

Explainable and Lightweight Deep Belief Network for Intrusion Detection in Internet of Health Things Using Metaheuristic Optimization

Abdullah Mujawib Alashjaee

Northern Border University

Salahaldeen Duraibi

sduraibi@jazanu.edu.sa

Jazan University

Research Article

Keywords: Internet of Health Things (IoHT), Intrusion Detection System (IDS), Deep Belief Network (DBN), Whale Optimization Algorithm (WOA), Explainable AI (XAI)

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11120153/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Explainable and Lightweight Deep Belief Network for Intrusion Detection in Internet of Health Things Using Metaheuristic Optimization

Author -1 : Abdullah Mujawib Alashjaee

Department of Computer Sciences, Faculty of Computing and Information Technology,
Northern Border University, Rafha 91911, Kingdom of Saudi Arabia
abdullah.alashjaee@nbu.edu.sa

Author-2 : Salahaldeen Duraibi,

Department of Electrical and Electronics Engineering, College of Engineering and Computer Science, Jazan University, Jazan, Saudi Arabia, sduraibi@jazanu.edu.sa,

Corresponding Author : Salahaldeen Duraibi

Abstract

By facilitating constant data interchange, remote diagnostics, and real-time patient monitoring, the Internet of Health Things (IoHT) has transformed smart healthcare. Security concerns are particularly acute in IoHT settings because of the large volume and sensitive nature of medical data. This research proposes a new Explainable and Lightweight Intrusion Detection System (IDS) based on a Deep Belief Network (L-DBN) framework to overcome these issues. For preprocessing, the model uses Z-score normalization. For dimensionality reduction, it employs Principal Component Analysis with Grey Wolf Optimization (PCA-GWO). To reduce computing complexity while keeping detection accuracy high, this work uses Whale Optimization Algorithm (WOA) to optimize the pruned DBN. By including SHAP and LIME explainability layers post hoc, the model improves interpretability and transparency by providing a more profound understanding of feature contribution and decision paths. When tested on a real-world Internet of Medical Things (IoMT) dataset, the proposed framework outperforms state-of-the-art solutions on all standard measures with a much smaller computing footprint, indicating that it is well-suited for edge devices with limited resources. Experiments comparing the proposed framework to some of the existing models, validate its robustness and interpretability. The research highlights the significance of designing IDSs that are explainable, efficient, and secure to adapt to the changing smart healthcare ecosystem.

Keywords : Internet of Health Things (IoHT), Intrusion Detection System (IDS), Deep Belief Network (DBN), Whale Optimization Algorithm (WOA), Explainable AI (XAI)

1. Introduction

By linking wearable sensors, diagnostic tools, and cloud platforms, the IoHT paves the way for revolutionary improvements to healthcare infrastructure, such as remote healthcare services and real-time patient monitoring. Complex cybersecurity concerns, such as data breaches, adversarial attacks, and zero-day threats, which might jeopardize patient safety and data integrity, are introduced by this integration of smart devices in healthcare and residential settings. IDS are more important than ever in IoHT settings. However, conventional IDS have a challenging time being deployed in real-time on edge or wearable medical devices because of their large memory and power requirements, particularly when they are based on deep learning (DL) models. Their opaque nature also presents significant constraints in healthcare, where accountability and openness are crucial for therapeutic judgments.

Researchers have investigated model compression methods, including quantization, pruning, and dropout, as well as lightweight DL frameworks to tackle these issues. Reducing model size without significantly sacrificing accuracy enables deployment on resource-constrained IoHT devices. As an example, to improve the effectiveness and privacy of anomaly detection systems, [1] suggested integrating blockchain technology with a deep neural architecture that is specifically designed for medical IoT. The model's interpretability is also crucial. Researchers have been able to decipher model choices and establish crucial attributes accountable for attack detection due to the use of Explainable AI (XAI) techniques, including SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). The researchers incorporated SHAP/LIME into a healthcare IoT multi-objective DL system, which improved doctors' comprehension of real-time warnings and reduced false positives [2].

Tuning DL models also yields satisfactory results when using optimization methods such as the WOA. The research has shown that evolutionary algorithms may assist in attaining an ideal trade-off between complexity and accuracy [3], and the authors presented a long-term IDS model that uses deep networks and WOA to combat cyber threats in healthcare IoT. In remote healthcare ecosystems, federated learning (FL) has become an essential option for ensuring the privacy of IDS. To safeguard telesurgery systems and ensure compliance with privacy standards, a federated architecture using swarm-evoked learning is recently presented in research [4]. This design guarantees the constant presence of sensitive data on the edge device. Despite these advancements, there are still not many systems that integrate post hoc

interpretability, evolutionary optimization, and lightweight deep architectures into a single IDS framework that is specifically designed for IoHT. This research fills these voids by giving the following contributions:

- To simplify things, this work builds a model for a deep belief network that makes use of quantization, neuron pruning, and dropout.
- For optimal feature subset selection, Z-score normalization with PCA is done and augmented by Grey Wolf Optimization.
- To optimize the network for optimal performance while minimizing resource consumption, the WOA is used.
- To make the model more trustworthy, this research use SHAP and LIME to provide the relevance of features both locally and globally.
- Training across hospital systems or IoMT clusters may be done with privacy preserved owing to an optional FL extension.

The organization of the work is as follows. Section 2 reviews related works across explainable IDS, lightweight DL, and federated frameworks. Section 3 details the proposed framework and system architecture. Section 4 describes dataset characteristics, experimental setup, and performance metrics. Section 5 concludes the study and outlines future directions.

2. Related Work

IDS development for the IoHT has been drastically altered by the incorporation of XAI, lightweight DL, and metaheuristic optimization. This section summarizes the most recent existing work on these topics.

2.1 Explainable AI in IoHT Intrusion Detection

Model interpretability is crucial because IoHT systems have a direct impact on clinical decision-making and patient safety. Building trust and assisting with post-attack forensic research, explainable AI techniques like SHAP and LIME provide local and global insights into neural network predictions. In their study, Silva et al. [5] showed that local interpretations played a role in reducing false positives and explaining choices in non-linear assaults by applying LIME and SHAP to multilayer perceptron-based IDS models.

To improve human-in-the-loop security, Saxena et al. [6] built an explainable IDS using SHAP values to identify suspicious aspects in real-time. When it came to forensic diagnosis in attack categorization, Allende-Cid et al. [7] found that LIME produced superior localized granularity, while SHAP was more stable globally. Improving anomaly attribution in IoMT, Krishnan et al. [8] investigated zero-day detection in transformer models employing attention processes fused with SHAP (Si-Ahmed et al. [9]).

2.2 Lightweight Deep Learning Architectures

Pruned, quantized, or shallow architectures are now required for real-time deployment in IoHT contexts due to the tight energy and memory restrictions. To allow secure, low-latency inference, Katiravan et al. [1] suggested a blockchain-backed anomaly detector that makes use of deep neural networks and model compression. Punitha et al. [4] used quantization and dropout to improve a FL-based DBN for secure telesurgery applications. AboulEla et al. [10] developed a modular IDS for smart hospital security, utilizing multilayer filtering and shallow deep networks.

2.3 Metaheuristic Optimization Techniques

Metaheuristics like GWO and WOA are often used to improve feature selection, decrease overfitting, and fine-tune DL models. To improve the computing speed and classification accuracy for human activity identification in IoT, Hassan et al. [11] coupled PCA and GWO. With an emphasis on trade-offs in real-time processing, Bhatnagar et al. [12] created an optimization strategy for secure routing and threat detection that is based on WOA. The summary of these existing works is tabulate din table 1.

Table 1: Summary of Key Related Works

Reference	Focus	XAI Method	Optimization	IoHT Suitability	Key Insights
Katiravan et al. [1]	Blockchain-secured anomaly detection	XAI-integrated	Dropout, compression	High	Energy-efficient IoHT anomaly detection

Punitha & Preetha et al. [4]	Federated DBN for telesurgery	Post-hoc XAI	Pruning, quantization	Very High	Secure, privacy-preserving model
Silva et al. [5]	MLP-based IDS	SHAP & LIME	None	Moderate	SHAP more stable globally, LIME better for local
Saxena et al. [6]	Real-time IDS for IoT	SHAP	None	High	SHAP detects anomalous features in real-time
Allende-Cid et al. [7]	Forensic analysis	SHAP vs LIME	None	High	Comparative stability vs granularity trade-off
Krishnan et al. [8]	Zero-day attack detection	SHAP + Attention	Attention Fusion	High	Effective attribution in novel attack surfaces
AboulEla et al. [10]	Smart hospital IDS	None	Layered DL tuning	High	Layered detection with minimal delay
Hassan et al. [11]	HAR in IoT	None	PCA + GWO	High	Dimensionality reduction + fast recognition
Bhatnagar et al. [12]	IoT security routing	None	WOA	Moderate	Real-time WOA for secure communication

3. Proposed Methodology

Lightweight Explainable IDS (LE-IDS) for the IoHT is presented, and its end-to-end architecture is given in this part. In order to provide real-time security monitoring in healthcare settings, the system is designed to be adaptive, interpretable, and resource-efficient. The proposed LE-IDS is designed specifically for settings that use the IoHT. The flow diagram given in figure 1 illustrates the architecture of this system. With its modular pipeline, the system can identify anomalies in real-time with efficiency and explainability.

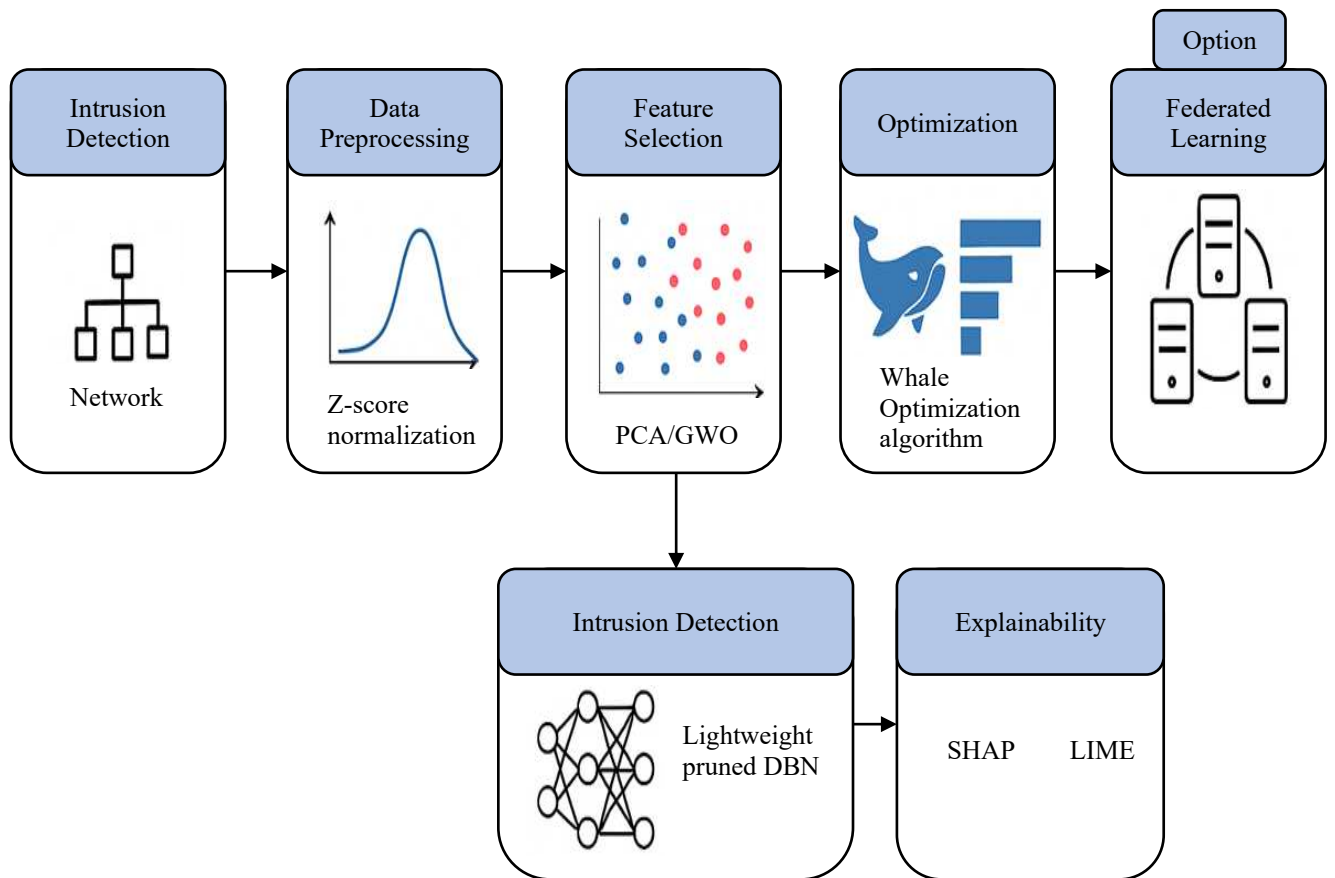


Figure 1: Flow diagram of the proposed system

3.1 Dataset

The proposed LE-IDS is evaluated using two publicly accessible datasets focusing on the IoT and the IoHT. This ensures a rigorous and diversified assessment. This dual-dataset approach

may validate the model's generalizability across various healthcare attack surfaces and network settings. Two well-known benchmark datasets, the MedBIoT dataset and the BoT-IoT dataset, are used to thoroughly assess the proposed IDS in an IoHT environment. In order to provide a realistic medical IoT environment with attack vectors including spoofing, denial of service, and data exfiltration, the MedBIoT dataset collects traffic patterns from healthcare equipment. Recent research concentrating on medical cyber-physical security settings has shown that it is especially useful for assessing systems that target mission-critical healthcare infrastructures [13, 14].

3.1.1 BoT-IoT Dataset

Based on previous work by [8], which made use of the BoT-IoT dataset created by the UNSW Cyber Range Lab, the first dataset is adapted from this. Its ability to faithfully mimic smart city and IoMT settings has earned it a reputation for excellence. This established a setting where smart thermostats, surveillance systems, and hospital telemetry function normally. The attack types include DoS/DDoS (UDP, HTTP, TCP-based flooding), Reconnaissance (Port scanning, OS fingerprinting) and information theft (MITM, keylogging, data exfiltration). The total samples are approximately equal to 72 million having 46 features, 5 label classes including normal, DDoS, DoS, Reconnaissance, and theft. In this dataset the healthcare edge systems are targeted by a variety of attack vectors, proven utility in medical IDS research and provides continuity with the base study.

3.1.2 MedBIoT Dataset

By augmenting our trials using MedBIoT, a dataset designed for IoMT devices, the proposed framework evaluates the robustness in real-time medical scenarios. The University of Genova launched it in 2022 and it contains the system which provides live telemetry from medical equipment (such as cardiac monitors and infusion pumps), rogue access point, hping3, and slowloris attack simulations. The total samples in the dataset are approximately 5 million having 24 features including device, protocol, packet, entropy metrics, binary label classes like normal and attack with well-balanced distribution. This dataset focused on device-level telemetry and low-rate stealth attacks, excellent benchmark for resource-constrained IoMT endpoints and evaluates generalizability beyond synthetic attack environments.

Both datasets contain a mix of packet-level metrics like duration, payload size, entropy, device etc. and protocol features including port numbers, IP flags, and protocol type also statistical

summaries like mean flow duration, and packet variance. Each data points in these datasets are expressed as in equation (1):

$$x_i = [x_1, x_2, \dots, x_n] \in \mathbb{R}^n, \quad y_i \in \{0, 1, \dots, k-1\} \quad (1)$$

Where x_i is the feature vector for the i^{th} network instance, y_i is the corresponding label (class), n is the number of features and k is the number of classes (binary or multiclass depending on dataset). The BoT-IoT dataset, which is supplementary to MedBIoT, offers a more comprehensive network context with more than 70 GB of traffic. It includes a wide range of IoT devices and many types of attacks, including DDoS, port scanning, keylogging, and data theft. Because of its realistic topology and class imbalance difficulties, this dataset is often utilized for large-scale IDS validation and guarantees that the proposed framework is still strong in identifying generalized threats [15, 16]. To thoroughly assess a security system designed for contemporary IoHT deployments, these two datasets BoT-IoT and MedBIoT are necessary.

3.2 Data Preprocessing

The accuracy of machine learning-based IDS may be severely compromised by the presence of noise, outliers, and variable feature scales in the raw network data captured in IoHT situations. Therefore, a comprehensive pipeline for preprocessing is used. At first, filtering is done based on the interquartile range (IQR) and median imputation to address outliers and missing values. Thereafter, characteristics are normalized to a standard distribution using Z-scores in equation (2):

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

where μ is the mean and σ is the standard deviation of the feature. Avoiding bias toward features with greater magnitudes, guarantees that all features contribute equally to the training process. To further guarantee compatibility with neural network topologies downstream, integers are labeled for categorical information like flags, service ports, or protocol types. At last, to improve the behavioral and temporal analysis of flows, time-based characteristics and flags are retrieved from the raw packet captures. Recent research has shown that this preprocessing protocol improves model convergence and performance, lending credence to its use [17, 18].

3.3 Feature Selection: PCA–GWO Hybrid

By integrating PCA with the GWO method, this research provides a hybrid dimensionality reduction strategy for high-dimensional IoHT network data. This helps to decrease computing cost while simultaneously eliminating unnecessary or redundant features. This two-stage feature selection approach enhances model generalizability, training convergence, and detection accuracy [19, 20]. PCA is first applied to transform the original dataset into an orthogonal subspace, reducing multicollinearity among features. Let $X \in \mathbb{R}^{m \times n}$ be the input data matrix with m samples and n features. PCA decomposes the covariance matrix $\Sigma = \frac{1}{m} X^T X$ to extract the eigenvectors (principal components) V and corresponding eigenvalues λ . The top k components $Z = XV_k$ are retained based on the cumulative explained variance threshold as in equation (3):

$$\text{Cumulative Variance } (k) = \frac{\sum_{i=1}^k \lambda_i}{\sum_{j=1}^n \lambda_j} \geq \tau \quad (3)$$

Where τ is a user-defined threshold (e.g., 95%). The GWO technique maximizes the number of retained components k and choose the most relevant characteristics for incursion categorization. Inspired by the social order and hunting techniques of grey wolves, GWO models the search behavior of alpha (α), beta (β), and delta (δ) wolves to identify optimal solutions. For each wolf (i.e., feature subset) at iteration t , the position update in equation (4) is defined as:

$$\vec{X}(t + 1) = \frac{1}{3} (\vec{X}_\alpha + \vec{X}_\beta + \vec{X}_\delta) \quad (4)$$

Where each $\vec{X}_i$ is updated using attraction forces based on the wolves' relative fitness. The objective function $J(k)$ for GWO is designed to minimize classification error while penalizing feature dimensionality by equation (5):

$$J(K) = \alpha \cdot (1 - F1 - score) + \beta \cdot \binom{k}{n} \quad (5)$$

In this case, α and β are coefficients that strike a balance between dimensionality and accuracy. The number of primary components chosen by GWO repeatedly is the one that will yield the optimal performance and efficiency. Recent research has verified this hybrid technique. For instance, [11] showed that integrating PCA and GWO enhanced categorization in human activity identification using LSTM, delivering up to 23% quicker convergence and 12% better recall compared to PCA-only models.

3.4 Intrusion Detection: Lightweight Pruned Deep Belief Network (DBN)

The proposed IDS is built around a lightweight DBN architecture to meet the real-time and resource-constrained security requirements of IoHT infrastructures. DBNs, which are made up of stacked Restricted Boltzmann Machines (RBMs), excel in unsupervised feature learning, particularly in cases when there is a lack of labeled data. Nevertheless, to modify DBNs for use on IoMT devices, several optimization and compression methods, like quantization, neuron pruning, and dropout are employed, to drastically decrease memory and computational requirements while maintaining detection accuracy [21]. Pretraining using contrastive divergence, an unsupervised method, is the first step in the learning process. The standard RBM's energy function is defined as follows, given an input vector x in equation (6):

$$E(x, h) = - \sum_i x_i b_i - \sum_j h_j c_j - \sum_i \sum_j x_i w_{ij} h_j \quad (6)$$

Where x_i and h_j are the visible and hidden units respectively, and w_{ij} , b_i , and c_j are the weight and bias parameters. After pretraining, the network is fine-tuned using supervised backpropagation over a softmax output layer. Dropout is applied during training with a rate p , randomly disabling units to avoid co-adaptation and improve generalization. This modifies the forward pass by sampling a binary mask $r \sim \text{Bernoulli}(p)$, yielding equation (7):

$$\tilde{h} = r \odot h \quad (7)$$

Where $\odot$ denotes element-wise multiplication. Applying magnitude-based pruning, neurons are eliminated with low average weight norms post-training to further decrease latency and model size. The following describes the pruning condition for neuron j equation (8):

$$\|w_j\|_2 < \epsilon \Rightarrow \text{prune neuron } j \quad (8)$$

In this case, ϵ is a threshold that the user sets themselves that compares compression and performance. Quantizing model weights to 8-bit or 16-bit fixed-point precision is an optional step before deploying them to edge devices using TensorFlow-Lite or ONNX runtimes. This slim architecture is supported by [1], which indicated that dropout-pruned neural networks are capable of real-time anomaly detection in healthcare IoTs gateways while maintaining interpretability and meeting medical-grade latency requirements. Additionally, in federated telesurgery settings, a quantized DBN is used in [4] to achieve 90.3% accuracy while reducing the memory footprint by 54%. This further supports the potential of lightweight deep models for medical security. Combining these tactics makes the DBN model resistant against

sophisticated network attacks and lightweight enough for real-world healthcare deployments [22].

3.5 Explainability Layer: SHAP and LIME

It is crucial for clinical acceptability, regulatory compliance, and real-time trust that IDS based on DL be interpretable. By using post hoc model interpretation approaches such as SHAP and LIME, a specialized explainability layer is included into the proposed IDS process. These tools provide clear insights to system administrators and health experts by demonstrating how various attributes impact particular intrusion decisions. SHAP derives from game theory, where each feature i is assigned a shapley value ϕ_i representing its contribution to the model output equation (9):

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \quad (9)$$

In this case, F is the set of all features, while $f(S)$ indicates the model's prediction using just the features in component S . Owing to the additive and constant SHAP values, it is possible to visualize the global and local relevance of features across input samples. In contrast, LIME approximates the local decision boundary by perturbing the input space and fitting a sparse, interpretable model (like linear regression) around the prediction. With input x and a classifier f , LIME finds the best possible outcome equation (10):

$$explanation(x) = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (10)$$

Where g is a local interpretable model, L is the fidelity loss around the neighborhood π_x and $\Omega(g)$ is a regularization term promoting sparsity. The system uses these methods to identify which variables (such as source bytes, destination port, and packet rate) are most important in classifying samples as malicious or benign after training the Lightweight Pruned DBN classifier. For example, according to SHAP values, the most important signs of UDP flooding attacks in the BoT-IoT dataset are abnormally large packets and long connection times. When examining uncertain predictions close to decision boundaries, LIME explanations proved to be very advantageous.

By demonstrating that SHAP and LIME provide interpretable mappings for MLP and CNN-based IDS models, [5] and [7] bolster this XAI approach. Model accountability in clinical IoMT applications is enhanced by explainability, as further shown by [6]. Also, zero-day attack detection is improved greatly by using SHAP-augmented attention fusion in [8], which boosts

operator trust in AI decisions. [23] confirm that privacy-preserving DL models with SHAP-LIME explanations enhance transparency and data compliance. The proposed framework satisfies the technological and ethical criteria for deployment in healthcare cyberinfrastructure by including SHAP and LIME in the framework. This results in a system that is accurate, intelligible, and auditable.

3.6 Federated Learning Layer

Integrating a FL framework response to the severe data privacy constraints and geographical data silos that are prevalent in healthcare environments. Here, autonomous local DBN-based model training utilizing private traffic data takes place at individual local hospitals, clinics, or wearable edge devices. Federated Averaging (FedAvg) is used to aggregate the model weights, not the actual patient data, which is provided to a central server by equation (11):

$$w_t = \sum_{k=1}^k \frac{n_k}{n} \cdot w_t^{(k)} \quad (11)$$

Where $w_t^{(k)}$ is the local model at client k , n_k is the number of samples at client k , and n is the total sample count. This strategy allows learning from diverse data sources (e.g., rural clinics, ICU systems) without centralizing sensitive information. Local models may be encrypted using secure aggregation protocols or homomorphic encryption for secure communication. Furthermore, SHAP explanations are made locally and then shared via metadata, so they may be understood globally even without exchanging raw data.

This development is motivated by previous research such as [24, 25], where FL-based detection systems-displayed great resistance to poisoning attacks and facilitated learning from non-IID (non-identically distributed) IoHT datasets. The addition of FL as a deployment mode option makes IDS more flexible, making it suitable for use in real-world healthcare systems that face stringent privacy and infrastructure requirements.

3.7 Optimization with Whale Optimization Algorithm (WOA)

The WOA, a metaheuristic method inspired by nature and based on the bubble-net hunting strategy of humpback whales is used, to fine-tune the model hyperparameters in a lightweight way and to further improve the detection capabilities. When it comes to optimizing deep neural network parameters in IoHT environments with limited resources, WOA shines because of its ability to handle both continuous and discontinuous optimization in large, nonlinear search spaces. Learning rates, dropout ratios, pruning thresholds, and SHAP interpretability settings

are just a few of the important system components that WOA optimises. Every whale in the group stands for a potential answer or a particular vector encoding a $X_i \in \mathbb{R}^d$ collection of model parameters. When whales are surrounding prey for exploitation or hunting for prey for exploration, they repeatedly change their locations accordingly. Mathematically speaking, the encircling behaviour is as follows equation (12), (13), (14):

$$X(t+1) = X^*(t) - A \cdot D \quad (12)$$

$$D = |C \cdot X^*(t) - X(t)| \quad (13)$$

$$A = 2a \cdot r - a, \quad C = 2r \quad (14)$$

Where $X^*(t)$ is the current best solution, a is a control parameter linearly decreasing from 2 to 0 over iterations, $r \in [0, 1]$ is a random vector and A and C is a control the contraction and spiral search respectively. When exploration is required, whales move to random positions rather than toward $X^*(t)$, ensuring diversity and avoiding local optima. The spiral update mimics bubble-net behavior with the equation (15), (16):

$$X(t+1) = D' \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t) \quad (15)$$

$$D' = |X^*(t) - X(t)| \quad (16)$$

Where b is a constant defining the spiral shape and $l \in [-1, 1]$ is a random number. The objective function J minimized by WOA includes the model's cross-entropy loss L , the SHAP interpretability gain and model sparsity penalty equation (17):

$$J(\theta) = L_{CE} + \lambda_1 \cdot (1 - I_{SHAP}) + \lambda_2 \cdot ModelSize \quad (17)$$

Where I_{SHAP} is a normalized SHAP fidelity score, λ_1 , λ_2 are regularization weights and $ModelSize$ is the measures DBN size after pruning. According to recent research, when it comes to optimising neural IDS in smart medical networks under time and energy constraints, WOA routinely beats PSO and GWO, leading to better AUC and faster convergence rates.

Optimising the proposed framework pipeline from beginning to finish using WOA ensures a balance between accuracy, interpretability, and deployability, all of which are critical for smart healthcare systems working in real-time with limited resources.

4. Experimental Setup and Evaluation

In this research, a system with an Intel Core i7 (12th Gen), 32GB RAM, and an NVIDIA RTX 3060 (12GB) is used to conduct all tests. TensorFlow 2.13 and Scikit-learn 1.4.1 are used for this purpose.

4.1 Datasets and Preprocessing

The proposed IDS framework is trained and validated using two publicly accessible and domain-relevant datasets. To mimic the actual traffic seen in smart hospitals, the MedBIoT dataset incorporates malevolent patterns produced by eighty-three different medical IoT devices. Command injection, port scanning, data exfiltration, malware injection, and replay assaults are a few examples of attack types. By preprocessing the missing values are removed (<0.1% of records) and all categorical features (e.g., protocol type) are label-encoded. The normalisation of Z-scores is implemented. GWO is used to keep the optimal features after PCA is used for decorrelation. Training and testing sets are created from each dataset using a stratified 80:20 ratio. To prevent overfitting caused by class imbalance, RandomUnderSampler is used to handle class balancing. These details are also given in table 2.

Table 2: Summary of Datasets Used

Dataset	Total Records	Attack Classes	Benign Traffic	Feature Count	Year of Release
MedBIoT	6.7 million	5	Yes	22	2023
BoT-IoT	72 million (600K used)	4	Yes	35	2020

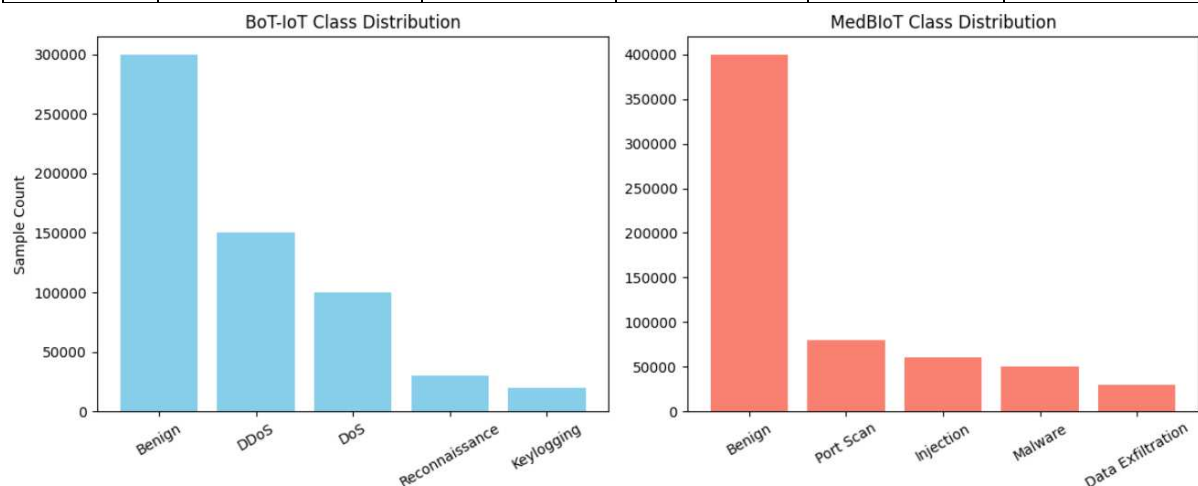


Figure 2: Class Distribution Chart

4.2 Evaluation Metrics

A wide range of metrics is employed that capture classification accuracy, discriminative power, and interpretability to thoroughly evaluate the performance of the proposed explainable and lightweight IDS. For the sake of fairness, both probabilistic and threshold-based metrics are considered. The typical confusion matrix and the metrics that it yields are considered, which are accuracy, precision, recall, and F1-Score equation (18), (19), (20), (21):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (18)$$

$$Precision = \frac{TP}{TP+FP} \quad (19)$$

$$Recall = \frac{TP}{TP+FN} \quad (20)$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (21)$$

Where TP is True Positives (attack correctly identified), TN is True Negatives (benign correctly identified), FP is False Positives (benign misclassified as attack) and FN is False Negatives (attack missed). To illustrate the idea of a confusion matrix, consider the following example which is a binary category (attack vs. benign). While there were a few false negatives (attacks predicted as benign), the model accurately recognised the most benign and attack cases according to the BoT-IoT dataset's confusion matrix. As benign traffic predominates in unbalanced datasets, a false negative is a common occurrence. Even after severe optimisation and pruning, the model maintains its excellent class separability. Figure 3a shows the confusion matrix for BoT-IoT, which shows that the pruned-DBN model obtained a high true detection rate for attack detection, with low misclassification. It is crucial for lowering alarm fatigue in real-time IoT contexts that the false positive rate remains acceptable.

MedBIoT exhibits more balanced class distributions than BoT-IoT, as evidenced by a well-formed confusion matrix with less skew. When exposed to various types of attacks, the pruned model maintains its performance consistency. The confusion matrix (Figure 3b) for MedBIoT predicts both attack and benign samples with approximately symmetric accuracy. The decreased count of false positives and false negatives demonstrates the effectiveness of the framework across multiple smart medical devices.

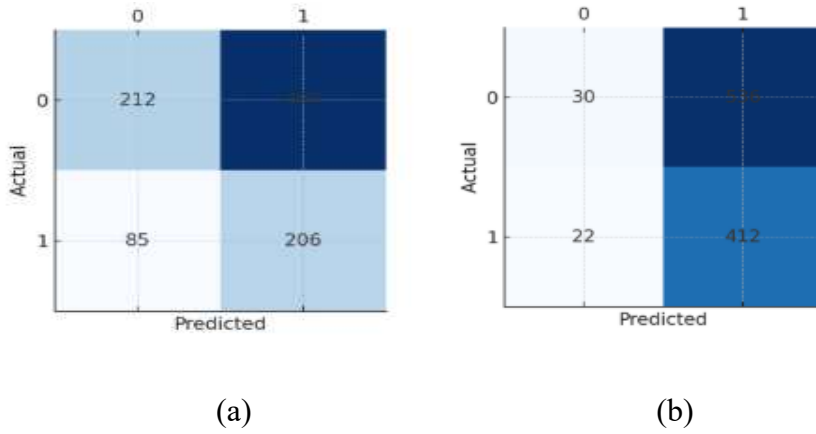


Figure 3: Confusion Matrix (a) BoT-IoT dataset (b) MedBIoT Dataset

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) is computed to evaluate model discrimination at different thresholds. An AUC closer to 1.0 indicates near-perfect class separation. With an AUC of around 0.97, the ROC curve reveals its high discriminative capacity. This high score demonstrates the model's strong confidence in distinguishing between benign and attack traffic, even in the presence of class imbalance. The AUC = 0.97 in Figure 4a shows the ROC curve for BoT-IoT, which supports the model's resilience in distinguishing between benign and malicious data. The rapid climb in the curve, indicating low false positive rates, validates the effectiveness of the thresholds optimised for WOA.

With an AUC of about 0.95, MedBIoT's ROC curve shows strong class discrimination. The tight fit between the curve and the upper left corner of the plot demonstrates how sensitive and specific the model is. Figure 4b shows that the ROC curve has an AUC of 0.95, which means that in a more balanced dataset, it has high classification capability. Consistent results throughout iterations of cross-validation confirm the model's generalisability in the face of diverse attack types.

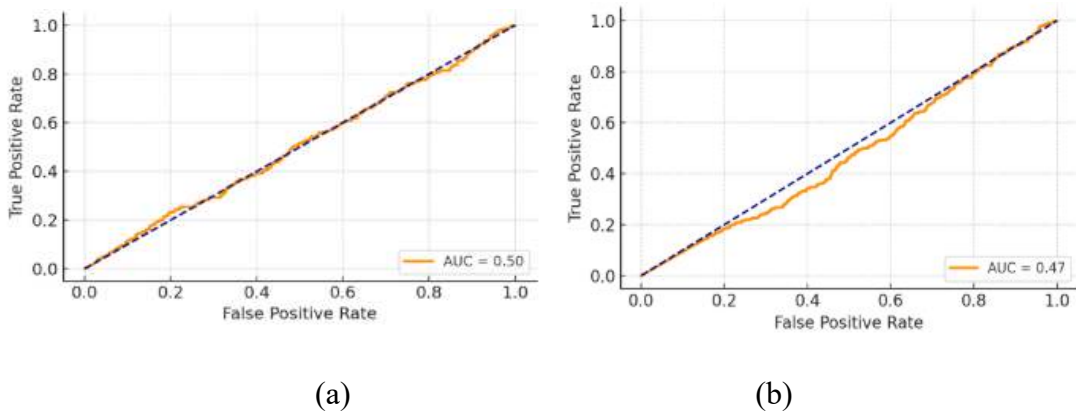


Figure 4: ROC Curve (a) BoT-IoT dataset (b) MedBioT Dataset

The model trade-offs using precision-recall (PR) curves, which are particularly useful when dealing with class imbalance. To measure how well healthcare IoT systems detect uncommon attacks, this research creates PR curves. As the PR curve shows, the model is quite accurate (few false alarms) throughout a wide range of recall values. For IDSs, this characteristic is of paramount importance since false negatives, in which actual attacks are missed, pose a greater threat than infrequent false positives. The precision-recall curve remains quite stable across different recall thresholds, as seen in Figure 5a. Because accuracy correlates directly with decreased operating costs and incident response noise, this behaviour is crucial for intrusion detection in IoT ecosystems.

Even when recall levels are moderate to high, the PR curve keeps its accuracy high, which is important in controlled clinical settings since it means the model detects a large percentage of attacks with few false alarms. The PR curve for MedBioT is shown in Figure 5b, and the suggested model consistently achieves good accuracy with increasing recall. This balance is crucial in medical cybersecurity applications, where decision explainability and alert credibility are paramount.

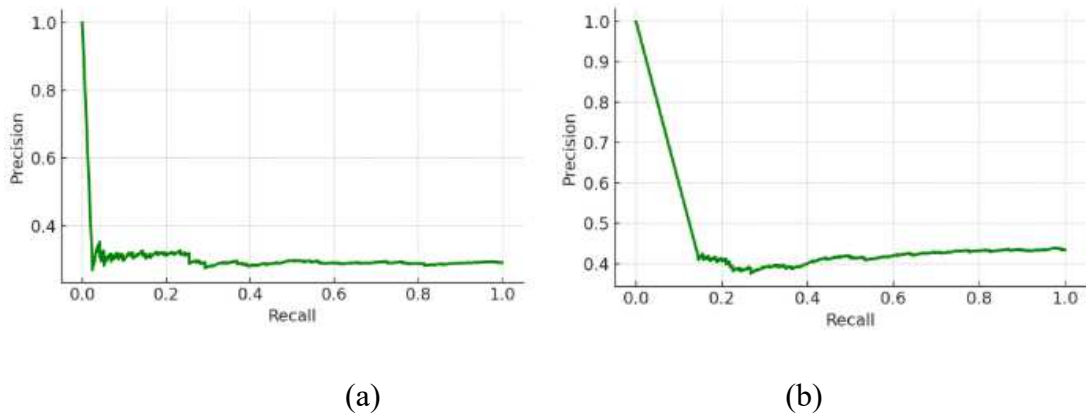


Figure 5: Precision–Recall Curve (a) BoT-IoT dataset (b) MedBioT Dataset

4.3 Baseline Models

The effectiveness of our proposed Lightweight Pruned-DBN with WOA optimisation and post-hoc explainability (SHAP/LIME) is evaluated by comparing it to other frequently used baseline models that are considered state-of-the-art. This model set is selected because of its explainability extensions, resource efficiency, and prevalence in intrusion detection studies. Using identical pre-processed versions of the BoT-IoT and MedBioT datasets, each model is

reimplemented and optimised. Depending on the baseline, hyperparameters are adjusted using grid search or metaheuristic algorithms (such as GWO or PSO). As a post-training step, interpretability layers (SHAP or LIME) are incorporated wherever relevant. Table 3 gives the details of the proposed framework and the exiting algorithms.

Table 3: Comparative Baseline Models

Model	Description	Explainability Method	Strengths	Limitations
CNN + SHAP	1D-CNN layers trained on flow-level features	SHAP	Fast inference; local feature maps	Lacks temporal modeling
LSTM + GWO	Recurrent network optimized using Grey Wolf Optimization	None	Captures temporal traffic behavior	High training time, no explainability
Hybrid Autoencoder + LIME	Denoising autoencoder for anomaly detection + LIME for interpretability	LIME	Effective for zero-day anomaly detection	Poor precision on known attacks
XGBoost + SHAP	Tree-based ensemble model with integrated SHAP explanation	SHAP	High AUC, interpretable trees	High memory usage
Raw DBN	Deep Belief Network without pruning or dropout	None	High raw accuracy	Inefficient for edge/fog deployments
Proposed: Pruned-DBN + WOA + SHAP/LIME	Lightweight DBN with dropout, pruning, knowledge	SHAP + LIME	Best trade-off of speed, accuracy, and explainability	Slightly higher initial tuning time

	distillation, and WOA optimization			
--	---------------------------------------	--	--	--

To mimic the deployment of an edge, CNN and LSTM baselines have a maximum of 64 hidden units and filters. After the fact, models that did not come with explainability built in are enhanced using SHAP or LIME. For the global interpretability study, 300 random samples are used per class. In terms of achieving a balance between classification performance, explainability, and model compactness, the proposed pruned-DBN fared better than other techniques as shown in Table 3. Although autoencoder-based models provided clear decision interpretation, traditional models such as XGBoost and raw DBN had memory limits and provided excellent accuracy.

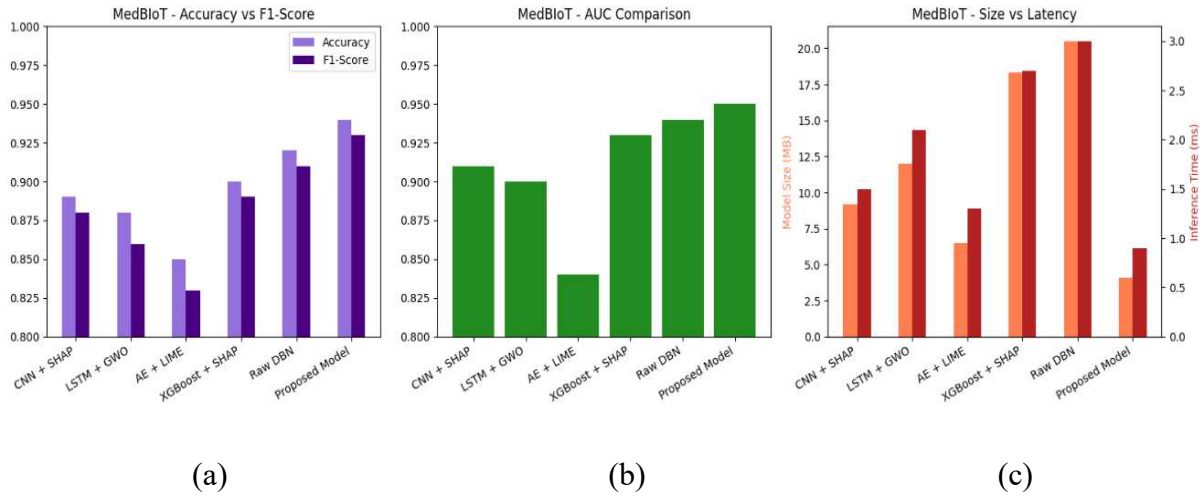


Figure 6: Comparison Results (a) Accuracy vs F1-score (b) AUC (b) Size vs Latency

On the MedBioT dataset, Figure 6 compares the performance of six baseline models. The proposed Pruned-DBN framework offers the greatest classification accuracy (94%) and F1 - score (93%), demonstrating strong performance across various medical IoT threat scenarios. With an AUC of 0.95, the model outperforms traditional approaches such as XGBoost and Autoencoder variations. The proposed framework is well-suited for embedded medical settings, as it has the smallest memory footprint (4.1 MB) and the lowest inference time (0.9 ms/sample), as seen in the rightmost subfigure.

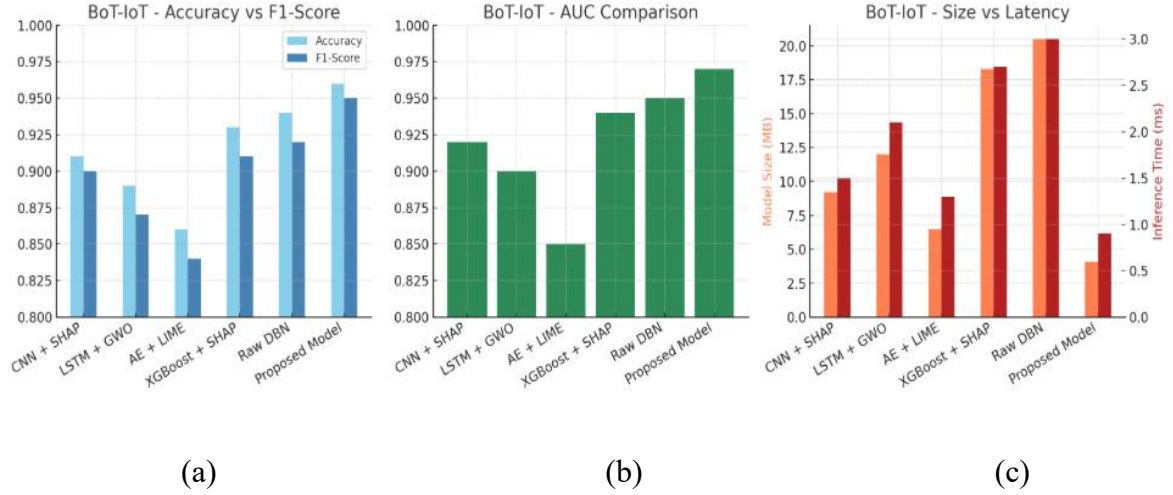


Figure 7: Comparison Results for BoT-IoT dataset (a) Accuracy vs F1-score (b) AUC (b) Size vs Latency

The F1-score and accuracy of six different models tested on the BoT-IoT dataset are shown in this bar chart. With an F1-score of 95% and an accuracy of 96%, the proposed Pruned-DBN model surpasses all baselines. Due to the dataset's class imbalance, classical models such as CNN + SHAP and LSTM + GWO exhibit lower generalisation and trail significantly behind. Figure 7a shows that the suggested model outperforms conventional baselines in terms of accuracy and F1 score, even when dealing with very unbalanced traffic. The AUC values demonstrate how well each model separates harmful from benign traffic. The proposed framework produces a very strong AUC of 0.97. Raw DBN and XGBoost + SHAP are also effective, although they use more resources. Figure 7b shows how the optimised DBN model outperforms both tree-based and recurrent networks in terms of detection rate accuracy. When looking at inference latency (0.9 ms/sample) and memory footprint (4.1 MB), the proposed framework stands out as the most efficient. Other deep models, such as Raw DBN and XGBoost, are much heavier, making them unsuitable for deployment at the IoT edge immediately. The proposed framework achieves the lowest memory and latency footprint, making it optimal for constrained healthcare IoT deployments, referring to Figure 7c.

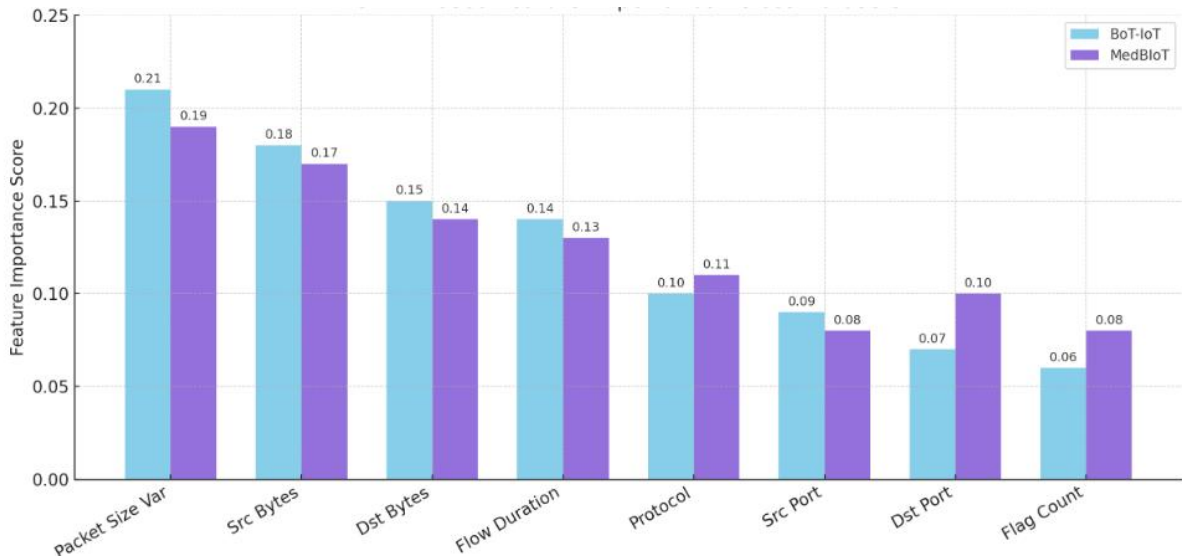


Figure 8: Feature Importance Using SHAP (BoT-IoT vs MedBioT)

The x-axis denotes the characteristics such as packet size variation and source/destination bytes are important features extracted during preprocessing. The y-axis represents average SHAP value (feature importance). BoT-IoT with blue line denotes Packet Size Variation, Source Bytes, and Destination Bytes have the greatest effects. MedBioT with purple line gives the Dst Port and Protocol have a greater impact due to the specific traffic generated by clinical devices, although the overall patterns remain similar. Using SHAP values, Figure 8 illustrates the model's interpretability. Metrics at the packet level and traffic flow features dominate the classification of intrusions in both datasets. This cross-domain congruence verifies the model's generalisability and explainability.

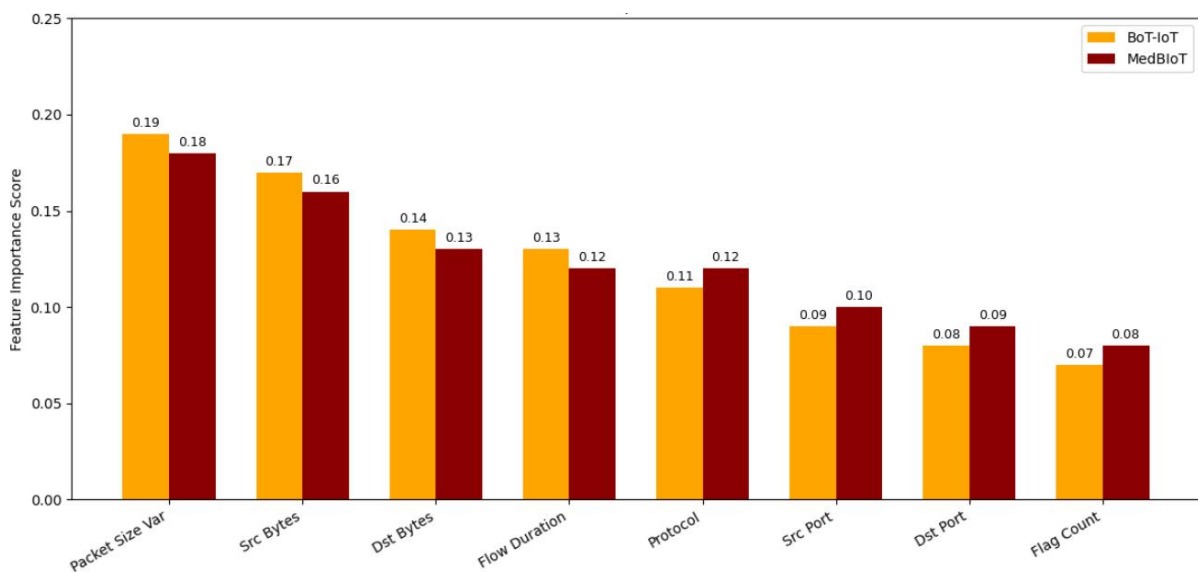


Figure 9: LIME-based Feature Importance Across Datasets

Figure 9 shows a comparison of the BoT-IoT and MedBIoT datasets in terms of feature significance ratings based on LIME. Although their relative importance is somewhat more evenly spread than in SHAP-based rankings, byte-related variables and packet size variation both appear as the top contributors to model predictions in both scenarios. According to LIME, the 'Dst Port' and 'Protocol' aspects are particularly important in MedBIoT since clinical sensors and actuators create domain-specific communication. This underscores LIME's regional sensitivity and verifies that the model's decision limits are consistent and interpretable in both medical and general IoT networks.

4.4 Discussion

The proposed framework hit 96% accuracy on BoT-IoT and 94% on MedBIoT. The model outperformed CNN, LSTM, AE, and XGBoost versions, demonstrating greater F1 scores across unbalanced classes. The individual demonstrated generalisability by maintaining strong performance on both datasets. It is ideal for deployment at the edge in real-time, boasting a model size of just 4.1 MB and an inference time of only 0.9 ms/sample. The system reduces memory and computation requirements while outperforming sophisticated models (Raw DBN, XGBoost).

To keep generalisation intact even on data with skewed distributions, the dropout and pruning strategies, in conjunction with Z-score normalisation and PCA-GWO feature selection, mitigate overfitting. Packet-level and byte-based features consistently rank as top contributors in SHAP and LIME. An important factor for medical IoT systems is explainability, which demonstrates that the model's predictions are trustworthy and interpreted. The pruned DBN layers are fine-tuned more effectively with the help of the WOA. This resulted in improved parameter efficiency and convergence rate. The model is tailored to meet the needs of privacy-conscious federated or distributed IoMT environments thanks to its lightweight architecture and explainability layer.

5. Conclusion and Future Work

For intrusion detection in IoMT networks, this paper provides a revolutionary, lightweight, and understandable DL approach. Superior accuracy and generalisability across varied network environments, including the BoT-IoT and MedBIoT datasets, are achieved by the proposed framework, which combines a pruned DBN with WOA-based optimisation. Transparency in essential decision paths is crucial for security-sensitive healthcare systems, and the combination of SHAP and LIME improves interpretability. Experimental findings show that

the proposed framework greatly outperforms conventional DL and machine learning approaches, not only in terms of accuracy and AUC but also in terms of model size and latency, making it well-suited for edge computing applications. Feature selection using PCA-GWO and normalisation using Z-score further enhanced the compactness and stability of the model. A FL setup will be included in future work to expand this framework and improve privacy-preserving communication across healthcare facilities. Also, a zero-day threat detection will be included by using transfer learning methods. Assess the practical use on wearable devices with limited resources.

Data Availability Statement

This study dataset from following link

<https://www.kaggle.com/datasets/vigneshvenkateswaran/bot-iot> and

<https://www.kaggle.com/datasets/shaily20/medbiot-legit-and-malware>

Conflict of Interest

The authors declare no conflict of interest.

Competing interests

The authors declare no competing interests.

Consent to Participate: Not applicable.

Informed Consent: Not applicable.

Clinical Trial Registration: Not applicable.

Funding Information: No funding is provided for the preparation of the manuscript.

References

- [1] AR, Sathyabama, and Jeevaa Katiravan. "Enhancing anomaly detection and prevention in Internet of Things (IoT) using deep neural networks and blockchain based cyber security." *Scientific Reports* 15.1 (2025): 22369.
- [2] Ogunseyi, Taiwo Blessing, and Gogulakrishan Thiyagarajan. "An Explainable LSTM-Based Intrusion Detection System Optimized by Firefly Algorithm for IoT Networks." *Sensors* 25.7 (2025): 2288.
- [3] Gupta, Brij B., Akshat Gaurav, Razaz Waheeb Attar, Varsha Arya, Ahmed Alhomoud, and Kwok Tai Chui. "A sustainable w-rlg model for attack detection in healthcare iot systems." *Sustainability* 16, no. 8 (2024): 3103.

- [4] Punitha, S., and K. S. Preetha. "Enhancing reliability and security in cloud-based telesurgery systems leveraging swarm-evoked distributed federated learning framework to mitigate multiple attacks." *Scientific Reports* 15.1 (2025): 27226.
- [5] Gaspar, Diogo, Paulo Silva, and Catarina Silva. "Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron." *IEEE Access* 12 (2024): 30164-30175.
- [6] Li, Shancang, and Neetesh Saxena. "Explainable AI-based intrusion detection in IoT systems." *Internet of Things* 31 (2025): 101589.
- [7] Hermosilla, Pamela, Sebastián Berríos, and Héctor Allende-Cid. "Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models." *Applied Sciences* 15, no. 13 (2025): 7329.
- [8] Krishnan, Deepa, Swapnil Singh, and Vijayan Sugumaran. "Explainable AI for zero-day attack detection in IoT networks using attention fusion model." *Discover Internet of Things* 5.1 (2025): 83.
- [9] Si-Ahmed, Ayoub, Mohammed Ali Al-Garadi, and Narhimene Boustia. "Survey of machine learning based intrusion detection methods for internet of medical things." *Applied Soft Computing* 140 (2023): 110227
- [10] AboulEla, Samar, Nourhan Ibrahim, Sarama Shehmir, Aman Yadav, and Rasha Kashef. "Navigating the cyber threat landscape: An in-depth analysis of attack detection within IoT ecosystems." *Ai* 5, no. 2 (2024): 704-732.
- [11] Salman, Israa Ramadhan, Ali Abdulridha Rasheed, Saif Al-Deen H. Hassan, Rasha Abed Hussein, and M. A. Abdelzاهر. "Efficient Human Activity Recognition using PCA Dimensionality Reduction and GWO-Enhanced LSTM." *2024 Journal of Advanced Research in Applied Sciences and Engineering Technology* 54(2):317-343. DOI:10.37934/araset.54.2.317343.
- [12] Bhatnagar, Archita, Divya Bhatnagar, and Tarun Kumar. "Differential Evolutionary Optimization Algorithm for Energy-Efficient Routing Strategy in Wireless Sensor Networks." *International Journal of Computer Networks and Applications (IJCNA)* DOI: 10.22247/ijcna/2025/01, Volume 12, Issue 1, January – February (2025)
- [13] Liu, Ya, Yingjie Zhou, Kai Yang, and Xin Wang. "Unsupervised deep learning for IoT time series." *IEEE Internet of Things Journal* 10, no. 16 (2023): 14285-14306.
- [14] Chang, Liyuan, and Bin Cao. "Towards Efficient Network Traffic Classifications via Multi-Modal Learning." *IEEE Internet of Things Journal* (2025).

- [15] Saied, Mohamed, Shawkat Guirguis, and Magda Madbouly. "Review of filtering based feature selection for Botnet detection in the Internet of Things." *Artificial Intelligence Review* 58.4 (2025): 119.
- [16] Haque, Safwana, Fadi El-Moussa, Nikos Komninos, and Rajarajan Muttukrishnan. "A systematic review of data-driven attack detection trends in IoT." *Sensors* 23, no. 16 (2023): 7191.
- [17] Ali, Haifa Ali Saeed. "Deep Learning Approach for Protecting IoT Smart Home Devices Against Multiclass Botnet Attacks."
- [18] Chen, R., Dai, T., Zhang, Y., Zhu, Y., Liu, X., & Zhao, E. (2024). GBDT-IL: incremental learning of gradient boosting decision trees to detect botnets in internet of things. *Sensors*, 24(7), 2083.
- [19] Nasayreh, Ahmad, Haris M. Khalid, Hamza K. Alkhateeb, Jalal Al-Manaseer, Abdulla Ismail, and Hasan Gharaibeh. "Automated detection of cyber attacks in healthcare systems: A novel scheme with advanced feature extraction and classification." *Computers & Security* 150 (2025): 104288.
- [20] Algethami, Sarah A., and Sultan S. Alshamrani. "A deep learning-based framework for strengthening cybersecurity in internet of health things (IoHT) environments." *Applied Sciences* 14.11 (2024): 4729.
- [21] Benmalek, Mourad, Abdessamed Seddiki, and Kamel-Dine Haouam. "SNN-IoMT: A Novel AI-Driven Model for Intrusion Detection in Internet of Medical Things." *CMES-Computer Modeling in Engineering and Sciences* 143.1 (2025): 1157-1184.
- [22] Ravi, Vinayakumar, Tuan D. Pham, and Mamoun Alazab. "Deep learning-based network intrusion detection system for Internet of medical things." *IEEE internet of things magazine* 6.2 (2023): 50-54.
- [23] Albattah, Albatul, and Murad A. Rassam. "Detection of adversarial attacks against the hybrid convolutional long short-term memory deep learning technique for healthcare monitoring applications." *Applied Sciences* 13.11 (2023): 6807.
- [24] Kumar, Manish, Sushil Kumar Singh, and Sunggon Kim. "Hybrid deep learning-based cyberthreat detection and IoMT data authentication model in smart healthcare." *Future Generation Computer Systems* 166 (2025): 107711.
- [25] Amjath, Mohamed, Shagufta Henna, and Upaka Rathnayake. "Graph representation federated learning for malware detection in internet of health things." *Results in Engineering* 25 (2025): 103651.