electronics

IMPACT FACTOR 2.6

CITESCORE 6.1

Article

An Explainable AI Approach for Interpretable Cross-Layer Intrusion Detection in Internet of Medical Things

Michael Georgiades and Faisal Hussain

Special Issue

IoT for Healthcare and Wellbeing: Trends, Challenges, and Applications, 2nd Edition

Edited by

Dr. Francisco Luna-Perejón, Dr. Lourdes Miró Amarante, Prof. Dr. Francisco Gómez-Rodríguez, Dr. Javier Civit Masot and Dr. Luis Muñoz

MDPI

https://doi.org/10.3390/electronics14163218

electronics

MDPI

Article

An Explainable AI Approach for Interpretable Cross-Layer Intrusion Detection in Internet of Medical Things

Michael Georgiades [1,2,*,†] and Faisal Hussain [3,†]

[1] ISLab, Department of Computer Science, Neapolis University, 8042 Paphos, Cyprus
[2] Research & Development Lab, Infostrada Communications, 8021 Paphos, Cyprus
[3] Department of Computing, National University of Modern Languages, Multan Campus, Multan 60000, Pakistan; faisal.hussain.engr@gmail.com
* Correspondence: m.georgiades@nup.ac.cy
† These authors contributed equally to this work.

Abstract

This paper presents a cross-layer intrusion detection framework leveraging explainable artificial intelligence (XAI) and interpretability methods to enhance transparency and robustness in attack detection within the Internet of Medical Things (IoMT) domain. By addressing the dual challenges of compromised data integrity, which span both biosensor and network-layer data, this study combines advanced techniques to enhance interpretability, accuracy, and trust. Unlike conventional flow-based intrusion detection systems that primarily rely on transport-layer statistics, the proposed framework operates directly on raw packet-level features and application-layer semantics, including MQTT message types, payload entropy, and topic structures. The key contributions of this research include the application of K-Means clustering combined with the principal component analysis (PCA) algorthim for initial categorization of attack types, the use of SHapley Additive exPlanations (SHAP) for feature prioritization to identify the most influential factors in model predictions, and the employment of Partial Dependence Plots (PDP) and Accumulated Local Effects (ALE) to elucidate feature interactions across layers. These methods enhance the system's interpretability, making data-driven decisions more accessible to nontechnical stakeholders. Evaluation on a realistic healthcare IoMT testbed demonstrates significant improvements in detection accuracy and decision-making transparency. Furthermore, the proposed approach highlights the effectiveness of explainable and cross-layer intrusion detection for secure and trustworthy medical IoT environments that are tailored for cybersecurity analysts and healthcare stakeholders.

Keywords: Internet of Medical Things (IoMT); explainable AI (XAI); interpretability; cross-layer intrusion detection; SHAP; ALE; PDP; CoAP; MQTT; healthcare security; cybersecurity; healthcare systems

Academic Editors: Francisco Gómez-Rodríguez, Francisco Luna-Perejón, Lourdes Miró Amarante, Javier Civit Masot and Luis Muñoz

Received: 31 May 2025
Revised: 4 August 2025
Accepted: 12 August 2025
Published: 13 August 2025

Citation: Georgiades, M.; Hussain, F. An Explainable AI Approach for Interpretable Cross-Layer Intrusion Detection in Internet of Medical Things. *Electronics* **2025**, *14*, 3218. https://doi.org/10.3390/electronics14163218

1. Introduction

The rapid evolution and widespread adoption of Internet of Things (IoT) devices have transformed various sectors such as agriculture, energy, transportation, and healthcare by improving operational efficiency and enabling real-time decision-making [1]. However, as the IoT systems become increasingly embedded in critical infrastructures, they also introduce significant cybersecurity vulnerabilities due to insufficiently robust security mechanisms [1,2].

In the healthcare sector, where IoT devices are commonly deployed as part of the Internet of Medical Things (IoMT), the consequences of cyberattacks are particularly severe. A report by the European Union Agency for Cybersecurity (ENISA) revealed that 53% of all documented cyberattacks in the EU between January 2021 and March 2023 were directed at healthcare systems, with ransomware, data theft, intrusion, and Denial of Service (DoS) attacks being the most prevalent [2–4]. Similarly, in the United States, approximately 25% of surveyed healthcare institutions experienced nine to fifteen cyberattacks involving IoT and IoMT devices between 2020 and 2022, with an additional 24% reporting four to eight incidents during the same period [5].

IoMT devices, which support applications such as patient monitoring, diagnostics, and treatment delivery, rely on lightweight communication protocols like Message Queuing Telemetry Transport (MQTT) and Constrained Application Protocol (CoAP). While these protocols are efficient, they lack robust security features, making IoMT systems highly vulnerable to cyber threats. These threats typically manifest in two forms: application-layer attacks, where malicious actors manipulate biosensor data to disrupt diagnostics or treatment, and network-layer attacks, such as Distributed Denial of Service (DDoS), that disrupt communication and compromise critical healthcare operations [6–8].

Efforts to mitigate these vulnerabilities have leveraged advancements in artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), alongside statistical and data-driven approaches. These techniques have significantly improved intrusion and anomaly detection systems by identifying complex patterns and behaviours in IoMT environments through behavioural analysis [6,7,9]. However, many of these advancements remain influenced by early work such as Lee and Stolfo [10], which popularized the use of flow-level statistics over raw packet data. While this abstraction improves efficiency, it often discards protocol-specific or semantic details, such as MQTT payload entropy or biosensor manipulations, that are critical in modern IoMT attacks.

Moreover, despite their success, these AI-based models often operate as opaque "black boxes" providing little insight into their decision-making processes. This lack of interpretability poses a significant challenge, particularly in high-stakes domains like healthcare, where trust and accountability are critical [11–13]. Recent studies such as [14] applied SHAP and LIME to MLP-based intrusion detection in IoT, emphasizing their critical role in aiding cybersecurity experts' interpretability. However, once again, the majority of these solutions continue to utilise flow-level statistics, limiting their visibility into protocol-specific and semantic-layer anomalies.

The tradeoff between interpretability and accuracy further exacerbates this challenge. Analysts frequently prioritise accuracy, favouring advanced models such as Convolutional Neural Networks (CNNs) and Random Forests for their superior predictive power. However, these models lack the transparency of simpler, interpretable models like Decision Trees or Logistic Regression, which may underperform in handling complex tasks [15]. In IoMT systems, where both accuracy and interpretability are essential, achieving this balance is critical to ensuring reliable threat detection while fostering trust among healthcare professionals and stakeholders.

Efforts to address these challenges have increasingly focused on Explainable AI (XAI) techniques. The XAI aims to enhance safety, transparency, and trustworthiness in understanding critical decisions made by AI-based solutions, particularly in sensitive domains like healthcare [12,16–18]. Traditional methods, such as feature importance scores and model coefficients, offer limited insights into which features are most impactful. In contrast, XAI techniques provide more detailed, human-understandable explanations that bridge the gap between model complexity and interpretability. Beyond technical advancements, XAI also addresses critical ethical considerations, such as fairness, compliance, and respect

for individual autonomy. McDermid et al. emphasise that XAI supports normative goals like accountability and enables explanations to trace back to human decisions made during the development lifecycle [19]. Similarly, Alam et al. highlight the role of XAI in mitigating bias, ensuring regulatory compliance, and fostering patient-centred care in healthcare applications [20]. Emerging frameworks even explore large language models (LLMs) for generating human-readable XAI narratives from SHAP outputs [21].

Popular XAI methods like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) go beyond basic feature importance by offering model-agnostic and localized insights, making them intuitive and accessible to stakeholders. Additionally, techniques such as Partial Dependence Plots (PDP), Accumulated Local Effects (ALE), Individual Conditional Expectation (ICE), Principal Component Analysis (PCA), dimensionality reduction, and clustering further enhance explainability by simplifying complex data structures and providing powerful visualization tools. These methods not only uncover underlying patterns in the data but also make the behaviour of AI models more transparent and interpretable [18]. Such approaches are invaluable for domain experts, including cybersecurity professionals who may not necessarily have expertise in AI [22]. Recent comparative evaluations confirm that SHAP outperforms LIME in certain IoT contexts in terms of feature stability and local fidelity [23]. Moreover, federated variants of XAI have emerged for privacy-preserving IDS in IoT, integrating SHAP explanations with decentralized architectures [24]. Even advanced quantum neural networks now incorporate SHAP to enhance transparency in IDS on hardware platforms like IonQ [25].

XAI holds particular significance in high-stakes environments like healthcare, where understanding feature interactions and contributions to model decisions can enable more informed and trusted outcomes. These capabilities are essential for fostering trust among stakeholders and ensuring the safe and effective deployment of AI-based solutions in healthcare applications [26–28].

However, while XAI methods promise enhanced transparency and accountability, current methodologies often fall short in providing actionable and clinically relevant explanations. This underscores the necessity of rigorous validation to prevent potentially harmful misinterpretations, particularly in sensitive healthcare environments. IoMT systems, for instance, face two primary cybersecurity challenges. At the application layer, attackers may compromise data integrity in medical diagnostics by subtly altering biosensor values during transmission through lightweight protocols like MQTT or CoAP, leading to critical misdiagnoses or inappropriate treatments. At the network layer, attacks such as Denial of Service (DoS) and Distributed Denial of Service (DDoS) disrupt communication between devices, overwhelming network resources and compromising the operation of IoMT devices reliant on edge-based networks for vital functions [8].

To address these challenges, this work applies cutting-edge methods in the field of IoMT cybersecurity to enhance the interpretability and performance of network-based attack detection. By integrating advanced XAI techniques, the approach combines cross-layer data from biosensors and network traffic, providing a comprehensive solution for mitigating cyber threats in IoMT systems. This work builds on recent work by the authors in the field of IoMT [29–31]. To the best of our knowledge, this study is the first to explore the cross-layer distinction between altered medical data at the application level and network-layer disruptions, fully leveraging XAI to advance security and interpretability in healthcare environments.

1.1. Research Contributions

The proposed framework addresses IoMT security challenges through the following key contributions:

1. **Application of KMeans Clustering and PCA to MQTT-Based Network Traffic**: The framework incorporates K-Means clustering combined with Principal Component Analysis (PCA) to analyze MQTT-related network traffic. This approach identifies key patterns and addresses limitations of existing methods in real-world IoMT environments, providing a foundation for enhanced threat detection.
2. **SHAP-Based Feature Selection for Improved Interpretability and Performance**: SHAP decision plots are utilized within the framework to identify the most influential features across various attack scenarios. This feature selection method enhances model interpretability, reduces complexity, and improves overall model performance.
3. **Integration of Cross-Layer Data for Holistic Transparency:** The framework integrates network traffic data with biosensor readings, offering a unified cross-layer perspective. This holistic approach improves explainability, making the framework accessible to both technical and non-technical stakeholders. It enhances transparency and trust in IoMT security measures.
4. **A Six-Stage Methodology for Interpretability:** This work outlines a six-stage methodology, including data collection, preprocessing, feature selection, model training, performance comparison, and explainability for users. The methodology supports repeatability and is adaptable to various XAI techniques beyond the methods presented in this study.

1.2. Paper Structure

The remainder of this paper is organized as follows: Section 2 provides a literature review of existing work in IoMT security and cyberattack detection. Section 3 details the proposed framework's six-stage methodology for detecting and preventing cyberattacks in IoMT environments. Section 4 describes the experimental setup used to evaluate the framework's effectiveness. Section 5 discusses the results of applying the proposed framework, emphasizing its contributions to transparency, interpretability, and performance in IoMT security. Finally, Section 6 concludes the article and outlines potential directions for future research.

2. Related Work

The recent advancements in research have revealed that explainable AI (XAI) is a promising tool for the explainability and interpretability of ML-based intrusion detection models. Ravi and Yu [32] et al. demonstrated the potential of XAI to enhance the interpretability of anomaly detection models through advanced techniques such as convolutional auto-encoders, underscoring the necessity of developing robust, interpretable models that significantly enhance user trust and model transparency, which are crucial elements in patient-centric applications. Similarly, Ghassemi et al. [33] highlighted the critical pitfalls of current XAI methodologies in healthcare. The authors emphasized the need for more rigorous validation processes to prevent misinterpretations that could have detrimental consequences in clinical settings.

As discussed earlier, AI-based approaches for intrusion detection are also gaining increased prominence in the field of IoT security. Bovenzi et al. [6] proposed a deep learning-based approach for unsupervised early anomaly detection in IoT environments, employing advanced architectures such as packet-level processing and ensemble learning to enhance performance and robustness, especially in the face of adversarial attacks like Label Flipping poisoning. Similarly, Meidan et al. [34] proposed N-BaIoT, a novel network-based anomaly

detection method for IoT devices, by utilizing deep autoencoders trained on statistical features extracted from benign traffic to detect botnet attacks such as those from Mirai and BASHLITE with high accuracy and low false positive rates. Likewise, Nascita et al. [35] applied advanced ML and DL techniques to classify attacks in IoT networks, using state-of-the-art DL architectures such as 1D-CNN (Convolutional Neural Network), hybrid 2D-CNN+LSTM (Long Short-Term Memory), and multimodal MIMETIC. The authors also compared their performance against traditional ML methods to achieve accurate and early detection of IoT-based cyberattacks.

Recent approaches further leverage optimization techniques for feature selection. For instance, Subramani and Selvi [36] proposed a multi-objective particle swarm optimization (PSO) method to enhance feature selection for SVM classifiers in wireless IoT environments, demonstrating improved detection accuracy. A complementary perspective is offered by Ozkan-Okay et al. [37], who introduced a statistical feature selection strategy aimed at improving attack classification accuracy through multi-class ensemble learning. These methods emphasize that selecting discriminative features is pivotal for ensuring reliable and interpretable IDS performance.

Wang et al. [38] proposed a structured XAI framework for IDSs, applying SHAP to improve the interpretability of intrusion detection models. Their approach integrates local and global explanations to clarify how feature contributions influence both specific decisions and general model trends.

Kök et al. [18] highlighted the critical role of XAI in improving transparency, trust, and security. By offering clear explanations for decisions made by IoT devices, XAI enhances transparency and accountability, which is especially important in healthcare settings where such decisions can have significant impacts on individuals. Likewise, Sharma et al. [39] and Kalutharage et al. [40] proposed DL models for IoT networks by leveraging XAI techniques like LIME and SHAP to provide explanations for various types of network attacks. These methods not only improved the reliability and interpretability of AI-driven intrusion detection systems but also enhance transparency and trust by making these systems more robust and user-friendly. Recently, research by Dadkhah et al. [41] underscored the importance of realistic datasets for developing advanced IDS tailored to healthcare devices. By simulating various cyberattacks on IoMT devices, their work provided a foundation for enhancing the security posture of healthcare environments. Similarly, Moustafa et al. [22] presented a comprehensive survey of XAI techniques specifically tailored for anomaly-based IDS in IoT networks. They also highlighted the current challenges and suggested future research directions to enhance the transparency, trust, and interpretability of cybersecurity measures in these environments.

Wu et al. [26] proposed a three-layer next-generation consumer electronics architecture in compliance with the IEEE 2668 and emphasized the urgent need for defining standardized frameworks for next-generation consumer electronics. Using cross-layer information to enhance security and trust is an emerging concept; however, with the advent of ML technologies and increased computational speed, there is now a significant opportunity for improvement. The concept of cross-layer intrusion detection, where data from multiple network layers (e.g., physical, MAC, network) is integrated, has been explored to enhance the detection of sophisticated attacks. Thamilarasu et al. [42] developed a cross-layer intrusion detection framework that effectively detected attacks by incorporating information from various network layers. Building on this concept, Poongothai et al. [43] applied ML techniques to cross-layer data in Mobile Ad-hoc Networks (MANETs) to improve detection accuracy while reducing computational overhead. The authors in [27,44] proposed to measure trust in IoT devices using a two-level approach that collected information from both the network traffic behavior (low-level) of IoT devices and the application layer (high-level)

to establish a comprehensive trust metric, which continuously monitors and adjusts the trust based on these two perspectives.

Table 1 presents a comparative overview of cross-layer and XAI-based intrusion detection models, each addressing parts of the interpretability and visibility challenges in IoT and healthcare networks. However, none of these works simultaneously leverage both application-layer (e.g., biosensor or semantic data) and network-layer information within an XAI-enhanced framework. Common limitations include reliance on global post-hoc explanations, omission of semantic temporal context, or the absence of biosensor-driven insight. Additionally, several models focus solely on flow-level or aggregate traffic features, overlooking finer-grained protocol-layer content such as MQTT semantics or payload entropy. Other recent methods have adopted federated or ensemble-based approaches yet still lack visibility into cross-layer dynamics essential in healthcare-grade IDS deployments. Explainability is often introduced as an afterthought, with limited real-time interpretability or integration into the model design itself. Despite notable advances, the application of explainable, cross-layer IDS architectures in the IoMT context remains underexplored. This work addresses that gap by proposing an XAI-driven intrusion detection solution that fuses network-layer traffic data with application-layer biosensor values, providing interpretable, transparent, and robust threat detection at the healthcare edge.

Table 1. Cross-layer and XAI-based IDS models and their limitations.

Citation	Approach	Limitation
Fatema et al. (2025) [24]	Federated XAI IDS with SHAP	Lacks MQTT-layer interpretability or entropy-based inspection
Rajkumar et al. (2025) [25]	QNN-based IDS with SHAP on BoT-IoT dataset	Limited to SHAP flow-level insights; lacks application-layer and MQTT feature visibility
Gaspar et al. (2024) [14]	SHAP/LIME applied to MLP for IoT-IDS	Lacks biosensor-level or protocol-layer specificity
Nascita et al. (2024) [13]	SHAP-based model simplification for IoT anomaly detection	Static feature analysis only; no application-layer modeling
Sharma et al. (2024) [39]	DL-based IDS with LIME and SHAP explainability	Post hoc explanations only; lacks cross-layer integration
Kalutharage et al. (2023) [40]	XAI-enhanced ensemble IDS for IoT networks	Does not incorporate biosensor input; lacks temporal alignment
Keshk et al. (2023) [45]	LSTM-based IDS with SHAP, PDP, ICE, PFI (SPIP toolkit)	Global explanations only; no cross-layer data fusion
Subramani et al. (2023) [36]	Multi-objective PSO-based feature selection with SVM classifier	No integration of packet-level entropy or protocol-specific insight; lacks XAI explanations
Ravi et al. (2021) [32]	CAE-based anomaly detection with XAI overlay	Image-centric focus; no integration with network or biosensor data
Poongothai et al. (2018) [43]	Cross-layer ML-based IDS in MANETs	Cross-layer approach; lacks explainability and healthcare focus

3. Materials and Methods

The proposed methodology to enhance transparency and trust using XAI for network-based attack detection in IoMT consists of six major stages. These stages include data collection of raw MQTT and CoAP traffic, dataset pre-processing, features selection, model training, performance comparison, and XAI, as illustrated in Figure 1. The workflow of the proposed cross-layer methodology consists of six steps for enhancing the transparency and trust using XAI, as illustrated in Figure 2, particularly showing the application of the proposed work in IoMT. The overall methodology adopted in this study, detailing each

stage and the related steps from data collection to explainability for users, is described in the following sections:

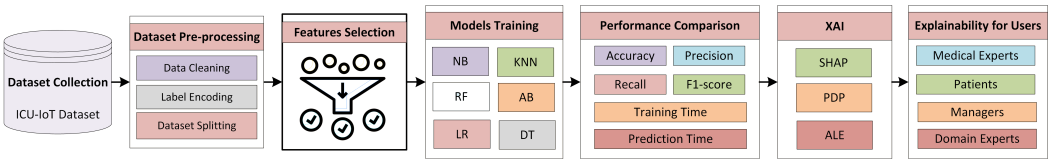

Figure 1. Proposed cross-layer methodology for enhancing transparency and trust using explainable AI (XAI) in IoMT attack detection.

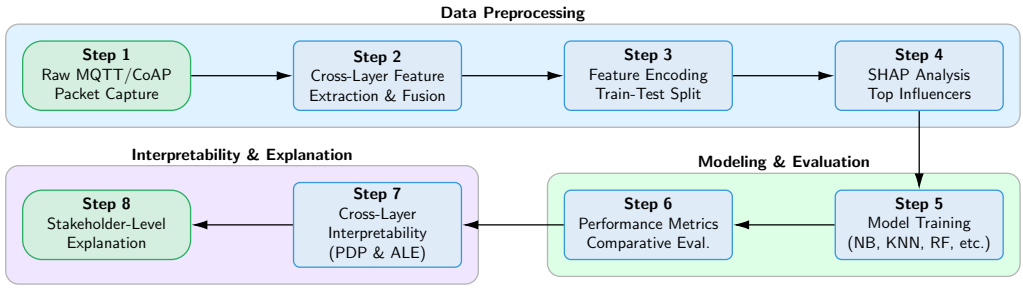

Figure 2. Workflow of the proposed cross-layer methodology for enhancing transparency and trust using explainable AI (XAI) in IoMT attack detection.

3.1. Data Collection and Preprocessing

This study utilizes data from the ICU-IoT dataset [46], which contains emulated real-world healthcare IoT environments under both normal and attack scenarios including SlowITe, malformed data, and MalariaDoS attacks. The dataset collection process focused on capturing raw packet-level data from MQTT and CoAP communication, including payloads, protocol-specific headers, and metadata fields (e.g., topic, QoS, retain, dup). These raw fields were extracted and transformed into structured features for learning purposes, providing deeper semantic and behavioral cues than traditional packet-level summaries.

The data preprocessing steps involved packet parsing, data cleaning, normalization, and label encoding for categorical protocol attributes (e.g., mqtt.msgtype, mqtt.topic). Dataset splitting into training and test sets followed this transformation.

Compared to conventional approaches that rely solely on statistical network features, this fine-grained, protocol-aware preprocessing enables more accurate and interpretable intrusion detection in the IoMT setting.

3.2. Features Selection

To optimize model performance, we employed SHAP, and specifically SHAP decision plots, to assess the contribution of each feature to the model's predictions. As an illustrative example in our study, we used logistic regression trained on a 70:30 stratified train–test split to compute SHAP values and examine feature importance across scenarios. Based on this approach, we selected 13 features that consistently appeared in the top 20 influential features across all four scenarios (Benign, Malformed, MalariaDoS, SlowITe), as determined by SHAP Decision Plots and the LinearExplainer.

Given a dataset X with n features and a set of scenarios $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$, the SHAP value ϕ_j for a feature j under scenario s is computed using Equation (1):

$$\phi_j^s = \sum_{S \subseteq \{1,\ldots,n\} \setminus \{j\}} \frac{|S|!(n - |S| - 1)!}{n!} [f_s(S \cup \{j\}) - f_s(S)] \quad (1)$$

where

- S is a subset of feature indices excluding j,

- $f_s(S)$ denotes the model's output when only the features corresponding to the indices in S are present, for a given scenario $s \in \mathcal{S}$.

For each scenario $s \in \mathcal{S}$, we compute SHAP values ϕ_j^s for all features and rank them to extract the top 20 most influential features. Let $T_{20}(s)$ denote this set for scenario s. We then compute the intersection of these sets across all four scenarios as shown in Equation (2):

$$F_{13} = T_{20}(s_1) \cap T_{20}(s_2) \cap T_{20}(s_3) \cap T_{20}(s_4) \tag{2}$$

Here, F_{13} represents the set of 13 features that were repeatedly ranked among the top 20 for all scenarios, including both attack and benign contexts. The persistence of these 13 features across conditions underscores their relevance and supports their inclusion in downstream modeling stages.

3.3. Model Training

Using the selected features, we trained six ML models, including Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), AdaBoost (AB), Logistic Regression (LR), and Decision Tree (DT). In our case, we employed a 70:30 stratified train–test split to ensure balanced class representation during training and evaluation. The training process involved optimizing each model to effectively generalize across different attack scenarios while maintaining high accuracy and efficiency. This model training was followed by a comprehensive evaluation to ensure that the selected features contributed positively to the model's performance.

3.4. Performance Comparison

We conducted a performance comparison across the trained models over commonly used evaluation metrics such as accuracy, precision, recall, F1-score, training time, and prediction time. This comparison provided insights into the trade-offs between different models and the effectiveness of the selected features. The results demonstrated that SHAP-based feature selection not only improved model interpretability but also significantly reduced computational overhead, making the models more suitable for real-time deployment in healthcare IoT systems.

3.5. Cross-Layer XAI and Interpretability

To further enhance model transparency and cross-layer interpretability, we incorporated XAI techniques such as Partial Dependence Plots (PDP) and Accumulated Local Effects (ALEs), with a specific focus on bridging insights between application-layer (e.g., MQTT fields) and network-layer (e.g., frame lengths) features.

Partial Dependence Plots (PDP): PDPs were used to visualize how specific features, such as `frame.len` and `mqtt.msg`, influence the detection of network-based attacks. The partial dependence of a feature j, like `frame.len`, is calculated as shown in Equation (3):

$$\hat{f}_{\text{PD}}(\texttt{frame.len}) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(\texttt{frame.len}, \mathbf{x}_{i,-\texttt{frame.len}}) \tag{3}$$

For two interacting features, `frame.len` and `mqtt.msg`, the joint partial dependence is defined in Equation (4):

$$\hat{f}_{\text{PD}}(\texttt{frame.len}, \texttt{mqtt.msg}) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}\big(\texttt{frame.len}, \texttt{mqtt.msg}, \mathbf{x}_{i,-\{\texttt{frame.len},\texttt{mqtt.msg}\}}\big) \tag{4}$$

Accumulated Local Effects (ALEs): ALE plots were used to accurately estimate the impact of features, particularly in the presence of correlations. ALEs offer a more precise representation than PDPs, especially when features are not independent. The ALE for a single feature j, such as `frame.len`, is defined in Equation (5):

$$\hat{f}_{\text{ALE}}(\texttt{frame.len}) = \frac{1}{n_j} \sum_{i=1}^{n_j} \left[\hat{f}(z_{k,\text{upper}}, \mathbf{x}_{-j}) - \hat{f}(z_{k,\text{lower}}, \mathbf{x}_{-j}) \right] \tag{5}$$

For two features, `frame.len` and `mqtt.msg`, the joint ALE is given by Equation (6):

$$\begin{aligned} \hat{f}_{\text{ALE}}(\texttt{frame.len}, \texttt{mqtt.msg}) = & \frac{1}{n_k} \sum_{i=1}^{n_k} \Big[\hat{f}(z_{j,\text{upper}}, z_{k,\text{upper}}) \\ & - \hat{f}(z_{j,\text{lower}}, z_{k,\text{lower}}) \Big] \end{aligned} \tag{6}$$

These XAI techniques were crucial for understanding how individual, joint, and cross-layer features affected the model's predictions, thus providing deeper insights into the detection of network-based attacks in IoMT environments.

3.6. Explainability for Users

Finally, the XAI techniques accomplished explainability of the models to make the AI-driven decisions understandable for various stakeholders, including medical experts, patients, managers, and domain experts, e.g., data scientists and cybersecurity experts. By using SHAP, PDP, and ALE, we translated complex model decisions into interpretable insights. This transparency helped ensure that the insights derived from the models were not only accurate but also actionable for non-technical users, thereby enhancing trust and facilitating the adoption of AI technologies in healthcare IoT environments.

4. Experimental Setup

To effectively illustrate and analyze the significant cybersecurity risks posed by weakly secured IoT devices in healthcare environments, particularly in ICUs, we utilized a comprehensive testbed setup designed to emulate an IoT-based ICU. This setup generated and captured real-time network traffic from healthcare devices under both normal and attack conditions, providing a foundation for understanding how different types of cyberattacks, specifically Denial of Service (DoS), SlowITe, and Man-in-the-Middle (MiTM) malformed data attacks impact healthcare systems.

4.1. Testbed Setup and Data Collection

The testbed setup we used includes an IoT-based ICU scenario that was designed by [46] to monitor and analyze network traffic as illustrated in Figure 3. The authors in [46] designed an IoT healthcare use case to generate and capture real-time network traffic of normal healthcare devices. The authors [46] also executed different types of cyberattacks on healthcare devices and captured the cyber-attack traffic as well. The IoT healthcare use case was designed using the IoT-Flock tool [47] which can emulate normal and attack of IoT devices over a real-time network as per specifications included in the device profile.

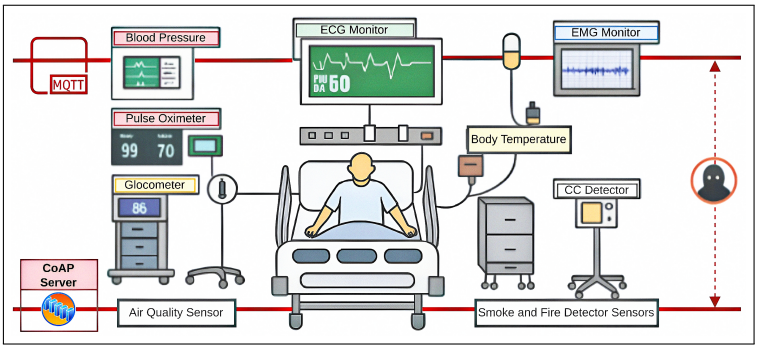

Figure 3. Illustration of the IoT-based ICU environment setup using the IoT-Flock emulator [47], featuring patient monitoring devices (e.g., ECG, pulse oximeter using MQTT protocol) and environmental sensors (e.g., air temperature, CO using CoAP protocol), capturing both normal and cyber-attack traffic for comprehensive analysis of network security.

The authors [46] categorized IoT healthcare devices into two types, i.e., environment monitoring devices and patient monitoring devices. The environment monitoring devices include seven IoT sensors that collect environment-related information like air temperature, air humidity, atmospheric pressure, etc. On the other hand, the patient monitoring devices include nine IoT sensors and actuators like infusion pump, blood pressure sensor, body temperature sensor, etc., as illustrated in Figure 3.

Each of the environment and patient monitoring devices has a well-defined data profile and time profile, based on which the device transmits data over a real-time network using the IoT-Flock tool [47]. The data profile mainly contains range of values that an IoT device can send in a real IoT healthcare use case. For example, a digital device has 2 possible binary values, i.e., 0 or 1, whereas an analogue device can send a value from a range of values; for example, a body temperature sensor can send any value between 0 F to 120 F.

Similarly, the time profile of an IoT device mainly contains time interval values after which an IoT device transmits data over a network in real IoT healthcare use cases. For example, a body temperature sensor transmits data after every 10 min, whereas an air temperature sensor transmits data after every 2 s.

Using IoT-Flock [47], we emulated a variety of cyberattacks within a healthcare IoT environment to analyze their impact and identify potential security vulnerabilities. The attacks included Slowlris attacks, which target server resources by maintaining many open but inactive connections; malformed packet attacks, where intentionally incorrect or unexpected packet values were used to exploit network vulnerabilities; and DoS attacks, designed to overwhelm the network or specific systems, potentially in a healthcare monitoring scenario. Additionally, we emulated MQTT-based attacks such as message spoofing, flooding, and Quality of Service (QoS) manipulation, targeting the IoT communication protocols commonly used in healthcare devices. The testbed also included Man-in-the-Middle (MiTM) attacks, where communications between devices could be intercepted or altered, affecting the reliability of sensor data in critical healthcare settings. These simulations provided a comprehensive understanding of how various cyber threats could impact healthcare IoT systems, offering valuable insights for enhancing security measures.

4.2. Leveraging ICU Network Data for Enhanced Attack Detection and Prediction

The ICU network data encompasses a broad range of attributes, such as `frame.time_delta`, `ip.src`, `tcp.flags`, and `mqtt.msgtype`, providing a comprehensive view of traffic patterns and device interactions. Unlike traditional network monitoring tools that rely on static rules and predefined signatures, our approach employs ML models to dynamically analyze this data, enabling the detection of both known and emerging threats in real-time.

By learning from historical data, ML models can identify normal and anomalous patterns, such as detecting SYN flood attacks through unusual `tcp.flags` sequences or predicting security breaches by recognizing subtle deviations in features like `mqtt.kalive`.

The inclusion of features like `mqtt.msg`, which contains the actual payload of MQTT messages, allows for deeper insights into transmitted content, enabling the identification of anomalies or potential breaches through unexpected data patterns. The proposed XAI-based approach will further enhance this process by providing interpretability, allowing non-technical domain experts to understand and act on the insights generated by ML models. This integration of domain expertise with ML-driven analysis and XAI ensures a more proactive and robust approach to intrusion detection, surpassing the limitations of traditional tools by effectively identifying and predicting both known and novel threats within the ICU environment.

5. Results and Discussion

5.1. Contribution I: Initial Attack Identification Using K-Means Clustering with PCA

K-means clustering via PCA is a powerful technique for dimensionality reduction and unsupervised learning [48]. We utilised the ICU-IoT dataset [46] in which network traffic was generated using the IoTFlock tool [47] across various scenarios or categories. These categories represent normal, malariaDoS, slowite, and malformed traffic types. The PCA is applied to this dataset to reduce the feature space while retaining the most significant variance. Mathematically, the PCA transforms the original feature space X into a new set of orthogonal components $Z = XA$, where A is the matrix of eigenvectors of the covariance matrix of X. The application of the PCA results in two principal components which are then used as inputs for K-Means clustering. The K-Means algorithm partitions the data into k clusters by minimizing the within-cluster sum of squares, defined as $\sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$, where C_i is the set of points in cluster i and μ_i is the centroid of cluster i. The resulting clustering outcome is shown in Figure 4.

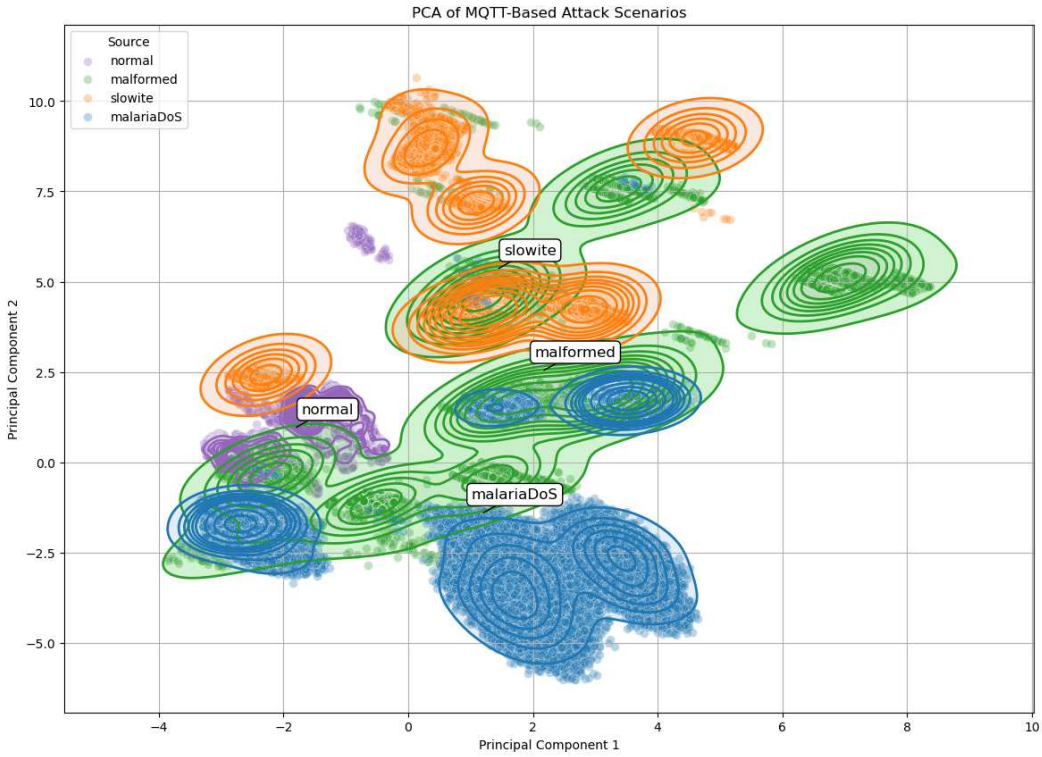

Figure 4. ICU PCA K-Means clustering with new traffic types including normal, malariaDoS, slowite, and malformed.

PCA combined with K-Means clustering not only helps visualize and separate the different MQTT-based attack types but also enhances explainability and interpretability in the context of IoMT security. By reducing the dimensionality of the data and identifying clusters, this method provides an initial identification of potential attacks, making it easier for analysts to interpret and investigate suspicious patterns in network traffic. This approach is particularly valuable for distinguishing between normal and anomalous activities, even in complex and overlapping scenarios.

While the application of K-Means clustering with PCA in the initial attack identification provides a useful starting point for categorizing traffic into distinct clusters, it inherently lacks the depth of insight necessary for a comprehensive understanding of the underlying attack mechanisms. This method primarily focuses on reducing the feature space and identifying broad patterns within the data, but it does not provide the granular level of interpretability needed to trace the specific features or interactions contributing to these patterns, as we will see in the following subsections.

5.2. Contribution II: Leveraging Explainable AI for Model Optimization

We applied SHapley Additive exPlanations (SHAP) to identify the most influential features in detecting network-based attacks within an IoT healthcare system. SHAP helped us pinpoint the key features driving model decisions, enhancing interpretability and enabling us to streamline the model by removing less significant features. This resulted in a more efficient and effective detection system.

5.2.1. SHAP Decision Plot Analysis

We used SHAP decision plots to visualize how each feature contributed to model predictions across four scenarios: Normal Operation, Malformed Data, MalariaDoS Attack, and SlowITe Attack (see Figure 5). These plots allowed us to observe the cumulative effect of feature values on model outputs, helping to identify features with consistently high or low impact.

For example, during Normal Operation, features like `tcp.hdr_len` and `tcp.srcport` had a moderate impact. In the Malformed Data scenario, `tcp.ack` and `tcp.hdr_len` were critical, while `frame.time_relative` and `tcp.ack` were key for detecting MalariaDoS Attacks. In the SlowITe Attack scenario, `ip.src` and `ip.dst` were the most influential. Importantly, MQTT features such as `mqtt.msgtype`, `mqtt.hdrflags`, and `mqtt.msg` consistently emerged as significant across all scenarios, underscoring their critical role in detecting various network-based attacks.

These insights enabled us to refine the feature set by prioritizing consistently high-impact variables, such as MQTT message types and header flags, which in turn improved detection accuracy and interpretability across all evaluated attack scenarios.

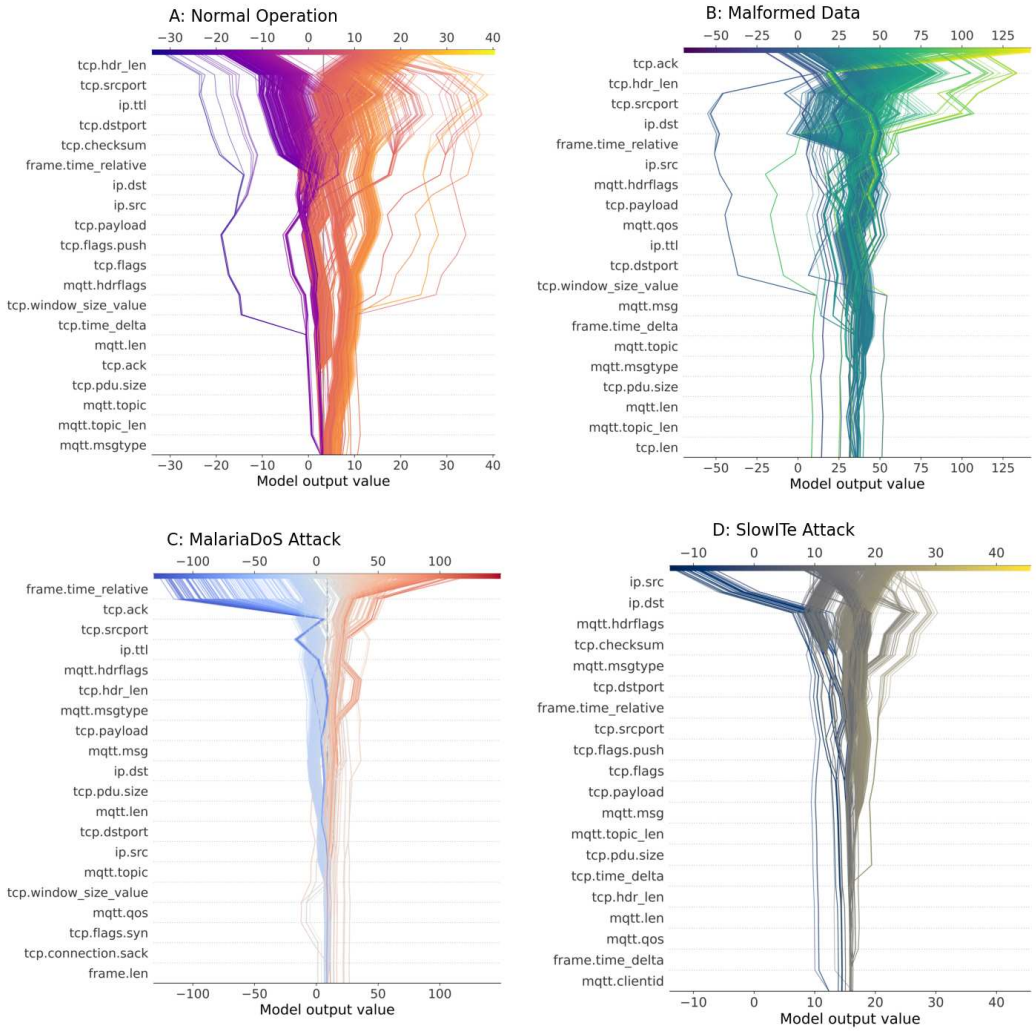

Figure 5. SHAP decision plots for the Logistic Regression model on IoMT healthcare data showing (**A**) Normal or Benign Operation, (**B**) Malformed Data, (**C**) MalariaDoS, and (**D**) SlowITe Attack scenarios using [8]. Warmer colors (e.g., yellow, orange, red) represent features contributing positively to the model's output value, while cooler colors (e.g., blue, purple) represent features contributing negatively. The magnitude indicates the strength of the contribution toward the predicted outcome.

5.2.2. Identifying Candidate Features Across Scenarios

In our analysis, we observed that certain network and application features consistently played a crucial role across all attack scenarios, as highlighted in the SHAP decision plots and summarized in Table 2. Features such as `ip.src`, `ip.dst`, `tcp.srcport`, and `tcp.hdr_len` from the network layer, alongside MQTT application-level features like `mqtt.msgtype`, `mqtt.hdrflags`, and `mqtt.msg`, were present across all scenarios. These features were pivotal in detecting various forms of network-based attacks in the IoT healthcare system.

Given their consistent presence and significant impact on the model's ability to differentiate between normal operations and various attack conditions, we strategically selected 13 key features for our models. This selection was guided by their high SHAP values across all attack scenarios, ensuring that the final model retained its accuracy while becoming more efficient and interpretable. Even though feature selection utilised a LinearExplainer, it proved to be beneficial across all types of classifiers.

Table 2. The table shows the most influential features across all attack scenarios, based on SHAP decision plots and the LinearExplainer. Highlighted features reflect cross-layer relevance, including `frame.time_relative` (green), `ip.src` and `ip.dst` (orange), as well as consistently highly important features from TCP (`tcp.*`, red) and MQTT (`mqtt.*`, blue). This underscores the need for cross-layer protocol integration in intelligent intrusion detection for IoT environments.

Feature	A: Normal	B: Malformed	C: Malaria	D: SlowITe
frame.len			•	
frame.time_delta		•		•
frame.time_relative	•	•	•	•
ip.dst	•	•	•	•
ip.src	•	•	•	•
ip.ttl	•	•	•	
mqtt.clientid				•
mqtt.hdrflags	•	•	•	•
mqtt.len	•	•	•	•
mqtt.msg		•	•	•
mqtt.msgtype	•	•	•	•
mqtt.qos		•	•	•
mqtt.topic	•	•	•	
mqtt.topic_len	•	•		•
tcp.ack	•	•	•	
tcp.checksum	•			•
tcp.connection.sack			•	
tcp.dstport	•	•	•	•
tcp.flags	•			•
tcp.flags.push	•			•
tcp.flags.syn			•	
tcp.hdr_len	•	•	•	•
tcp.len		•		
tcp.payload	•	•	•	•
tcp.pdu_size	•	•	•	•
tcp.srcport	•	•	•	•
tcp.time_delta	•			•
tcp.window_size_value	•	•	•	

5.2.3. Performance Evaluation with Optimized Feature Selection

In this section, we discuss the performance of various classifiers using three different feature sets: all features, 13 randomly selected features, and the top features identified through SHAP (SHapley Additive exPlanations). This analysis allows us to evaluate the trade-off between model complexity and performance, highlighting how SHAP-based feature selection can enhance model efficiency without sacrificing accuracy.

The comparison of performance metrics, namely Precision, Recall, Accuracy, F1-score, Training Time, and Prediction Time, across classifiers is illustrated in Figure 6. These classifiers include NB, KNN, RF, AB, LR, and DT classifier.

Precision, Recall, and Accuracy: Across most classifiers, using the SHAP-selected features resulted in performance that was comparable to using all 52 features. Notably, classifiers like KNN and RF achieved nearly perfect scores across all three metrics when utilizing the SHAP-selected features, demonstrating the effectiveness of this feature optimization method. Selecting 13 features at random (tested several times), however, did not compare, resulting in lower Precision, Recall, Accuracy, and F1-score.

F1-score: The F1-score, which balances Precision and Recall, further confirms the advantage of using SHAP-selected features. Classifiers such as RF and DT maintained high F1-scores, showcasing their robustness even with a reduced feature set.

Training and Prediction Time: Referring to plot 5 and plot 6 in (the last row of) Figure 6, which illustrate training and prediction time comparisons, it is evident that SHAP-based feature selection significantly reduces both training and prediction times across various classifiers. For some classifiers, these times were reduced by a large factor, particularly in resource-constrained environments where computational efficiency is crucial.

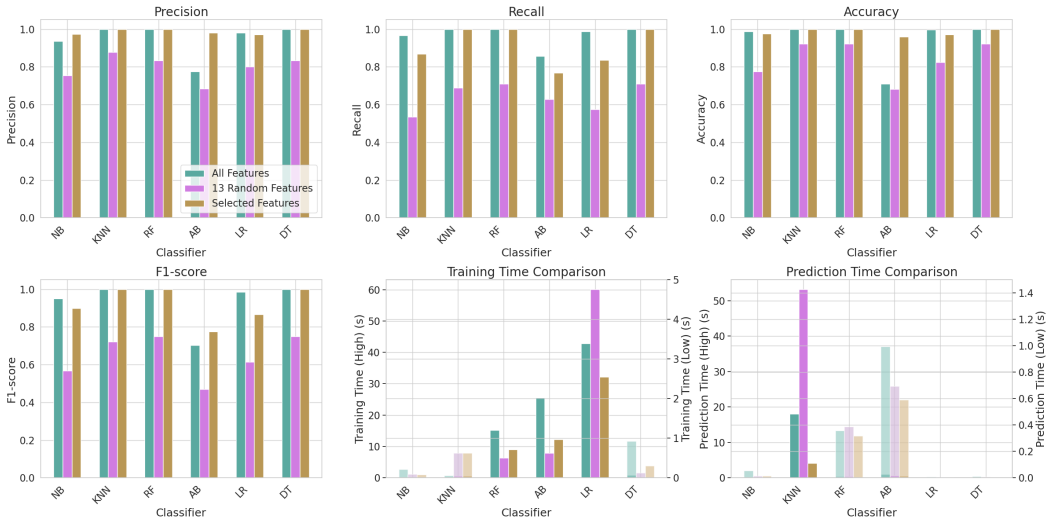

Figure 6. Performance metrics comparison across different classifiers and feature sets: All Features, 13 Random Features, and SHAP-Selected Features.

The results have demonstrated how SHAP-based feature selection enhances model efficiency by reducing training and prediction times, lowering complexity, and making the models more suitable for real-time applications in dynamic, resource-constrained environments like healthcare IoT, where rapid, reliable, and interpretable decision-making is essential.

5.3. Contribution III: Cross Layer (Biosensor and Network) Feature Interaction and Influence Using XAI

This section presents an in-depth analysis of the interactions between biosensor data and network features using XAI techniques. By applying methods like Partial Dependence Plots (PDP) and Accumulated Local Effects (ALEs), we can portray a clearer understanding of how these cross-layer features influence the detection of network-based attacks.

5.3.1. Partial Dependence Plot (PDP)

Partial Dependence Plots (PDPs) are model-agnostic visualization tools used to interpret how one or two input features influence a machine learning model's predictions. By marginalizing over the effects of all other features, PDPs provide a clear view of the average impact that selected feature(s) have on the predicted outcome, making them especially useful for interpreting complex models such as ensembles.

To explore how the interaction between network traffic and biosensor data influences predictions, we employed 2D Partial Dependence Plots (PDPs) using a Random Forest classifier trained on the ICU-IoT dataset [46]. This approach enhances transparency and interpretability in security-critical applications.

Specifically, we focused on the interaction between two critical features: `frame.len` (the length of network frames) and `mqtt.msg` (the MQTT message type). By analyzing these features together, we can better understand how specific combinations affect the model's ability to detect network-based attacks on healthcare edge devices.

The axes in the PDP reflect standardized feature values, which allows for a normalized interpretation across different scales. This means that values are expressed in terms of their deviation from the mean, making the influence of both features on the model output directly comparable. This setup supports the identification of patterns that would be difficult to observe in raw feature space.

The 2D PDP in Figure 7 illustrates how the variation in both `frame.len` and `mqtt.msg` influences the model's predictions. For instance, the Random Forest model exhibits elevated attack prediction probabilities when certain ranges of `frame.len` are paired with specific `mqtt.msg` values, indicating potential anomalies or malicious behavior. This dual-layer analysis of network and sensor data is essential for achieving transparency in the model's decision-making process, which in turn fosters consumer trust, particularly in sensitive healthcare applications.

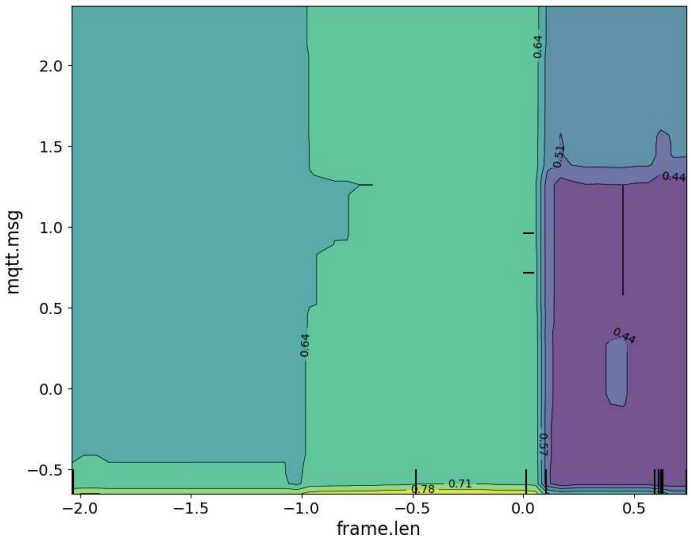

Figure 7. A 2D Partial Dependence Plot (PDP) illustrating the joint effect of `frame.len` and `mqtt.msg` on the predicted probability of a network-based attack, as learned by the Random Forest classifier. The x- and y-axis values are standardized feature values (i.e., zero mean and unit variance), where, for example, -0.5 corresponds to half a standard deviation below the feature mean. Color shading represents the predicted probability of an attack, with lighter colors indicating higher probabilities and darker colors indicating lower probabilities. Contour lines mark probability levels, and short vertical tick marks along each axis denote the marginal distribution of observed values in the ICU-IoT dataset [46].

5.3.2. Analyzing Feature Interactions with Accumulated Local Effects (ALEs)

ALE plots (Accumulated Local Effects) are a model-agnostic interpretability method that quantifies how individual features, or combinations of features, affect model predictions. Unlike Partial Dependence Plots (PDPs), ALE plots estimate the local effect of a feature by averaging changes in predictions over small intervals, which reduces bias from correlated features and provides more faithful explanations.

We generated the 2D Accumulated Local Effects (ALEs) plot in Figure 8 using a Random Forest classifier trained on standardized features from the ICU-IoT dataset [46]. This plot captures the localized, marginal interaction between `frame.len` and `mqtt.msg` and their impact on the predicted probability of a network-based attack.

The color gradient visually reveals interaction zones where the combined influence of the two features either amplifies or suppresses the model's likelihood of predicting an attack. Red areas signify a higher predicted probability, while blue areas indicate a lower probability.

Axis values represent standardized feature ranges (i.e., zero mean and unit variance), allowing for consistent interpretation across different features. This analysis helps pinpoint localized regions where certain combinations are especially influential, enhancing the interpretability and reliability of the model's output.

ALE plots are particularly advantageous in our context because they mitigate the limitations of PDPs by focusing on local effects, avoiding unrealistic feature combinations, and providing a more accurate depiction of feature interactions. As a result, this contributes to the development of XAI systems that are both performant and trustworthy, essential qualities for safe deployment in healthcare environments.

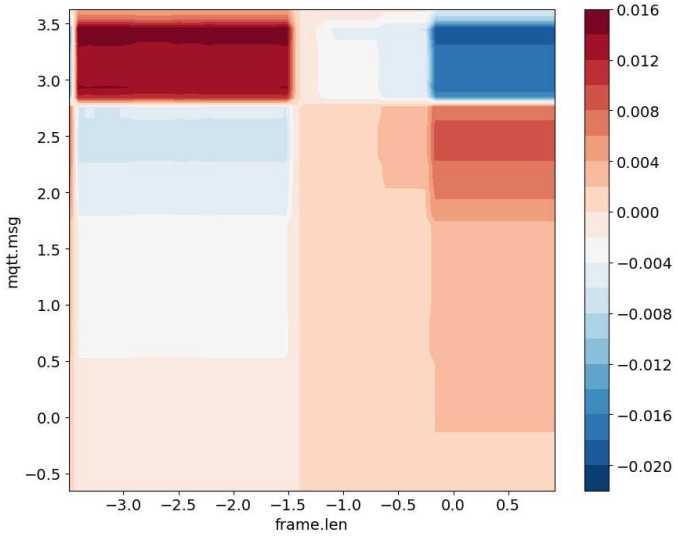

Figure 8. A 2D Accumulated Local Effects (ALEs) Plot visualizing the marginal interaction between `frame.len` and `mqtt.msg` on the predicted probability of attack using a Random Forest classifier. The x- and y-axes represent standardized feature values (i.e., zero mean and unit variance), where, for instance, −1.0 denotes one standard deviation below the mean. The color gradient illustrates how joint deviations in the two features locally affect the model output: red regions correspond to a stronger positive influence on attack predictions, while blue areas indicate a negative (suppressive) effect. The contour pattern provides insight into nonlinear interactions between packet length and MQTT payloads in the ICU-IoT dataset [46].

5.4. Challenges and Opportunities for Stakeholders

Despite the interpretability gains achieved through SHAP, PDP, and ALE visualizations, several practical and conceptual considerations emerge regarding how different stakeholders engage with these outputs.

From the perspective of technical experts, such as cybersecurity analysts or clinical IT administrators, our framework presents clear opportunities. SHAP-based rankings allow these users to prioritize a reduced set of highly influential features (e.g., `mqtt.msgtype`, `tcp.payload`, `ip.dst`) for real-time monitoring, thereby reducing computational overhead while maintaining model accuracy. Similarly, PDP and ALE visualizations facilitate in-depth inspection of feature interactions, helping experts diagnose potential edge-device spoofing, traffic anomalies, or cross-layer irregularities. For instance, a simultaneous rise in temperature readings and anomalous `mqtt.msgtype` may signal adversarial manipulation rather than a clinical event. These insights support both system-level diagnostics and model refinement in complex IoMT settings.

However, for non-technical stakeholders, such as medical personnel or healthcare administrators, these forms of explanation can pose usability challenges. In particular, SHAP, PDP, and ALE methods, while technically robust, are not always intuitive to interpret for users unfamiliar with machine learning, potentially limiting their effectiveness in time-

sensitive clinical or operational environments. SHAP plots and PDP/ALE graphs, while informative, require familiarity with ML concepts like marginal effects, model dependence, or feature contributions, knowledge that non-experts may not possess. Furthermore, the actionable takeaway from "feature X increases the probability of class Y" may be unclear when the end-user's primary concern is whether a patient, device, or subsystem is under attack, and what steps to take in response.

To this end, emerging research in explainable security highlights the value of integrating natural language generation (NLG) techniques to provide user-friendly, context-aware justifications for model predictions [49,50]. Such methods translate technical inference outputs into concise explanations that align with stakeholder roles and operational contexts. For example, instead of presenting a SHAP plot, the system might generate an alert stating, "Device 12 may be compromised; anomalous traffic pattern inconsistent with expected sensor behavior detected; recommendation: verify MQTT source integrity."

6. Conclusions and Future Work

In this work, we presented a comprehensive approach leveraging XAI to address critical challenges of transparency, interpretability, and performance for cross-layer intrusion detection in IoMT environment. By integrating multiple techniques such as K-Means clustering with PCA for data categorization, SHAP for feature prioritization, and interpretability tools like PDP and ALE for feature interaction analysis, this study offered a structured methodology to tackle both technical and ethical challenges in IoMT security. The proposed approach was operationalized through a six-stage methodology, encompassing data collection, preprocessing, feature selection, model training, performance comparison, and explainability for users. These well-defined stages not only supported repeatability but also provided flexibility for adapting the methodology to other contexts or alternative methods while ensuring its broad applicability beyond the presented use case. The key contributions of this research mainly include the identification of influential features across various attack scenarios–such as SlowITe, MalariaDoS, and malformed data attacks–enabling accurate and computationally efficient threat detection. We further used cross-layer data from biosensors and network traffic to ensure comprehensive interpretability addressing dual-layer vulnerabilities unique to IoMT systems. Furthermore, the experimental results manifested a reliable healthcare IoMT testbed by highlighting significant improvements in model accuracy, transparency, and stakeholder trust.

By enhancing transparency and delivering actionable insights, this methodology also addressed a few critical ethical considerations, including fairness, accountability, and respect for individual autonomy. These capabilities are very useful to foster confidence in AI-driven security solutions, aligning them with regulatory and ethical standards and promoting their safe and effective adoption in healthcare environments. The contributions of this work not only present a systematic approach for cybersecurity analysts to design and evaluate robust detection systems but also bridge the gap between complex AI-driven solutions and their practical usability for the domain experts. By providing interpretable insights into attack detection mechanisms, this work equipped healthcare professionals and stakeholders with the tools to understand and respond effectively to emerging cyber threats in IoMT environments. This will not only foster a confidence in AI-driven systems while aligning technological advancements with real-world needs, ultimately enhancing the security and resilience of critical healthcare infrastructures.

Future efforts will explore the application of this methodology across multiple datasets as well as protocols such as CoAP, and a wider variety of healthcare settings and additional attack scenarios to further enhance the robustness and adaptability of the proposed approach. Moreover, enhancing cross-layer integration and supporting real-time imple-

mentation will also be critical in maintaining performance in dynamic IoMT contexts. In parallel, incorporating natural language generation (NLG) strategies can complement our XAI visual toolkit by increasing accessibility, supporting explainability for diverse user profiles, and enhancing the overall trust and usability of AI-driven security in medical IoT environments.

Author Contributions: Conceptualization, M.G.; methodology, F.H. and M.G.; software, M.G.; validation, M.G.; formal analysis, M.G.; investigation, M.G.; resources, F.H.; data curation, F.H.; writing—original draft preparation, M.G. and F.H.; writing—review and editing, M.G. and F.H.; visualization, M.G.; supervision, M.G. and F.H.; project administration, M.G.; funding acquisition, M.G. and F.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The research and APC were partly supported through the internal budgets of Neapolis University Pafos and Infostrada Communications, which employed Michael Georgiades during the time of this research.

Informed Consent Statement: Not applicable. This study did not involve human participants directly. The data used are publicly available and fully anonymized, as provided by the IEEE Dataport-hosted IoT Healthcare Security Dataset. Therefore, no informed consent was required.

Data Availability Statement: The dataset used in this study, titled IoT Healthcare Security Dataset, is publicly available on IEEE Dataport at https://dx.doi.org/10.21227/9w13-2t13 (accessed on 28 May 2025). The dataset was published by Faisal Hussain et al. in 2021 and supports the results reported in this manuscript. To support reproducibility and community adoption, we have released a public GitHub repository under the GPL-3.0 license for future code updates and extensions [51].

Conflicts of Interest: The author Michael Georgiades was employed by the company Infostrada Communications during part of this research. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DL	Deep Learning
ICE	Individual Conditional Expectation
IoMT	Internet of Medical Things
ML	Machine Learning
PCA	Principal Component Analysis
SHAP	SHapley Additive exPlanations
ALE	Accumulated Local Effects
DoS	Denial of Service
DDoS	Distributed Denial of Service
IDS	Intrusion Detection System
IoT	Internet of Things
MQTT	Message Queuing Telemetry Transport
PDP	Partial Dependence Plot
XAI	Explainable Artificial Intelligence

References

1. Kilincer, I.F.; Ertam, F.; Sengur, A.; Tan, R.S.; Acharya, U.R. Automated detection of cybersecurity attacks in healthcare systems with recursive feature elimination and multilayer perceptron optimization. *Biocybern. Biomed. Eng.* **2023**, *43*, 30–41. [CrossRef]
2. Nemec Zlatolas, L.; Welzer, T.; Lhotska, L. Data breaches in healthcare: Security mechanisms for attack mitigation. *Clust. Comput.* **2024**, 27, 8639–8654. [CrossRef]

3. Theocharidou, M.; Ifigeneia Lella, E. ENISA Threat Landscape: Health Sector. 2023. Available online: https://www.enisa.europa.eu/publications/health-threat-landscape (accessed on 12 August 2024).
4. ENISA Threat Landscape for DoS Attacks. 2023. Available online: https://www.enisa.europa.eu/publications/enisa-threat-landscape-for-dos-attacks (accessed on 12 August 2024).
5. Petrosyan, A. U.S. Healthcare Breadth of IoT and IoMT Cyberattacks 2020–2022. 7 November 2023. Available online: https://www.statista.com/statistics/1421177/us-healthcare-cyberattacks-iot-iomt/ (accessed on 19 August 2024).
6. Bovenzi, G.; Aceto, G.; Ciuonzo, D.; Montieri, A.; Persico, V.; Pescapé, A. Network anomaly detection methods in IoT environments via deep learning: A Fair comparison of performance and robustness. *Comput. Secur.* **2023**, *128*, 103167. [CrossRef]
7. Al-Hawawreh, M.; Hossain, M.S. A privacy-aware framework for detecting cyber attacks on internet of medical things systems using data fusion and quantum deep learning. *Inf. Fusion* **2023**, *99*, 101889. [CrossRef]
8. Hussain, F.; Abbas, S.G.; Shah, G.A.; Pires, I.M.; Fayyaz, U.U.; Shahzad, F.; Garcia, N.M.; Zdravevski, E. A framework for malicious traffic detection in IoT healthcare environment. *Sensors* **2021**, *21*, 3025. [CrossRef]
9. Javeed, D.; Gao, T.; Kumar, P.; Jolfaei, A. An explainable and resilient intrusion detection system for industry 5.0. *IEEE Trans. Consum. Electron.* **2023**, *70*, 1342–1350. [CrossRef]
10. Lee, W.; Stolfo, S.J. Data Mining Approaches for Intrusion Detection. In Proceedings of the 7th USENIX Security Symposium, San Antonio, TX, USA, 26–29 January 1998.
11. Kalakoti, R.; Bahsi, H.; Nõmm, S. Improving IoT Security With Explainable AI: Quantitative Evaluation of Explainability for IoT Botnet Detection. *IEEE Internet Things J.* **2024**, *11*, 18237–18254. [CrossRef]
12. Zhang, Z.; Al Hamadi, H.; Damiani, E.; Yeun, C.Y.; Taher, F. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access* **2022**, *10*, 93104–93139. [CrossRef]
13. Nascita, A.; Carillo, R.; Giampetraglia, F.; Iacono, A.; Persico, V.; Pescapé, A. Interpretability and Complexity Reduction in IoT Network Anomaly Detection Via XAI. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), Seoul, Republic of Korea, 14–19 April 2024; IEEE: New York, NY, USA, 2024; pp. 325–329.
14. Gaspar, D.; Silva, P.; Silva, C. Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access* **2024**, *12*, 30164–30175; Erratum in *IEEE Access* **2024**, *7*. [CrossRef]
15. Assis, A.; Dantas, J.; Andrade, E. The performance-interpretability trade-off: A comparative study of machine learning models. *J. Reliab. Intell. Environ.* **2025**, *11*, 1. [CrossRef]
16. Hulsen, T. Explainable artificial intelligence (XAI): Concepts and challenges in healthcare. *AI* **2023**, *4*, 652–666. [CrossRef]
17. Ahmed, M.; Zubair, S. Explainable artificial intelligence in sustainable smart healthcare. In *Explainable Artificial Intelligence for Cyber Security: Next Generation Artificial Intelligence*; Springer: Cham, Switzerland, 2022; pp. 265–280.
18. Kök, I.; Okay, F.Y.; Muyanlı, Ö.; Özdemir, S. Explainable artificial intelligence (XAI) for internet of things: A survey. *IEEE Internet Things J.* **2023**, *10*, 14764–14779. [CrossRef]
19. McDermid, J.A.; Jia, Y.; Porter, Z.; Habli, I. Artificial intelligence explainability: The technical and ethical dimensions. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200363. [CrossRef] [PubMed]
20. Alam, M.N.; Kaur, M.; Kabir, M.S. Explainable AI in Healthcare: Enhancing transparency and trust upon legal and ethical consideration. *Int. Res. J. Eng. Technol.* **2023**, *10*, 1–9.
21. Khediri, A.; Slimi, H.; Yahiaoui, A.; Derdour, M.; Bendjenna, H.; Ghenai, C.E. Enhancing machine learning model interpretability in intrusion detection systems through shap explanations and llm-generated descriptions. In Proceedings of the 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), El Oued, Algeria, 24–25 April 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
22. Moustafa, N.; Koroniotis, N.; Keshk, M.; Zomaya, A.Y.; Tari, Z. Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 1775–1807. [CrossRef]
23. Uysal, I.; Kose, U. Analysis of Network Intrusion Detection via Explainable Artificial Intelligence: Applications with SHAP and LIME. In Proceedings of the 2024 Cyber Awareness and Research Symposium (CARS), Grand Forks, ND, USA, 28–29 October 2024; IEEE: New York, NY, USA, 2024; pp. 1–6.
24. Fatema, K.; Dey, S.K.; Anannya, M.; Khan, R.T.; Rashid, M.M.; Su, C.; Mazumder, R. Federated XAI IDS: An explainable and safeguarding privacy approach to detect intrusion combining federated learning and SHAP. *Future Internet* **2025**, *17*, 234. [CrossRef]
25. Rajkumar, K.; Shalinie, S.M. SHAP-based Intrusion Detection in IoT Networks Using Quantum Neural Networks on IonQ Hardware. *J. Parallel Distrib. Comput.* **2025**, *204*, 105133. [CrossRef]
26. Wu, C.K.; Cheng, C.T.; Uwate, Y.; Chen, G.; Mumtaz, S.; Tsang, K.F. State-of-the-art and research opportunities for next-generation consumer electronics. *IEEE Trans. Consum. Electron.* **2022**, *69*, 937–948. [CrossRef]
27. Macedo, E.L.; Delicato, F.C.; de Moraes, L.F.; Fortino, G. Assigning trust to devices in the context of consumer IoT applications. *IEEE Consum. Electron. Mag.* **2022**, *13*, 12–21. [CrossRef]

28. Khan, W.Z.; Aalsalem, M.Y.; Khan, M.K.; Arshad, Q. Data and privacy: Getting consumers to trust products enabled by the Internet of Things. *IEEE Consum. Electron. Mag.* **2019**, *8*, 35–38. [CrossRef]
29. Georgiades, M.; Hussain, F.; Christodoulou, L. Backdoor Adversarial Machine Learning Attack on Graph Convolutional Networks for IoMT Traffic Misclassification. In Proceedings of the International Conference on Innovations in Computing Research, London, UK, 25–27 August 2025; Springer: Cham, Switzerland, 2025; pp. 174–184.
30. Christodoulou, L.; Chari, A.; Georgiades, M. AI-enhanced healthcare IoT system: Advanced ML detection and classification algorithms for real-time cardiovascular monitoring. In Proceedings of the 2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), Abu Dhabi, United Arab Emirates, 29 April–1 May 2024; IEEE: New York, NY, USA, 2024; pp. 440–449.
31. Georgiades, M.; Christodoulou, L.; Chari, A.; Wang, K.; Ho, K.H.; Hou, Y.; Chai, W.K. Federated Learning for Early Cardiac Anomaly Prediction in Cross-Silo IoMT Environments. In Proceedings of the 2025 21st International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), Lucca, Italy, 9–11 June 2025; Presented at DCOSS-IoT.
32. Ravi, A.; Yu, X.; Santelices, I.; Karray, F.; Fidan, B. General frameworks for anomaly detection explainability: Comparative study. In Proceedings of the 2021 IEEE International Conference on Autonomous Systems (ICAS), Montreal, QC, Canada, 11–13 August 2021; IEEE: New York, NY, USA, 2021; pp. 1–5.
33. Ghassemi, M.; Oakden-Rayner, L.; Beam, A.L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **2021**, *3*, e745–e750. [CrossRef]
34. Meidan, Y.; Bohadana, M.; Mathov, Y.; Mirsky, Y.; Shabtai, A.; Breitenbacher, D.; Elovici, Y. N-baiot—Network-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Comput.* **2018**, *17*, 12–22. [CrossRef]
35. Nascita, A.; Cerasuolo, F.; Di Monda, D.; Garcia, J.T.A.; Montieri, A.; Pescape, A. Machine and deep learning approaches for IoT attack classification. In Proceedings of the IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Virtual, 2–5 May 2022; IEEE: New York, NY, USA, 2022; pp. 1–6.
36. Subramani, S.; Selvi, M. Multi-objective PSO based feature selection for intrusion detection in IoT based wireless sensor networks. *Optik* **2023**, *273*, 170419. [CrossRef]
37. Ozkan-Okay, M.; Samet, R.; Aslan, Ö.; Kosunalp, S.; Iliev, T.; Stoyanov, I. A novel feature selection approach to classify intrusion attacks in network communications. *Appl. Sci.* **2023**, *13*, 11067. [CrossRef]
38. Wang, M.; Zheng, K.; Yang, Y.; Wang, X. An explainable machine learning framework for intrusion detection systems. *IEEE Access* **2020**, *8*, 73127–73141. [CrossRef]
39. Sharma, B.; Sharma, L.; Lal, C.; Roy, S. Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach. *Expert Syst. Appl.* **2024**, *238*, 121751. [CrossRef]
40. Kalutharage, C.S.; Liu, X.; Chrysoulas, C.; Pitropakis, N.; Papadopoulos, P. Explainable AI-based DDOS attack identification method for IoT networks. *Computers* **2023**, *12*, 32. [CrossRef]
41. Dadkhah, S.; Neto, E.C.P.; Ferreira, R.; Molokwu, R.C.; Sadeghi, S.; Ghorbani, A.A. CICIoMT2024: Attack Vectors in Healthcare devices-A Multi-Protocol Dataset for Assessing IoMT Device Security. *Internet Things* **2024**, *28*, 101351. [CrossRef]
42. Thamilarasu, G.; Balasubramanian, A.; Mishra, S.; Sridhar, R. A cross-layer based intrusion detection approach for wireless ad hoc networks. In Proceedings of the IEEE International Conference on Mobile Adhoc and Sensor Systems Conference, Washington, DC, USA, 7 November 2005; IEEE: New York, NY, USA, 2005; pp. 861–868.
43. Poongothai, T.; Jayarajan, K. Intrusion Detection System for Mobile Ad Hoc Networks using Cross Layer and Machine Learning Approach. *Int. J. Comput. Appl.* **2018**, *179*, 5–13. [CrossRef]
44. Macedo, E.L.; Silva, R.S.; de Moraes, L.F.; Fortino, G. Trust Aspects of Internet of Things in the Context of 5G and Beyond. In Proceedings of the 2020 4th Conference on Cloud and Internet of Things (CIoT), Niterói, Brazil, 7–9 October, 2020; IEEE: New York, NY, USA, 2020; pp. 59–66.
45. Keshk, M.; Koroniotis, N.; Pham, N.; Moustafa, N.; Turnbull, B.; Zomaya, A.Y. An explainable deep learning-enabled intrusion detection framework in IoT networks. *Inf. Sci.* **2023**, *639*, 119000. [CrossRef]
46. Hussain, F.; Abbas, S.G.; A. Shah, G.; Pires, I.M.; Fayyaz, U.U.; Shahzad, F.; Garcia, N.M.; Zdravevski, E. IoT Healthcare Security Dataset. **2021**. [CrossRef]
47. Ghazanfar, S.; Hussain, F.; Rehman, A.U.; Fayyaz, U.U.; Shahzad, F.; Shah, G.A. IoT-Flock: An Open-source Framework for IoT Traffic Generation. In Proceedings of the 2020 International Conference on Emerging Trends in Smart Technologies (ICETST), Karachi, Pakistan, 26–27 March 2020; IEEE: New York, NY, USA, 2020; pp. 1–6.
48. Ding, C.; He, X. K-means clustering via principal component analysis. In Proceedings of the Twenty-First International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 29.
49. Li, L.; Zhang, Y.; Chen, L. Personalized prompt learning for explainable recommendation. *ACM Trans. Inf. Syst.* **2023**, *41*, 1–26. [CrossRef]

50. Breve, B.; Cimino, G.; Deufemia, V. Hybrid prompt learning for generating justifications of security risks in automation rules. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–26. [CrossRef]
51. Georgiades, M. crosslayer_xai_ids: An Explainable AI Approach to enhance Interpretability and Transparency for Intrusion Detection in Internet of Medical Things. 2025. Available online: https://github.com/mgeorgiades/crosslayer_xai_ids (accessed on 12 July 2025)