



Edge-Optimized Deep Neural Network for Real-Time Intrusion Detection in Telehealth Systems: A Pruning and Quantization Approach

Okuta, Joseph A.^{1*}, Agu, Edward O.¹, and Giliki Jerison¹

¹Department of Computer Science, Faculty of Computing and Information System, Federal University Wukari, Nigeria.

*Corresponding Author

Okuta, Joseph A.

Email: joe.okuta@gmail.com

Article History

Received: 23.06.2026

Revised: 08.08.2026

Accepted: 02.09.2026

Published: 21.09.2026

Abstract: The growing reliance on telehealth and Internet of Medical Things (IoMT) technologies has improved healthcare accessibility but exposed patient data and clinical operations to cybersecurity threats, including unauthorized access, spoofing, and data alteration attacks. Conventional cloud-based intrusion detection systems (IDS) are unsuitable for telehealth environments because they introduce latency, bandwidth overhead, and privacy risk. This study proposes an Edge-Optimized Deep Neural Network integrated with Marine Predators Algorithm feature selection (EO-DNN-MPA) for real-time intrusion detection in telehealth systems. The model was trained and evaluated on the WUSTL-EHMS-2020 healthcare cybersecurity dataset following preprocessing, class balancing, and feature optimization that reduced dimensionality from 36 to 16 features. Three model configurations were evaluated: the original EO-DNN-MPA, a pruned version, and a pruned-and-quantized version. The original model achieved 95.33% accuracy, 93.15% precision, 97.85% recall, 95.44% F1-score, 99.29% AUC, and an inference time of 1.1101 ms. The pruned-and-quantized model maintained comparable detection performance, 95.12% accuracy, 93.20% precision, 97.34% recall, 95.22% F1-score, and 99.25% AUC, while reducing inference time to 0.6679 ms, a 39.8% latency reduction. The proposed model outperformed comparable state-of-the-art approaches evaluated on the same dataset, including federated DNN, ensemble boosting, and explainable XGBoost methods. The results demonstrate that combining edge-aware architectural design, pruning, and quantization with deep learning yields an efficient, privacy-preserving, and low-latency intrusion detection solution suitable for real-time telehealth security applications.

Keywords: Intrusion Detection; Edge Computing; Deep Neural Network; Telehealth; Model Pruning; Quantization; Internet of Medical Things; Healthcare Cybersecurity.

This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/) (CC BY-NC 4.0)

Cite this article:

Okuta, J. A., Agu, E. O., and Jerison, G., (2026). Edge-Optimized Deep Neural Network for Real-Time Intrusion Detection in Telehealth Systems: A Pruning and Quantization Approach. *ISAR Journal of Science and Technology*, 4(9), 45-49.

Introduction

Cloud computing, wireless networks, and the Internet of Things (IoT) have revolutionized healthcare delivery through telehealth, allowing for long-distance health care and communication, electronic health record sharing, and continuous patient monitoring (Kong et al., 2019). This growth has, however, added a number of cybersecurity challenges since the telehealth infrastructure is very distributed and includes sensors, wearables, gateways, and mobile devices, which are potential attack surfaces (Mahmood et al. 2022). Cyber attacks in such systems can lead to breaches of patient privacy and, in extreme instances, misdiagnosis or even non-treatment (Aldaej et al., 2023). Most of the conventional Intrusion Detection Systems (IDS) are built for centralized or cloud-based systems, where data is continuously sent to remote servers, leading to high latency, bandwidth usage and privacy leakage (Li et al., 2023). The challenge of edge computing is to move the decision-making closer to the edge, thereby allowing localized, real-time intrusion detection with minimal latency and maximizing the preservation of privacy (Huynh & Nguyen, 2024). Although Deep Neural Networks (DNNs) have greatly improved the performance of IDS, their high computational demands have posed challenges for deployment on resource-limited edge devices, prompting methods like pruning (He & Xiao, 2023) to eliminate redundant connections and quantization (Weng, 2021) to decrease the accuracy of model weights and inference time. This study aims to tackle these challenges by proposing a new method called an Edge-Optimized Deep Neural Network integrated with Marine Predators Algorithm-based feature selection (EO-DNN-MPA) for telehealth intrusion detection, and to investigate the impact of structured pruning and quantization on telehealth intrusion detection performance and inference efficiency. A goal is to create an EO-DNN model for real-time intrusion detection and classification of telehealth network traffic, to deploy the model in an edge-computing-aware framework to guarantee low latency detection, and to assess the model's performance using accuracy, precision, recall, F1-score, AUC, and inference time.

Methods

A. Dataset and Preprocessing

The study dataset was WUSTL-EHMS-2020 with 44 features, 16,318 samples, and 14,272 samples representing normal activity and 2,046 samples representing attack activity, representing 87.5% and 12.5% respectively. The dataset was initially cleaned, then categorical encoded, Min-Max normalized and class balanced using the Synthetic Minority Oversampling Technique (SMOTE) to build a new attack category with 14,265 samples. Next, the dimensionality of the features was reduced to 16, a small

number of discriminative features, by applying the feature selection technique called Marine Predators Algorithm (MPA) to 36 features, in order to provide a compact and discriminative input representation to the classification model.

B. Model Architecture

The Edge-Optimized Deep Neural Network (EO-DNN) was designed as a feed-forward multilayer neural network consisting of the input layer, a number of hidden layers and output classification layer. The feature vector obtained by the MPA-based selection was reduced to 16 dimensional and passed to the input layer. All the hidden layers applied a nonlinear transformation with a weight matrix, bias vector and Rectified Linear Unit (ReLU) activation function, which for edge processors with limited resources is of low computational complexity and helps to avoid the vanishing gradients. To enhance generalization and training stability, we added batch normalization and dropout regularization. The output layer used a sigmoid activation function and the classification was performed using binary cross-entropy loss, in the case of binary intrusion classification.

C. Edge-Aware Optimization Strategy

The EO-DNN added a number of optimizations to support deployment on resource constrained edge nodes, including telehealth gateways and embedded medical devices. Network depth and width were limited by architectural simplicity, requiring only a certain number of floating-point operations. Structured magnitude-based pruning was performed, which eliminated the weight connections whose absolute values were lower than a given value, so that the network was not dense, thereby reducing the number of operations and the storage space required to maintain the network while maintaining its ability to make predictions. Post training quantization converted the floating point weights into low precision integer weights of 8 bits, which significantly reduced the amount of memory required and increased the speed of inference on edge accelerators. In some cases, the Leaky-ReLU activation was employed as a simple alternative, since neurons do not get “inactive” during training.

D. Training, Validation, Evaluation.

The dataset was stratified and split into training (70%), validation (15%) and testing (15%) sets to ensure that the distribution of the classes remains the same in all three subsets. The model was developed in Python with TensorFlow/Keras, NumPy and Pandas for numerical processing and manipulation, Scikit-learn for model evaluation and Matplotlib/Seaborn for visualization in Anaconda. The following three model configurations were tested: EO-DNN-MPA (Original DNN-MPA), Pruned EO-DNN-MPA, and Pruned and Quantized EO-DNN-MPA.

The performance metrics were analyzed in terms of accuracy, precision, recall, F1-score, Area Under the ROC Curve (AUC), classification threshold and inference time, based on the confusion matrix of true positive, true negative, false positive and false negative.

Results and Discussions

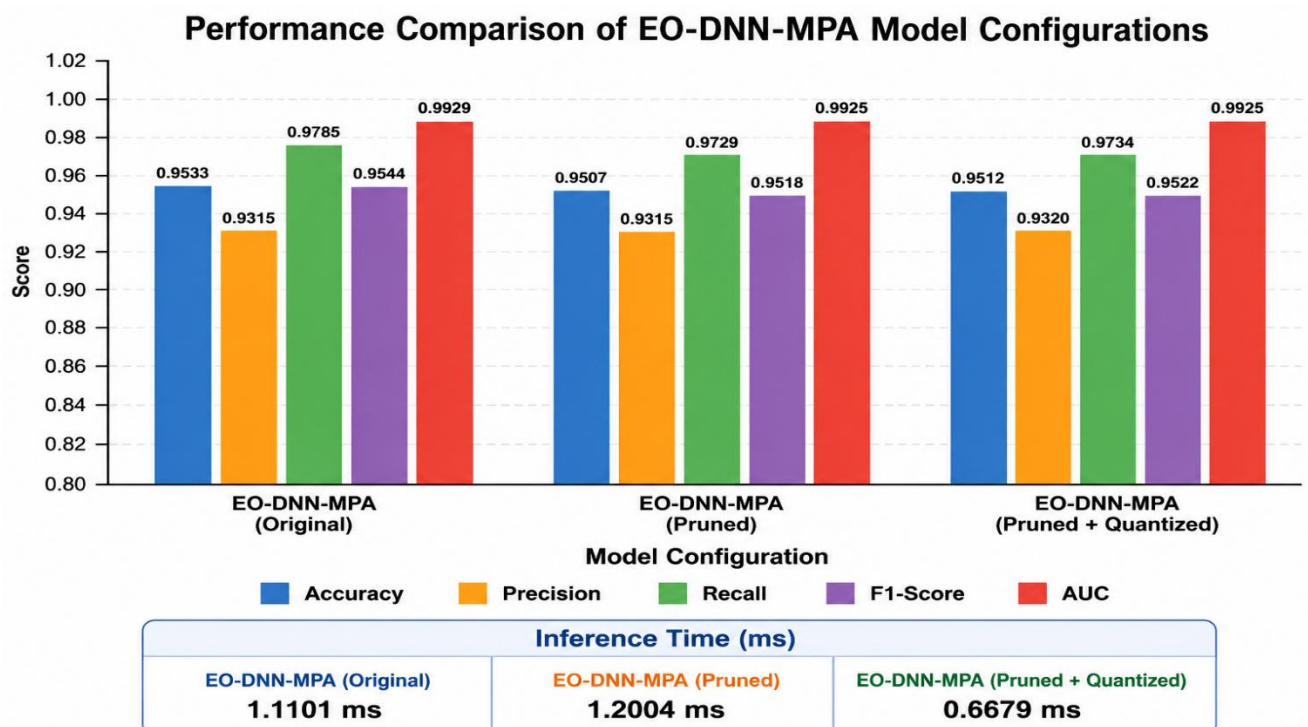
A. Model Performance Comparison

Table 1 presents the performance of the three EO-DNN-MPA configurations. The original model achieved the

highest overall performance, with 95.33% accuracy, 93.15% precision, 97.85% recall, 95.44% F1-score, and 99.29% AUC, at an inference time of 1.1101 ms. The high recall indicates that the model correctly identified the large majority of intrusions, minimizing false negatives, an important property for healthcare networks where undetected attacks can compromise patient safety.

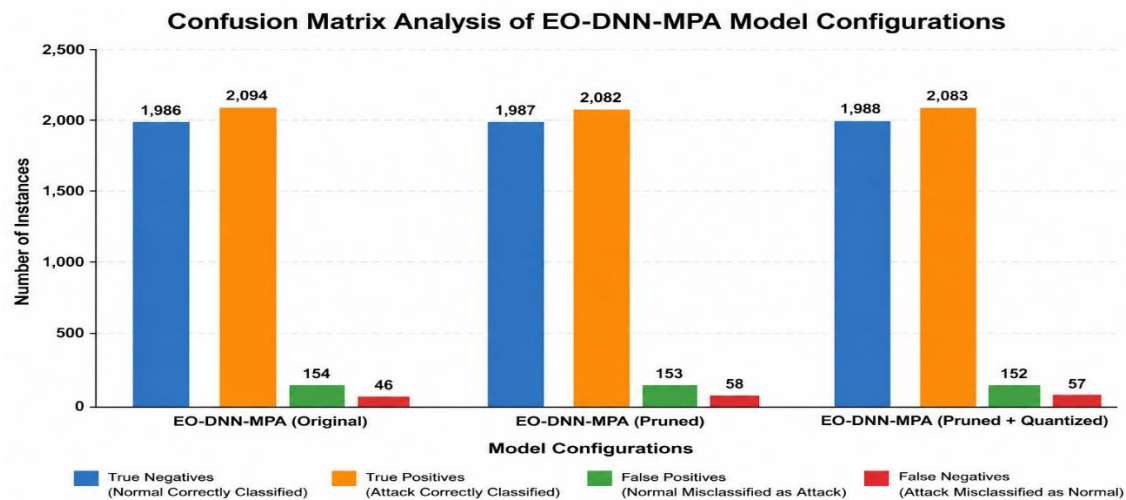
Table 1. Performance Comparison of EO-DNN-MPA Model Configurations

Model	Accuracy	Precision	Recall	F1-Score	AUC	Inference (ms)
EO-DNN-MPA (Original)	0.9533	0.9315	0.9785	0.9544	0.9929	1.1101
EO-DNN-MPA (Pruned)	0.9507	0.9315	0.9729	0.9518	0.9925	1.2004
EO-DNN-MPA (Pruned + Quantized)	0.9512	0.9320	0.9734	0.9522	0.9925	0.6679



The accuracy dropped just 0.26% from the original model to 95.07% after pruning, but there was not a significant drop in AUC to 99.25%. This means that parameters removed had little influence on the overall predictive ability. The pruned and quantized configuration achieved the best accuracy (95.12%), precision (93.20%), recall (97.34%) and F1 score (95.22%) with only 39.8% reduction of inference time from 1.1101ms to 0.6679ms. In all three cases, pruning and quantization maintained accuracy of more than 95%, F1-scores of more than 95% and AUC scores of more than 99% without significantly degrading detection performance, substantiating the fact that these techniques greatly enhance the computational efficiency of the system.

B. Confusion Matrix Analysis

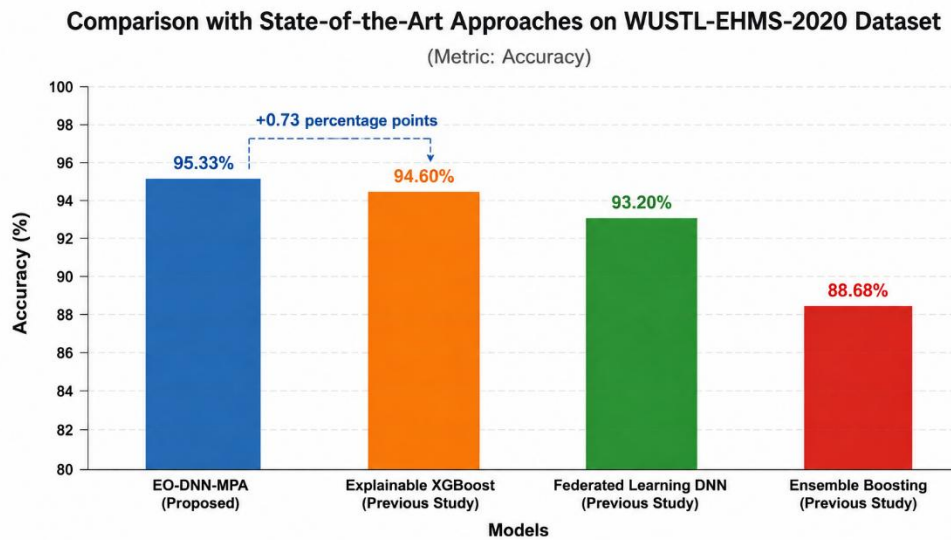


C. Comparison with State-of-the-Art Approaches

A proposed EO-DNN-MPA framework for intrusion detection was compared with previous studies on intrusion detection on the same WUSTL-EHMS-2020 data set. The proposed model was able to achieve an accuracy of 95.33%, which is higher than explainable XGBoost (94.60%), federated learning DNN (93.20%) and ensemble boosting (88.68%). While this increase is relatively small (0.73 percentage points), it could represent a significant advantage in healthcare intrusion detection applications where small improvements in detection accuracy are directly realized as greater security of patient information and better identification of cyber threats. **Conclusions** This study proposes and tests an Edge-Optimized Deep Neural Network combined with Marine Predators Algorithm feature selection (EO-DNN-MPA) to filter out intrusions in real-

The original EO-DNN-MPA model correctly classified 1,986 normal instances and 2,094 attacks, resulting in 154 false positives and 46 false negatives. After the removal of redundant connections, 1,987 normal instances and 2,082 attacks were correctly classified, highlighting a slight rise in the number of attacks missed due to the redundant connections being removed, 153 false positives and 58 false negatives. The pruned and quantized model correctly identified 1,988 normal instances and 2,083 attacks, while giving 152 false positive and 57 false negative in the normal traffic classification, which is a little better than the pruned-only model. Similarly, in all three configurations, false negatives were low (46-58 instances), reflecting robust attack-detection capabilities for real-time telehealth use.

time telehealth systems. The proposed model demonstrated high detection performance, with the original model reaching 95.33% accuracy and 99.29% AUC, and the pruned and quantized model reducing the inference speed by about 39.8% with minimal performance degradation. The results show that the integration of edge-aware architectural design, structured pruning, and quantization with deep learning offers an efficient, private, and real-time intrusion detection solution that can be applied to real-time applications in telehealth and IoMT and can outperform similar state-of-the-art methods tested on the same benchmark dataset. Future work should confirm the proposed model based on real-time healthcare network traffic, and explore the integration of federated learning for decentralized privacy-preserving training and empirical evaluation on various hardware configurations on the edge.



Acknowledgements

The author acknowledges the supervision and guidance received from the Department of Computer Science, Federal University Wukari, during the course of this research.

References

1. Aldaej, A., Ahanger, T. A., & Ullah, I. (2023). Deep Learning-Inspired IoT-IDS Mechanism for Edge Computing Environments. *Sensors*, 23(24), 9869.
2. He, Y., & Xiao, L. (2023). Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5), 2900-2919.
3. Huynh Trong, T., & Nguyen Hoang, T. (2024). Effective Multi-Stage Training Model for Edge Computing Devices in Intrusion Detection. *arXiv preprint*.
4. Kasongo, S. M. (2023). A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework. *Computer Communications*, 199, 113-125.
5. Kong, L., Sadatsafavi, H., & Verma, R. (2019). Usage and impact of information and communication technologies in healthcare delivery. *Service Science*, 11(3), 172-188.
6. Li, Y., Wang, T., & Zhao, K. (2023). Privacy-preserving intrusion detection for telehealth using federated edge learning. *IEEE Internet of Things Journal*, 10(5), 4221-4235.
7. Mahmood, S., Rehman, S., & Naeem, M. (2022). Cybersecurity Threats and Countermeasures in IoMT-Based Healthcare Systems. *IEEE Access*, 10, 112345-112358.
8. Weng, O. (2021). Neural network quantization for efficient inference: A survey. *arXiv preprint arXiv:2112.06126*.
9. Mosaiyebzadeh, F., Pouriyeh, S., Han, M., Liu, L., Xie, Y., Zhao, L., & Batista, D. M. (2025). Privacy-Preserving Federated Learning-Based Intrusion Detection System for IoHT Devices. *Electronics*, 14(1), 67.
10. Alsolami, T., Alsharif, B., & Ilyas, M. (2024). Enhancing cybersecurity in healthcare: Evaluating ensemble learning models for intrusion detection in the Internet of Medical Things. *Sensors*, 24(18), 5937.
11. Alshammari, A., Alotaibi, M., & colleagues. (2025). Explainable XGBoost framework for intrusion detection in Internet of Medical Things healthcare systems. *Scientific Reports*.