 99% accuracy with a CNN-BiLSTM-DNN model on Edge-IIoTset, and Yin et al. [12] incorporated attention into CNN-BiLSTM for class-imbalanced abnormal-traffic detection. These studies motivate the hybrid backbone used here, but they do not by themselves address federated IoMT training, node-level balancing, and SHAP explanations in one process.

2.2. Cybersecurity in IoMT Networks

IoMT-specific intrusion detection requires datasets that reflect medical or health-oriented devices. Hady et al. [13] combined patient biometric information with network-flow information in a healthcare testbed and reported AUC improvements of up to 25% over single-modality alternatives across four classical machine-learning algorithms. Ahmed et al. [14] released ECU-IoHT, a dataset designed for cyberattacks on Internet of Health Things devices. It contains four attack categories and is the main healthcare-specific dataset used in this paper.

Broader IoT and IIoT datasets provide additional attack diversity. Edge-IIoTset includes 14 attack types grouped into five threat categories across a heterogeneous IoT testbed [15]. ToN-IoT covers nine attack families across network traffic, Windows logs, and Linux telemetry [16], whereas CICIoT2023 contains 33 attack types generated using 105 real IoT devices [17]. Beyond the datasets themselves, Ghourabi [1] evaluated centralized LightGBM and Transformer models across multiple intrusion-detection datasets. Hassan et al. [18] provided a broader IoMT perspective by reviewing intrusion-detection datasets, federated learning, blockchain-based approaches, and open research directions.

2.3. Federated Learning for IDS

Federated learning allows participating nodes to train on local data without transferring their raw training samples to a central server. Instead, model updates are exchanged with an aggregation component. Olanrewaju-George and Pranggono [19] evaluated federated AutoEncoder and deep neural-network classifiers using N-BaIoT data collected from nine IoT devices. Their experiments showed that federated models could achieve performance comparable to the centralized alternatives considered in their study while retaining the training samples at the individual clients. Mahmud et al. [20] proposed a FedAvg-based privacy-preserving IDS for cyber-physical systems and reported more than 90% accuracy for DoS, DDoS, data-injection, and ransomware scenarios. Yang et al. [21] addressed imbalanced IIoT traffic using a model that combines attention, BiGRU, Inception-CNN, and hybrid sampling. Their evaluation included Edge-IIoTset, CICIDS2017, and CICIoT2023.

Among the reviewed studies, SECIoHT-FL [3] is the closest to the present framework because it combines federated learning, differential privacy, and SHAP in an IoHT intrusion-detection setting and reports 95.48% accuracy. The present work differs in two aspects relevant to the proposed process. SMOTE is applied independently within each federated node before local model training, and the evaluation considers both ECU-IoHT and a subset of CICIoT2023.

2.4. Class Imbalance and Explainability in IDS

Severe class imbalance can make aggregate accuracy difficult to interpret because strong performance on majority traffic may coexist with poor detection of minority attacks. Sayegh et al. [22] addressed this problem by combining SMOTE with an LSTM-based classifier on CICIDS2017, NSL-KDD, and UNSW-NB15, reporting improved detection of minority classes. Fan et al. [23] compared four class-balancing strategies across five classifiers using 10-fold cross-validation. In their experiments, SMOTE with XGBoost produced the strongest result among the evaluated combinations, although the reported accuracy remained around 75%.

Interpretability is another concern for intrusion-detection models used in healthcare environments. Hafid et al. [24] applied SHAP to XGBoost and logistic-regression models for interpretable and cost-aware IoMT intrusion detection. Georgiades and Hussain [25] used SHAP together with Partial Dependence Plots and Accumulated Local Effects in a cross-layer IoMT IDS. These papers show how post-hoc explanations can expose the contribution of individual features to IDS predictions. Among the studies reviewed here, however, none combines node-level SMOTE, federated model training, a CNN-BiLSTM-Attention classifier, and SHAP within the same IoMT intrusion-detection process.

3. Research Gaps and Motivation

Table 2 summarizes five gaps identified from the reviewed literature. Ghourabi [1] provides a strong centralized benchmark but depends on pooled training data and does not include a class-balancing or explanation layer. The reviewed federated IDS studies [3], [19], [20] address distributed training, but node-level SMOTE applied independently before local optimization was not identified in the reviewed IoMT-specific federated IDS literature. Existing balancing studies [6], [22], [23] apply oversampling in centralized settings, while Yin et al. [12] addresses imbalance through an attention-based deep architecture rather than node-level federated oversampling.

These observations motivate a combined design with three practical objectives: keep raw samples local during training, reduce the effect of severe minority-class imbalance within each node, and provide feature-level explanations after global training. The framework is therefore evaluated on ECU-IoHT and CICIoT2023, with the computational limitation on CICIoT2023 reported explicitly rather than treating the subset as a full-dataset experiment.

Table 2. *Research gaps and proposed solutions.*

Gap	Limitation in Existing Papers	How This Paper Addresses It
G1	Ghourabi [1] uses centralized training, which requires data pooling and may conflict with institutional, legal, or privacy restrictions on cross-institutional sharing in healthcare settings.	A simulated federated learning framework using FedAvg is evaluated across distributed virtual hospital nodes so that raw training samples remain local while model weights are aggregated.
G2	Ghourabi [1] does not explicitly address the severe class imbalance in ECU-IoHT, which can reduce sensitivity to rare attack classes despite high aggregate accuracy.	At each federated node, SMOTE is applied to the local data before training, ensuring minority attack classes are adequately represented.
G3	Ghourabi [1] employs black-box LightGBM and BERT-Transformer models that lack interpretability or feature-level reasoning, which healthcare security practitioners need.	SHAP-based explanations are added to provide feature-level interpretation of selected predictions in the simulated IoMT healthcare setting.
G4	The reviewed federated IDS studies do not jointly evaluate the same combination of IoMT-specific data, node-level balancing, and the CNN-BiLSTM-Attention architecture used in this paper; several evaluate different IoT/CPS datasets or different privacy mechanisms [3], [19], [20].	The framework is evaluated on ECU-IoHT and an evaluated CICIoT2023 subset using the same federated simulation and CNN-BiLSTM-Attention backbone.
G5	Node-level SMOTE applied independently before local federated optimization was not identified among the reviewed IoMT/IoT IDS studies. The reviewed balancing studies use centralized oversampling [6], [22], [23] or alternative imbalance-handling mechanisms [12].	Node-level SMOTE is applied locally within each federated participant before local training and aggregation; this procedure distinguishes the evaluated design from the reviewed studies.

4. Background

4.1. Federated Learning and FedAvg

Federated learning trains a shared model across clients without moving their raw samples to a central dataset [20]. At communication round r, each client k trains locally and returns updated model weights $\mathbf{w}_k^{(r+1)}$. FedAvg computes the next global model as a sample-weighted average:

$$\mathbf{w}_{\text{global}}^{(r+1)} = \sum_{k=1}^{K} \frac{n_k}{N} \mathbf{w}_k^{(r+1)}, \tag{1}$$

where n_k is the number of training samples at client k and $N = \sum_{k=1}^{K} n_k$. The global weights are then redistributed to the clients for the next round. Only model parameters are exchanged in the simulated setup used in this paper.

4.2. CNN-BiLSTM Architecture

A one-dimensional CNN extracts local relationships among adjacent positions in the input representation, while a BiLSTM models dependencies in both forward and backward directions [8], [10]. For an LSTM cell, the state update is

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1}), \tag{2}$$

where f_t and i_t are the forget and input-gate outputs, c_{t-1} is the previous cell state, x_t is the current input, and h_{t-1} is the previous hidden state. The proposed network places a CNN before the BiLSTM so that local interactions in the input representation are encoded before bidirectional sequence modeling.

4.3. Attention Mechanism

The attention layer assigns an importance score to each BiLSTM hidden state rather than weighting every sequence position equally [12]. The score is

$$\alpha_t = \text{softmax}\left(v^\top \tanh(W_a h_t)\right), \tag{3}$$

where h_t is the BiLSTM hidden state and W_a and v are learned parameters. The weighted hidden-state representation is passed to the dense classifier.

4.4. SMOTE-Based Oversampling

SMOTE creates synthetic minority samples by interpolating between a minority example and one of its nearest neighbors [22]. For minority sample x_i and selected neighbor x_{z_i}, the synthetic sample is

$$x_{\text{new}} = x_i + \lambda \left(x_{z_i} - x_i\right), \qquad 0 \le \lambda \le 1. \tag{4}$$

In this framework, SMOTE is applied locally to each node's training partition. Test samples are not oversampled.

4.5. SHAP-Based Explainability

SHAP assigns a contribution value to each feature using Shapley values from cooperative game theory [25]. For feature i and feature set F, the Shapley value is

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f(S \cup \{i\}) - f(S) \right]. \tag{5}$$

The resulting values support local explanations for individual predictions and global summaries across the explained sample set.

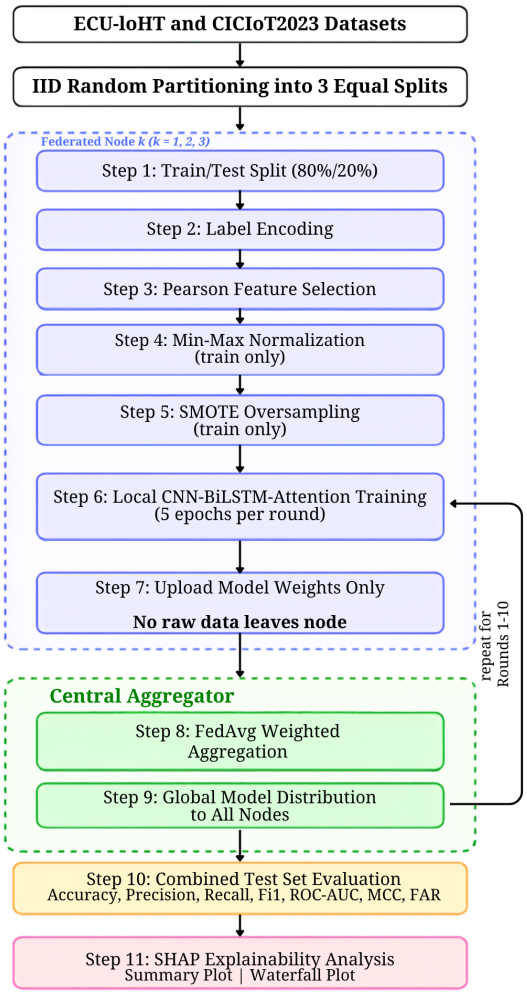

Figure 1: *End-to-end process of the federated intrusion detection framework*

5. Proposed Methodology

5.1. Dataset Description

Two public datasets are used to cover healthcare-specific IoHT attacks and broader IoT attack diversity. ECU-IoHT [14] contains 111,207 records: 23,453 normal and 87,754 malicious samples. The malicious traffic includes Smurf Attack (77,920 instances), Nmap Port Scan (6,836), ARP Spoofing (2,359), and Denial-of-Service (DoS) Attack (639). The resulting ratio between DoS Attack and Smurf Attack is approximately 1:122, providing a demanding setting for minority-class detection.

CICIoT2023 [17] was generated from 105 real IoT devices and contains 33 attack types across seven categories: DDoS, DoS, Reconnaissance, Web-based, Brute Force, Spoofing, and Mirai attacks. The standard CICIoT2023 CSV

release contains 46,686,579 records distributed across 169 CSV files. Because of the memory limits described in Section 6, the experiments use 10 of those 169 files rather than the complete dataset. The evaluated files include examples from all seven attack domains. Together, the two datasets are used to evaluate the process in Figure 1.

5.2. Data Preprocessing

Each dataset is divided into three equal IID partitions, one for each virtual hospital node. Within each node, a stratified 80%/20% train/test split is created first. Categorical features, including protocol type and connection state where applicable, are label encoded into numeric form. A fixed Pearson-correlation-based feature-selection configuration is then applied consistently so that all nodes use the same input dimensionality. Min-Max normalization parameters are fitted on each node's training portion and then applied to its held-out test portion. SMOTE is applied only to the local training data. This ordering is the one shown in Figure 1; no test samples are oversampled.

5.3. Federated Learning Setup

Three virtual hospital nodes simulate a distributed IoMT deployment. Federated training is executed sequentially on a single Google Colab machine rather than on physically distributed hardware, consistent with the simulation approach described in [20]. Each node trains a local CNN-BiLSTM-Attention model and sends only the model weights to the central aggregator. Raw training samples remain local in the simulated workflow. The experiment does not implement differential privacy, secure aggregation, or cryptographic protection of model updates; therefore, the privacy claim in this paper is limited to raw-data locality rather than a formal guarantee against information leakage from shared weights.

The aggregator applies Eq. (1). The resulting global weights are returned to all nodes before the next local-training phase. Training is run for 10 communication rounds, with 5 local epochs per round. This setup evaluates federated optimization behavior without claiming the communication latency or systems overhead of a real multi-hospital deployment.

5.4. Node-Level SMOTE Oversampling

The ECU-IoHT class distribution is strongly imbalanced, with 639 DoS Attack samples and 77,920 Smurf Attack samples. SMOTE is therefore applied independently to the local training partition at each node after preprocessing and before local model fitting. Synthetic samples are generated using Eq. (4). The test partition is unchanged, so evaluation remains on the original class distribution.

Applying SMOTE locally is consistent with the raw-data-local design because centralized oversampling would require pooled samples. In the reviewed literature, centralized oversampling appears in [6], [22], [23], while the same node-level procedure was not identified in the reviewed federated IoMT IDS studies. The current experiments, however, do not include a federated no-SMOTE ablation; therefore, the isolated causal contribution of SMOTE cannot be quantified from the reported results alone.

5.5. CNN-BiLSTM-Attention Architecture

At each node, the preprocessed and SMOTE-balanced training data are passed to the local CNN-BiLSTM-Attention network. A 1D convolutional layer uses 64 filters, kernel size 3, and ReLU activation. Max pooling with pool size 2 reduces the intermediate feature dimension. A dropout layer with rate 0.3 is then applied.

The pooled representation is processed by a BiLSTM layer with 128 hidden units. In this manuscript, the BiLSTM is described as modeling bidirectional dependencies in the ordered input representation; no claim is made that the sequence axis necessarily corresponds to elapsed packet time. The attention mechanism in Eq. (3) produces a weighted context representation. A fully connected dense layer with 64 units and ReLU activation precedes the softmax output layer for multi-class classification. The model uses Adam with a learning rate of 0.001, batch size 32, and 5 local epochs per communication round.

5.6. SHAP Explainability Integration

After 10 communication rounds, SHAP KernelExplainer is applied to the final global model. KernelExplainer is used because it is model-agnostic and can be applied to the custom attention architecture. The background distribution is summarized by k-means using 500 test instances, and SHAP values are computed for 50 selected test samples. The manuscript reports a global SHAP summary plot and a local waterfall plot for one selected prediction. In the reported SHAP summary, Length and Protocol are the two displayed features with the largest effects for the No Attack class.

5.7. Evaluation Metrics

Seven metrics are reported. ECU-IoHT and CICIoT2023 are evaluated as multi-class classification problems. Let C denote the confusion matrix for K classes, where C_{ij} is the number of samples with true class i predicted as class j, and

let $N = \sum_{i,j} C_{ij}$. Overall accuracy is

$$\text{Accuracy} = \frac{\sum_{i=1}^{K} C_{ii}}{N}. \tag{6}$$

For each class i, one-vs-rest precision and recall are

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \tag{7}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}, \tag{8}$$

and

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \, \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}. \tag{9}$$

The reported aggregate Precision, Recall, and F1 values are support-weighted averages,

$$M_{\text{weighted}} = \sum_{i=1}^{K} \frac{n_i}{N} M_i, \tag{10}$$

where n_i is the support of class i and M_i is the corresponding per-class metric. ROC-AUC is reported using a support-weighted one-vs-rest aggregation across classes. For multi-class MCC, the generalized confusion-matrix form is used:

$$\text{MCC} = \frac{cs - \sum_{k=1}^{K} p_k t_k}{\sqrt{\left(s^2 - \sum_{k=1}^{K} p_k^2\right)\left(s^2 - \sum_{k=1}^{K} t_k^2\right)}}, \tag{11}$$

where $c = \sum_k C_{kk}$, $s = \sum_{i,j} C_{ij}$, $p_k = \sum_i C_{ik}$ is the number predicted as class k, and $t_k = \sum_j C_{kj}$ is the number whose true class is k. FAR is first computed one-vs-rest for each class,

$$\text{FAR}_i = \frac{FP_i}{FP_i + TN_i}, \tag{12}$$

and the reported aggregate FAR is the support-weighted average $\sum_i (n_i/N)\,\text{FAR}_i$.

Table 3 summarizes the dataset statistics used in the experiments, and Table 4 lists the main simulation and hyperparameter settings.

Table 3. *Dataset statistics.*

Dataset	Full Records	Normal	Malicious	Attack Types	Data Evaluated Here	Source
ECU-IoHT	111,207	23,453	87,754	4	Full dataset	Ahmed et al. [14]
CICIoT2023	46,686,579	1,098,195	45,588,384	33 (7 groups)	10 of 169 CSV files; per-round subsampling described in Sec. 6.1	Neto et al. [17]

Table 4. *Simulation and hyperparameter settings.*

Parameter	Value
Number of Federated Nodes	3
Communication Rounds	10
Local Epochs per Round	5
CNN Filters	64
CNN Kernel Size	3
BiLSTM Units	128
Dropout Rate	0.3
Dense Layer Units	64
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Train/Test Split	80/20
SMOTE Applied	Yes (node-level, training only)
Partitioning Strategy	IID

6. Experimental Results

6.1. Experimental Setup

All experiments were executed in Google Colab using Python 3. The CNN-BiLSTM-Attention model was implemented with TensorFlow and Keras. Scikit-learn was used for preprocessing and the LightGBM baseline, imbalanced-learn was used for node-level SMOTE, and SHAP KernelExplainer was used for the explanation analysis. The reported Colab instance used an NVIDIA T4 GPU. Federated learning was simulated by assigning three equal partitions to three virtual hospital nodes and executing local training sequentially on one machine.

Three baselines, LightGBM, BERT-Transformer, and BiLSTM, were trained using the same preprocessed data splits and testing conditions as the proposed model. Table 4 lists the main hyperparameters. The BERT-Transformer result is retained as an implementation-specific baseline on the numerical traffic representation used in these experiments; it is not a direct reproduction of Ghourabi [1] and should not be interpreted as a like-for-like evaluation of the original BERT-based process. For CICIoT2023, memory limits in the free Colab environment restricted the experiment to 10 of 169 CSV files, with per-round subsampling of 50,000 instances per node. Ten files correspond to 5.9% of the available CSV files, but not necessarily 5.9% of all records. The evaluated files include all seven attack groups, so the results are reported as a computationally constrained subset experiment rather than as a complete CICIoT2023 benchmark.

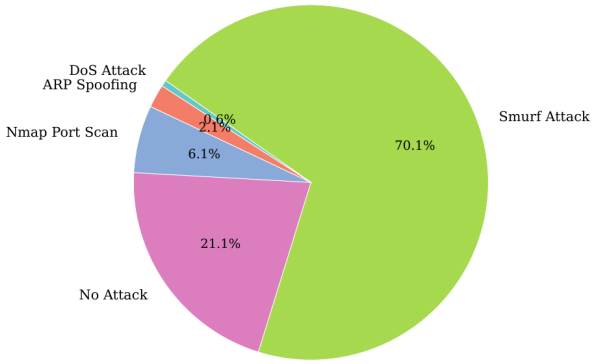

Figure 2: *ECU-IoHT class distribution before node-level SMOTE.*

Table 5. *Overall performance comparison on ECU-IoHT.*

Model	Accuracy	Precision	Recall	F1	FAR	ROC-AUC	MCC
LightGBM (Baseline)	0.9612	0.9795	0.9612	0.9673	0.0041	0.9976	0.9182
BERT-Transformer (Baseline)	0.2109	0.0445	0.2109	0.0735	0.2109	0.5000	0.0000
BiLSTM (Baseline)	0.9412	0.9788	0.9412	0.9550	0.0044	0.9976	0.8790
FL-CNN-BiLSTM-Attention	**0.9402**	**0.9786**	**0.9402**	**0.9540**	**0.0043**	**0.9975**	**0.8771**

Table 6. *Overall performance comparison on the evaluated CICIoT2023 subset.*

Model	Accuracy	Precision	Recall	F1	FAR	ROC-AUC	MCC
LightGBM (Baseline)	0.3129	0.3546	0.3129	0.2697	0.0626	0.6375	0.2798
BERT-Transformer (Baseline)	0.0008	0.0000	0.0008	0.0000	0.0008	0.5000	0.0000
BiLSTM (Baseline)	0.6303	0.6515	0.6303	0.5786	0.0329	0.9402	0.6192
FL-CNN-BiLSTM-Attention	**0.6213**	**0.7691**	**0.6213**	**0.5714**	**0.0301**	**0.9410**	**0.6173**

6.2. Performance Comparison

Table 5 compares the four models on ECU-IoHT. The proposed FL-CNN-BiLSTM-Attention model achieved 94.02% accuracy, weighted F1-score of 0.9540, ROC-AUC of 0.9975, and MCC of 0.8771. LightGBM obtained 96.12% accuracy and 0.9673 F1-score, while BiLSTM obtained 94.19% accuracy and 0.9550 F1-score. The implementation-specific BERT-Transformer configuration reached 21.09% accuracy on the numerical traffic representation used here; because this is not a direct reproduction of [1], the value is reported only as an internal baseline. The 2.10 percentage-point accuracy gap between LightGBM and the proposed model should be interpreted alongside the different training settings: LightGBM is centralized, whereas the proposed model is trained through federated aggregation.

Figure 2 shows the overall ECU-IoHT class distribution before balancing, while Figure 3 shows the balanced local training distribution after node-level SMOTE. In the full ECU-IoHT dataset, the DoS Attack class contains 639 samples against 77,920 Smurf Attack samples, approximately 1:122. The proposed model reports DoS Attack recall of 0.91 and precision of 0.11. This is a high-recall, low-precision outcome for the rare class. The result is consistent with the intended purpose of oversampling, but because a no-SMOTE federated ablation is not reported, the magnitude of improvement attributable specifically to SMOTE remains unresolved.

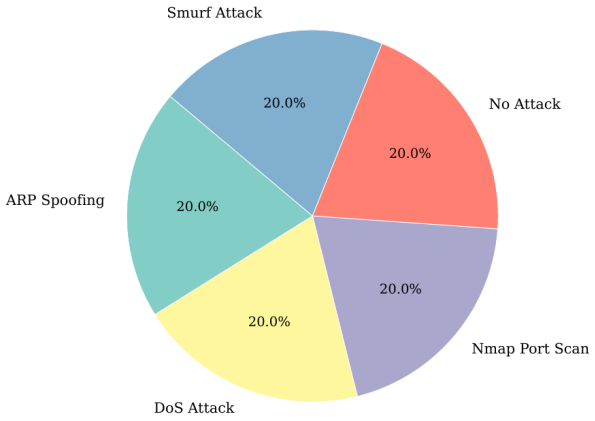

Figure 3: *ECU-IoHT class distribution after node-level SMOTE.*

The proposed model obtained a FAR of 0.0043 on ECU-IoHT, close to the LightGBM FAR of 0.0041. Figure 4 compares the FAR values of the four models. Ghourabi [1] reported approximately 99% accuracy, with performance approaching 100% in some centralized settings. In comparison, the present federated model achieved 94.02% accuracy under the simulated distributed setup used in this study. These results are not treated as a direct system-level comparison because the training procedures, data-processing processes, and experimental implementations differ.

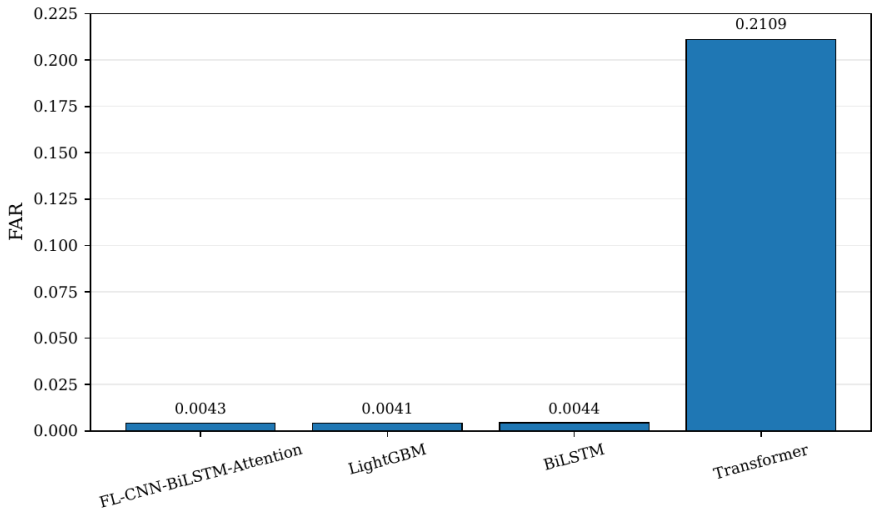

Figure 4: *False Alarm Rate comparison on ECU-IoHT; lower values are better.*

Table 8 presents the global model accuracy over 10 federated communication rounds. On ECU-IoHT, accuracy increased from 14.74% in Round 1 to 93.69% in Round 3 and reached 94.37% in Round 10. Figure 5 shows the corresponding round-wise trajectory. The curve exhibits rapid improvement during the early rounds followed by a relatively stable region in this run. Repeated independent runs would be required to quantify the variability of this convergence behavior.

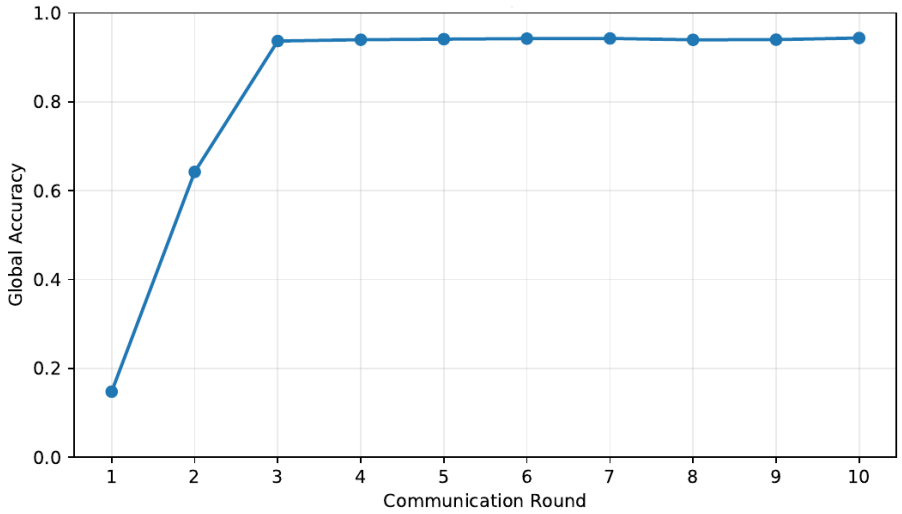

Figure 5: *ECU-IoHT global model accuracy across federated communication rounds.*

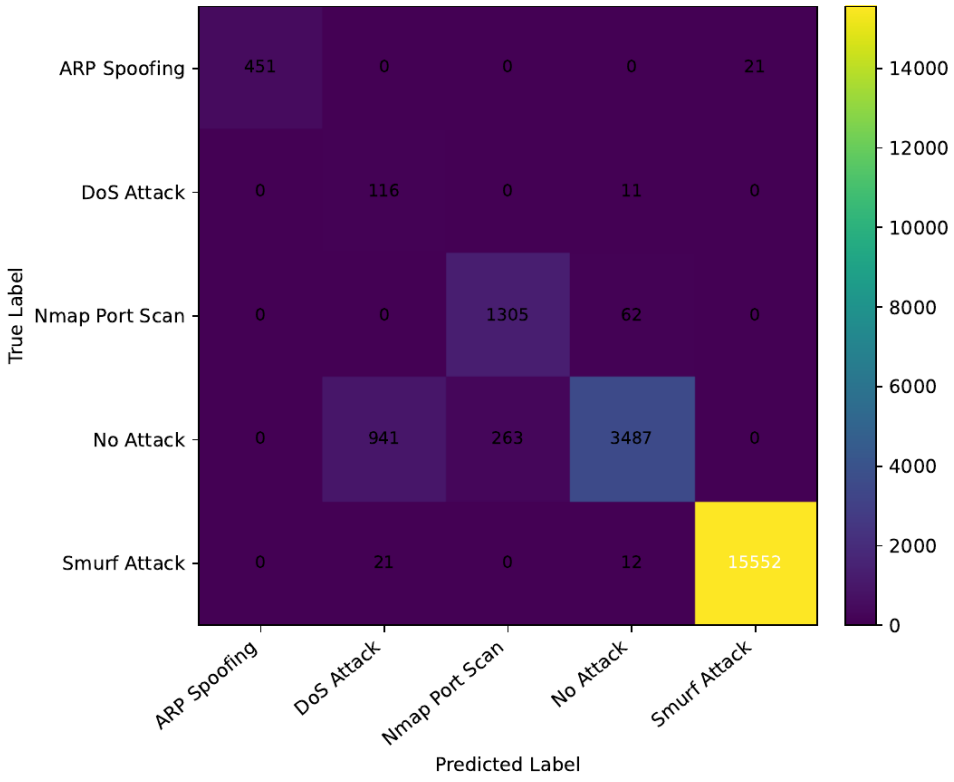

Figure 6: *Confusion matrix of the FL-CNN-BiLSTM-Attention model on ECU-IoHT.*

Table 7. *Per-class detection results of the proposed model on ECU-IoHT.*

Attack Class	Precision	Recall	F1-score	Support
No Attack (Normal)	0.98	0.74	0.84	4,691
Smurf Attack	1.00	1.00	1.00	15,585
Nmap Port Scan	0.83	0.95	0.89	1,367
ARP Spoofing	1.00	0.96	0.98	472
DoS Attack	**0.11**	**0.91**	**0.19**	**127**
Weighted Avg	0.98	0.94	0.95	22,242

Figure 6 shows the distribution of predictions across the ECU-IoHT classes, while Table 7 reports the corresponding precision, recall, F1-score, and support. The DoS Attack class reaches a recall of 0.91 but a precision of only 0.11. The model therefore detects most of the DoS instances in the test set, but this sensitivity is accompanied by a substantial number of false-positive predictions for the rare class.

On the evaluated CICIoT2023 subset, the proposed model achieved 62.13% accuracy, precision of 0.7691, and ROC-AUC of 0.9410, as reported in Table 6. LightGBM obtained 31.29% accuracy, whereas BiLSTM reached 63.03%. Although BiLSTM produced slightly higher accuracy, the proposed model obtained the highest precision among the compared models in this experiment: 0.7691 compared with 0.3546 for LightGBM and 0.6515 for BiLSTM. Its FAR was 0.0301, compared with 0.0626 for LightGBM. Because the evaluation used only 10 of the 169 CICIoT2023 CSV files, these results describe the tested subset and should not be generalized to the complete CICIoT2023 dataset.

Table 8. *Global model performance across federated communication rounds on ECU-IoHT.*

Round	Global Accuracy
1	0.1474
2	0.6422
3	0.9369
4	0.9398
5	0.9411
6	0.9423
7	0.9427
8	0.9396
9	0.9400
10	0.9437

The SHAP summary and waterfall plots are presented in Figsures 7-8. The explanation uses 50 selected test samples with a k-means background constructed from 500 test instances. Within the displayed SHAP summary, Length and Protocol have the largest effects for the No Attack class. The waterfall plot complements this global view by showing how the individual feature contributions move one selected prediction relative to the explainer baseline.

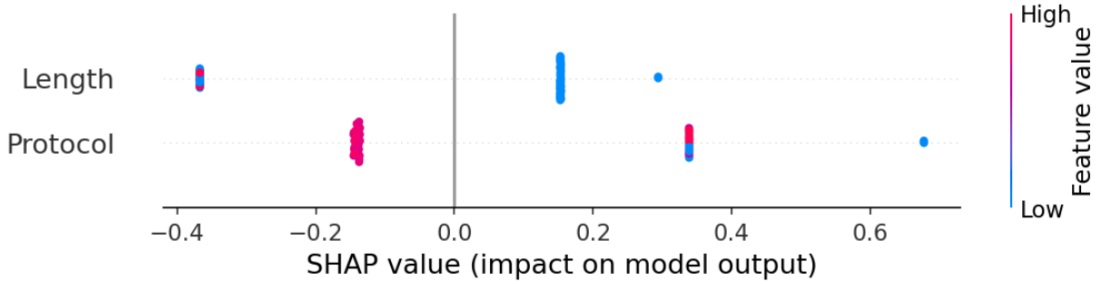

Figure 7: *SHAP summary plot for the No Attack class on ECU-IoHT.*

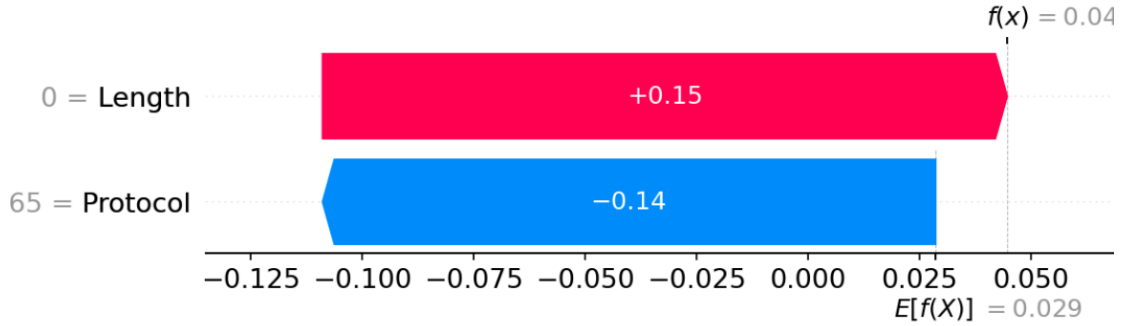

Figure 8: *SHAP waterfall plot for a single ECU-IoHT prediction.*

6.3. Discussion

The ECU-IoHT results show that the federated CNN-BiLSTM-Attention model achieves competitive aggregate performance relative to the centralized baselines considered in this study. The architecture combines local convolutional feature extraction, bidirectional recurrent modeling, and attention. However, the current evaluation does not include architectural ablations that isolate the contribution of these components. The results therefore support the complete model and training process as evaluated, but do not establish which individual component is responsible for the final performance.

The DoS Attack result illustrates the effect of extreme class imbalance. The class begins with only 639 samples, compared with 77,920 Smurf Attack samples, yet reaches a recall of 0.91 after training. Its precision of 0.11 indicates the false-positive cost associated with this high sensitivity. Sayegh et al. [22] discuss similar recall-precision effects when aggressive oversampling is applied to extreme minority classes. Future evaluation could therefore consider decision-threshold tuning, cost-sensitive learning, and a direct no-SMOTE ablation to determine whether the current DoS recall can be maintained while reducing false-positive predictions.

Several limitations define the scope of these findings. First, the federated process is simulated sequentially on a single machine rather than deployed across independent hospital infrastructures. Second, the IID client partitioning used in the current evaluation does not represent the non-IID traffic distributions that are likely to occur across different healthcare institutions. Third, the CICIoT2023 evaluation uses only 10 of the 169 available CSV files because of Colab memory limitations. Fourth, SHAP KernelExplainer is evaluated on 50 test samples because of its computational cost. Fifth, no differential privacy or secure aggregation is implemented, so shared model weights may still expose information and the privacy claim is limited to raw-data locality. Sixth, the implementation-specific BERT baseline is not a direct reproduction of [1]. Finally, the current experiment does not report repeated-seed uncertainty, statistical significance tests, or component-level ablations. These limitations do not invalidate the reported measurements, but they restrict how broadly the results should be generalized.

7. Conclusion and Future Work

This paper evaluated a federated intrusion-detection framework for IoMT healthcare networks in which raw training samples remain local to simulated nodes. The framework combines a CNN-BiLSTM-Attention classifier, FedAvg aggregation, node-level SMOTE, and SHAP KernelExplainer. In the simulated federated setup, only model weights are exchanged, while raw training data remain at the three virtual nodes. On ECU-IoHT, the proposed model achieved 94.02% accuracy, 0.9975 ROC-AUC, 0.0043 FAR, and 0.91 recall for the rare DoS Attack class. On the evaluated CICIoT2023 subset, it achieved 62.13% accuracy and 0.0301 FAR, compared with 31.29% accuracy and 0.0626 FAR for LightGBM in the same evaluated subset experiment. The CICIoT2023 result is based on 10 of 169 CSV files and should therefore be treated as a computationally constrained subset result. No formal update-level privacy guarantee is claimed because differential privacy and secure aggregation were not implemented.

Future work should deploy the framework across physically distributed infrastructure, evaluate non-IID client partitions, evaluate all 169 CICIoT2023 CSV files on dedicated hardware, and add repeated-seed, no-SMOTE, and architecture-component ablations. These experiments would separate the contributions of local oversampling, the hybrid network architecture, and federated aggregation more clearly while providing a stronger estimate of deployment behavior.

Author Declarations

Author Contributions (CRediT): Aleeha Azher: Conceptualization, Methodology, Software, Validation, Formal Analysis, Visualization. Muhammad Usman: Writing - Original Draft, Writing - Review & Editing. All authors have read and approved the submitted version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this paper are publicly available as described in [14] and [17]. No new datasets were generated during this research.

Acknowledgments: The authors have no acknowledgments to declare.

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

Ethics Statement: Not applicable.

Declaration of Generative AI: During the preparation of this manuscript, the authors used ChatGPT to assist with language editing.

References

[1] A. Ghourabi, "A security model based on LightGBM and Transformer to protect healthcare systems from cyberattacks," *IEEE Access*, vol. 10, pp. 48 890-48 903, 2022, doi: 10.1109/ACCESS.2022.3172432.

[2] World Health Organization, "WHO reports fivefold increase in cyber attacks, urges vigilance," Apr. 23, 2020. Accessed: Aug. 16, 2026. [Online]. Available: WHO news release.

[3] F. Mosaiyebzadeh, S. Pouriyeh, M. Han, L. Liu, Y. Xie, L. Zhao, and D. M. Batista, "Privacy-preserving federated learning-based intrusion detection system for IoHT devices," *Electronics*, vol. 14, no. 1, Art. no. 67, 2025, doi: 10.3390/electronics14010067.

[4] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 3146-3154.

[5] D. Jin, Y. Lu, J. Qin, Z. Cheng, and Z. Mao, "SwiftIDS: Real-time intrusion detection system based on LightGBM and parallel intrusion detection mechanism," *Computers & Security*, vol. 97, Art. no. 101984, 2020, doi: 10.1016/j.cose.2020.101984.

[6] J. Liu, Y. Gao, and F. Hu, "A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM," *Computers & Security*, vol. 106, Art. no. 102289, 2021, doi: 10.1016/j.cose.2021.102289.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, Minneapolis, MN, USA, Jun. 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.

[8] J. Gao, "Network intrusion detection method combining CNN and BiLSTM in cloud computing environment," *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 7272479, 2022, doi: 10.1155/2022/7272479.

[9] A. Halbouni, T. S. Gunawan, M. H. Habaebi, M. Halbouni, M. Kartiwi, and R. Ahmad, "CNN-LSTM: Hybrid deep neural network for network intrusion detection system," *IEEE Access*, vol. 10, pp. 99 837-99 849, 2022, doi: 10.1109/ACCESS.2022.3206425.

[10] S. Sadhwani, M. A. H. Khan, R. Muthalagu, P. M. Pawar, and K. Suresh, "A hybrid BiLSTM-CNN approach for intrusion detection for IoT applications," *Scientific Reports*, vol. 16, Art. no. 155, 2026, doi: 10.1038/s41598-025-29079-y.

[11] B. A. Agbor, B. U.-A. Stephen, P. Asuquo, U. O. Luke, and V. Anaga, "Hybrid CNN-BiLSTM-DNN approach for detecting cybersecurity threats in IoT networks," *Computers*, vol. 14, no. 2, Art. no. 58, 2025, doi: 10.3390/computers14020058.

[12] J. Yin, B. Hou, J. Dai, and Y. Zu, "A CNN-BiLSTM method based on attention mechanism for class-imbalanced abnormal traffic detection," in *Proc. 2024 6th International Conference on Computer Vision and Deep Learning (CVDL)*, Changsha, China, Jan. 2024, pp. 1-6, doi: 10.1145/3653781.3653807.

[13] A. A. Hady, A. Ghubaish, T. Salman, D. Unal, and R. Jain, "Intrusion detection system for healthcare systems using medical and network data: A comparison study," *IEEE Access*, vol. 8, pp. 106 576-106 584, 2020, doi: 10.1109/ACCESS.2020.3000421.

[14] M. Ahmed, S. Byreddy, A. Nutakki, L. F. Sikos, and P. Haskell-Dowland, "ECU-IoHT: A dataset for analyzing cyberattacks in Internet of Health Things," *Ad Hoc Networks*, vol. 122, Art. no. 102621, 2021, doi: 10.1016/j.adhoc.2021.102621.

[15] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40 281-40 306, 2022, doi: 10.1109/ACCESS.2022.3165809.

[16] T. M. Booij, I. Chiscop, E. Meeuwissen, N. Moustafa, and F. T. H. den Hartog, "ToN_IoT: The role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 485-496, Jan. 2022, doi: 10.1109/JIOT.2021.3085194.

[17] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment," *Sensors*, vol. 23, no. 13, Art. no. 5941, 2023, doi: 10.3390/s23135941.

[18] S. R. Hassan, M. U. Tanveer, S. Prajapat, and M. Shabaz, "A comprehensive survey on intrusion detection in Internet of Medical Things: Datasets, federated learning, blockchain, and future research directions," *ICT Express*, vol. 11, no. 6, pp. 1291-1310, 2025, doi: 10.1016/j.icte.2025.11.005.

[19] B. Olanrewaju-George and B. Pranggono, "Federated learning-based intrusion detection system for the Internet of Things using unsupervised and supervised deep learning models," *Cyber Security and Applications*, vol. 3, Art. no. 100068, 2025, doi: 10.1016/j.csa.2024.100068.

[20] S. A. Mahmud, N. Islam, Z. Islam, Z. Rahman, and S. T. Mehedi, "Privacy-preserving federated learning-based intrusion detection technique for cyber-physical systems," *Mathematics*, vol. 12, no. 20, Art. no. 3194, 2024, doi: 10.3390/math12203194.

[21] K. Yang, J. Wang, and M. Li, "An improved intrusion detection method for IIoT using attention mechanisms, BiGRU, and Inception-CNN," *Scientific Reports*, vol. 14, Art. no. 19339, 2024, doi: 10.1038/s41598-024-70094-2.

[22] H. R. Sayegh, W. Dong, and A. M. Al-Madani, "Enhanced intrusion detection with LSTM-based model, feature selection, and SMOTE for imbalanced data," *Applied Sciences*, vol. 14, no. 2, Art. no. 479, 2024, doi: 10.3390/app14020479.

[23] Z. Fan, S. Sohail, F. Sabrina, and X. Gu, "Sampling-based machine learning models for intrusion detection in imbalanced dataset," *Electronics*, vol. 13, no. 10, Art. no. 1878, 2024, doi: 10.3390/electronics13101878.

[24] A. Hafid, M. Rahouti, and M. Aledhari, "Optimizing intrusion detection in IoMT networks through interpretable and cost-aware machine learning," *Mathematics*, vol. 13, no. 10, Art. no. 1574, 2025, doi: 10.3390/math13101574.

[25] M. Georgiades and F. Hussain, "An explainable AI approach for interpretable cross-layer intrusion detection in Internet of Medical Things," *Electronics*, vol. 14, no. 16, Art. no. 3218, 2025, doi: 10.3390/electronics14163218.