Trustworthy Network Intrusion Detection Systems for IoMT: Explainability and Optimization

Donato D'Ambrosio, Alfredo Nascita, Antonio Montieri, Antonio Pescapè
dona.dambrosio@studenti.unina.it, {alfredo.nascita, antonio.montieri, pescape}@unina.it

***Abstract*—The rapid growth of the Internet of Medical Things (IoMT) has significantly increased the exposure of healthcare infrastructures to cyber threats, making effective and trustworthy Network Intrusion Detection Systems (NIDSs) essential. While NIDSs based on Deep Learning (DL) models achieve high detection performance, their black-box nature limits transparency, hinders trust, and complicates their adoption in safety-critical environments, such as healthcare. This work investigates the role of eXplainable Artificial Intelligence (XAI) in IoMT intrusion detection beyond pure post-hoc interpretation. Specifically, we analyze DL-based NIDS using SHapley Additive exPlanations (SHAP) to characterize feature relevance and model decision behavior, and we evaluate the consistency of explanations across different model architectures. Furthermore, we leverage SHAP as a design tool to guide input reduction, enabling the development of lightweight models capable of earlier intrusion detection from shorter traffic sequences. This optimization improves detection timeliness while reducing computational complexity. We evaluate the proposed DL-based NIDS using a holistic approach that encompasses detection performance, computational complexity, and reliability. Our results demonstrate that XAI can effectively support not only the interpretation but also the design of more transparent, reliable, and efficient NIDSs for IoMT environments.**

***Index Terms*—eXplainable AI, Network Intrusion Detection Systems, Network Attack Traffic, Internet of Medical Things**

I. Introduction

The rapid proliferation of the *Internet of Medical Things (IoMT)* has facilitated the interconnection of heterogeneous medical devices, sensors, and applications, supporting advanced healthcare services such as remote monitoring and real-time data analysis. However, this pervasive connectivity significantly expands the attack surface of healthcare infrastructures, exposing them to cyber threats that can compromise data confidentiality, system availability, and patient safety. In this context, *Network Intrusion Detection Systems (NIDSs)* play a pivotal role in monitoring network traffic and detecting malicious activities in IoMT environments.

Machine Learning (ML) and, in particular, Deep Learning (DL) techniques have demonstrated strong performance in network traffic classification [1] and intrusion detection [2], leveraging their ability to model complex and high-dimensional patterns. Nevertheless, these models are typically deployed as black-box predictors, masking the rationale behind their decisions. This lack of transparency undermines trust, complicates incident response, and raises operational concerns in safety-critical domains such as healthcare.

ISBN 978-3-903176-82-9 © 2026 IFIP

Recently, *eXplainable Artificial Intelligence (XAI)* has emerged as a key paradigm to mitigate these transparency issues. Among the various interpretability techniques, feature-attribution methods, such as *SHapley Additive exPlanations (SHAP)*, have gained significant traction due to their ability to quantify the impact of individual input variables on model outcomes [3]. In the NIDS domain, these methods are primarily employed for post-hoc interpretation of predictions to identify the most relevant traffic features associated with specific attack vectors. However, current research often treats explainability as a diagnostic end-goal rather than an architectural driver. Existing approaches primarily focus on interpretability in isolation, seldom exploiting XAI as a tool to support model design, improve computational efficiency, or accelerate detection timeliness. Moreover, the inter-model consistency of these explanations and probabilistic reliability of models remain largely underexplored.

In this work, we investigate the role of XAI in IoMT intrusion detection beyond post-hoc interpretation. Specifically, we analyze model behavior through SHAP-based explanations, evaluate the consistency of feature importance across different architectures, and assess model reliability through calibration. Furthermore, we leverage explainability to guide input selection and reduction, enabling the design of lightweight models capable of earlier intrusion detection from shorter traffic sequences.

This work provides the following contributions:

- We investigate the use of SHAP to interpret DL-based NIDSs in IoMT, analyzing input relevance and characterizing model decision behavior.
- We propose a comprehensive evaluation framework that integrates detection performance, computational complexity, and probabilistic reliability, alongside a cross-model evaluation of SHAP feature importance consistency across different DL architectures and attack classes.
- We leverage XAI to drive input reduction, designing lighter models capable of earlier intrusion detection to improve responsiveness in time-sensitive IoMT scenarios.

The rest of the manuscript is structured as follows. Sec. II reviews related work. Sec. III outlines our methodology for trustworthy NIDS. Sec. IV describes the experimental setup, while Sec. V presents the evaluation results. Lastly, Sec. VI ends our paper and presents possible future research.

II. Related Work

In recent years, the adoption of ML and DL techniques for NIDSs to secure IoMT environments has significantly increased, driven by the proliferation of connected medical devices and the critical sensitivity of healthcare data. Early works in IoT intrusion detection demonstrate the effectiveness of DL approaches, such as convolutional and fully connected neural networks [4]. However, as highlighted in [5], NIDS performance is highly domain-dependent, with IoMT introducing unique challenges related to traffic heterogeneity and data criticality.

Several approaches have been proposed to address this critical task. Lightweight, deployable solutions based on classical ML are investigated for practical IoMT scenarios in [6–8]. More recently, Alsharaiah et al. [9] propose a transformer-based framework for spoofing detection, while Sohail et al. [10] show that ensemble methods like XGBoost achieve competitive results on the `CIC-IoMT24` dataset [11]. Other works explore advanced modeling strategies, including cross-layer traffic analysis [12] and sequential architectures to capture temporal dependencies in network traffic [13].

Concurrently, XAI has been increasingly adopted to improve the transparency of DL-based NIDSs. In the related literature, XAI predominantly acts as a post-hoc layer to explain predictions, analyze feature relevance, and validate decision patterns [14], particularly in medical contexts where trustworthiness is paramount [15, 16]. Feature attribution methods, such as SHAP and LIME, are widely used to provide both local and global explanations [3]. Various studies apply these techniques to network traffic classification. The works in [2, 17, 18] combine multiple XAI methods to analyze feature contributions in IoT datasets. Specifically for IoMT, Alsharaiah et al. [9] and Sharma and Shambharkar [19] employ SHAP to identify the most influential features for intrusion detection on the `CIC-IoMT24` dataset. Explainability has also been extended to more complex settings, including cross-layer traffic analysis [12] and temporal modeling, where feature relevance evolves across packet sequences [13].

Despite these advances, current literature predominantly confines XAI to a passive diagnostic role, aiming primarily to interpret model outputs. The potential of explainability to actively engineer model design, particularly for reducing computational complexity and accelerating detection, remains largely untapped. Moreover, critical dimensions such as the consistency of explanations across distinct models and reliability are rarely investigated. Finally, prior research typically evaluates NIDSs through isolated lenses, neglecting the inherent trade-offs between trustworthiness, reliability, detection accuracy, and operational timeliness [1]. To bridge these gaps, this work introduces an XAI-driven framework for intrusion detection in IoMT positioned at the intersection of model analysis and optimization. Rather than limiting XAI to interpreting NIDS decisions, we leverage SHAP attributions to dynamically guide input sequence reduction. Differently from prior approaches based on feature selection or dimensionality reduction, our method targets the temporal dimension by reducing the input biflow sequence length. By coupling this architectural refinement with a holistic, multi-view evaluation (encompassing detection performance, temporal and spatial overhead, and calibration), this work demonstrates how explainability can practically shape the design of lighter, faster, and highly reliable NIDSs tailored to the strict requirements of healthcare networks.

III. Trustworthy NIDS

A. Deep Learning-based Attack-Traffic Classification

In this work, we formalize the problem of intrusion detection in IoMT environments as a multi-class attack-traffic classification task. Formally, given a traffic object, defined as an aggregation of packets sharing common properties, the goal is to assign each instance to one of L distinct classes, representing either benign traffic or a specific attack category.

To this end, we employ state-of-the-art DL classifiers, exploring diverse architectural paradigms to model the underlying traffic patterns. Particular attention is devoted to constructing a robust input representation, avoiding data biases that could artificially inflate predictive performance and, consequently, lead to misleading conclusions [2].

B. Assessing Input Importance through SHAP

We interpret model predictions using feature-attribution methods that assign an importance score e_m to each feature of the input x, such that the sum of attributions approximates the model output $F(x)$. We focus on local, post-hoc explanations, analyzing model behavior in the neighborhood of each instance. Specifically, we compute Shapley values via (expected) *Gradient SHAP* [20], where e_m quantifies the contribution of the m-th input feature to the predicted probability $\hat{p}(x)$. Positive values act as evidence supporting the prediction, whereas negative values contrast it, indicating a decrease in the model's confidence. Global explanations are obtained by aggregating normalized SHAP values across correctly classified samples, computing the median attribution to characterize the model's overall decision behavior.

To evaluate explanations' consistency, we conduct a cross-model analysis comparing the Top-k most important features identified for different DL architectures. The similarity between these feature sets is measured using the *Dice-Sørensen index*:

$$DS(M_1, M_2) = \frac{|\text{Top}_k(\text{SHAP}_{M_1}) \cap \text{Top}_k(\text{SHAP}_{M_2})|}{k} \quad (1)$$

where $\text{Top}_k(\text{SHAP}_{M_i})$ denotes the set of the k most relevant features according to the global SHAP ranking of model M_i. The metric (expressed as a percentage) ranges in $[0, 100]\%$, with higher values indicating stronger agreement. To ensure a fair comparison, this analysis is limited to samples correctly classified by all the models under evaluation.

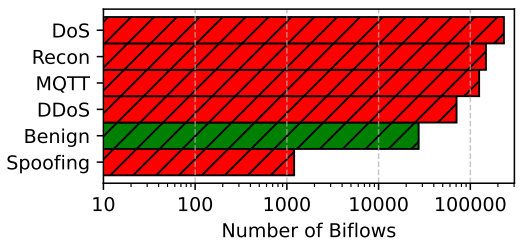

Fig. 1: Number of biflows per traffic class in the CIC-IoMT24 dataset.

IV. Experimental Setup

A. IoMT Dataset Description

All experiments are conducted on the `CIC-IoMT24` dataset [11], a recent benchmark for cybersecurity in IoMT environments. The dataset comprises network traffic generated across heterogeneous communication technologies (e.g., MQTT, Bluetooth, and Wi-Fi), organized into a `Benign` class representing legitimate traffic and five attack macro-categories: `DDoS`, `DoS`, `MQTT`, `Recon`, and `Spoofing`. The class distribution in terms of the number of biflows[1] is depicted in Fig. 1. The dataset is originally provided in PCAP format; traffic traces are preprocessed to extract the following six per-packet header features for the first 10 packets of each biflow: (i) the packet length in bytes (PL); (ii) the packet direction $\in \{-1, 1\}$ (DIR); (iii) the TCP window size (WIN), set to zero for non-TCP packets; (iv) the inter-arrival time (IAT), representing the time elapsed since the arrival of the previous packet; (v) the time-to-live (TTL); and (vi) the TCP flags (FLG), encoding the binary vector of control flags as a decimal integer. These features capture both statistical and protocol-level characteristics of the network traffic, providing a compact yet highly informative representation for intrusion detection.

B. Deep Learning Architectures

To realize the IoMT NIDS, we evaluate two complementary DL architectures, a Convolutional Neural Network (CNN) and a Long Short-Term Memory (LSTM), which capture distinct structural facets of network traffic. Each input biflow is encoded as a fixed-length sequence of 10 packets (with 6 features per packet as described in Sec. IV-A), applying zero-padding where necessary. Both models are trained for 200 epochs with a batch size of 64.

CNN. This architecture is designed to extract localized spatial patterns by processing the input as a 2D packet-feature map. It consists of two sequential composite convolutional blocks encompassing convolution, ReLU activation, max-pooling, and batch normalization, followed by a fully connected layer (with 200 neurons) and the final softmax classifier.

LSTM. This architecture captures temporal dependencies by modeling the sequential evolution of network traffic as a time series. It employs a unidirectional LSTM layer whose final hidden state serves as a comprehensive sequence representation, which is subsequently fed to a fully connected layer (with 200 neurons) and the final softmax. All experiments were performed on virtual machines equipped with 24 vCPU cores and 48 GB of RAM.

[1]A bidirectional flow or *biflow* is defined as a set of packets sharing the same quintuple (i.e., transport-level protocol, source and destination IP addresses, and ports) regardless of the direction of communication.

C. Evaluation Framework and Metrics

To provide a comprehensive assessment aligned with our holistic approach, alongside the SHAP-based feature attribution, we analyze the proposed NIDS across three core dimensions: detection performance, computational complexity, and model reliability (via calibration analysis). Specifically, SHAP is employed to derive both per-class and global importance scores (as detailed in Sec. III-B), supporting the cross-model analysis of feature relevance across different architectures and guiding model optimization.

Detection Performance. Classification capabilities are assessed using standard metrics, namely *Accuracy*, *Precision*, *Recall*, and *F1-score*. To ensure robust estimates, all reported values are averaged over five independent runs.

Computational Complexity. To evaluate deployment feasibility in time-sensitive IoMT scenarios, we track three efficiency metrics: (i) the *waiting time* required to assemble the input sequence; (ii) the *training time*; and (iii) the *number of trainable parameters*, serving as a proxy for memory footprint and inference overhead.

Model Reliability. We assess model reliability via *calibration*, which measures the consistency between predicted probabilities and observed outcomes. Let $\hat{P} \in [0, 1]$ denote the confidence assigned to the predicted class, and let $\hat{Y}$ and Y denote the predicted and true labels, respectively. A model is perfectly calibrated if $\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) = p, \forall p \in [0, 1]$. Deviations are quantified using the *Expected Calibration Error (ECE)*. The probability interval $[0, 1]$ is partitioned into B disjoint bins $\{I_b\}_{b=1}^{B}$. For each bin, we define empirical accuracy $\mathrm{acc}(I_b) = \frac{1}{|I_b|} \sum_{i \in I_b} 1(\hat{Y}_i = Y_i)$ and average confidence $\mathrm{conf}(I_b) = \frac{1}{|I_b|} \sum_{i \in I_b} \hat{P}_i$, where $|I_b|$ is the number of samples in bin I_b. The ECE is then computed as $\mathrm{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{n} |\mathrm{acc}(I_b) - \mathrm{conf}(I_b)|$, where n is the total number of samples. Lower ECE values indicate better calibration and more reliable probabilistic predictions.

V. Experimental Results

In this section, we present the experimental results. First, Sec. V-A compares the CNN and LSTM models in terms of detection performance. Then, Sec. V-B provides a SHAP-based explainability analysis at both per-class and global levels. Finally, Sec. V-C discusses the SHAP-guided NIDS optimization, evaluating its impact on efficiency, detection timeliness, and model reliability via calibration analysis.

A. Detection Performance Overview

As summarized in Tab. I, both the CNN and LSTM architectures achieve comparable and effective performance across all

TABLE I: Detection performance metrics.

Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
CNN	96.13 ± 0.02	93.48 ± 0.43	90.65 ± 0.10	91.56 ± 0.24
LSTM	96.07 ± 0.02	93.12 ± 0.24	90.53 ± 0.11	91.32 ± 0.10

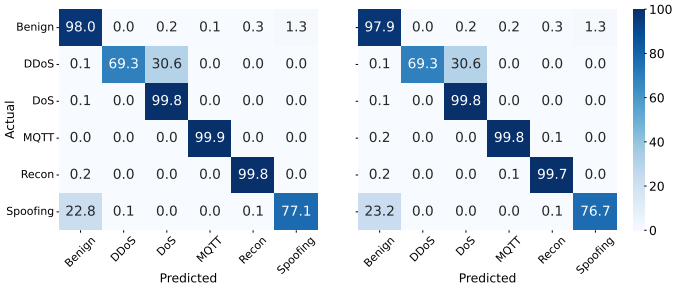

Fig. 2: Normalized confusion matrices for NIDS based on CNN (left) and LSTM (right) models.

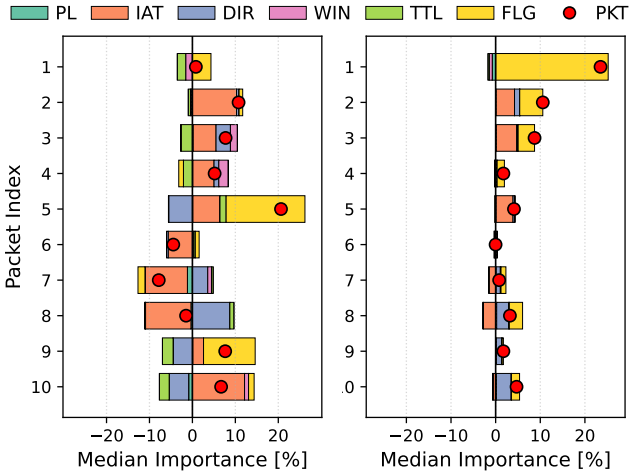

Fig. 3: Explainability results for the Benign class: comparison between CNN (left) and LSTM (right). Per-packet importance is obtained by aggregating per-feature contributions.

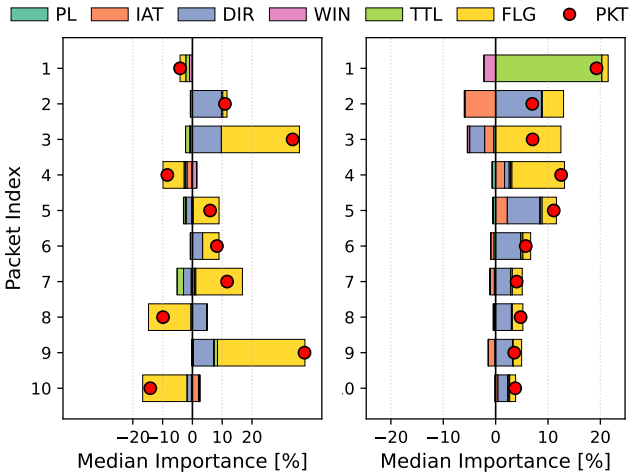

Fig. 4: Explainability results for the DoS class: comparison between CNN (left) and LSTM (right). Per-packet importance is obtained by aggregating per-feature contributions.

evaluation metrics, with the CNN exhibiting only a marginal gain in the average F1-score (91.56% vs. 91.32%).

A granular class-wise inspection via the normalized confusion matrices (Fig. 2) reveals common behavioral patterns between the two models. Both architectures almost perfectly identify `DoS`, `MQTT`, and `Recon` attacks, achieving recalls nearing 100%. Conversely, a notable misclassification pattern emerges for `DDoS` traffic: more than 30% of these biflows are erroneously classified as `DoS` by both architectures. This confusion is largely attributable to the inherent class imbalance (as depicted in Fig. 1), combined with the strong structural similarities between the two attack vectors. The `Spoofing` class also exhibits a lower recall (approx. 77%), being heavily penalized by its limited representation in the training set. Notably, nearly all misclassified `Spoofing` instances are erroneously attributed to `Benign` traffic. This suggests the presence of overlapping header-feature patterns, making these specific malicious biflows particularly stealthy and difficult to isolate from legitimate traffic.

B. SHAP-based Explainability Analysis

Per-Class Feature Attribution. Figure 3 illustrates the SHAP-based explainability results for the `Benign` class, comparing the CNN and LSTM models. For each packet position within the input sequence, the stacked horizontal bars represent the median contribution of individual features, while the red marker (PKT) denotes the aggregated packet-level importance, computed by algebraic summing feature contributions within the same packet. Positive values indicate evidence supporting the `Benign` prediction, whereas negative values act as counter-evidence, potentially reducing the model's confidence.

From a packet-level perspective, the two architectures exhibit markedly different behaviors. The CNN shows an irregular distribution of importance across the sequence, characterized by a prominent positive peak at the fifth packet and several packets contributing both positively and negatively (e.g., packets 6, 7, and 8 act mostly as counter-evidence). This suggests that the CNN relies on localized patterns that may emerge at various positions within the biflow. In contrast, the LSTM assigns the highest importance to the first packet, followed by a sharply decreasing trend, with successive packets contributing only marginally to the final NIDS decision. This behavior directly reflects the LSTM's tendency to prioritize early sequential/temporal information.

At the feature level, both models consistently identify the TCP flags (FLG) as the most influential feature, particularly in the most relevant packet positions (e.g., packet 5 for the CNN and packet 1 for the LSTM). Other features, such as IAT and DIR, provide additional contributions, although their impact is more distributed across the sequence. The coexistence of positive and negative feature contributions within the same packet further highlights that individual packets often contain a mix of supporting and conflicting signals. Overall, while both

models rely on a similar subset of core features, they differ fundamentally in how this importance is distributed across the packet sequence, highlighting the complementary nature of spatial (CNN) and temporal (LSTM) network traffic modeling.

Figure 4 shows analogous explainability results for the DoS class, selected as a representative attack category. As in the previous analysis, feature contributions are reported both at the packet level (red markers for aggregated importance) and at the feature level (colored horizontal bars).

From a packet-level perspective, the architectures confirm their distinct processing paradigms. The CNN exhibits an irregular and alternating distribution of importance, characterized by positive peaks at specific positions (e.g., packets 3 and 9) followed by counter-evidence at adjacent packets (e.g., packets 4, 8, and 10). This confirms that the CNN extracts discriminative attack signatures from localized spatial patterns rather than sequential ordering. Conversely, the LSTM maintains a clear tendency to assign higher importance to the initial packets, followed by a decreasing trend, confirming the prominence of early temporal dynamics in its decision-making process, even for DoS traffic.

At the feature level, both models rely on a similar subset of features but differ in how importance is distributed across packets. For the CNN, FLG dominates both positively and negatively, followed by DIR. Conversely, for the LSTM, TTL is the most influential feature in the first packet, acting as a key DoS discriminator and differing from its behavior on Benign traffic. In later packets, the LSTM also relies on FLG and DIR, while IAT contributes with mixed impact. Finally, compared to the Benign class, the DoS attack shows more pronounced feature contributions (up to 40% median importance in the CNN), reflecting stronger and more distinctive traffic patterns.

Cross-Model Explanation Consistency. To quantitatively assess the alignment of the learned representations and related explanations, we compute their consistency across models using the Dice-Sørensen index (Sec. III-B) with $k = 10$. Overall, the results indicate that CNN and LSTM share approx. $50-60\%$ of their most influential features. This confirms that, despite architectural differences, both models capture partially consistent patterns in the traffic, identifying a common subset of features as highly informative for attack-traffic classification. However, this agreement fluctuates significantly across classes, dropping drastically for Spoofing. This behavior is likely linked to its limited sample size: in data-scarce regimes, models struggle to extract universally robust representations, defaulting instead to architecture-specific heuristics that lead to divergent explanations. Ultimately, while a baseline global consistency exists, structural differences remain a dominant factor in challenging and underrepresented attack scenarios.

Global Feature Attribution. The global explainability analysis (Fig. 5) provides further insights into the overall distribution of feature importance across packets. For the LSTM architecture, the aggregated packet-level importance (red markers) exhibits a clear decreasing trend. The initial packets overwhelmingly drive the decision, with successive packets

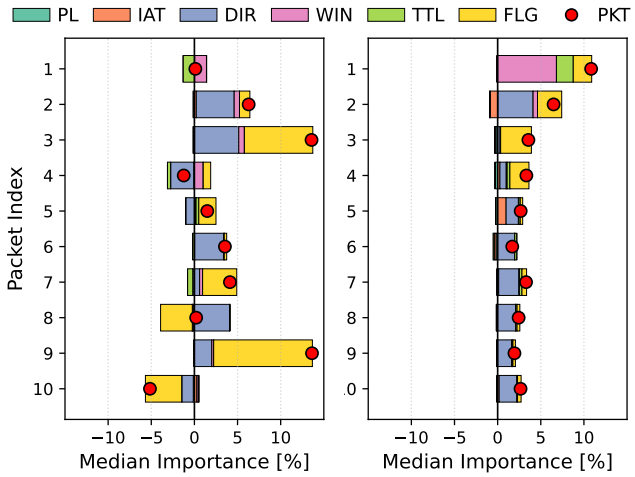

Fig. 5: Global explainability results: comparison between CNN (left) and LSTM (right). Per-packet importance is obtained by aggregating per-feature contributions.

having a progressively marginal (although positive) impact. Conversely, the CNN maintains a more irregular distribution of importance at the global level. While specific late-sequence packets (e.g., packet 9) still show non-negligible importance, the overall contribution of trailing packets is generally limited, and in some cases even negative, indicating that they may act as counter-evidence rather than supporting the detection.

These observations suggest that the most relevant information for classification is densely concentrated in the early portion of the biflow, with diminishing contributions from later packets. This insight motivates the exploration of more compact input representations, as detailed in the next section.

C. SHAP-Guided Model Optimization

The SHAP analysis reveals that the most relevant information for intrusion detection is concentrated in the early portion of the packet sequence, while later packets provide a progressively limited contribution to the prediction. This suggests that the original input length may be partially redundant. Guided by this observation, we design optimized models by reducing the input sequence length from 10 to 5 packets, while retaining the original DL architectures. This modification is not arbitrary; rather, it is directly driven by the explainability analysis, which pinpointed the most informative portion of the input sequence.

TABLE II: Comparison between CNN and LSTM models with different input lengths (i.e., 10 vs. 5 packets per biflow).

	CNN		*LSTM*	
Metric	*10 pkts*	*5 pkts*	*10 pkts*	*5 pkts*
F1-score [%]	91.6	91.4	91.3	91.2
Training Time [h]	5.8	5.0	2.0	1.3
# trainable params	700k	400k	300k	50k
ECE [%]	0.1	2.0	0.1	0.1

The comparison between the original and optimized NIDSs is reported in Tab. II. Results show that reducing the input size does not significantly affect detection performance, as both CNN and LSTM maintain comparable F1-scores. This indicates that the reduced input is sufficient to preserve discriminative capabilities. Crucially, while detection performance remains stable, reducing input length improves computational efficiency. Specifically, the training time for the CNN decreases from 5.8 hours to 5 hours, while the LSTM training time is almost halved, dropping from 2 hours to 1.3 hours. Moreover, processing fewer packets directly translates to a reduced data collection delay: the average waiting time required to assemble the input drops from 31 ms to 12.6 ms. This acceleration enables earlier intrusion detection and improves responsiveness in time-sensitive IoMT cybersecurity scenarios. Additionally, the overall memory footprint is substantially reduced: the CNN trainable parameters drop from 700k to 400k, while the LSTM experiences an even more drastic reduction, from 300k to 50k parameters, making it particularly suitable for resource-constrained environments. From a reliability perspective, the LSTM maintains a low ECE of 0.1% even after input reduction, whereas the CNN shows a larger degradation in calibration (with ECE from 0.1% to 2.0%), mainly due to underconfidence in the 80–100% confidence range, although the model remains reliable. This indicates that SHAP-guided optimization affects architectures differently, with recurrent models showing greater robustness in preserving accurate confidence despite truncated inputs. Overall, these findings demonstrate that XAI can be effectively leveraged not only for post-hoc interpretation but also as an actionable tool to guide model optimization. This approach facilitates the design of lightweight and timely NIDSs without compromising their defensive efficacy.

VI. Conclusion

In this work, we investigated the application of XAI to enhance the trustworthiness of DL-based NIDSs in IoMT environments. We evaluated two DL architectures, namely CNN and LSTM, for attack-traffic classification, demonstrating that both models achieve effective and comparable performance across standard metrics. Moving beyond mere predictive accuracy, we demonstrated how SHAP-based explanations can provide insights into their decision-making processes, uncovering attack-specific feature importance patterns. Furthermore, we showed that explainability can be leveraged also to actively guide NIDS optimization. Specifically, SHAP-guided input reduction enabled the design of lightweight NIDSs with reduced input size and fewer trainable parameters. We demonstrated that this optimization successfully preserves both detection performance and model reliability (assessed through calibration), while drastically reducing training time and data collection delays for earlier intrusion detection. Ultimately, we underline that deploying a NIDS that is simultaneously accurate, explainable, reliable, and timely is paramount for ensuring robust security in critical IoMT scenarios. As future work, we will validate the proposed framework across diverse datasets and real-world IoMT deployments, investigating cross-domain generalization via domain adaptation. quantifying the trade-off between detection performance, computational efficiency, and timeliness, and assess models and explanations robustness under adversarial conditions. Furthermore, we will explore advanced architectures (e.g., Transformers) and class-aware training strategies, to address class imbalance and improve classification performance and calibration. Finally, we will study the integration of lightweight and explainable NIDSs into operational platforms using digital twin environments for realistic, safe, and scalable testing.

References

[1] G. Aceto *et al.*, "AI-Powered Internet Traffic Tlassification: Past, Present, and Future," *IEEE Commun. Mag.*, vol. 62, no. 9, pp. 168–175, 2023.

[2] A. Nascita *et al.*, "Machine and Deep Learning Approaches for IoT Attack Classification," in *IEEE Infocom Wkshps*, 2022, pp. 1–6.

[3] A. Nascita *et al.*, "A Survey on Explainable Artificial Intelligence for Internet Traffic Classification and Prediction, and Intrusion Detection," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 5, pp. 3165–3198, 2024.

[4] B. Sharma *et al.*, "Explainable Artificial Intelligence for Intrusion Detection in IoT Networks: a Deep Learning Based Approach," *Expert Systems with Applications*, vol. 238, p. 121751, 2024.

[5] J. Doménech *et al.*, "Evaluating and Enhancing Intrusion Detection Systems in IoMT: The Importance of Domain-Specific Datasets," *Internet of Things*, vol. 32, p. 101631, 2025.

[6] A. Altrad, "IoT Medical Network Security System Based Explainable AI Model," in *IEEE Icit*, 2025, pp. 404–409.

[7] Y. Hosain *et al.*, "XAI-XGBoost: an Innovative Explainable Intrusion Detection Approach for Securing Internet of Medical Things Systems," *Scientific Reports*, vol. 15, no. 1, p. 22278, 2025.

[8] S. A. Memon *et al.*, "Explainable Intrusion Detection for Internet of Medical Things," in *IC3K*, 2023, pp. 40–51.

[9] M. A. Alsharaiah *et al.*, "An Explainable AI-Driven Transformer Model for Spoofing Attack Detection in Internet of Medical Things (IoMT) Networks," *Discover Applied Sciences*, vol. 7, no. 5, p. 488, 2025.

[10] F. Sohail *et al.*, "Explainable Boosting Ensemble Methods for Intrusion Detection in Internet of Medical Things (IoMT) Applications," in *IEEE ICODT2*, 2024, pp. 1–8.

[11] S. Dadkhah *et al.*, "CICIoMT2024: a Benchmark Dataset for Multi-Protocol Security Assessment in IoMT," *Internet of Things*, vol. 28, p. 101351, 2024.

[12] M. Georgiades *et al.*, "An Explainable AI Approach for Interpretable Cross-Layer Intrusion Detection in Internet of Medical Things," *Electronics*, vol. 14, no. 16, p. 3218, 2025.

[13] I. A. Khan *et al.*, "XSRU-IoMT: Explainable Simple Recurrent Units for Threat Detection in Internet of Medical Things Networks," *Future Generation Computer Systems*, vol. 127, pp. 181–193, 2022.

[14] W. Liu *et al.*, "Explainable AI for Medical Image Analysis in Medical Cyber-Physical Systems: Enhancing Transparency and Trustworthiness of IoMT," *IEEE J. Biomed. Health Inform.*, 2023.

[15] J. Dutta *et al.*, "Next Generation Healthcare with Explainable AI: IoMT-Edge-Cloud Based Advanced Ehealth," in *IEEE Globecom*, 2023, pp. 7327–7332.

[16] H. Raza *et al.*, "An IoMT-Enabled Smart Healthcare Model to Monitor Elderly People using Explainable Artificial Intelligence (EAI)," *Journal of NCBAE*, vol. 1, no. 2, pp. 16–22, 2022.

[17] M. Keshk *et al.*, "An Explainable Deep Learning-Enabled Intrusion Detection Framework in IoT Networks," *Inf. Sciences*, vol. 639, p. 119000, 2023.

[18] K. P. Sharma *et al.*, "Interpretable Intrusion Detection for IoT Environments using a Self-Attention-Based Explainable AI Framework," *Scientific Reports*, vol. 15, no. 1, p. 39937, 2025.

[19] N. Sharma *et al.*, "Multi-Attention DeepCRNN: an Efficient and Explainable Intrusion Detection Framework for Internet of Medical Things Environments," *Knowledge and Information Systems*, vol. 67, no. 7, pp. 5783–5849, 2025.

[20] S. M. Lundberg *et al.*, "A Unified Approach to Interpreting Model Predictions," *NeurIPS*, vol. 30, 2017.