Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot

Research article

Evaluating and enhancing intrusion detection systems in IoMT: The importance of domain-specific datasets

Jordi Doménech [a,b,*], Olga León [a], Muhammad Shuaib Siddiqui [b], Josep Pegueroles [a]

[a] *Universitat Politècnica de Catalunya (UPC), Barcelona, 08034, Spain*
[b] *i2CAT, Barcelona, 08034, Spain*

ARTICLE INFO

Keywords:
Internet of Medical Things
Intrusion detection systems
Machine learning
Dataset Shift
Dataset optimization
CICIoMT2024
CICIoT2023

ABSTRACT

The emergence of the Internet of Medical Things (IoMT) is revolutionizing healthcare delivery, but also introducing critical challenges to cybersecurity and patient safety. Intrusion Detection Systems (IDSs) enhanced by Machine Learning (ML) have emerged as a powerful solution to identify cyberattacks in these environments. However, existing studies often rely on general IoT datasets, potentially limiting their applicability in IoMT-specific scenarios. This study addresses these limitations by comparing the performance of ML models trained on a general IoT dataset (CICIoT2023) and an IoMT-specific dataset (CICIoMT2024) to demonstrate the importance of domain-specific data. Our findings reveal substantial drops of up to 66.87% in the F1-score when models trained on one dataset are tested on the other. Furthermore, the study critiques key dataset design choices in CICIoMT2024, and proposes baseline optimization techniques including uniform windowing, proper train-validation-test splits, adjustments in temporal dependencies for time series data, and improved dataset balancing. By applying these techniques, we observe significant improvements in IDS performance in comparison to other approaches, with scores of 0.9985 in model accuracy. The findings show the necessity of using IoMT-specific datasets and carefully designed preprocessing techniques to build robust IDSs tailored to the unique demands of medical IoT environments.

1. Introduction

The Internet of Medical Things (IoMT), which includes connected medical devices, portable devices, implantables, health monitoring systems, therapeutic tools, and health applications, has transformative potential in healthcare by enabling remote patient monitoring, enhanced diagnostics, and efficient care delivery [1]. Despite its benefits, the rapid adoption of IoMT in healthcare has introduced critical cybersecurity challenges. In a recent study, the European Union Agency for Network and Information Security (ENISA) reported that during 2023 there was an increase of 55% of cybersecurity attacks in the healthcare sector, with IoMT devices being prime targets [2]. The vulnerabilities of these devices pose significant risks to patient safety and data privacy, making IoMT security a priority in emerging cybersecurity frameworks [3].

Intrusion Detection Systems (IDSs), defined as software or hardware that recognizes and responds to cybersecurity attacks autonomously, play a vital role in detecting and responding to sophisticated cyber threats that traditional detection systems struggle to identify [4]. In fact, Artificial Intelligence (AI)-based IDSs, which integrate Machine Learning (ML) techniques, have emerged

[*] Corresponding author at: Universitat Politècnica de Catalunya (UPC), Barcelona, 08034, Spain.
E-mail addresses: jordi.domenech.fons@upc.edu, jordi.domenech@i2cat.net (J. Doménech), olga.leon@upc.edu (O. León), shuaib.siddiqui@i2cat.net (M.S. Siddiqui), josep.pegueroles@upc.edu (J. Pegueroles).

https://doi.org/10.1016/j.iot.2025.101631
Received 18 March 2025; Received in revised form 25 April 2025; Accepted 27 April 2025
Available online 14 May 2025

as a crucial cybersecurity technology, leveraging data analysis to identify patterns and enhance the detection of cyber threats. ML has been successfully applied in the Internet of Things (IoT) to various domains beyond security, including agriculture and mobile computing, which demonstrate the potential of this technology in resource-constrained environments [5,6]. Furthermore, the application of ML to IDSs arises as a new approach that can give promising results in improving detection accuracy and reducing detection time [7]. However, recent studies [8,9] indicate that current AI-based IDSs are not optimized for realistic IoMT environments, leading to biased detection outcomes. Many existing IDS models focus on general IoT contexts, lacking the specificity required for IoMT environments and their unique security needs [10–13]. A key reason for this limitation is the extensive dependency on general IoT datasets, which may not accurately capture the attack patterns, network behavior, and threat landscape specific to medical IoT devices. This raises concerns about the applicability of IoT-based IDSs in healthcare settings, where security failures can directly impact patient safety.

To address this gap, it is necessary to determine whether general IoT datasets can adequately represent IoMT environments. Validation is required to evaluate whether there are significant differences between IoT and IoMT contexts, as using general IoT data to train IDSs for IoMT may be inadequate. In this study, we explore this assumption using two prominent data sets from the Canadian Institute of Cybersecurity (CIC): CICIoT2023, representing general IoT, and CICIoMT2024, specifically simulating IoMT environments [14,15]. While this study does not present an exhaustive comparison of all IoT and IoMT datasets, the selected datasets provide a representative and domain-specific foundation for evaluating the need for tailored datasets in medical IoT environments. This focus allows us to investigate whether developing IoMT-specific datasets, instead of only relying on general IoT datasets, is critical for enhancing IDS performance in healthcare settings. Although cross-evaluation has been explored in general IoT IDSs [16], our study is the first to conduct a systematic cross-domain generalization analysis between general IoT and IoMT. We utilize the extension context approach [17] to quantify how even a small inclusion of IoMT-specific data in training significantly improves IDS performance. To the best of our knowledge, no prior studies have applied this specific approach to assess the generalization and transferability of intrusion detection models between IoT and IoMT environments.

Additionally, although the CICIoMT2024 dataset is currently the most comprehensive and realistic dataset simulating IoMT environments, its data preprocessing techniques are not yet optimal. Enhancing preprocessing techniques in the CICIoMT2024 dataset could improve the realism of use cases and cyberattack detection performance in healthcare while reducing detection time. Although a recent study [18] proved that the development of AI-based IDSs in this dataset can lead to promising results, further work should be done to optimize AI models' performance in these critical environments. Such optimization techniques will serve as a foundation for future investigations, enabling the construction of more accurate and efficient AI-based IDSs.

In this context, the present study aims to enhance the understanding and development of IDSs tailored to IoMT environments by addressing key challenges in dataset representation, data preprocessing, and baseline optimization. Specifically, this study evaluates the adequacy of general IoT datasets versus IoMT-specific datasets and provides foundational techniques for optimizing IDSs in realistic healthcare scenarios. To achieve this, the study will:

1. Compare the CICIoT2023 and CICIoMT2024 datasets to determine the suitability of general IoT datasets for representing IoMT-specific environments by analyzing the generalization capabilities of ML models trained on one dataset and tested on the other.
2. Analyze the CICIoMT2024 dataset limitations by identifying and proposing solutions to its challenges, including issues such as inconsistencies in packet windowing, lack of proper validation splits, temporal correlation in data, and imbalanced traffic classes.
3. Conduct experiments using the CICIoMT2024 dataset to correctly evaluate IDS performance using the baseline optimization techniques proposed, thereby providing a benchmark for future AI-based IDS research in IoMT environments.

The remainder of this paper is structured as follows: Section 2 reviews related work on IoT and IoMT-specific IDSs, highlighting the current gaps in ML approaches. Section 3 describes the two datasets used in this study, focusing on their relevance and suitability for IoT and IoMT security. Section 4 provides a comparative analysis of IoT and IoMT scenarios, examining the representational differences that impact IDS training. In Section 5, we discuss the key limitations of the CICIoMT2024 dataset and propose baseline optimization techniques to enhance IDS performance. Finally, Section 6 concludes the paper, summarizing key findings and proposing directions for future research in robust IoMT intrusion detection systems.

2. Related work

In this section, we explore research on attack detection in IoT and IoMT environments, providing an overview of current IDSs used in these scenarios and the specific challenges introduced by IoMT. Moreover, we discuss current optimization techniques in IDS to improve performance.

2.1. IDS on IoT and IoMT environments

Intrusion Detection Systems play a critical role in securing IoT environments against cyber threats. Numerous studies have leveraged ML techniques to enhance detection accuracy and efficiency. For instance, Thereza and Rimili [19] developed models, such as decision trees, K-nearest neighbors, random forests, and Naïve Bayes, for identifying DDoS attacks on IoT networks, achieving 100% accuracy on the CICIoT2023 dataset. Similarly, Kumar et al. [20] evaluated different ML models to identify attacks using 34, 8, and 2 classes, reaffirming the efficacy of machine learning algorithms in enhancing the security of IoT devices. Studies such as Shafin et al. [21] relied on Deep Learning (DL)-based models for attack detection. They developed a hybrid DL model based on advanced neural networks using Convolutional Neural Networks (CNN) and Long-Short Term Memory (LSTM) to detect obfuscated memory malware attacks in IoT devices, outperforming existing models in terms of detection speed, reaching results of up to 0.255ms/sample. Despite these advancements, most IDS research in IoT focuses on generic IoT devices, often overlooking critical differences between IoT and domain-specific IoT, such as the Internet of Medical Things. This limitation highlights the significant challenge of Dataset Shift, wherein the statistical properties of the training dataset (IoT) differ substantially from those of the target application (IoMT) [22]. This shift raises concerns about the generalizability of IDS models when applied to IoMT environments, which involve unique real-time constraints and higher stakes due to patient safety concerns [8].

Guida et al. [16], in his cross-evaluation study, emphasizes the difficulties of ML models to generalize among three general IoT datasets: IoT-23 [23], IDS2018 [24], and KITSUNE [25]. The authors observed drops in performance of up to 76% in the F1-score metric. To address this problem, Apruzzese et al. [17] promoted the idea of cross-evaluating intrusion detection systems by defining the extension context: that is, the usage of existing labeled data from different networks to reduce the performance drops when generalizing the IDS to different environments. Similarly, Cantone et al. [26] applied data visualization techniques to four datasets acquired from different networks. The authors of this study proved that training ML models on a single dataset is not sufficient to create robust IDSs capable of generalizing to different environments. The study reported the presence of anomalies in the datasets' composition that directly impacted the generalizability performance of the classifiers to new scenarios. The work from Li et al. [27] also tackled the cross-domain problem for network traffic classification. The authors proposed a novel framework to classify cross-domain traffic by using unsupervised domain adaptation, achieving higher performance compared to other methods.

IoMT introduces additional complexity to generalization due to its integration with healthcare workflows and its reliance on life-critical devices. Studies such as Maimó et al. [28] and Hady et al. [29] demonstrate high performance in using ML algorithms for Ransomware and Man-in-the-Middle attack detection in real clinical environments, with accuracies of 99.97% and 92.71%, respectively. However, these studies only focus on a subset of attacks, leaving a significant gap in addressing the full spectrum of cybersecurity threats. IoMT remains underexplored in IDS literature, with most works repurposing IoT solutions that can fail to capture the specific characteristics of real-world healthcare environments. For instance, Nandy et al. [30] described an attack scenario about an edge-centric IoMT framework based on an Empirical Intelligent Agent using Swarm-NN strategy for attack detection. Although the paper proposes an IoMT-specific solution, the authors use the ToN-IoT dataset [31] (i.e., a general IoT dataset) for testing the framework. Similarly, Binbusayyis et al. [32] investigate five different ML algorithms for intrusion detection in IoMT networks, yet they use a general IoT dataset to evaluate the performance of the IDS. Unfortunately, there is still no research available comparing general IoT and IoMT-specific environments to validate the performance of IDS across these domains. This study will address this research gap by comparing the CICIoT2023 and CICIoMT2024 datasets.

2.2. Optimizing IDS for enhanced performance

To improve the performance and efficiency of AI-based IDSs, various optimization techniques have been proposed in the literature. Feature selection algorithms, such as Mutual Information Feature Selection (MIFS), have proved their ability to reduce dataset dimensionality and improve computational efficiency in IoMT environments [33]. Furthermore, Priya et al. [12] showed that combining multiple feature selection techniques, such as Principal Component Analysis (PCA) and Grey-Wolf Optimization (GWO), can lead to significant performance gains. Specifically, their approach resulted in a 15% improvement in accuracy and a 32% reduction in time complexity.

Addressing class imbalance is another critical aspect of optimizing IDS performance. Techniques such as undersampling the majority class and oversampling the minority class have proven effective for enhancing model accuracy [34,35]. For instance, Ayoub et al. [36] proposed an IDS tailored for remote healthcare, employing an oversampling approach based on the Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP). Their method achieved higher performance results in comparison with other IDS solutions that did not use oversampling methods.

Hyperparameter optimization is equally important in tailoring IDS performance. Fine-tuning hyperparameters enables models to adapt to specific environments, outperforming random search methods [37]. For instance, Masum et al. [38] demonstrated that hyperparameter optimization could improve IDS accuracy by 5.49% and F1-score by 6.04%, emphasizing its importance in achieving optimal model configurations.

Finally, despite the recent introduction of the CICIoMT2024 dataset, several studies in the literature have already utilized it to enhance the performance of AI-based IDS in IoMT environments. Lucia Hernandez-Jaimes et al. [39] built an IDS for detecting ransomware-spreading behavior based on Nilsimsa fingerprinting and ML, achieving an accuracy and F1-score of 98.37% and 98.59%, but only using a 2-class approach. Moreover, the authors in [40] used a Convolutional Neural Network (CNN) approach for attack detection, achieving perfect accuracy of 99% in binary, categorical, and multiclass classification tasks. However, the weighted approach they employ in calculating the performance metrics is unrealistic in imbalanced datasets such as the CICIoMT2024, providing biased results regarding precision, recall, and F1-score. An unweighted mean, which does not take into consideration label imbalance, should be utilized to calculate the performance metrics in these types of scenarios.

Regardless of the extensive application of these techniques in general IoT IDS research, their adaptation to IoMT remains limited. Few studies investigate how dataset balancing impacts the detection of IoMT-specific attack patterns, presenting a notable research gap. Additionally, advanced preprocessing methods such as feature selection and hyperparameter optimization remain underexplored in datasets like CICIoMT2024. Addressing these gaps is critical for improving IDS performance in realistic IoMT environments.

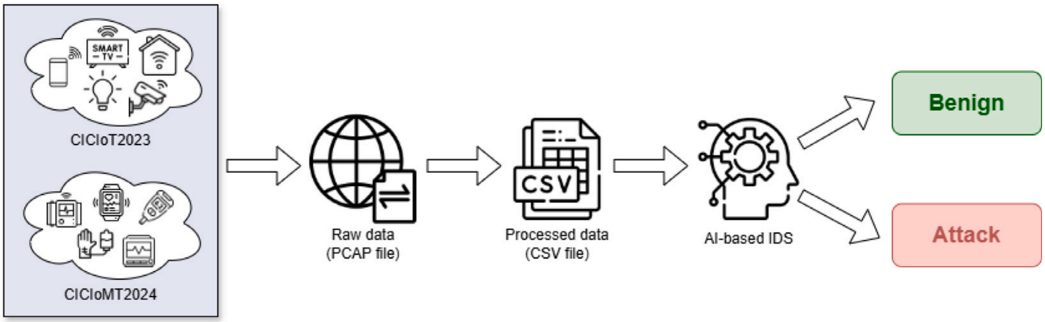
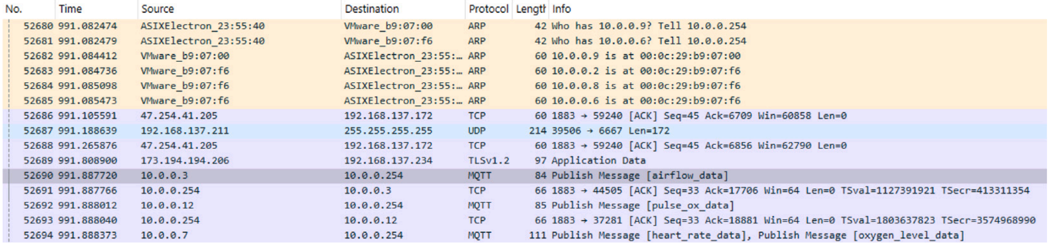


Fig. 1. Data collection workflow for IDS evaluation.



No.	Time	Source	Destination	Protocol	Length	Info
52680	991.082474	ASIXElectron_23:55:40	VMware_b9:07:00	ARP	42	Who has 10.0.0.9? Tell 10.0.0.254
52681	991.082479	ASIXElectron_23:55:40	VMware_b9:07:f6	ARP	42	Who has 10.0.0.6? Tell 10.0.0.254
52682	991.084412	VMware_b9:07:00	ASIXElectron_23:55:…	ARP	60	10.0.0.9 is at 00:0c:29:b9:07:00
52683	991.084736	VMware_b9:07:f6	ASIXElectron_23:55:…	ARP	60	10.0.0.2 is at 00:0c:29:b9:07:f6
52684	991.085098	VMware_b9:07:f6	ASIXElectron_23:55:…	ARP	60	10.0.0.8 is at 00:0c:29:b9:07:f6
52685	991.085473	VMware_b9:07:f6	ASIXElectron_23:55:…	ARP	60	10.0.0.6 is at 00:0c:29:b9:07:f6
52686	991.105591	47.254.41.205	192.168.137.172	TCP	60	1883 → 59240 [ACK] Seq=45 Ack=6709 Win=60858 Len=0
52687	991.188639	192.168.137.211	255.255.255.255	UDP	214	39506 → 6667 Len=172
52688	991.265876	47.254.41.205	192.168.137.172	TCP	60	1883 → 59240 [ACK] Seq=45 Ack=6856 Win=62790 Len=0
52689	991.808900	173.194.194.206	192.168.137.234	TLSv1.2	97	Application Data
52690	991.887720	10.0.0.3	10.0.0.254	MQTT	84	Publish Message [airflow_data]
52691	991.887766	10.0.0.254	10.0.0.3	TCP	66	1883 → 44505 [ACK] Seq=33 Ack=17706 Win=64 Len=0 TSval=1127391921 TSecr=413311354
52692	991.888012	10.0.0.12	10.0.0.254	MQTT	85	Publish Message [pulse_ox_data]
52693	991.888040	10.0.0.254	10.0.0.12	TCP	66	1883 → 37281 [ACK] Seq=33 Ack=18881 Win=64 Len=0 TSval=1803637823 TSecr=3574968990
52694	991.888373	10.0.0.7	10.0.0.254	MQTT	111	Publish Message [heart_rate_data], Publish Message [oxygen_level_data]

Fig. 2. Example of raw data (PCAP file).

3. Description of datasets

To study the particularities of medical environments in IoT scenarios, several dataset options were available for analysis. The authors in [41] provide a list of existing IoMT datasets. Among the most complete and recent are CICIoMT2024 and IoMT-TrafficData. As we discuss in Section 6, combining these two datasets could support broader and more general studies. However, this work focuses on the CICIoMT2024 dataset. The availability of CICIoT2023 and CICIoMT2024 datasets, which have similar characteristics in terms of attributes, testbed design, and data generation methodology, allows for a meaningful comparison between general IoT and IoMT environments. Despite some limitations, these datasets are considered sufficiently comprehensive to be representative of their respective domains for the purpose of this study.

The CICIoT2023 and CICIoMT2024 datasets, developed by the University of New Brunswick and presented in [14,15], respectively, serve as comprehensive benchmarks for evaluating IDSs in IoT and IoMT environments. The CICIoT2023 dataset, created in 2023, represents a general IoT operational environment and includes a wide array of IoT-specific attacks. Recognizing the distinct characteristics and challenges of IoMT environments, the University introduced the CICIoMT2024 dataset in 2024 to simulate a domain-specific IoMT scenario. While these datasets provide valuable resources for intrusion detection research, it remains uncertain whether general IoT datasets like CICIoT2023 adequately represent IoMT environments or if domain-specific datasets like CICIoMT2024 are indispensable for achieving effective IDS performance in healthcare. This section provides an overview of these datasets and highlights their key differences.

As mentioned above, both datasets were generated using an IoT topology, although their environments and device compositions differ. The CICIoT2023 dataset replicates a smart home environment with 105 IoT devices, whereas the CICIoMT2024 dataset focuses on healthcare applications, utilizing 40 devices commonly employed in medical care and monitoring. Among these, 25 devices are physical, while the remaining 15 are simulated using the IoT Flock framework [42]. These differences in device types and communication protocols are summarized in Table 1.

In both datasets, all devices were connected to the Internet through various interconnection devices, such as access points (APs), hubs, switches, and routers, facilitating communication across the network. Data collection was performed in a controlled environment using a Gigamon Network Tap, a hardware tool capable of monitoring and storing network packets without interfering with regular network operations. The overall data collection workflow is illustrated in Fig. 1, which outlines the process from raw packet capture (PCAP files) to feature extraction and classification using AI-based IDS. To provide further insight into the raw data structure, Fig. 2 presents a sample of diverse captured packets before feature extraction.

The datasets include both benign and malicious traffic. Benign traffic reflects legitimate network activity, while malicious traffic was generated by executing a variety of attacks using Raspberry Pi devices. For feature extraction, network packets were processed to derive 47 features in the CICIoT2023 dataset using the DPKT package [43]. A similar approach was applied to the CICIoMT2024 dataset, resulting in 45 attributes. These attributes largely overlap with those in CICIoT2023 but with three features omitted, and one new feature added, resulting in 44 attributes that both datasets have in common. In Table A.9 of Appendix we can see a detailed

Table 1
Comparison of devices included in CICIoT2023 and CICIoMT2024 datasets.

Protocol	Device type	CICIoT2023	CICIoMT2024
WiFi	Camera	✓	✓
	Bulb, smart plug	✓	✗
	Baby monitor	✗	✓
ZigBee	Camera	✓	✗
	Bulb, plug	✓	✗
	Motion, flood, door and window sensor	✓	✗
Bluetooth	Camera, oxygen and heart rate monitors	✗	✓
MQTT	Camera, spirometer, EMG	✗	✓
	Infusion pump, glucose sensor	✗	✓

Table 2
Comparison of attacks included in CICIoT2023 and CICIoMT2024 datasets.

Attack Type	Attack subtype	CICIoT2023	CICIoMT2024
DDoS	TCP, UDP, SYN, ICMP flood	✓	✓
	ACK, UDP and ICMP fragmentation, SlowLoris	✓	✗
	RSTFIN, PSHACK, HTTP, SynonymousIP flood	✓	✗
DoS	TCP, UDP, SYN, ICMP flood	✓	✓
Reconnaissance	OS, Port and Vulnerability scan, Ping sweep	✓	✓
	Host discovery	✓	✗
Web-based	SQL, Command injection	✓	✗
	Backdoor malware, Uploading attack	✓	✗
	Cross-site scripting, Browser hijacking	✓	✗
Spoofing	ARP spoofing	✓	✓
	DNS spoofing	✓	✗
Brute force	Dictionary	✓	✗
Mirai	UDPPlain, GREIP and Greeth flood	✓	✗
MQTT	DoS connect and Publish flood, Malformed data	✗	✓
	DDoS connect and Publish flood	✗	✓

comparison of these attributes. Finally, the authors from the CICIoT2023 and CICIoMT2024 datasets averaged network packets using varying window sizes of 10 packets (i.e., for non-large-scale attacks), and 100 packets (i.e., for large-scale attacks such as DDoS, DoS, and Mirai). This resulted in various CSV files available for researchers in a labeled format.

A comparison of the attacks simulated in each dataset is provided in Table 2. These attacks were carefully selected by Neto et al. [14] and Dadkhah et al. [15] to represent real-world scenarios commonly encountered in IoT and IoMT environments. In fact, the CICIoT2023 dataset includes 33 attacks grouped into 7 categories, reflecting its focus on a wide range of general IoT applications and corresponding threats. In contrast, the CICIoMT2024 dataset contains 18 attacks categorized into 5 groups, specifically designed to capture the unique challenges and vulnerabilities of IoMT environments. While the number of attacks in CICIoMT2024 is smaller, this aligns with the narrower scope of IoMT, where certain attack methods are more critical due to their direct impact on patient safety and healthcare operations [8]. To ensure a comprehensive analysis, we included common attacks provided in each dataset (attack subtype) and grouped them into categories (attack type) for better comparability. These differences in attack diversity and classification further reflect the distinct focuses of the datasets and underline the importance of tailoring IDSs to the specific needs of IoMT environments.

As illustrated in Fig. 3, the number of attack samples varies significantly between the two datasets. Certain attack types, such as DDoS and DoS, exhibit a substantially higher frequency in CICIoT2023, whereas other attacks, including MQTT-based threats, are only present in CICIoMT2024. Furthermore, some attack types, such as Reconnaissance and Spoofing, are shared across both datasets but with different sample distributions.

4. Comparison of IoT and IoMT scenarios

The primary objective of this section is to evaluate whether general IoT datasets can adequately represent IoMT-specific environments for the development of IDSs. To achieve this, we analyze the generalization capabilities of ML models trained on IoT environments and tested them on IoMT scenarios.

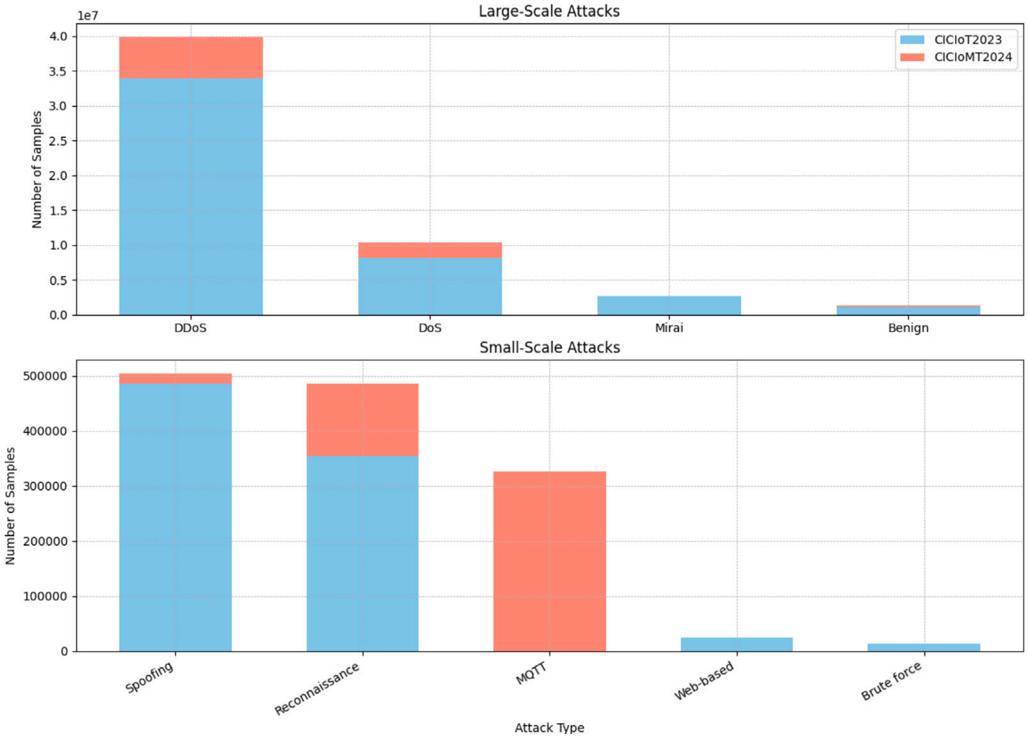


Fig. 3. Attack distribution of the number of samples per attack type in both datasets.

Table 3
Summary of the hyperparameters used by each ML Model.

ML Model	Hyperparameters
Decision Tree (DT)	{criterion=‘gini’, max_depth=10}
Random Forest (RF)	{n_estimators=10}
XGBoost (XGB)	{objective=‘multi:softprob’, max_depth=6, learning_rate=0.3, subsample=0.8, colsample_bytree=0.8, eval_metric=‘mlogloss’}
Feedforward Neural Network (FNN)	{ hidden_layer_sizes=(64,64,64), optimizer=‘adam’, activation=‘relu’, learning_rate=0.001, beta_1=0.9, beta_2=0.999, epsilon=1e−07, early_stopping=True, monitor=‘val_loss’, patience=5}

4.1. Experimental setup

4.1.1. Datasets

The analysis uses the two publicly available datasets described in Section 3: CICIoT2023 [14], which simulates general IoT environments, and CICIoMT2024 [15], the most comprehensive dataset representing IoMT-specific environments. We focus on their key characteristics, including feature sets, attack types, and contextual differences.

4.1.2. ML models

We employ three widely known tree-based ML classifiers for intrusion detection: Decision Tree (DT), Random Forest (RF), and eXtreme Gradient Boosting (XGB). Moreover, we incorporate a Feedforward Neural Network (FNN) to expand the scope of the study and highlight any performance difference that a Deep Learning (DL)-based model can have in our scenario. These ML algorithms based on trees and neural networks are selected for their proven performance in attack detection on IoT and IoMT environments [18–20,44,45]. The hyperparameters used from the scikit-learn library are detailed in Table 3. They were selected based on common configurations observed in similar IDS literature [18,38]. For instance, a $max\_depth <= 10$ was chosen to avoid overfitting in three-based models, while the learning rate for XGBoost was set to 0.3, a value frequently observed to achieve fast convergence in intrusion detection tasks. Furthermore, the hyperparameters used for Random Forest and Feedforward Neural Network are similar to those presented in the original work by the University of New Brunswick to facilitate robust comparisons [14,15].

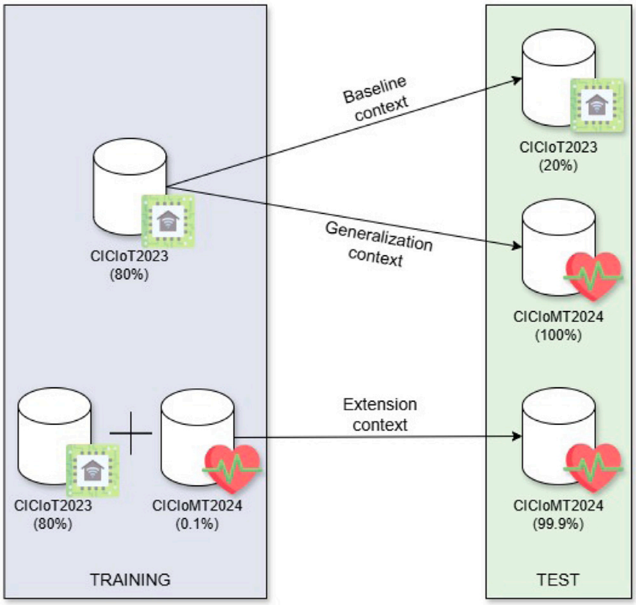


Fig. 4. Comparison among the Baseline, Generalization and Extension contexts.

4.1.3. *Evaluation metrics*

Several evaluation metrics exist in the literature for evaluating AI-based IDS in attack detection scenarios, which are based on True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [46]. The metrics used in this study include: Accuracy, which measures the proportion of correctly predicted instances out of the total number of instances. Precision, for quantifying the proportion of correct positive predictions. Recall, for measuring the proportion of actual positive instances that the model correctly identifies. F1-score, which is the harmonic mean of precision and recall, and provides a single metric that balances both measures. And, ROC-AUC score, which is the area under the Receiver Operating Characteristic (ROC) curve; that is the plot of the true positive rate (TPR) against the false positive rate (FPR).

4.1.4. *Experimental procedure*

In this study, we consider three contexts where the experimental evaluation is performed [16,17]:

- **Baseline Context:** This is the context more investigated in the current literature; models are trained and tested on the same environment to establish a baseline performance.
- **Generalization Context:** This context aims to assess whether the IDS can detect attacks from other datasets, that is, other environments. Thus, the model is trained on the baseline environment, and tested on a new environment not yet seen by the IDS.
- **Extension Context:** The extension context tries to minimize the low detection performance of the model when tested on a new environment (i.e., generalization context). To this aim, a few attack instances originating from the new environment are included in the training set for extending the detection capabilities of the IDS.

Fig. 4 illustrates the experimental workflow carried out for comparing general IoT and IoMT-specific datasets using the baseline, generalization, and extension contexts. For the baseline context, the CICIoT2023 dataset is used to train and test the ML models, following a training–testing split of 80%–20%. In the generalization context, the CICIoMT2024 dataset is entirely used for testing the models while being trained in the CICIoT2023 dataset. Finally, the extension context uses the CICIoT2023 for training and 0.1% of the CICIoMT2024 dataset, since we want to extend the detection capabilities of the IDS by adding augmented knowledge of the IoMT environment.

For consistency, the datasets are preprocessed identically to standardize feature distribution and ensure comparability across scenarios. A total of 44 features are used, which represent common network features between both datasets, as seen in Table A.9 of Appendix.

This experimental setup leverages the concept of Dataset Shift [22] to systematically quantify the representational differences between general IoT and IoMT-specific environments. General IoT datasets typically encompass diverse applications, such as smart homes, cities, industries, and healthcare, which exhibit substantial variability in communication protocols, device types, and data patterns compared to the specialized requirements of IoMT environments. Research on IDS in healthcare emphasizes that ignoring these domain-specific factors can result in significant performance gaps, underscoring the need for specialized datasets for IoMT [8]. To validate this hypothesis, three cross-evaluation experiments are designed:

Table 4
Performance analysis of Cross-Evaluation Experiment 1.

	Context	ML model	F1-score	ROC-AUC
5 classes	Baseline	DT	0.9344	0.9994
		RF	**0.9429**	0.9920
		XGB	0.9419	**0.9996**
		FNN	0.8595	0.9982
	Generalization	DT	0.4370	0.6587
		RF	0.3889	0.7989
		XGB	0.3749	0.8059
		FNN	**0.5491**	**0.8199**
2 classes	Baseline	DT	0.9733	0.9995
		RF	**0.9773**	0.9993
		XGB	0.9762	**0.9996**
		FNN	0.9499	0.9699
	Generalization	DT	0.6482	0.7534
		RF	0.7147	0.8977
		XGB	0.5523	0.9825
		FNN	**0.9491**	**0.9842**

- **Cross-Evaluation Experiment 1:** Analysis of equivalent attack scenarios in both datasets to evaluate ML model performance when generalizing to environments containing identical attacks (i.e., model training on CICIoT2023 and testing on CICIoMT2024 with same attacks). This first experiment represents the use case where an organization using IoT devices faces the same attacks as a healthcare entity using IoMT devices.
- **Cross-Evaluation Experiment 2:** Evaluation of healthcare-specific attack scenarios to reflect realistic conditions with attacks unique to the CICIoMT2024 dataset (i.e., model training on CICIoT2023 and testing on CICIoMT2024, using the same attacks as in Experiment 1, along with a few attacks exclusive to CICIoMT2024). This second experiment represents the use case in which a healthcare entity using IoMT devices faces attacks different from an organization using IoT devices.
- **Cross-Evaluation Experiment 3:** Application of the extension context to test the impact of augmenting training data with malicious traffic from the IoMT domain (i.e., model training on CICIoT2023 and with some data of CICIoMT2024, and tested on CICIoMT2024). This last experiment represents the use case where the healthcare entity, having realized poor detection performance due to its specific attacks, adds malicious traffic from the IoMT environment to the IoT environment to extend the capabilities of the IDS.

The experiments will utilize the attack instances detailed in Table 2. In Experiment 1, shared attacks between the datasets will be used, categorized into two classes (Benign and Malicious) and five classes (Benign, DDoS, DoS, Reconnaissance, and ARP Spoofing). Experiments 2 and 3 will extend this analysis by incorporating MQTT-specific attacks, which are specific from the CICIoMT2024 dataset, providing a more comprehensive evaluation of domain-specific generalization.

4.2. Experimental results and discussion

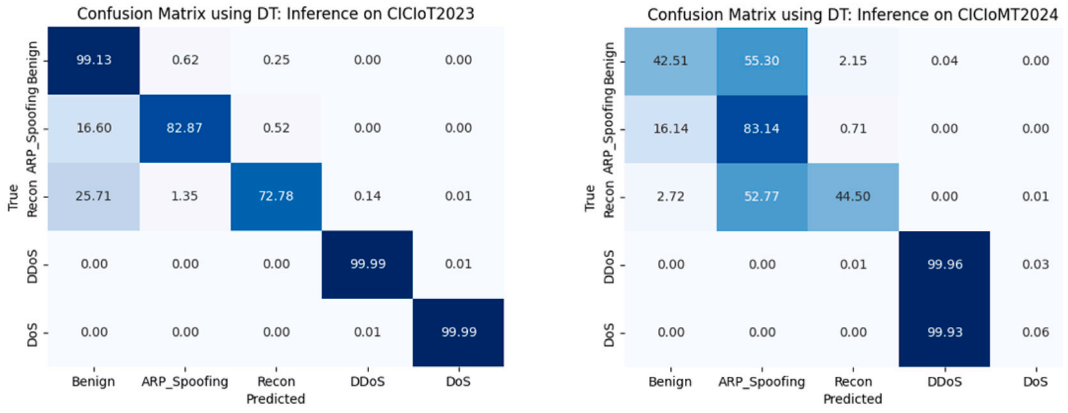
The results from the Cross-Evaluation Experiments 1, 2, and 3 are illustrated in this section. The F1-score and the ROC-AUC score are selected as the most appropriate metrics for comparing IoT and IoMT datasets. The F1-score is chosen due to its ability to balance precision and recall, making it robust against the class imbalances present in these datasets. Unlike accuracy, which can be misleading in scenarios with skewed class distributions, the F1-score provides a more reliable measure by harmonizing the trade-off between correctly identifying malicious traffic (precision) and minimizing undetected threats (recall). Similarly, the ROC-AUC score complements this evaluation by quantifying the overall performance of the models across all classification thresholds. It provides a comprehensive view of the trade-offs between true positive and false positive rates, enabling a more detailed comparison between IoT and IoMT datasets.

4.2.1. Cross-evaluation experiment 1: Same attacks

Table 4 describes the results from the first cross-evaluation experiment. The results from the 5-class and 2-class scenarios are presented, considering both the baseline and generalization contexts.

In the baseline context, the F1-scores for DT, RF, XGB, and FNN trained and tested on the 5-class IoT dataset were 0.9344, 0.9429, 0.9419, and 0.8595, respectively. These results, combined with high ROC-AUC scores, demonstrate the models' robustness and their ability to accurately classify both benign and malicious traffic across multiple attack types when the training and testing environments are consistent. This reflects optimized feature learning specific to the IoT dataset, aligning with prior studies that highlight the efficacy of ML models in detecting attacks in IoT environments [47].

However, the performance clearly decreased when models were tested on the IoMT dataset, with reductions of 49.74% (DT), 55.4% (RF), 56.7% (XGB), and 31.04% (FNN) in the F1-scores. The FNN exhibited the least decrease in performance, being the model with the greatest generalization capabilities in this context. These findings underscore the inability of models trained on



(a) DT confusion matrix using 5 classes in the baseline context

(b) DT confusion matrix using 5 classes in the generalization context

Fig. 5. Confusion matrices of the Cross-Evaluation Experiment 1 for multiclass classification.

general IoT data to generalize effectively to IoMT environments, even for equivalent attack scenarios. The confusion matrices in Figs. 5(a) and 5(b) depict the main differences between the baseline and generalization contexts. Note that, in the generalization context, the models struggle to identify benign traffic and reconnaissance attacks, while failing to distinguish between DDoS and DoS attacks.

In the simplified 2-class scenario (benign vs. attack), the F1-scores for DT, RF, XGB, and FNN in the baseline context were 0.9733, 0.9773, 0.9762, and 0.9499, respectively. These results, together with the high ROC-AUC scores, highlight the models' strong performance in binary classification within the IoT dataset, consistent with existing literature results [48].

However, when tested on the IoMT dataset, the F1-scores of the tree-based models dropped to 0.6482 (DT), 0.7147 (RF), and 0.5523 (XGB). While the performance degradation is less significant compared to the 5-class scenario, it still indicates suboptimal generalization to the IoMT environment. Among the three models, Random Forest showed the smallest performance drop of 26.26%, underscoring its relative robustness in this context. The confusion matrices in Figs. 6(a) and 6(b) further reveal that, despite high baseline performance, the models exhibit significant challenges in distinguishing between benign and malicious traffic when generalized to the IoMT environment. In fact, 71.22% of the benign traffic is misclassified and detected as an attack, which alters the regular performance of the IDS.

On the other hand, the DL-based model showed promising results for the 2-class scenario in the generalization context. The FNN only experienced a drop of 0.08% in the F1-score, underscoring that DL-based models can have high robustness when generalized to different environments in a simplified 2-class scenario (benign vs. attack). The confusion matrices in Figs. 6(c) and 6(d) reveal this small performance drop, with only 10.99% of the benign traffic misclassified.

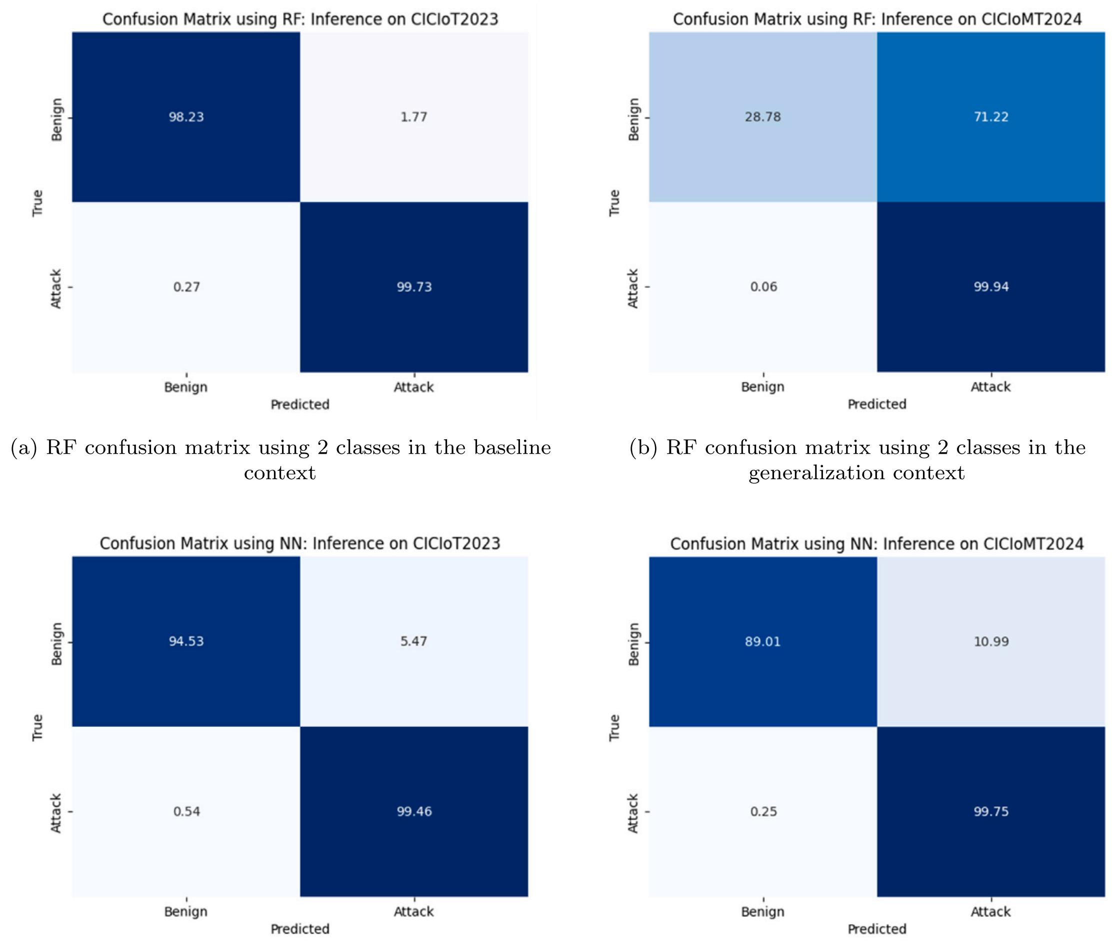
4.2.2. Cross-evaluation experiment 2 and 3: Different attacks

In this subsection, we provide the results and discussion for Cross-Evaluation Experiments 2 and 3, including the baseline, generalization, and extension contexts for the 5-class and 2-class scenarios (Table 5).

For the 5-class scenario in the baseline context, the F1-scores for DT, RF, XGB, and FNN remained consistently high at 0.9344, 0.9429, 0.9419, and 0.8595, respectively. However, in the generalization context, the F1-scores dropped significantly to 0.3541 (DT), 0.2742 (RF), 0.2905 (XGB), and 0.4122 (FNN), representing reductions of 58.03%, 66.87%, 65.14%, and 44.73% respectively. These results highlight the severe limitations of these models in detecting IoMT-specific attack scenarios when trained only on general IoT data. This strong decline emphasizes that IoMT environments exhibit unique traffic patterns and attack characteristics not adequately captured by general IoT datasets. The confusion matrix in Fig. 7(a) further reveals that the model misclassified several threats. Specifically, the MQTT attack was misclassified, which is the attack from the IoMT environment that is not incorporated into the IoT dataset.

The introduction of the extension context, where IoMT traffic was integrated into the training set, considerably improved the F1-scores of the tree-based models to 0.8338 (DT), 0.8540 (RF), and 0.8015 (XGB). While still below the baseline performance, these results prove the potential of hybrid training datasets to enhance detection capabilities for IoMT-specific scenarios. The great improvement in performance compared to the generalization context suggests that incorporating even limited IoMT-specific data can significantly mitigate the loss in generalization performance (Fig. 7(b)). However, the FNN performance in the extension context did not improve as much as the tree-based models, with 0.5289 for the F1-score, and 0.8636 for the ROC-AUC score. This lower performance may be attributed to its simple NN architecture, which limits its capacity to learn complex patterns.

In the simplified 2-class scenario, the baseline F1-scores were again strong at 0.9733, 0.9773, 0.9762, and 0.9499 for DT, RF, XGB, and FNN, respectively. However, in the generalization context, the F1-scores from the tree-based models dropped to 0.6473 (DT), 0.7137 (RF), and 0.5526 (XGB). While these results represent less significant declines compared to the 5-class scenario, they



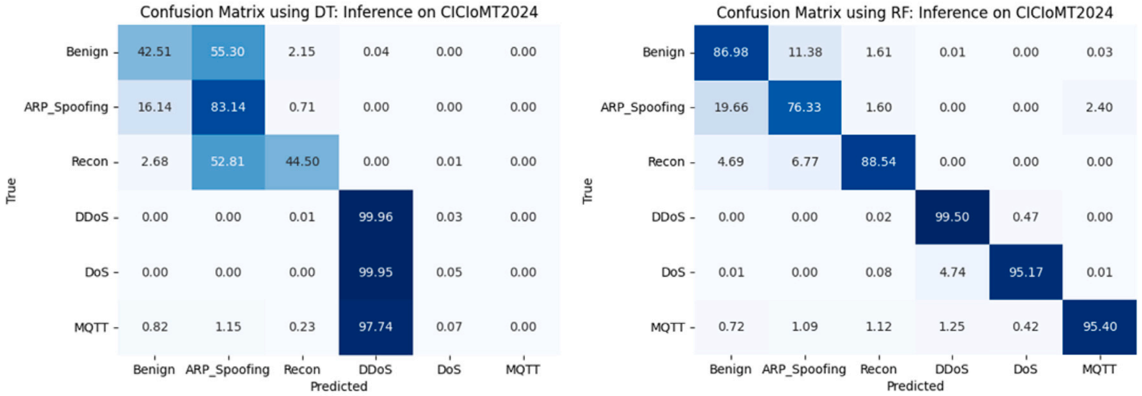
(a) RF confusion matrix using 2 classes in the baseline context

(b) RF confusion matrix using 2 classes in the generalization context

(c) FNN confusion matrix using 2 classes in the baseline context

(d) FNN confusion matrix using 2 classes in the generalization context

Fig. 6. Confusion matrices of the Cross-Evaluation Experiment 1 for binary classification.



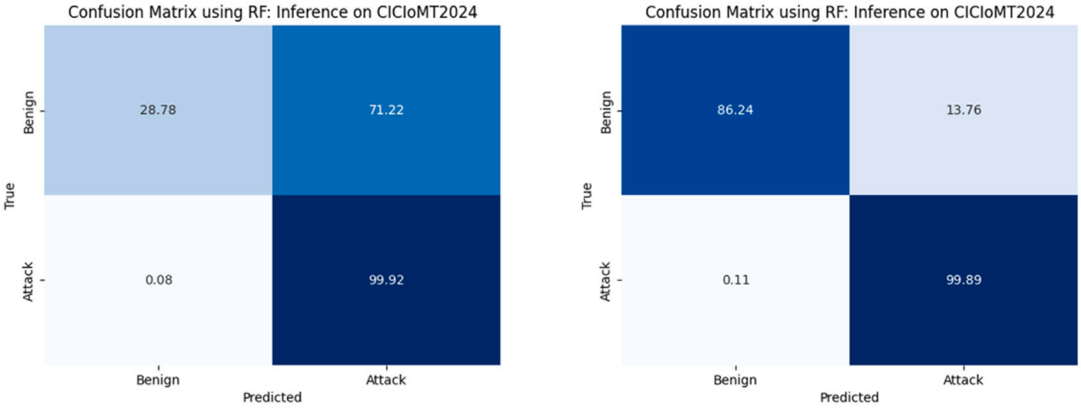
(a) DT confusion matrix using 5 classes in the generalization context

(b) RF confusion matrix using 5 classes in the extension context

Fig. 7. Confusion matrices of the Cross-Evaluation Experiments 2 and 3 for multiclass classification.

Table 5
Performance analysis of Cross-Evaluation Experiment 2 and 3.

	Context	ML model	F1-score	ROC-AUC
5 classes	Baseline	DT	0.9344	0.9994
		RF	**0.9429**	0.9920
		XGB	0.9419	**0.9996**
		FNN	0.8595	0.9982
	Generalization	DT	0.3541	0.6211
		RF	0.2742	0.7463
		XGB	0.2905	0.7184
		FNN	**0.4122**	**0.7873**
	Extension	DT	0.8338	**0.9978**
		RF	**0.8540**	0.9910
		XGB	0.8015	0.9725
		FNN	0.5289	0.8636
2 classes	Baseline	DT	0.9733	0.9995
		RF	**0.9773**	0.9993
		XGB	0.9762	**0.9996**
		FNN	0.9499	0.9699
	Generalization	DT	0.6473	0.7535
		RF	0.7137	0.8968
		XGB	0.5526	0.9664
		FNN	**0.9272**	**0.9865**
	Extension	DT	0.9500	**0.9992**
		RF	**0.9515**	0.9925
		XGB	0.9429	0.9990
		NN	0.9298	0.9886



(a) RF confusion matrix using 2 classes in the generalization context

(b) RF confusion matrix using 2 classes in the extension context

Fig. 8. Confusion matrices of the Cross-Evaluation Experiments 2 and 3 for binary classification.

still indicate suboptimal performance in distinguishing between benign traffic and IoMT-specific attacks, as illustrated in Fig. 8(a). Note that RF exhibited the smallest performance reduction (26.36%), consistent with its observed robustness in similar studies [49]. On the other hand, the DL-based model (FNN) did not suffer any important degradation in their F1-score and ROC-AUC score, being again the most robust model in the generalization context of 2 classes.

Under the extension context, the F1-scores improved to 0.9500 (DT), 0.9515 (RF), 0.9429 (XGB), and 0.9298 (FNN). These values approach the baseline performance, demonstrating the efficacy of incorporating IoMT-specific attack instances into the training data, specifically in the 2-class scenario. This recovery, depicted in Fig. 8(b), highlights the practical utility of augmenting general IoT datasets with domain-specific data to enhance model generalization across diverse environments.

The results from the experiments verify that general IoT datasets are insufficient for training effective IDSs for IoMT environments. The significant drops in F1-scores and ROC-AUC scores, particularly in the generalization context, highlight the importance of IoMT-specific datasets to address the unique challenges posed by healthcare environments, ensuring robust and reliable intrusion detection systems. Nevertheless, the FNN demonstrated significant generalization capabilities compared to the tree-based models, specifically for the 2-class scenario. For the 5-class scenario, more complex neural network architectures might be explored to improve generalization capabilities. Additionally, the extension context offers a promising avenue for improving model adaptability.

Still, the ultimate solution lies in creating dedicated IoMT datasets that reflect the complexity and variability of real-world traffic and threats.

4.3. Limitations of the study

While the results of this study are promising, several limitations should be considered:

- This work relies exclusively on the CICIoT2023 and CICIoMT2024 datasets. Although these datasets are well-designed and realistic, they may not capture the full range of behaviors and characteristics present in real-world IoT and IoMT deployments. However, real-world datasets that fully capture the complexity and diversity of these networks are unavailable yet. Future studies should include additional available datasets to validate the generalizability of the findings.
- The analysis is limited to the set of network features common to both datasets. However, this shared feature set may overlook important characteristics specific to each environment. Developing a standardized preprocessing framework to extract a consistent set of features from diverse datasets would improve cross-dataset comparisons and support more robust model evaluation.
- The datasets primarily use MQTT as the communication protocol, which limits protocol diversity. Other widely used IoT protocols, such as CoAP, are not represented and should be considered in future work to better reflect real-world IoMT conditions.
- Although the datasets include a range of devices and attack types, they may not fully reflect the scale and complexity of large healthcare networks. Real-world hospital environments may involve more heterogeneous devices, layered infrastructures, and sophisticated attack scenarios.

Whereas addressing these limitations would help improve the reliability and applicability of our study, it is important to note that they do not compromise the validity of the core findings. The proposed methodology and observations remain relevant and valuable for advancing intrusion detection in IoMT environments.

5. Optimizing IDS for enhanced performance in IoMT

As shown in the previous section, IoMT datasets are essential for evaluating IDSs in IoMT environments. The CICIoMT2024 dataset [15] represents a significant advancement in capturing the specific cybersecurity requirements of IoMT devices for attack detection. However, when using this dataset for research, certain creation and modeling aspects need careful adjustments to ensure robust intrusion detection systems. In this section, we discuss the key limitations of the dataset and propose baseline techniques to enhance IDS performance with a proper methodology. Moreover, the results from applying these techniques are discussed and compared with previous approaches. These guidelines might serve as a foundation for researchers aiming to design optimized IDS tailored to IoMT-specific environments.

5.1. Critiques of the CICIoMT2024 dataset and modeling practices

Although the main conclusions are still valid when considering results presented in [15], we have to notice that some important methodological aspects have to be considered when using this dataset. Next, we highlight four paths for improvement.

5.1.1. Varied window sizes for attack traffic

The authors from the CICIoMT2024 dataset adopt a varying windowing strategy for computing the network attributes described in Table A.9. In fact, they utilize a specific packet window for capturing DDoS and DoS attacks compared to other traffic types. Traffic from these two attack types is grouped into windows of 100 packets, whereas the rest use 10-packet windows. While this approach was intended to mitigate the extensive traffic of these attacks, it introduces inconsistencies in the statistical properties of the dataset that can lead to different detection rates [50,51].

Changing window sizes alters the underlying distribution of features, complicating model training and evaluation. In practical deployments, IDS models require a fixed window size for prediction to segment the dataset into fixed time intervals, enabling consistent feature extraction and analysis. The window size must be carefully selected to balance granularity and computational efficiency, ensuring sufficient detail for intrusion detection without overwhelming the processing pipeline. To ensure consistency, we recommend using uniform window sizes across all traffic types, as this approach standardizes the representation of traffic patterns through the dataset. Future work should focus on optimizing window size selection to capture diverse attack patterns. Additionally, we encourage dataset authors to adopt this uniform windowing approach when preprocessing raw network traffic data into features, as it enhances the reliability and comparability of IDS model evaluations.

5.1.2. Lack of proper validation splits

In the modeling section of the original paper, the authors split the dataset into training and testing subsets but did not incorporate a validation set for model selection. This practice can lead to overfitting and unreliable performance metrics, as the test set should only be used for final evaluation.

A more rigorous approach is to divide the dataset into three subsets: training, validation, and test. Models should be trained on the training set, optimized using the validation set, and only then evaluated on the test set. This ensures that model selection is not biased by information from the test set, leading to more reliable generalization metrics [52].

Table 6
Average validation results employing optimization techniques in the CICIoMT2024 using 6 classes and Random Undersampling strategy.

ML model	Accuracy	Precision	Recall	F1-score
DT	0.9900	0.8872	0.9660	0.8946
RF	0.9950	**0.9276**	0.9756	0.9436
XGB	**0.9976**	0.9270	**0.9834**	**0.9470**
FNN	0.7353	0.7503	0.8065	0.7324

Table 7
Average validation results employing optimization techniques in the CICIoMT2024 using 6 classes and SMOTE Oversampling strategy.

ML model	Accuracy	Precision	Recall	F1-score
DT	0.9966	0.9096	0.9695	0.9256
RF	**0.9985**	**0.9706**	0.9729	**0.9716**
XGB	0.9983	0.9433	**0.9806**	0.9585
FNN	0.8057	0.8156	0.8431	0.7888

5.1.3. Temporal correlation and shuffling

IoMT traffic data often exhibits temporal correlations due to its sequential nature. While most machine learning models assume that data instances are independently and identically distributed (IID), temporal dependencies can lead to biased predictions [53]. Although shuffling is a common preprocessing step to randomize data order, it is unclear whether this was explicitly performed in CICIoMT2024.

Tree-based models, such as Random Forests and Gradient Boosting, are less affected by local correlations, but neural networks can generalize poorly if temporal dependencies are not mitigated. Based on available parameters in the CICIoMT2024 scripts, it appears shuffling was applied to neural network training. We recommend explicitly documenting and performing shuffling for all models to reduce potential biases and ensure reproducibility.

5.1.4. Dataset imbalance

Class imbalance is a recurring challenge in the majority of publicly available datasets for attack detection, including the CICIoMT2024 dataset. In fact, DDoS and DoS attacks dominate the dataset, and other attack types, such as MQTT attacks, Reconnaissance attacks, ARP spoofing attacks, and even Benign traffic, are underrepresented. This imbalance alters model predictions toward the majority classes, leading to poor detection rates for minority attacks.

To address this, we recommend employing oversampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE) [54], or undersampling methods to balance the dataset. These approaches ensure that minority attack types are adequately represented, improving the model's ability to generalize across all traffic categories [55].

5.2. Results and discussion

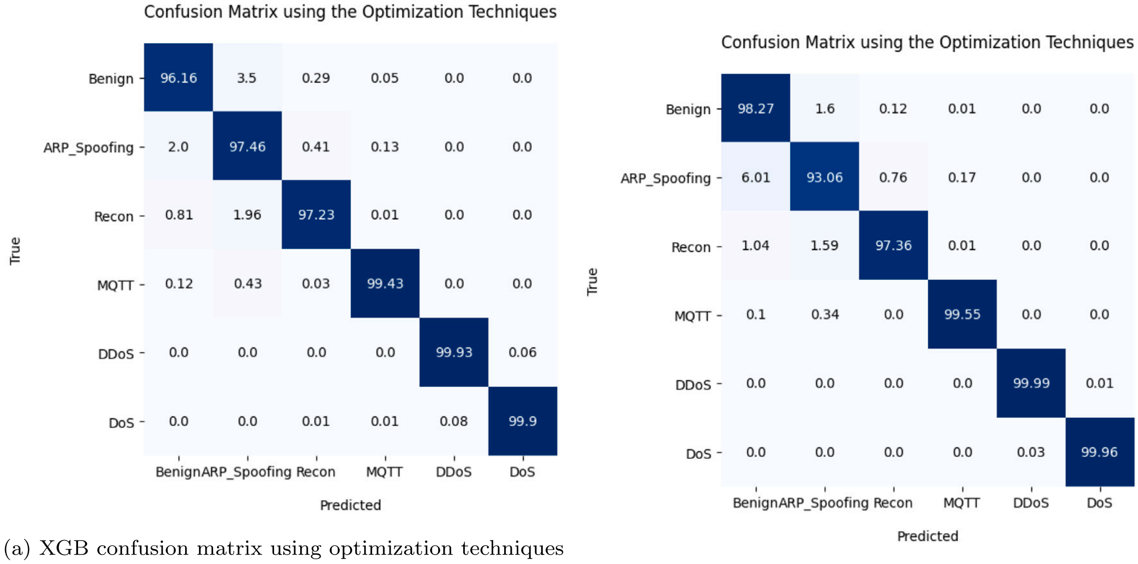
The application of the optimization techniques described earlier yielded significant improvements in the performance of IDSs for IoMT environments, as detailed in Tables 6 and 7. To address dataset imbalance, we employ and compare two different strategies: (1) An undersampling strategy by randomly reducing the number of instances in all classes—except for the minority class (ARP Spoofing)—to achieve uniform class distribution; (2) The SMOTE method for oversampling the minority classes (i.e., ARP Spoofing, Recon and Benign) until reaching the same amount of samples as the MQTT class. Moreover, we employ a train-validation-test split of 64%-16%–20% to support robust model optimization while maintaining a sufficiently large test set for unbiased performance evaluation and reducing the risk of underfitting. This configuration ensures that 80% of the data is allocated to model development (training and validation), with the remaining 20% reserved exclusively for testing. Finally, a shuffling strategy is applied to ensure that the order of data is randomized for each ML model, avoiding temporal correlations among samples.

For encoding categorical data, one-hot encoding was utilized for Decision Tree, Random Forest, and Feedforward Neural Network models, while label encoding was applied to the Extreme Gradient Boosting classifier. A min–max scaler was used for data normalization, ensuring consistency across features. The hyperparameters selected for each model are detailed in Table 3. Additionally, a k-fold cross-validation approach with $k = 5$ was adopted. This approach mitigates biases and ensures robust evaluation of the machine learning classifiers.

Among the validated models, tree-based classifiers demonstrated high-performance results. In fact, the XGB was the most effective ML classifier for attack detection in the CICIoMT2024 dataset when combined with the Random Undersampling strategy. It achieved remarkable performance results, with an accuracy of 0.9976, a precision of 0.9270, a recall of 0.9834, and an F1-score of 0.9470. Standard deviations were minimal, being ±0.0001, ±0.0013, ±0.0003, and ±0.0011, respectively, and, being ±0.0257, the maximum deviation observed. On the other hand, the RF model achieved outstanding results when combined with the SMOTE Oversampling strategy, leading to performance metrics of 0.9985 in accuracy, 0.9706 in precision, 0.9729 in recall, and 0.9716 in F1-score. The standard deviations were also minimal, being ±0.0000, ±0.0013, ±0.0008, and ±0.0005, respectively. The maximum deviation

Table 8
Comparison of our results in the test split with the best results on CICIoMT2024.

IDS solution	Accuracy	Precision	Recall	F1-score
Dadkhah et al. [15]	0.7350	0.7130	0.7350	0.6760
Proposed IDS	0.9985	0.9691	0.9712	0.9700



(a) XGB confusion matrix using optimization techniques and Random Undersampling

(b) RF confusion matrix using optimization techniques and SMOTE Oversampling.

Fig. 9. Comparison of the confusion matrices of XGB and RF models when using optimization techniques with Undersampling and Oversampling strategies, respectively.

observed among the ML models was ±0.0046. These results confirm consistent model behavior across splits and state that RF is the best model when combined with the SMOTE Oversampling strategy.

In contrast, the FNN model exhibited significantly lower performance, being the least effective model in the study. These findings suggest that this type of FNN model may not be well-suited for attack detection in scenarios like those represented by the CICIoMT2024 dataset, and more complex neural network architectures are needed.

Furthermore, Table 8 illustrates the performance of the RF in the test split, and how our proposed IDS demonstrates a major improvement in comparison with the previous approach proposed by the authors of the CICIoMT2024 dataset. The increase of 26.35% in accuracy and 29.40% in the F1-score emphasizes the importance of correctly applying preprocessing techniques in the dataset for improved realism and enhanced performance in IoMT environments.

The impact of these improvements is particularly evident in the confusion matrices presented in Figs. 9 and 10. Minority classes, such as ARP Spoofing, which previously exhibited low detection rates, show considerable improvements in classification accuracy. These improvements are directly related to the balancing strategy used, which ensured that minority classes were adequately represented during training. Nevertheless, the ARP Spoofing class detection performance decreased when applying the SMOTE Oversampling strategy (Fig. 9(b)), in comparison with the Random Undersampling strategy (Fig. 9(a)). This can be due to SMOTE inaccuracies in generating synthetic ARP Spoofing samples: If the ARP Spoofing class is close in feature space to another class, the model can misclassify the newly generated samples as another class.

These findings underscore the critical role of effective preprocessing and optimization in developing robust IDS for IoMT environments, not only improving overall performance metrics but also addressing challenges such as imbalanced datasets.

6. Conclusions and future work

The study presented a comprehensive analysis of AI-based IDSs in the context of IoT and IoMT environments, emphasizing the need for domain-specific optimization to enhance cybersecurity in healthcare settings. The comparison between the CICIoT2023 and CICIoMT2024 datasets revealed significant performance drops of up to 66.87% in the F1-score when models trained on one dataset were tested on the other. This underscores the inherent differences between IoT and IoMT environments, particularly in terms of attack patterns, traffic characteristics, and device-specific vulnerabilities. The findings validate the hypothesis that IoMT-specific datasets are essential for developing effective IDS tailored to healthcare scenarios. The study also demonstrated that the

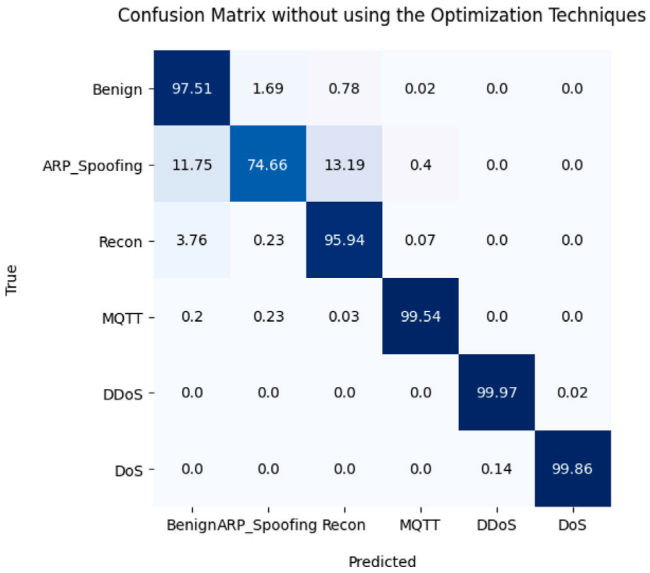


Fig. 10. XGB confusion matrix without using optimization techniques.

FNN exhibited robustness in generalization contexts, specifically in the 2-class scenario. Future validation of our approach using other IoMT-specific datasets, such as IoMT-TrafficData, would enhance the findings. Moreover, combining different IoMT datasets to create a unified database could improve model robustness and support broader generalization across diverse healthcare scenarios.

Additionally, the implementation of targeted optimization techniques for the CICIoMT2024 dataset demonstrated substantial improvements in IDS performance, achieving an accuracy of 99.85% and an F1-score of 97.00%, representing a 26.35% and 29.40% improvement, respectively, over the baseline methods proposed by the authors of the CICIoMT2024 dataset. The findings emphasize the critical need for rigorous data handling in AI-based IDS research to develop real and effective solutions in the field. Finally, this study provides a strong foundation for future research aimed at enhancing IDS for IoMT environments. Future work could explore more advanced neural network architectures, such as FNN with a higher number of layers, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Transformer-based models, and optimize their hyperparameters to improve their robustness and generalization capabilities across datasets and attack classes. Furthermore, considering the application of Federated Learning could provide a privacy-preserving approach to collaboratively train IDS models across different IoMT infrastructures, mitigating the limitations of centralized datasets. Tuning strategies, such as Grid Search or Bayesian Optimization, could also help identify optimal configurations in AI-based IDSs, while addressing class imbalance through oversampling techniques like Generative Adversarial Networks (GANs) could further enhance model performance. Moreover, incorporating feature selection methods, such as Principal Component Analysis (PCA) or Particle Swarm Optimization (PSO), can improve interpretability, computational efficiency, and detection accuracy. These advancements would contribute to developing more reliable and scalable IDSs solutions, ensuring the security and safety of IoMT-based healthcare systems.

CRediT authorship contribution statement

Jordi Doménech: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Olga León:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Conceptualization. **Muhammad Shuaib Siddiqui:** Writing – review & editing, Supervision, Funding acquisition. **Josep Pegueroles:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the predoctoral program AGAUR-FI, Spain ajuts Joan Oró with grant number 2024 FI-1 00643 and by the Chair of Cybersecurity called CARISMATICA, Spain. This funding is backed by the Secretariat of Universities and Research of the Department of Research and Universities of the Generalitat of Catalonia, Spain, as well as the European Social Plus Fund and the funds from the Recovery, Transformation, and Resilience, financed by the European Union (Next Generation), under the auspices of the INCIBE. Moreover, the authors thank Oriol López Petit for advice on experimental design and statistical analysis.

Appendix. Network attributes comparison

This appendix provides a clear comparison between the network attributes that can be found in the CICIoT2023 and CICIoMT2024 datasets. In fact, CICIoT2023 and CICIoMT2024 have in common a total of 44 network attributes. However, *tx*, *flow duration*, and *urg_count* attributes are part of the CICIoT2023 dataset but not from the CICIoMT2024 dataset. Moreover, the *IGMP* attribute is used in CICIoMT2024 but not in CICIoT2023, as described in Table A.9.

Table A.9
Attributes in CICIoT2023 and CICIoMT2024 datasets.

No. 2023	Feature	Description	No. 2024
1	ts	Timestamp	✗
2	flow duration	Duration of packet’s flow	✗
3	Header Length	Header Length	1
4	Protocol type	IP, UDP, TCP...(integer)	2
5	Duration	TTL	3
6	Rate	Rate of packet transmission in a flow	4
7	Srate	Rate of outbound packet transmission in a flow	5
8	Drate	Rate of inbound packet transmission in a flow	6
9	fin_flag	Packet has flag set to 1	7
10	syn_flag	Packet has flag set to 1	8
11	rst_flag	Packet has flag set to 1	9
12	psh_flag	Packet has flag set to 1	10
13	ack_flag	Packet has flag set to 1	11
14	ece_flag	Packet has flag set to 1	12
15	cwr_flag	Packet has flag set to 1	13
16	ack_count	Number of packets with ack flag set	14
17	syn_count	Number of packets with syn flag set	15
18	fin_count	Number of packets with fin flag set	16
19	urg_count	Number of packets with urg flag set	✗
20	rst_count	Number of packets with rst flag set	17
21	HTTP	Application protocol is http	18
22	HTTPS	Application protocol is https	19
23	DNS	Application protocol is dns	20
24	Telnet	Application protocol is telnet	21
25	SMTP	Application protocol is smtp	22
26	SSH	Application protocol is ssh	23
27	IRC	Application protocol is irc	24
28	TCP	Application protocol is tcp	25
29	UDP	Application protocol is udp	26
30	DHCP	Application protocol is dhcp	27
31	ARP	Link-layer protocol is arp	28
32	ICMP	Network-layer protocol is icmp	29
33	IP	Network-layer protocol is IP	31
34	LLC	Link-layer protocol is LLC	32
35	Tot sum	Sum of packet lengths within flow	33
36	Min	Min packet length in flow	34
37	Max	Max packet length in flow	35
38	AVG	Average packet length in flow	36
39	Std	Standard deviation of packet length in flow	37
40	Tot size	Packet’s length	38
41	IAT	Time difference with previous packet	39
42	Number	Number of packets in the flow	40
43	Magnitude	Average length	41
44	Radius	Variance of length	42
45	Covariance	Covariance of length	43
46	Variance	Ratio of the variance	44
47	Weight	Number of incoming packets/outcoming packets	45
✗	IGMP	Protocol is igmp	30

Data availability

Data will be made available on request.

References

[1] D. Dimitrov, Medical Internet of Things and big data in healthcare, Heal. Inform. Res. 22 (2016) 156, http://dx.doi.org/10.4258/hir.2016.22.3.156.

[2] European Union Agency for Cybersecurity, M. Theocharidou, I. Lella, ENISA Threat Landscape: Health Sector (January 2021 to March 2023), European Network and Information Security Agency, 2023, http://dx.doi.org/10.2824/163953.

[3] European Union Agency for Cybersecurity, R. Mattioli, A. Malatras, Foresight Cybersecurity Threats For 2030 - Update 2024: Extended report, European Network and Information Security Agency, 2024, http://dx.doi.org/10.2824/349493.

[4] A. Heidari, M.A.J. Jamali, Internet of Things intrusion detection systems: a comprehensive review and future directions, Clust. Comput. 26 (2023) 3753–3780, http://dx.doi.org/10.1007/s10586-022-03776-z.

[5] A.A. Khan, M. Faheem, R.N. Bashir, C. Wechtaisong, M.Z. Abbas, Internet of things (IoT) assisted context aware fertilizer recommendation, IEEE Access 10 (2022) 129505–129519.

[6] A.A. Khan, M. Driss, W. Boulila, G.A. Sampedro, S. Abbas, C. Wechtaisong, Privacy preserved and decentralized smartphone recommendation system, IEEE Trans. Consum. Electron. 70 (1) (2023) 4617–4624.

[7] I.H. Sarker, Machine learning: Algorithms, real-world applications and research directions, SN Comput. Sci. 2 (3) (2021) 160, http://dx.doi.org/10.1007/s42979-021-00592-x.

[8] J. Doménech, I.V. Martin-Faus, S. Mhiri, J. Pegueroles, Ensuring patient safety in IoMT: A systematic literature review of behavior-based intrusion detection systems, Internet Things 28 (2024) 101420, http://dx.doi.org/10.1016/j.iot.2024.101420, URL https://www.sciencedirect.com/science/article/pii/S2542660524003615.

[9] S.B. Weber, S. Stein, M. Pilgermann, T. Schrader, Attack detection for medical cyber-physical systems-A systematic literature review, IEEE Access 11 (2023) 41796–41815, http://dx.doi.org/10.1109/ACCESS.2023.3270225.

[10] G. Nagarajan, M. Margala, S. Shankar S, P. Chakrabarti, R. Minu, A trust-centric approach to intrusion detection in edge networks for medical internet of thing ecosystems, Comput. Electr. Eng. 115 (2024) 109129, http://dx.doi.org/10.1016/j.compeleceng.2024.109129, URL https://www.sciencedirect.com/science/article/pii/S0045790624000570.

[11] G. Zachos, I. Essop, G. Mantas, K. Porfyrakis, J.C. Ribeiro, J. Rodriguez, An anomaly-based intrusion detection system for internet of medical things networks, Electronics 10 (21) (2021) http://dx.doi.org/10.3390/electronics10212562, URL https://www.mdpi.com/2079-9292/10/21/2562.

[12] R.M.S. Priya, P.K.R. Maddikunta, M. Parimala, S. Koppu, T.R. Gadekallu, C.L. Chowdhary, M. Alazab, An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture, Comput. Commun. 160 (2020) 139–149, http://dx.doi.org/10.1016/j.comcom.2020.05.048.

[13] P. Kumar, G.P. Gupta, R. Tripathi, An ensemble learning and fog-cloud architecture-driven cyber-attack detection framework for IoMT networks, Comput. Commun. 166 (2021) 110–124, http://dx.doi.org/10.1016/j.comcom.2020.12.003, URL https://www.sciencedirect.com/science/article/pii/S0140366420320090.

[14] E.C.P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, A.A. Ghorbani, CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment, Sensors 23 (13) (2023) http://dx.doi.org/10.3390/s23135941, URL https://www.mdpi.com/1424-8220/23/13/5941.

[15] S. Dadkhah, E.C.P. Neto, R. Ferreira, R.C. Molokwu, S. Sadeghi, A.A. Ghorbani, CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT, Internet Things 28 (2024) 101351, http://dx.doi.org/10.1016/j.iot.2024.101351, URL https://www.sciencedirect.com/science/article/pii/S2542660524002920.

[16] C. Guida, A. Nascita, A. Montieri, A. Pescapé, Cross-evaluation of deep learning-based network intrusion detection systems, in: 2023 10th International Conference on Future Internet of Things and Cloud, FiCloud, 2023, pp. 328–335, http://dx.doi.org/10.1109/FiCloud58648.2023.00055.

[17] G. Apruzzese, L. Pajola, M. Conti, The cross-evaluation of machine learning-based network intrusion detection systems, IEEE Trans. Netw. Serv. Manag. 19 (4) (2022) 5152–5169, http://dx.doi.org/10.1109/tnsm.2022.3157344.

[18] A. Alarcón, O. León, J. Pegueroles, J. Doménech, Técnicas de Machine Learning para la detección de ciberataques en redes IoT médicas, in: XVIII Reunión Española sobre Criptología y Seguridad de la Información: XVIII RECSI, León 23-25 octubre 2024, Universidad de León, Servicio de Publicaciones, 2025, pp. 25–30.

[19] N. Thereza, K. Ramli, Development of intrusion detection models for IoT networks utilizing CICIoT2023 dataset, in: 2023 3rd International Conference on Smart Cities, Automation & Intelligent Computing Systems, ICON-SONICS, 2023, pp. 66–72, http://dx.doi.org/10.1109/ICON-SONICS59898.2023.10435006.

[20] A.G. Kumar, A. Rastogi, V. Ranga, Evaluation of different machine learning classifiers on new IoT dataset CICIoT2023, in: 2024 International Conference on Intelligent Systems for Cybersecurity, ISCS, 2024, pp. 1–6, http://dx.doi.org/10.1109/ISCS61804.2024.10581375.

[21] S.S. Shafin, G. Karmakar, I. Mareels, Obfuscated memory malware detection in resource-constrained IoT devices for smart city applications, Sensors 23 (11) (2023) http://dx.doi.org/10.3390/s23115348, URL https://www.mdpi.com/1424-8220/23/11/5348.

[22] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence, Dataset Shift in Machine Learning, MIT Press, 2009, pp. 1–38.

[23] S. Garcia, A. Parmisano, M.J. Erquiaga, IoT-23: A Labeled Dataset with Malicious and Benign IoT Network Traffic, Zenodo, 2020, http://dx.doi.org/10.5281/zenodo.4743746.

[24] I. Sharafaldin, A.H. Lashkari, A.A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, in: International Conference on Information Systems Security and Privacy, 2018, URL https://api.semanticscholar.org/CorpusID:4707749.

[25] Y. Mirsky, T. Doitshman, Y. Elovici, A. Shabtai, Kitsune: An ensemble of autoencoders for online network intrusion detection, 2018, arXiv:1802.09089, URL https://arxiv.org/abs/1802.09089.

[26] M. Cantone, C. Marrocco, A. Bria, Machine learning in network intrusion detection: A cross-dataset generalization study, IEEE Access 12 (2024) 144489–144508, http://dx.doi.org/10.1109/ACCESS.2024.3472907.

[27] D. Li, Q. Yuan, T. Li, S. Chen, J. Yang, Cross-domain network traffic classification using unsupervised domain adaptation, in: 2020 International Conference on Information Networking, ICOIN, 2020, pp. 245–250, http://dx.doi.org/10.1109/ICOIN48656.2020.9016470.

[28] L.F. Maimó, A.H. Celdrán, Á.L.P. Gómez, F.J.G. Clemente, J. Weimer, I. Lee, Intelligent and dynamic ransomware spread detection and mitigation in integrated clinical environments, Sensors (Switzerland) 19 (2019) http://dx.doi.org/10.3390/s19051114.

[29] A.A. Hady, A. Ghubaish, T. Salman, D. Unal, R. Jain, Intrusion detection system for healthcare systems using medical and network data: A comparison study, IEEE Access 8 (2020) 106576–106584, http://dx.doi.org/10.1109/ACCESS.2020.3000421.

[30] S. Nandy, M. Adhikari, M.A. Khan, V.G. Menon, S. Verma, An intrusion detection mechanism for secured IoMT framework based on swarm-neural network, IEEE J. Biomed. Heal. Inform. 26 (2022) 1969–1976, http://dx.doi.org/10.1109/JBHI.2021.3101686.

[31] N. Moustafa, A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets, Sustain. Cities Soc. 72 (2021) 102994, http://dx.doi.org/10.1016/j.scs.2021.102994, URL https://www.sciencedirect.com/science/article/pii/S2210670721002808.

[32] A. Binbusayyis, H. Alaskar, T. Vaiyapuri, M. Dinesh, An investigation and comparison of machine learning approaches for intrusion detection in IoMT network, J. Supercomput. 78 (2022) 17403–17422, http://dx.doi.org/10.1007/s11227-022-04568-3.

[33] M. Alalhareth, S.C. Hong, An improved mutual information feature selection technique for intrusion detection systems in the internet of medical things, Sensors 23 (2023) http://dx.doi.org/10.3390/s23104971.

[34] P. Mooijman, C. Catal, B. Tekinerdogan, A. Lommen, M. Blokland, The effects of data balancing approaches: A case study, Appl. Soft Comput. 132 (2023) 109853, http://dx.doi.org/10.1016/j.asoc.2022.109853, URL https://www.sciencedirect.com/science/article/pii/S1568494622009024.

[35] S. Susan, A. Kumar, The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent state of the art, Eng. Rep. 3 (4) (2021) e12298, http://dx.doi.org/10.1002/eng2.12298, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/eng2.12298, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.12298.

[36] S.-a. Ayoub, A.-G. Mohammed Ali, B. Narhimene, Enhanced intrusion detection system for remote healthcare, in: M.R. Senouci, S.Y. Boulahia, M.A. Benatia (Eds.), Advances in Computing Systems and Applications, Springer International Publishing, Cham, 2022, pp. 323–333.

[37] L. Yang, A. Shami, On hyperparameter optimization of machine learning algorithms: Theory and practice, Neurocomputing 415 (2020) 295–316, http://dx.doi.org/10.1016/j.neucom.2020.07.061, URL https://www.sciencedirect.com/science/article/pii/S0925231220311693.

[38] M. Masum, H. Shahriar, H. Haddad, M.J.H. Faruk, M. Valero, M.A. Khan, M.A. Rahman, M.I. Adnan, A. Cuzzocrea, F. Wu, Bayesian hyperparameter optimization for deep neural network-based network intrusion detection, in: 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 5413–5419, http://dx.doi.org/10.1109/BigData52589.2021.9671576.

[39] M. Lucia Hernandez-Jaimes, A. Martínez-Cruz, K. Alejandra Ramírez-Gutiérrez, E. Guevara-Martínez, Enhancing machine learning approach based on nilsimsa fingerprinting for ransomware detection in IoMT, IEEE Access 12 (2024) 153886–153897, http://dx.doi.org/10.1109/ACCESS.2024.3480889.

[40] A. Mohamadi, H. Ghahramani, S.A. Asghari, M. Aminian, Securing healthcare with deep learning: A CNN-based model for medical IoT threat detection, 2024, arXiv:2410.23306, URL https://arxiv.org/abs/2410.23306.

[41] J. Areia, I.A. Bispo, L. Santos, R.L.d.C. Costa, IoMT-TrafficData: Dataset and tools for benchmarking intrusion detection in internet of medical things, IEEE Access 12 (2024) 115370–115385, http://dx.doi.org/10.1109/ACCESS.2024.3437214.

[42] S. Ghazanfar, F. Hussain, A.U. Rehman, U.U. Fayyaz, F. Shahzad, G.A. Shah, IoT-flock: An open-source framework for IoT traffic generation, in: 2020 International Conference on Emerging Trends in Smart Technologies, ICETST, 2020, pp. 1–6, http://dx.doi.org/10.1109/ICETST49965.2020.9080732.

[43] D. Song, DPKT documentation, 2023, https://dpkt.readthedocs.io/en/latest/. (Accessed 22 June 2024).

[44] K. Gupta, D.K. Sharma, K. Datta Gupta, A. Kumar, A tree classifier based network intrusion detection model for internet of medical things, Comput. Electr. Eng. 102 (2022) 108158, http://dx.doi.org/10.1016/j.compeleceng.2022.108158, URL https://www.sciencedirect.com/science/article/pii/S0045790622004049.

[45] K.P. Vijayakumar, K. Pradeep, A. Balasundaram, M.R. Prusty, Enhanced cyber attack detection process for internet of health things (IoHT) devices using deep neural network, Processes 11 (4) (2023) http://dx.doi.org/10.3390/pr11041072, URL https://www.mdpi.com/2227-9717/11/4/1072.

[46] N. Elmrabit, F. Zhou, F. Li, H. Zhou, Evaluation of Machine Learning Algorithms for Anomaly Detection, in: 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), 2020, pp. 1–8, http://dx.doi.org/10.1109/CyberSecurity49315.2020.9138871.

[47] S. Haque, F. El-Moussa, N. Komninos, R. Muttukrishnan, A systematic review of data-driven attack detection trends in IoT, Sensors 23 (16) (2023) http://dx.doi.org/10.3390/s23167191, URL https://www.mdpi.com/1424-8220/23/16/7191.

[48] M.M. Khan, M. Alkhathami, Anomaly detection in IoT-based healthcare: machine learning for enhanced security, Sci. Rep. 14 (1) (2024) 5872, http://dx.doi.org/10.1038/s41598-024-56126-x.

[49] M. Moure-Garrido, C. Garcia-Rubio, C. Campo, Reducing DNS traffic to enhance home IoT device privacy, Sensors 24 (9) (2024) http://dx.doi.org/10.3390/s24092690, URL https://www.mdpi.com/1424-8220/24/9/2690.

[50] L. You, Construction of early warning mechanism of university education network based on the Markov model, Mob. Inf. Syst. 2022 (1) (2022) 7302623, http://dx.doi.org/10.1155/2022/7302623, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/7302623, URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/7302623.

[51] J. Khan, D.-W. Lim, Y.-S. Kim, Intrusion detection system CAN-bus in-vehicle networks based on the statistical characteristics of attacks, Sensors 23 (7) (2023) http://dx.doi.org/10.3390/s23073554, URL https://www.mdpi.com/1424-8220/23/7/3554.

[52] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, pp. 96–161, http://www.deeplearningbook.org.

[53] M.A. Lones, Avoiding common machine learning pitfalls, Patterns 5 (10) (2024) 101046, http://dx.doi.org/10.1016/j.patter.2024.101046.

[54] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.

[55] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, G. Bing, Learning from class-imbalanced data: Review of methods and applications, Expert Syst. Appl. 73 (2017) 220–239, http://dx.doi.org/10.1016/j.eswa.2016.12.035, URL https://www.sciencedirect.com/science/article/pii/S0957417416307175.