scientific reports

Check for updates

OPEN

XAI-XGBoost: an innovative explainable intrusion detection approach for securing internet of medical things systems

Yousif Hosain[1]✉ & Muhammet Çakmak[2]

The Internet of Medical Things (IoMT) has transformed healthcare delivery but faces critical challenges, including cybersecurity threats that endanger patient safety and data integrity. Intrusion Detection Systems (IDS) are essential for protecting IoMT networks, yet conventional models often struggle with class imbalance, lack interpretability, and are unsuitable for real-world deployment in sensitive healthcare settings. This study aims to develop an innovative, explainable IDS framework tailored for IoMT systems that ensures both high detection accuracy and model transparency. The proposed approach integrates a hybrid random sampling technique to mitigate class imbalance, Recursive Feature Elimination (RFE) for feature selection, and an optimized XGBoost classifier for robust attack detection. Explainable AI techniques, namely SHAP and LIME, are employed to provide global and local insights into model predictions, enhancing interpretability and trustworthiness. The system was evaluated using the WUSTL-EHMS-2020 dataset, which contains network flow and biometric data, achieving outstanding performance: 99.22% accuracy, 98.35% precision, 99.91% recall, 99.12% F1-score, and 100% ROC-AUC. The proposed framework outperforms several traditional Machine Learning (ML) models and state-of-the-art IDS approaches, demonstrating its robustness and suitability for practical healthcare environments. By integrating advanced ML with explainable AI, this work addresses the critical need for secure, interpretable, and high-performing IDS solutions in IoMT systems. The study concludes that explainability is not an optional feature but a fundamental requirement in healthcare cybersecurity, and the proposed framework represents a significant step towards safer and more accountable AI-driven security solutions for the IoMT ecosystem.

Keywords Internet of medical things (IoMT), Intrusion detection system (IDS), Explainable artificial intelligence (XAI), XGBoost classifier, Feature selection

Recently, the rapid advancement of internet technologies and computational paradigms has resulted in the emergence of the Internet of Things (IoT), an innovative technology that enables communication and information sharing among multiple objects or "things" over the internet. IoT ecosystem is crucial in creating smart environments, such as smart homes, smart cities, smart education, smart transportation, and smart healthcare systems[1,2]. The integration of the IoT within medical devices has led to the evolution of the Internet of Medical Things (IoMT)[3].

IoMT is a network of internet-connected medical devices that allow the exchange of data and information[4]. This network consists of various devices, including remote monitoring systems, wearable technology, and at-home diagnostic tools that collect and relay patient data to healthcare specialists[5]. The IoMT can revolutionize the healthcare sector by enhancing patient care, reducing healthcare expenses, and augmenting efficiency. IoMT devices enable real-time health monitoring for the patients, allowing healthcare specialists to respond promptly and prevent the progression of serious health conditions[6]. Furthermore, IoMT devices can alleviate the strain on healthcare systems by enabling in-home medical care, reducing hospital visits, and improving access to healthcare services[7].

Although the IoMT has offered several benefits, it also poses significant privacy and security concerns[8]. In contrast to other IoT domains, cybersecurity in IoMT is especially important because it directly affects human health and safety. A cyberattack on an IoMT device can cause serious problems, such as disabling medical alarms,

[1]Department of Computer Engineering, Karabuk University, Karabuk 78050, Turkey. [2]Faculty of Engineering and Architecture, Sinop University, Sinop, Turkey. ✉email: 2138166007@ogrenci.karabuk.edu.tr

delivering the wrong dose of medication, or showing false vital sign readings. These issues can delay treatment, mislead clinical decisions, or in severe cases, endanger the patient's life. Consequently, securing IoMT systems is not only about protecting data, but also about protecting human lives and ensuring safe medical care. Scholars have identified various attacks on IoMT systems, such as message alteration, eavesdropping, Man-in-the-Middle (MiTM) attacks, Denial of Service (DoS), and Distributed Denial of Service (DDoS) attacks, all of which can jeopardize patient security, compromise privacy, and disrupt the availability of critical healthcare systems[9].

Because IoMT systems are vulnerable, creating an effective Intrusion Detection System (IDS) is essential for keeping this network safe. IDS is a security mechanism specifically designed to detect unauthorized activities or breaches within a network, providing robust protection for IoMT infrastructure against a wide range of cyberattacks[10–12].

Nonetheless, the development of an effective IDS is accompanied by numerous complexities and challenges. Various aspects, such as data heterogeneity, the intricacy nature of IoMT systems, the vast volume of network traffic, and the constantly evolving nature of cyberattacks pose serious concerns[13]. Furthermore, traditional IDS methods have been shown to be insufficient in addressing the unique requirements and complexities of IoMT environments, highlighting the need for more innovative and robust solutions tailored to these systems[14].

Recently, Artificial intelligence (AI) techniques, particularly Machine Learning (ML) algorithms, have gained significant attention for enhancing the detection accuracy of IDS[15]. This is attributed to their capability to detect intricate and sophisticated cyberattacks. ML algorithms can leverage pre-labeled datasets to learn patterns, enabling them to detect attacks without the need for explicit programming instructions[16].

However, a major challenge in implementing ML algorithms, particularly in security-critical applications such as IDS for IoMT systems, lies in these models' opaque, black-box nature, which limits the transparency and interpretability of their decision-making processes. This ambiguity raises critical concerns about how ML-based IDS models make decisions, posing a significant drawback in such systems where comprehending the reasoning behind decisions is substantial for maintaining trust and ensuring accountability[17,18].

Explainable AI (XAI) strives to address this gap by enhancing transparency and offering valuable explanations of the decision-making processes within AI models[19]. Methods such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP) are among XAI's most commonly utilized techniques. The LIME technique simplifies a complex model by creating a smaller, interpretable surrogate model for each prediction, therefore offering local explanations. While the SHAP technique provides a consistent approach to measuring feature importance across various models, offering a comprehensive view on the behavior of the model. These two techniques serve to clarify the "black-box" nature of AI models, providing detailed feature-level insights that improve transparency and accountability in IDS models[20,21].

Despite the increasing adoption of IoMT technologies, securing such systems presents several critical challenges, including the heterogeneity of data sources (e.g., network traffic and biometric signals), the class imbalance in datasets, the resource constraints of IoMT devices, and the evolving and sophisticated nature of cyber threats. Furthermore, traditional IDS models often operate as black-box systems, lacking transparency and interpretability, which is a major concern in safety-critical domains such as healthcare. These limitations are particularly concerning in clinical environments, where both detection accuracy and decision transparency are vital for patient safety.

To address these challenges, this study proposes an intelligent and explainable IDS specifically designed for IoMT systems. Our approach utilizes the WUSTL-EHMS-2020 IoMT dataset, which comprises a combination set of network flow metrics and patient biometrics collected from a real-time Enhanced Healthcare Management System (EHMS) testbed. Our proposed methodology incorporates advanced preprocessing techniques and a hybrid Random sampling technique to address class imbalance, a common issue that can significantly affect models' performance. Feature selection is subsequently performed using Recursive Feature Elimination (RFE) to identify and retain the most relevant features from the dataset. An optimized XGBoost classifier is then used for accurate and efficient classification. We also use XAI techniques such as LIME and SHAP to provide detailed insights into the model's decision-making processes, enhancing transparency and trustworthiness while addressing the unique security challenges of IoMT systems.

The main contributions of this work are as follows:

- We develop a novel explainable intrusion detection framework specifically tailored for IoMT environments, integrating an optimized XGBoost classifier with SHAP and LIME to provide both high detection performance and model interpretability.
- A hybrid random sampling approach is proposed to effectively address the class imbalance issue by combining oversampling of minority instances with undersampling of majority instances, thereby improving learning balance and model fairness.
- The proposed model is thoroughly evaluated on the WUSTL-EHMS-2020 dataset, achieving outstanding performance with 99.22% accuracy, 98.35% precision, 99.91% recall, and a 99.12 F1-score. These results represent a significant improvement over several recent state-of-the-art IDS models developed for IoMT security, confirming the robustness and reliability of the proposed approach.

The rest of this paper is organized as follows: Sect. 2 discusses the existing works for IDS in IoMT systems. Section 3 provides a comprehensive overview of all ML algorithms and explainable AI techniques used in this study. Section 4 presents the proposed methodology. Section 5 presents and interprets the proposed IDS model's evaluation results. Finally, Sect. 6 concludes the study, summarizing the key contributions and suggesting future research directions.

Related work

The rapid growth of the IoMT has highlighted the urgent need for extensive research into effective privacy and security measures. ML and DL have emerged as essential methods for detecting cyberattacks within the IoMT system.

Swarna Priya et al.[14] introduced a hybrid IDS model to detect cyberattacks targeting IoMT systems. The suggested model integrates a hybrid Principal Component Analysis (PCA) method optimized by the Gray Wolf Optimization (GWO) algorithm for feature selection and employs the Deep Neural Network (DNN) algorithm for classification. The NSL-KDD dataset was used to test the suggested approach, and the evaluation results demonstrated that feature selection based on optimization techniques significantly improved the model's intrusion detection performance.

Larzek et al.[22] suggested an advanced IDS framework to address the growing cybersecurity challenges in IoMT systems. The proposed framework integrates the RFE technique with ML algorithms and combines Ridge regression with DL algorithms for feature selection. The authors used the WUSTL-EHMS-2020 dataset to validate the framework model, demonstrating high performance. Notably, the RFE-based Decision Tree (DT) model achieved outstanding results, with an accuracy of 97.85% and a remarkable False Alarm Rate (FAR) of 0.03.

Saheed et al.[23] proposed a robust hybrid intrusion detection model tailored for IoMT environments, employing Particle Swarm Optimization (PSO) for feature selection with a Deep Recurrent Neural Network (DRNN) and multiple ML algorithms, including Random Forest (RF), DT, K-Nearest Neighbor (KNN), and Ridge Classifier (RC), for classification. The NSL-KDD dataset was employed for experimental validation. Experimental results demonstrated that the combination of PSO and DRNN significantly enhanced detection capabilities, achieving a remarkable accuracy of 99.76%, demonstrating its strong potential for securing IoMT environments against diverse cyberattacks.

Similarly, Chaganti et al.[24] suggested a hybrid IDS model to secure IoMT systems against cyberattacks. The authors used PSO algorithm for feature selection and utilized the DNN algorithm for classification. The WUSTL-EHMS-2020 dataset was used to test the developed approach. The evaluation results show that combining PSO with DNN significantly enhances detection capabilities, achieving an impressive accuracy of 96%, showcasing its potential to strengthen the cybersecurity of healthcare systems.

Kulshrestha et al.[25] introduced an IDS model to detect cyberattacks targeting IoMT systems. The authors used the extra tree algorithm for feature selection and various ML algorithms, including Logistic Regression (LR), DT, Naïve Bayes (NB), bagging, RF, Adaboost, gradient boosting, XGBoost, and ensemble voting classifiers for classification. The ToN-IoT dataset was used to validate the suggested approach. Evaluation results confirm that the Adaboost classifier achieved the highest results compared to other classification algorithms.

Thamilarasu et al.[26] suggested a novel approach to securing IoMT devices from several threats using mobile agent-based IDS. The authors used the PCA technique for feature selection and various ML algorithms, including NB, KNN, Support Vector Machine (SVM), DT, and RF, to identify the intrusions. The authors utilized simulation-generated datasets based on a hospital network topology to validate the developed framework. The results indicate that the developed approach achieved high detection accuracy while maintaining minimal resource overhead.

Hady et al.[27] introduced an IDS approach to safeguard IoMT systems against cyberattacks. They created a dedicated healthcare testbed to capture network traffic and biometrics data, which resulted in creating the WUSTL-EHMS-2020 dataset. Evaluation results demonstrated a 7–25% improvement in attack detection performance, confirming the robustness of the developed model.

Gupta et al.[28] suggested an intrusion detection approach using a tree-based classifier to enhance the intrusion detection capabilities of IoMT systems. The suggested model leverages data augmentation techniques to balance the input dataset and employs dimensionality reduction to minimize dataset size, significantly enhancing the effectiveness and speed of the intrusion detection process. The WUSTL-EHMS-2020 dataset was employed to validate the developed framework. Evaluation results showed that the developed approach achieved an accuracy of 94.23% and an F1-score of 93.8%.

Nandy et al.[29] suggested a novel IDS-based Swarm-Neural Network approach for detecting cyberattacks inside edge-centric IoMT systems. The developed model is tailored to detect assaults during data transmission and effectively manage health data at the network edge, enhancing security and performance in healthcare systems. The authors used ToN-IoT dataset to validate the suggested model. Evaluation results demonstrated superior performance compared to traditional classification models, with an impressive detection accuracy of 99.5%.

Kumaar et al.[30] suggested a hybrid IDS model to safeguard IoMT systems against cyberattacks. The authors employed different ML and DL algorithms, including LR, DT, RF, XGBoost, and ImmuneNet. Three datasets, such as CIC-IDS 2017, CIC-IDS 2018, and Bell DNS 2021, were used to test the developed approach. Evaluation results show that ImmuneNet outperforms all other algorithms in detecting cyberattacks.

Kilincer et al.[31] introduced a novel IDS model using DL for automatically identifying cyberattacks in IoMT systems. The authors used the RFE technique for feature selection and the Multi-layer Perceptron (MLP) algorithm for classification. To further enhance performance, the hyperparameters of the MLP algorithm were fine-tuned using the grid search technique. The developed IDS was evaluated on different datasets, including ECU-IoHT, WUSTL-EHMS-2020, ICU, and ToN-IoT. The evaluation results showed that the combination of RFE and MLP significantly improved detection accuracy, achieving superior results compared to other intrusion detection methods.

Zachos et al.[32] introduced a robust IDS to improve the protection of IoMT systems. The proposed model leverages both host-based and network-based techniques to analyze log files and traffic data collected from IoMT devices and gateways, with a focus on minimizing computational costs. Various ML algorithms, including

LR, SVM, KNN, DT, and RF, were employed to detect intrusions in the IoMT system. The evaluation results showed that the KNN algorithm achieved the highest accuracy in detecting threats.

Ibrahim et al.[33] proposed HealthGuard, an IDS model using ensemble learning techniques to detect cyberattacks targeting IoMT systems. The proposed approach merges LR with KNN algorithms to enhance detection accuracy and robustness. The WUSTL-EHMS-2020 was used to validate the developed model. The developed model showed a detection accuracy of 92.5% and F1-score of 60.68%.

Ravi et al.[34] developed a hybrid IDS model for detecting cyberattacks in IoMT systems by combining Convolution Neural Network (CNN) and Long Short Term Memory (LSTM) algorithms. A global attention layer was employed to extract the most relevant features, while a cost-sensitive learning technique was utilized to address data imbalance issues. The experimental results demonstrated that the suggested framework achieved high detection accuracy and outperformed existing models.

Alhareth et al.[35] introduced a novel IDS model for identifying cyberattacks in IoMT systems. The authors employed the Logistic Redundancy Coefficient Gradual Upweighting MIFS (LRGU-MIFS) technique for feature selection and several ML and DL algorithms, including RF, LR, DT, SVM, and LSTM, for classification. The WUSTL-EHMS-2020 was used to test the developed model. The developed model demonstrated high performance, underscoring its effectiveness in detecting cyberattacks.

Manimurugan et al.[36] introduced an IDS model using a Deep Belief Network (DBN) to identify cyberattacks in healthcare systems. They employed the CICIDS 2017 dataset to test the developed approach, and the experimental results demonstrated its ability to accurately classify and detect a wide range of cyberattacks.

Alzubi et al.[37] proposed a robust intrusion detection framework tailored for the consumable edge-centric IoMT environment, aiming to enhance the real-time detection of cyber threats at the edge of healthcare networks. The framework integrates the strengths of CNN and LSTM architectures to form a blended DL model capable of accurately identifying both known and emerging intrusions during data transmission. The CSE-CIC-IDS 2018 dataset was used to evaluate the performance of the proposed framework. Experimental analysis revealed that the proposed hybrid model achieved a detection accuracy of 98.53%, significantly outperforming existing state-of-the-art techniques.

Bouke et al.[38] introduced a DL-based IDS model to address cybersecurity and data privacy challenges in the IoMT system. The authors employed an information gain technique for feature selection and a deep sequential algorithm for classification. The authors applied the SMOTE-ENN technique to address the class imbalance problem in the WUSTL-EHMS-2020 dataset. The developed model achieved superior performance compared to traditional methods, with a 99% detection accuracy and excellent precision, recall, F1 score, and ROC AUC, ensuring its reliability and effectiveness in securing IoMT systems.

Anjelin et al.[39] The authors suggested a hybrid IDS to identify cyberattacks targeting the IoMT system. They used the elephant herding optimization algorithm for feature selection and various ML models, including NB, SVM, KNN, DT, and RF, for classification. The NSL-KDD dataset was utilized to test the detection approach. The evaluation results demonstrated that the developed model achieved a high detection accuracy of 98.6%.

In summary, while existing IDS approaches for IoMT systems have shown promising results in terms of detection accuracy, several critical technical gaps remain unaddressed. Many prior works rely on ML and DL models, which, despite their effectiveness, often operate as black-box systems that lack transparency and explainability, an essential requirement in safety-critical healthcare environments. Additionally, most studies have not sufficiently tackled the issue of class imbalance in IoMT datasets, which can lead to biased models that fail to generalize effectively in real-world settings. The majority of existing IDS frameworks also focus primarily on accuracy metrics without considering interpretability, trust, and the practical deployment challenges of resource-constrained IoMT devices. Furthermore, some works lack validation on real IoMT traffic, relying instead on synthetic or simulated datasets that may not reflect the complexity of actual healthcare environments. There is also a limited exploration of combining robust classifiers with interpretable AI techniques to create explainable IDS frameworks that balance detection performance and transparency. These gaps highlight the need for a novel, explainable IDS approach that not only achieves high detection accuracy but also provides feature-level explanations to support clinical decision-making and enhance trust in AI-driven security systems. Motivated by these limitations, our study proposes an innovative IDS framework that integrates a hybrid random sampling technique for class imbalance, Recursive Feature Elimination for feature selection, an optimized XGBoost classifier for accurate detection, and explainable AI methods (SHAP and LIME) for model interpretability, specifically tailored for IoMT environments.

Background

This section provides a comprehensive overview of the ML algorithms and XAI techniques used in this paper:

Machine learning algorithms

Machine learning (ML) algorithms play a crucial role in intrusion detection by identifying patterns in network traffic and distinguishing between normal and malicious activities. This study explores a range of ML classifiers to establish a robust foundation for the proposed IDS model. The algorithms employed include both traditional and ensemble-based techniques, each offering unique advantages in handling classification tasks.

Logistic regression (LR) is a fundamental classification algorithm that estimates the probability of a given instance belonging to a particular class based on input features. It employs the logistic function, also known as the sigmoid function, to map real-valued inputs into a probability range between 0 and 1, as defined in Eq. (1). LR can be classified into two types: binary LR, employed when the target variable has two classes, and multinomial LR, applied when the dependent variable has more than two classes[22].

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

Support vector machine (SVM) is a robust classification technique that seeks to find the optimal hyperplane that separates different classes in a high-dimensional feature space. The main goal of SVM algorithms is to maximize the margin between classes, ensuring robustness against outliers and effectively handling complex decision boundaries[40].

K-nearest neighbour (KNN) is a simple and intuitive ML algorithm utilized for classification and regression tasks. The classification process involves two primary steps[25]. Firstly, the algorithm determines the k nearest neighbors of a new, unknown data point by calculating a distance metric or a similarity measure between the new point and the training dataset. Secondly, the unknown data point is classified according to a majority class among these k neighbors. The Euclidean distance metric is widely employed to compute the distance between data points, which is mathematically defined as follows:

$$d(a, b) = \sqrt{\sum_{i=1}^{k} (ai - bi)^2} \tag{2}$$

Where d is the distance, a and b are two data points with k features.

Decision tree (DT) is a popular non-parametric algorithm that solves both classification and regression problems. It consists of three primary components: root nodes, branch nodes, and leaf nodes. The internal node represents classification rules, the branch denotes the outcomes of these decisions, and the leaves correspond to the model's final predictions. Combining these components forms a hierarchical, tree-like structure. The construction process of DT commences by computing Information Gain (IG) scores for each attribute within the feature set. IG measures the decrease in impurity or uncertainty achieved by splitting the data on a specific attribute and is computed as follows:

$$IG(T, A) = H(T) - \sum_{v \in V} \left| \frac{Tv}{T} \right| H(Tv) \tag{3}$$

Here, H(T) represents the entropy of the dataset T, which quantifies the impurity or uncertainty in the data. The entropy is calculated as:

$$H(T) = -\sum_{i=1}^{n} pi \log_2(pi) \tag{4}$$

Where pi represents the ratio of instances belonging to class i, and n denotes the total number of classes. The attribute with the highest information gain is selected as the decision node, forming the root of the tree. This process is repeated iteratively by generating a sub-tree under the decision node until specific stopping conditions are satisfied, such as when all elements in a subgroup have the same value or when no distinguishable attribute remains. This iterative partitioning enables DT to effectively capture complex decision boundaries, making them highly capable of modeling complex, non-linear data relationships[41].

Random forest (RF) is an ensemble learning algorithm that aggregates the output of several DTs to produce robust and accurate predictions. It significantly enhances the performance and stability of the model compared to using a single DT. During the prediction process, each DT in the forest produces an individual prediction for a given sample. The final output is determined through an aggregation mechanism, such as a majority vote for classification scenarios[42].

Adaptive boosting (Adaboost) is a prominent boosting technique utilized to enhance the performance of weak classifiers. It achieves this by combining several weak classifiers to construct a single robust classifier. When a single classifier inaccurately classifies the input, the misclassified samples are forwarded to subsequent classifiers, therefore enhancing the total performance. Various types of classifiers can be employed, such as DT, LR, RF, etc. This algorithm is characterized by its ability to avoid overfitting, handle noisy data, and achieve superior performance across various classification tasks[24].

Gradient boosting is a strong ensemble learning classifier that integrates several weak learners, typically DTs, to build a powerful and precise classifier. This algorithm works by iteratively adding DTs, each attempting to correct its predecessor's residual errors. By combining the outputs from all the DTs, the final prediction is obtained[25].

Extreme gradient boosting (XGBoost) is a powerful and efficient ensemble learning algorithm utilized for solving classification and regression problems. XGBoost builds a set of DTs, where each new DT focuses on correcting the residual errors of the previous ones. The output from all DTs is aggregated to generate the final prediction. The core of XGBoost's efficiency lies in its regularized objective function, which achieves an optimal balance between predictive accuracy and the model's complexity:

$$L = \min_{\theta} \sum_{i=1}^{n} L(yi, \hat{y}i) + \sum_{k=1}^{k} \Omega(f_k) \tag{5}$$

where L is the total objective function, Loss (yi, ŷi) computes the error between the actual and the predicted values, while $\Omega(f_k)$ denotes the regularization of the kth model, which reduces overfitting by Mitigating model complexity[43]. XGBoost provides several advantages over traditional ML algorithms, including fast training times, superior prediction accuracy, and efficient handling of large and complex datasets. Its scalability, robustness

against noisy or missing data, and efficient handling of model complexity make it particularly well-suited for complex applications, such as IDS[44].

Explainable AI techniques: SHAP and LIME

ML-based IDS have proven highly effective in identifying anomalies and safeguarding systems from different types of cyberattacks. However, these models are commonly regarded as black boxes, raising concerns about their lack of trust and transparency. To address this issue, XAI provides methods to explain how complex ML models make decisions. By integrating XAI with IDS, administrators and security specialists can better understand the outputs generated by ML-based IDS, thereby enhancing trust and interpretability in cybersecurity applications. Our study employed XAI-based techniques, namely SHAP and LIME, to interpret and explain how the developed model makes decisions to protect IoMT systems.

SHAP

SHAP is a game theory-based technique aimed at interpreting the results produced by ML and DL models, explaining how each feature in the dataset affects the model's predictions[45]. It calculates the Shapley value, which represents the average contribution of each feature to the prediction by evaluating all possible feature combinations. The Shapley value is mathematically represented as:

$$\varphi i \; = \; \sum\nolimits_{S \subseteq N\{i\}} \frac{|S|!\,(\mathrm{X}\,|N| - |S| - 1)!}{|N|!}\,(v\,(S \cup \{i\}) - v\,(S)) \tag{6}$$

Where N represents the collection of all features, and v(S) is the value function that provides the prediction for a given subset of features S. The SHAP technique assigns a unique importance value to each feature, offering clear insights into its significance in the model's predictions[46].

LIME

LIME is an approach used to explain the predictions generated by ML and DL models, particularly models that are difficult to comprehend. It generates a simple local model that focuses on the specific instance that needs to be explained. It then generates synthetic data points around the instance and then uses the surrogate model to predict the outcomes of those points. This interpretable model then explains the original model's prediction for the specific instance. LIME helps users understand the underlying factors that influence the projections, providing a valuable view of the model's decision-making process. LIME constructs a local model L(x) that approximates the global model M(x) within a defined neighborhood N around the instance x. This relationship can be expressed mathematically as:

$$L\,(x) \approx \; M\,(x)\,, \;\; for\, x \in \; N \tag{7}$$

The main objective is to develop a model L(x) that is both interpretable and locally faithful to the behavior of M(x), enabling users to better understand the rationale behind the decision-making process of the model[47].

Proposed methodology

This section presents a detailed analysis of the suggested intelligent and explainable IDS framework, specifically designed to detect attacks and threats within IoMT systems effectively. Figure 1 shows a comprehensive overview of the proposed IDS framework, highlighting its structure and functionality. The discussion begins with describing

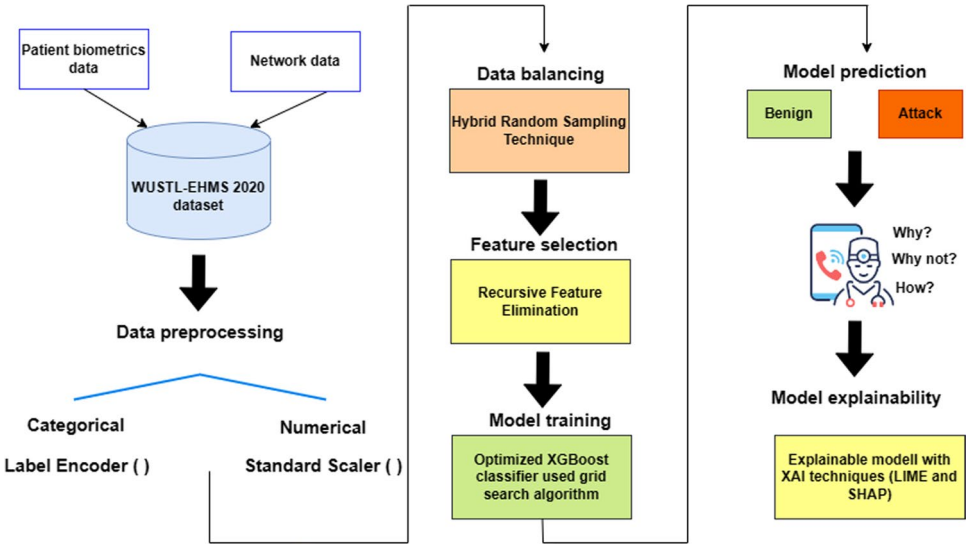

Fig. 1. The proposed intelligent and explainable IDS.

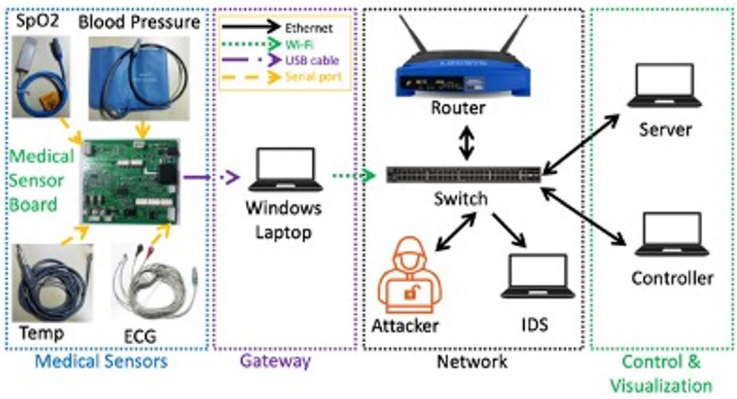

Fig. 2. EHMS testbed[29].

Measurement	Value
Dataset size	4.4 MB
Number of normal instances	14,272 (87.5%)
Number of attack instances	2046 (12.5%)
Total number of instances	16,318

Table 1. Statistical details about the WUSTL EHMS 2020 dataset.

the dataset used, followed by the data preprocessing and balancing techniques implemented to enhance the model's strength and handle class imbalances effectively. It further delves into the feature selection methods used to select the most significant features, thereby reducing model complexity and speeding up the training process. The section also outlines the training process of the IDS framework, highlighting the use of the XGBoost classifier for its ability to handle large datasets, handle imbalanced data, and provide superior performance. Finally, the section concludes with an extensive evaluation of the model's performance using various metrics, ensuring a thorough assessment of its efficacy in multiple scenarios. Furthermore, the proposed methodology incorporates advanced Explainable XAI techniques, providing valuable transparency into the model's decision-making process. These insights enable users and healthcare professionals to understand the model's behavior better, enhancing trust and facilitating practical implementation in real-world IoMT environments.

Dataset description

The WUSTL EHMS 2020 dataset employed in our study was generated through an augmented Real-Time Healthcare Monitoring System (EHMS) testbed, as illustrated in Fig. 2, providing a realistic and practical framework for our study. Biometric data from patients was captured by medical IoT sensors and transmitted via a gateway to a server for further analysis. The uniqueness of this dataset lies in its emphasis on common cyberattacks, including MiTM attacks, spoofing attacks, and data injection attacks. The WUSTL-EHMS 2020 dataset includes network traffic flows and patient biometric data, which have been stored as a CSV file produced by the ARGUS software (version 6.0). The dataset is publicly available and can be accessed at: https://www.cse.wustl.edu/~jain/ehms/index.html.

It contains 44 features, including 35 network flow features, 8 biometric features, and 1 dependent variable, classified as either attack or normal traffic[27]. The statistical details of the WUSTL-EHMS 2020 dataset are provided in Table 1.

Furthermore, we assumed that the dataset is representative of practical healthcare network conditions and attack patterns. The evaluation was performed in an offline setting using a standard desktop environment, and the system was not deployed on physical IoMT devices. Additionally, we assumed that the provided feature set captured sufficient behavioral patterns for accurate intrusion detection. These assumptions align with standard practices in machine learning-based IDS research and help ensure reproducibility.

Data preprocessing

The dataset undergoes preprocessing to extract useful insights, presented in a format that ensures compatibility with ML algorithms. This approach enhances the efficiency of model training and evaluation, thereby contributing to improved classification performance[48]. A detailed explanation of the data preprocessing steps is provided below.

- Handling missing values: The initial step in preprocessing the dataset involved addressing missing values. However, verification has proven that the dataset contained no missing values, eliminating the need to apply imputation techniques.

- Label encoding: The dataset employed in this study consists of both categorical and numerical values. Consequently, the categorical values were transformed into numerical representations utilizing the label encoding technique.
- Normalization: The feature values were normalized using the standard scaler technique to convert them to a common scale with a mean of 0 and a standard deviation of 1. Feature scaling is essential to improving the predictive accuracy of the model and maintaining robustness when handling features with varying numerical ranges. The Standard Scaler is defined mathematically as follows[49]:

$$Z = \frac{x - \mu}{\sigma} \tag{8}$$

where x represents the value of original feature, μ denotes the feature mean, and σ is the standard deviation. This technique standardizes feature values to prevent those with larger magnitudes from dominating the learning process, ensuring that all features contribute equally, regardless of their scale.

- Dataset splitting: To accurately evaluate the performance of developed model, the dataset was split into 80% for training and 20% for testing, a widely recommended division to avoid overfitting[50]. The training set was utilized to develop the model, while the testing set was reserved to test its performance on unseen data.

Data balancing approach

A significant data imbalance issue is observed in the WUSTL EHMS 2020 dataset, which we addressed through the application of a hybrid random sampling technique. This technique ensures a balanced class distribution by employing a hybrid approach that integrates oversampling of the minority class with undersampling of the majority class. An oversampling technique involves creating synthetic instances for the minority class (attack traffic), while undersampling focuses on selectively decreasing the majority class (normal traffic). This balanced representation enables the model to learn and recognize patterns effectively from both the normal and attack classes, enhancing its ability to detect anomalies while mitigating the risk of bias toward the majority class.

Let A1 represent the majority class and A2 the minority class. The developed Hybrid Random Sampling method can be defined using the following mathematical expression:

$$\mathrm{D} = \mathrm{Over\_sample(A2,\ b)}\ \cup\ \mathrm{Under\_sample(A1,\ b)}$$

In this formula:

- Over_sample (A2,b) denotes the oversampling process applied to the minority class A2, where b represents the balancing factor, determining the proportion of synthetic samples to be generated.
- Under_sample (A1,b) refers to the undersampling process applied to the majority class A1. This reduces the size of A1 to align it with the desired level b, ensuring a balanced contribution from both classes.

The union of these two processes produces a balanced dataset D, ensuring that the model is trained on fair representation of both classes. In this study, (b = 0.8) is adopted as the balancing factor, supported by empirical findings that demonstrate this value effectively reduces overfitting risk while achieving an optimal balance between the classes[51]. The decision to employ the Hybrid Random Sampling technique was motivated by its demonstrated ability to achieve balanced class distributions while avoiding unnecessary complexity. This approach minimizes the risk of over-generalization, making it an optimal and well-suited choice for this study. Before applying this technique, the dataset had a clear class imbalance, with 14,272 instances in the normal class and only 2046 instances in the attack class. This imbalance can cause the model to favor the majority class, greatly reducing its effectiveness in detecting and classifying the minority class, such as attack instances. Consequently, mitigating this issue is substantial to developing a strong and reliable IDS model. The hybrid random sampling technique was used to mitigate this imbalance, leading to significant improvements and a more balanced class distribution. After the data balancing process, the dataset was adjusted to have 14,271 normal samples and 11,417 attack samples, achieving a much more balanced class ratio. Notably, the number of normal instances was marginally decreased from 14,272 to 14,271, whereas the attack instances increased significantly from 2046 to 11,417. This balanced dataset ensures that the model can effectively detect both normal and attack behaviors, improving its overall performance, particularly its ability to identify and respond to attacks accurately. This balanced approach makes the IDS more reliable and capable of handling real-world IoMT environments, where detecting attacks is a critical priority. Figure 3 shows the effect of the proposed balancing technique on the class distribution.

Feature selection: recursive feature elimination

Feature selection is the process of selecting the key features from a dataset to improve the performance of ML models and reduce the risk of overfitting[52]. This study employs the Recursive Feature Elimination (RFE) wrapper method, a robust feature selection technique designed to identify the key features by recursively removing less important ones. The RFE technique begins with the entire set of features and systematically eliminates those with lower importance until the optimal number of features remains. Specifically, this study used RFE integrated with an RF estimator as a base model to identify the optimal set of features. The RF estimator leverages its ensemble-based structure to evaluate feature importance by calculating the reduction in impurity across its decision trees. This method assigns higher importance scores to features that significantly contribute to decision-making within the forest, providing a robust and interpretable metric of feature relevance. The RF estimator is highly suitable

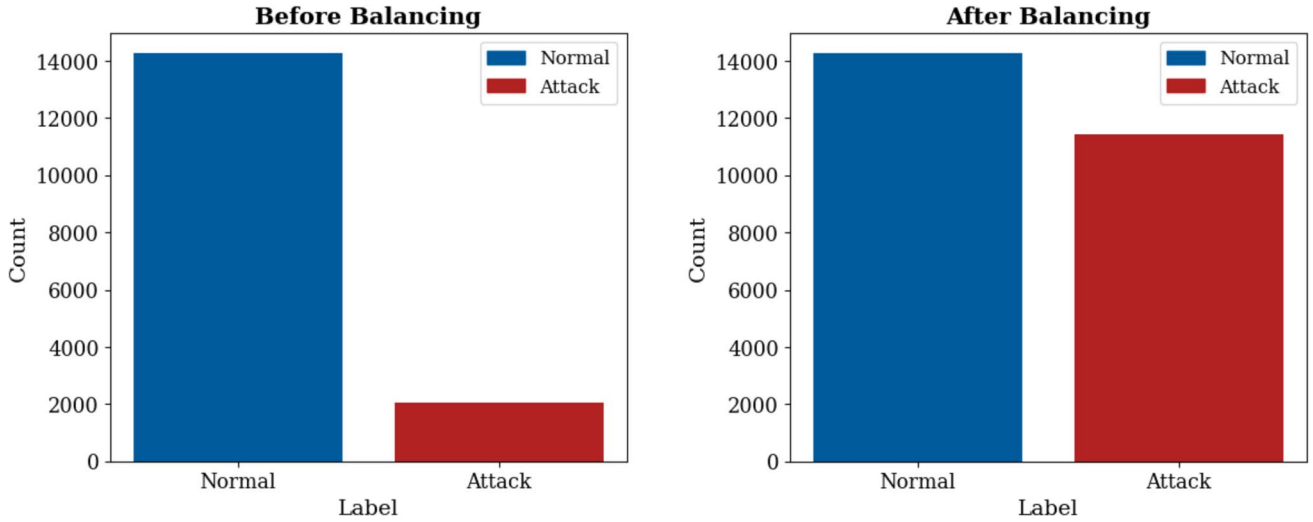

Fig. 3. Class distribution comparison before & after implementing the proposed balancing technique.

Feature selection method	Estimator model	Selected features
RFE	RF	Sport, SrcLoad, DstLoad, SIntPkt, DIntPkt, SrcBytes, DstBytes, TotBytes, DstJitter, sMinPktSz, sMaxPktSz, TotPkts, loss, pLoss, pSrcLoss, Dur, Load, Rate, Temp, SpO2, Pulse_ Rate, Resp_Rate, Heart_rate,, DIA, ST.

Table 2. List of selected features.

for this purpose due to its ability to capture complex, non-linear relationships and interactions between features, ensuring that the selected subset of features not only enhances predictive accuracy but also strengthens the ability of the model to generalize to new data. The combination of RFE and RF provides an efficient and scalable solution for high-dimensional datasets, making it ideal for applications such as intrusion detection. As a result, 25 optimal features were selected through this process, as listed in Table 2.

Model training and hyperparameter optimization

During the model training phase, an XGBoost classifier was employed, a high-performance ensemble learning algorithm recognized for its strength, scalability, and ability to handle complex classification tasks effectively. XGBoost was chosen for this study because of its exceptional ability to manage high-dimensional datasets, address missing values, and mitigate overfitting through built-in regularization techniques. These characteristics make it well-suited for intrusion detection in IoMT networks, which often involve challenges such as class imbalance, noisy data, and non-linear feature relationships.

A grid search algorithm was employed to optimize the hyperparameters of the XGBoost algorithm in order to maximize the model's performance and ensure optimal generalization. This method systematically evaluates a predefined range of hyperparameter values and extensively searches for the best combination that reduces the loss function and achieves superior results[53]. The key hyperparameters selected for tuning in the XGBoost algorithm included: the number of trees in the forest (n_estimators), The maximum depth of decision trees (max_depth), The proportion of samples used to grow each tree (Subsample) and the learning rate (learning_ rate) which controls how much impact each new DT has on the model. The Grid Search technique systematically assessed various XGBoost hyperparameter combinations using cross-validated performance metrics, ultimately selecting the configuration that delivered the best performance.

The mathematical expression of the grid search algorithm is as follows:

$$Optimal\ Parameters\ \alpha^{*} = arg \min_{\alpha \in A} \mathcal{L}(\alpha) \tag{9}$$

Where A represents all potential hyperparameter combinations, and L(α) denotes the loss function calculated using the cross-validation technique. By applying three-fold cross-validation, the algorithm evaluated each

hyperparameter configuration for its ability to balance performance and generalization, ultimately selecting the best-performing set.

Table 3 summarizes the hyperparameter search space and the optimal parameters identified during the model training process.

The comprehensive grid search process, as outlined, ensured the development of an XGBoost model with optimal hyperparameters, maximizing its performance in detecting intrusions within IoMT networks.

6 evaluation metrics

To ensure a comprehensive and reliable evaluation of the proposed IDS approach within the sensitive context of IoMT environments, we employed five widely recognized evaluation metrics: including accuracy, precision, recall, F1-score, and ROC-AUC. Each of these metrics serves a specific role in evaluating IDS performance:

Accuracy is the number of instances correctly detected out of the total instances. It is computed as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Where

- TP (True Positive): Number of attack instances correctly detected as attacks.
- FP (False Positive): Number of benign instances misclassified by the model as attack instances.
- TN (True Negative): Number of benign instances correctly identified as benign.
- FN (False Negative): Number of attack instances misclassified by the model as benign instances.

Precision evaluates the proportion of attack instances correctly detected as attacks to the total instances classified as attacks.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall is the number of attack instances correctly detected as attacks divided by the total actual attack instances.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

F1-Score: is the harmonic mean of precision and recall, measuring the trade-off between them. It is particularly useful for evaluating performance on imbalanced datasets.

$$\mathrm{F1-Score} = \frac{2^{*}\ \mathrm{Precision}^{*}\mathrm{Recall}}{\mathrm{precision} + \mathrm{Recall}} \tag{13}$$

ROC AUC evaluates the balance between TP and FP rates over different thresholds.

These metrics are particularly vital in healthcare-related IDS, where both missed detections and false alerts can have serious consequences. This multi-metric approach ensures a robust and fair assessment of the proposed model's effectiveness in real-world IoMT scenarios.

Experiment setup and result analysis

This section discusses the experimental setup and evaluates the performance of the ML models employed in this study. To demonstrate its effectiveness, it compares the developed IDS with state-of-the-art benchmark IoMT solutions. Additionally, the results of XAI techniques are presented, highlighting the interpretability and explainability of the proposed IDS model. These insights not only validate the IDS's performance but also emphasize its transparency, a critical factor in real-world IoMT network applications.

Experimental setup

The experiments were performed on a personal computer running Windows 7 Ultimate (64-bit operating system) with 8 GB of RAM and an Intel(R) Core (TM) i7-3537U processor at 2.00 GHz (up to 2.50 GHz). The proposed approach was developed using Jupyter Notebook, with Anaconda software (version 2024.10.1) serving as the platform for managing and executing Python code. The ML models were trained and tested using various Python libraries, Scikit-learn for model development, NumPy for numerical computations, Matplotlib for data visualization, and GridSearchCV for hyperparameter optimization.

Hyperparameter	Search space	Optimal value
n_estimators	50, 100, 200	200
max_depth	3, 5, 7	7
Subsample	0.7, 0.8, 1.0	0.7
learning_rate	0.01. 0.1, 0.2	0.2

Table 3. The tuned hyperparameter of XGBoost model.

Metric/class	Precision (%)	Recall (%)	F1-Score (%)	Support
Normal (0)	100	99	99	2874
Attack (1)	98	100	99	2264
Accuracy			99.22	5138
Macro avg	99	99	99	5138
Weighted avg	99	99	99	5138

Table 4. Evaluation results of the developed model.

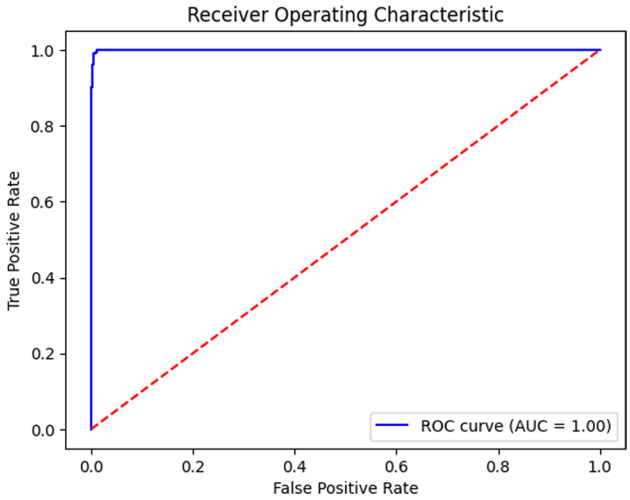

Fig. 4. ROC curve of the proposed IDS model.

Results and discussions

This subsection introduces a detailed analysis of the experimental results obtained through the rigorous implementation of the established research methodology.

Classification results

As summarized in Table 4, the proposed IDS model demonstrated outstanding classification performance, achieving an overall accuracy of 99.22%, confirming its capability to reliably distinguish between normal and attack traffic in IoMT environments. For normal traffic (Class 0), the model achieved a precision of 100% and a recall of 99%, reflecting its ability to correctly identify benign activity while minimizing false positives. In the case of attack traffic (Class 1), the precision and recall were 98% and 100%, respectively, demonstrating the model's strength in detecting malicious behavior with minimal false negatives. These balanced results led to an F1-score of 99% for both classes, which strongly indicates the model's robustness in handling both the majority (normal) and minority (attack) classes. This performance is particularly important in real-time healthcare applications, where both false alarms and missed detections can have serious consequences. The combination of high precision, recall, and F1-score for both classes confirms that the model performs well even under class-imbalanced conditions, and is thus well-suited for practical deployment in sensitive IoMT systems.

Receiver operating characteristic (ROC)

The ROC-AUC, illustrated in Fig. 4, further validated the model's performance with a perfect score of 100%. This ideal score demonstrates the model's ability to reliably differentiate between benign and attack traffic across varying classification thresholds, ensuring consistent performance under diverse conditions. The combination of high accuracy, balanced precision-recall metrics, and a perfect ROC-AUC score highlights the model's exceptional predictive confidence and reliability. These results confirm the suitability of suggested approach for real-world applications, particularly in the sensitive and dynamic environment of IoMT networks, where the prevention of security breaches and the protection of critical healthcare data are paramount.

Confusion matrix analysis

The confusion matrix presented in Fig. 5 offers additional insight into the classification capability of the proposed IDS model. The system successfully identified 2836 normal samples and 2262 attack samples, while misclassifying only 38 normal instances as attacks (false positives) and 2 attack instances as normal (false negatives). This low rate of misclassification underscores the model's reliability and its potential to work effectively in real-world scenarios, where even minor errors could compromise patient safety or the integrity of medical data.

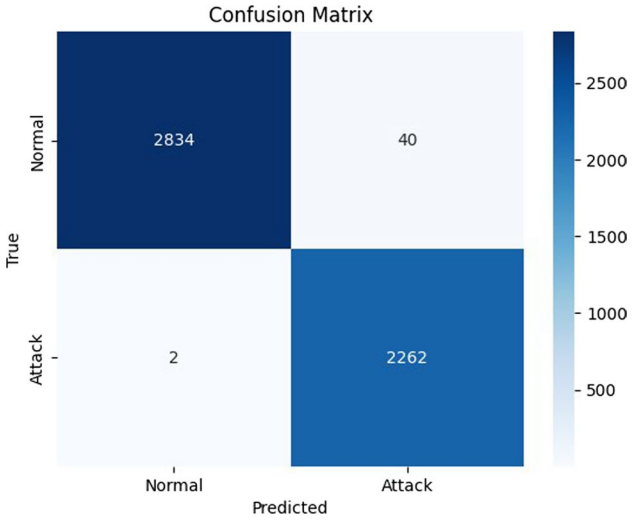

Fig. 5. The confusion matrix for the proposed IDS model.

Models	Accuracy	Precision	Recall	F1-Score
LR	75.82	90.56	50.41	64.85
SVM	77.12	97.80	49.15	65.13
KNN	96.54	92.69	99	96.17
DT	96.81	93.28	99.96	98.51
RF	98.89	97.96	99.56	98.57
Adaboost	83.83	84.36	77.62	80.84
Gradient boosting	98.32	97.76	98.45	98.11
The proposed model	**99.22**	**98.35**	**99.91**	**99.12**

Table 5. Comparison between the proposed model and various benchmarking ML algorithms. Bold indicate the highest performance values achieved by the proposed model when compared against standard machine learning algorithms. This highlights the superiority of our approach across all evaluation metrics (accuracy, precision, recall, and F1-score).

Comparative analysis with baseline ML models

To ensure a fair and comprehensive evaluation, the performance of the proposed model was compared with several benchmark machine learning algorithms, including LR, SVM, KNN, Decision Tree (DT), Random Forest (RF), AdaBoost, and Gradient Boosting. The comparison was conducted using key evaluation metrics, accuracy, precision, recall, and F1-score, widely used in IDS research. The results, summarized in Table 5, highlight the superior performance of the proposed XGBoost-based approach across all evaluation metrics. The model achieved the highest overall accuracy of 99.22%, outperforming all traditional baselines. Among the compared models, Random Forest and Gradient Boosting achieved the second- and third-highest accuracies at 98.89% and 98.32%, respectively, while Logistic Regression and SVM obtained notably lower accuracies at 75.82% and 77.12%. In terms of precision, the proposed model achieved 98.35%, slightly higher than RF (97.96%) and Gradient Boosting (97.76%). The recall score was particularly impressive at 99.91%, demonstrating the model's exceptional ability to detect actual attack instances. In comparison, RF achieved 99.56%, while other classifiers showed lower sensitivity. The model also achieved the highest F1-score of 99.12%, indicating a well-balanced performance between precision and recall. RF and Gradient Boosting followed with F1-scores of 98.57% and 98.11%, respectively. Although DT (98.51%) and KNN (96.17%) also performed well, they were still outperformed by the proposed framework. These findings collectively confirm that the proposed model offers significant improvements over conventional machine learning approaches, achieving consistent and outstanding results across all major performance indicators, which is crucial for effective and reliable intrusion detection in IoMT systems.

Benchmarking against State-of-the-art IDS approaches

To evaluate the effectiveness of our proposed IDS model, a comparison is conducted between the developed model in this study and the previous works that employed the same dataset. This comparison provides a robust context for assessing the advancements achieved in this study through the integration of XAI techniques with optimized XGBoost classifier. Table 6 compares the proposed model against existing works using key evaluation metrics such as accuracy, precision, recall, and F1-score. The results clearly demonstrate the superiority of the

Study	Method	Explainable IDS	Accuracy	Precision	Recall	F1-Score
Larzek et al. [22]	RFE-DT	X	97.85	96.50	86.29	–
Chaganti et al. [24]	PSO-DNN	X	96	96	96	95
Hady et al. [27]	ANN	X	90.42%	–	–	–
Gupta et al. [28]	Tree classifier	X	94.23	93.45	90.86	93.8
Kilincer et al. [31]	RFE-MLP	X	96.20	96.19	96.16	96.17
Ibrahim et al. [33]	Ensemble of LR + KNN	X	92.5	96.47	44.4	60.68
Alhareth et al. [35]	LRGU-MIFS	X	88.9	–	–	–
Bouke et al. [38]	Deep sequential model	X	99	99	99	99
The proposed model	XAI-bases XGBoost	√	**99.22**	**98.35**	**99.91**	**99.12**

Table 6. Comparison with the existing works. Bold values indicate that the proposed model outperforms existing state-of-the-art IDS frameworks that used the same dataset (WUSTL-EHMS-2020). This emphasizes the novelty and effectiveness of our framework in real-world IoMT intrusion detection scenarios.

Rank	Feature	LIME value
0	pSrcLoss <= −0.03	−0.178219
1	−0.11 < DIntPkt <= −0.08	−0.080925
2	pLoss <= −0.03	0.073339
3	SrcBytes <= −0.02	0.067476
4	Loss <= −0.03	−0.052451
5	sMaxPktSz <= −0.02	−0.048281
6	sMinPktSz <= −0.02	0.041562
7	−0.12 < DstJitter <= −0.08	0.037668
8	−0.84 < Sport <= −0.05	0.028903
9	Temp <= −0.44	−0.024749
10	−0.37 < Pulse_Rate <= 0.44	0.021428
11	Heart_rate > 0.54	0.018368
12	−0.07 < SIntPkt <= −0.06	−0.016326
13	TotPkts <= −0.05	−0.016207
14	−0.09 < Resp_Rate <= 0.59	0.014239

Table 7. Analysis of the LIME technique for enhancing model explainability.

proposed model, which achieved an exceptional performance with an accuracy of 99.22%, surpassing all previous works across all evaluation metrics. It is worth noting that this improved performance can be attributed to our comprehensive methodology employed in this study, which includes implementing hybrid random balancing techniques, feature selection using RFE technique, and a novel application of XAI to improve the interpretability of the model. In contrast to comparative studies, which did not use data balancing or exploit the capabilities of XAI techniques, our model highlights the critical impact of these strategies on improving the effectiveness and reliability of intrusion detection. The use of an optimized XGBoost classifier, combined with the hybrid random balancing technique and RFE-based feature selection, provides a robust and efficient framework for addressing the complexities associated with intrusion detection in IoMT systems. Furthermore, the integration of XAI techniques such as LIME and SHAP, allows for a comprehensive understanding of the model's decision-making process, thereby fostering trust and transparency in its predictions. This is a notable departure from other approaches mentioned, which predominantly depend on traditional ML models without providing insights into their operational mechanisms.

Model explainability

This section presents the results of applying the LIME and SHAP techniques to our proposed model, which demonstrated exceptional performance in our experiments. We examine the potential benefits of employing XAI techniques in the context of this study, providing detailed insights into the proposed model's decision-making processes. This analysis aims to enhance the model's transparency and interpretability while deepening our understanding of its capabilities.

The LIME analysis, shown in Table 7, is an important tool for understanding the complex decision-making processes of the XGBoost-based XAI model used in this study. LIME's output explains predictions for specific data points, helping us identify why the model predicted a particular instance as an "Attack" or "Normal." The LIME values tell us the direction (positive or negative) and strength of a feature's influence. The most striking observation is that network-level features dominate the negative LIME values, which push the prediction toward "Attack." For instance, pSrcLoss with a LIME value of −0.1782 is the strongest contributor to the "Attack"

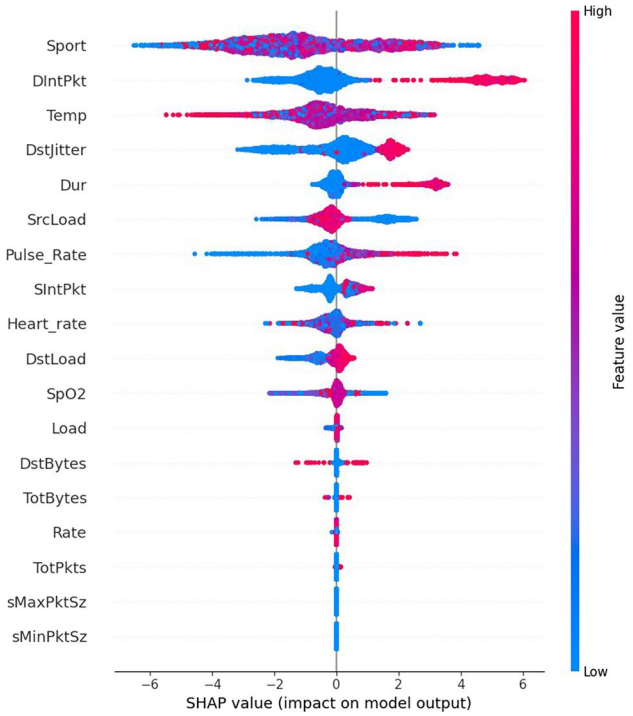

Fig. 6. Overview of the SHAP values for the proposed model.

prediction. This suggests that when source packet loss (pSrcLoss) drops below −0.03, it creates a strong anomaly in the data, signaling intrusion or abnormal behavior. This result makes sense in an IoMT environment, where stable network communication is critical, and packet loss may indicate a malicious attack disrupting the connection. Similarly, DIntPkt (delay between incoming packets) contributes significantly with a LIME value of -0.0809. When DIntPkt falls within the range −0.11–−0.08, it suggests an unusually short delay between packets, which is often a characteristic of network flooding or packet injection attacks. The model correctly identifies this behavior as anomalous. Another notable feature is Loss, which has a LIME value of −0.0524 when it falls below −0.03. High data loss in the network typically signals instability or a malicious entity overwhelming the system, reinforcing the model's classification of "Attack." On the positive side, physiological features such as Pulse_Rate and Resp_Rate push the prediction toward "Normal." For example, Pulse_Rate within the range −0.37–0.44 has a positive LIME value of 0.0214, indicating that pulse rates within this range are consistent with normal human health metrics. Similarly, Resp_Rate < = 0.59 contributes 0.0142 positively, suggesting that respiratory rates within this range align with expected behavior, counteracting any suspicion of an attack. These positive contributions show the complementary role of physiological features in distinguishing normal system behavior from anomalies. While packet-level anomalies may indicate attacks, the presence of stable health signals (pulse and respiratory rates) can help the model balance its decision-making process.

In contrast to LIME, the SHAP results provide a more comprehensive, global perspective of how features influence predictions across the entire dataset. The SHAP summary plot, as shown in Fig. 6, reveals both the magnitude of influence and the actual feature values contributing to predictions. One of the most important findings from the SHAP results is the dominant role of DIntPkt (inter-packet delay). SHAP shows that low values of DIntPkt (blue points) have a strong negative SHAP value, meaning they heavily contribute to "Attack" predictions. This aligns with the LIME results and highlights that unusually short delays between packets are consistently flagged as a sign of network-based intrusions. High DIntPkt values, on the other hand, show positive SHAP contributions (red points), reinforcing that a more normal packet delay aligns with "Normal" predictions. This feature's importance indicates that DIntPkt is a reliable metric for identifying network anomalies. Another critical feature is Sport (source port), which displays both positive and negative SHAP values. This variability suggests that certain ranges of source port values strongly correlate with anomalies, while others align with normal system behavior. For instance, specific low Sport values (blue) may signal a targeted or malicious packet transmission, whereas higher source port values (red) are seen as part of expected network communication. Temperature (Temp) is another feature with strong influence, as shown by its wide spread of SHAP values. Abnormal temperature values, whether low or high, are associated with negative SHAP values, pushing predictions toward "Attack." This is an important finding for IoMT systems, where unusual physiological signals, such as temperature irregularities, can be indicative of system malfunction, spoofing, or tampering. When temperature values fall within a normal range, they contribute positively, supporting "Normal" predictions. The physiological features, Pulse_Rate and Resp_Rate, again play a critical role in balancing predictions. SHAP shows that higher pulse rate values (red) contribute positively to the prediction, indicating normal behavior, while low pulse rates (blue) may push predictions toward "Attack" due to their abnormality. This mirrors the LIME results

and emphasizes that physiological data adds an important layer of insight in IoMT systems. By incorporating these signals, the model ensures that health-related anomalies are not overlooked, complementing the network-based features. One additional observation is the feature DstJitter, which measures packet jitter or variation in delay. SHAP values indicate that lower jitter values (blue) are associated with negative contributions, suggesting anomalies, while higher jitter values are seen as more normal. This reinforces the importance of network stability in detecting intrusions, as jitter can signal abnormal packet flows or delays caused by malicious activity.

Conclusion and future work

In this study, we proposed an explainable IDS framework tailored for IoMT environments. Our approach integrates an optimized XGBoost classifier with hybrid random sampling techniques to address class imbalance and a recursive feature elimination (RFE) strategy for feature selection. To ensure interpretability, we incorporated SHAP and LIME to provide both global and local explanations of model predictions, enabling transparency and trust in healthcare applications. The model was evaluated using the WUSTL-EHMS-2020 dataset, achieving superior performance with 99.22% accuracy, 98.35% precision, 99.91% recall, 99.12% F1-score, and 100% ROC-AUC, outperforming several standard machine learning models. These results highlight the potential of combining interpretable AI techniques with robust classifiers for cybersecurity in critical domains such as IoMT.

We emphasize that in healthcare contexts, explainability is not a trade-off but a fundamental requirement to ensure trust, accountability, and effective collaboration between AI systems and human experts. While high detection accuracy is crucial, the ability to understand and validate model decisions is equally vital for safe and ethical deployment.

Future work will focus on extending this approach to diverse IoT environments, testing on additional datasets, and exploring model optimization techniques (e.g., pruning, quantization) for deployment on resource-constrained devices. We also plan to investigate the integration of deep learning models, adversarial robustness techniques, and additional signals such as device behavior patterns (e.g., abnormal access times or communication rates), firmware integrity indicators (e.g., unauthorized updates or hash mismatches), and temporal usage anomalies (e.g., unexpected activity outside clinical hours) to further enhance detection capabilities. Furthermore, we plan to explore domain-specific explainable AI techniques such as counterfactual reasoning, causal inference, and healthcare ontologies to enhance interpretability in complex and high-stakes clinical scenarios. These enhancements aim to ensure the framework remains resilient and adaptable to future and evolving cybersecurity threats in IoMT ecosystems.

Data availability

The dataset used in this study is publicly available and can be accessed at the following https://www.cse.wustl.edu/~jain/ehms/index.html.

Received: 14 April 2025; Accepted: 17 June 2025
Published online: 01 July 2025

References

1. Ghosal, P., Das, D. & Das, I. Extensive survey on cloud-based IoT-healthcare and security using machine learning. In *2018 4th International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* 1–5 (IEEE, 2018).
2. Alzubi, O. A., Alzubi, J. A., Qiqieh, I. & Al-Zoubi, A. An IoT intrusion detection approach based on salp swarm and artificial neural network. *Int. J. Netw. Manag.* **35**, e2296 (2025).
3. Kumar, P., Gupta, G. P. & Tripathi, R. An ensemble learning and fog-cloud architecture-driven cyber-attack detection framework for IoMT networks. *Comput. Commun.* **166**, 110–124 (2021).
4. Ghubaish, A. et al. Recent advances in the internet-of-medical-things (IoMT) systems security. *IEEE Internet Things J.* **8**, 8707–8718 (2020).
5. Razdan, S. & Sharma, S. Internet of medical things (IoMT): Overview, emerging technologies, and case studies. *IETE Tech. Rev.* **39**, 775–788 (2022).
6. Hasan, M. K. et al. A novel resource oriented DMA framework for internet of medical things devices in 5G network. *IEEE Trans. Ind. Inform.* **18**, 8895–8904 (2022).
7. Kumar, P., Gupta, G. P. & Tripathi, R. A distributed ensemble design based intrusion detection system using fog computing to protect the internet of things networks. *J. Ambient Intell. Humaniz. Comput.* **12**, 9555–9572 (2021).
8. Sun, W. et al. Security and privacy in the medical internet of things: A review. *Secur. Commun. Netw.* **2018**, 5978636 (2018).
9. Yaqoob, T., Abbas, H. & Atiquzzaman, M. Security vulnerabilities, attacks, countermeasures, and regulations of networked medical devices-a review. *IEEE Commun. Surv. Tutor.* **21**, 3723–3768 (2019).
10. Rasool, R. U., Ahmad, H. F., Rafique, W., Qayyum, A. & Qadir, J. Security and privacy of internet of medical things: A contemporary review in the age of surveillance, botnets, and adversarial ML. *J. Netw. Comput. Appl.* **201**, 103332 (2022).
11. Rahman, M. S., Alabdulatif, A. & Khalil, I. Privacy aware internet of medical things data certification framework on healthcare blockchain of 5G edge. *Comput. Commun.* **192**, 373–381 (2022).
12. Chen, T. M., Blasco, J., Alzubi, J. & Alzubi, O. Intrusion detection. *Eng. Technol. Ref.* **1**, 1–9 (2014).
13. Bouke, M. A. & Abdullah, A. An empirical assessment of ML models for 5G network intrusion detection: A data leakage-free approach. *e-Prime Adv. Electr. Eng. Electron. Energy* **8**, 100590 (2024).
14. Sp, R. M. et al. An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. *Comput. Commun.* **160**, 139–149 (2020).
15. Alzubi, O. A. et al. Optimized machine learning-based intrusion detection system for fog and edge computing environment. *Electronics* **11**, 3007 (2022).
16. Hussain, F., Hussain, R., Hassan, S. A. & Hossain, E. Machine learning in IoT security: Current solutions and future challenges. *IEEE Commun. Surv. Tutor.* **22**, 1686–1721 (2020).
17. Machlev, R. et al. Explainable artificial intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy AI* **9**, 100169 (2022).
18. Love, P. E. D. et al. Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. *Adv. Eng. Inform.* **57**, 10224 (2023).

19. Kaadoud, I. C., Bennetot, A., Mawhin, B., Charisi, V. & Díaz-Rodríguez, N. Explaining Aha! moments in artificial agents through IKE-XAI: Implicit knowledge extraction for eXplainable AI. *Neural Netw.* **155**, 95–118 (2022).
20. Nazir, S., Dickson, D. M. & Akram, M. U. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Comput. Biol. Med.* **156**, 106668 (2023).
21. Saheed, Y. K. & Chukwuere, J. E. Xaiensembletl-iov: A new explainable artificial intelligence ensemble transfer learning for zero-day botnet attack detection in the internet of vehicles. *Results Eng.* **24**, 103171 (2024).
22. Lazrek, G., Chetioui, K., Balboul, Y., Mazer, S. & El Bekkali, M. An RFE/Ridge-ML/DL based anomaly intrusion detection approach for securing IoMT system. *Results Eng.* **23**, 102659 (2024).
23. Saheed, Y. K. & Arowolo, M. O. Efficient cyber attack detection on the internet of medical things-smart environment based on deep recurrent neural network and machine learning algorithms. *IEEE Access* **9**, 161546–161554 (2021).
24. Chaganti, R. et al. A particle swarm optimization and deep learning approach for intrusion detection system in internet of medical things. *Sustainability* **14**, 1–18 (2022).
25. Kulshrestha, P. & Vijay-Kumar, T. V. Machine learning based intrusion detection system for IoMT. *Int. J. Syst. Assur. Eng. Manag.* **15**, 1802–1814 (2024).
26. Thamilarasu, G., Odesile, A. & Hoang, A. An intrusion detection system for internet of medical things. *IEEE Access* **8**, 181560–181576 (2020).
27. Hady, A. A., Ghubaish, A., Salman, T., Unal, D. & Jain, R. Intrusion detection system for healthcare systems using medical and network data: A comparison study. *IEEE Access* **8**, 106576–106584 (2020).
28. Gupta, K., Kumar, D., Datta, K. & Kumar, A. A tree classifier based network intrusion detection model for internet of medical things ☆. *Comput. Electr. Eng.* **102**, 108158 (2022).
29. Network, S., Nandy, S., Adhikari, M. & Khan, M. A. An intrusion detection mechanism for secured IoMT framework based on. *IEEE J. Biomed. Heal. Informat.* **26**, 1969–1976 (2022).
30. Akshay Kumaar, M. et al. A hybrid framework for intrusion detection in healthcare systems using deep learning. *Front. Public Heal.* **9**, 1–18 (2022).
31. Firat, I., Ertam, F., Sengur, A., Tan, R. & Acharya, U. R. Automated detection of cybersecurity attacks in healthcare systems with recursive feature elimination and multilayer perceptron optimization. *Biocybern. Biomed. Eng.* **43**, 30–41 (2023).
32. Zachos, G., Essop, I., Mantas, G., Porfyrakis, K. & Ribeiro, J. C. An anomaly-based intrusion detection system for internet of medical things. *Electronics* **10**(21), 1–25 (2021).
33. Ibrahim, M., Al-Wadi, A. & Elhafiz, R. Security analysis for smart healthcare systems. *Sensors* **24**, 3375 (2024).
34. Ravi, V., Pham, T. D. & Alazab, M. Deep learning-based network intrusion detection system for internet of medical things. *IEEE Internet Things Mag.* **6**, 50–54 (2023).
35. Alalhareth, M. & Hong, S.-C. An improved mutual information feature selection technique for intrusion detection systems in the internet of medical things. *Sensors* **23**, 4971 (2023).
36. Manimurugan, S. & Al-mutairi, S. Effective attack detection in internet of medical things smart environment using a deep belief neural network. *IEEE Access* **8**, 77396 (2020).
37. Alzubi, J. A., Alzubi, O. A., Qiqieh, I. & Singh, A. Towards robust and efficient intrusion detection in IoMT: A deep learning approach addressing data leakage and enhancing model generalizability. *IEEE Trans. Consum. Electron.* **70**, 2049–2057 (2024).
38. Bouke, M. A., El Atigh, H. & Abdullah, A. Towards robust and efficient intrusion detection in IoMT: A deep learning approach addressing data leakage and enhancing model generalizability. *Multimed. Tools Appl.* https://doi.org/10.1007/s11042-024-19916-z (2024).
39. Praveena Anjelin, D. & Ganesh Kumar, S. An effective classification using enhanced elephant herding optimization with convolution neural network for intrusion detection in IoMT architecture. *Cluster Comput.* **5**, 12341 (2024).
40. Areia, J., Bispo, I. A., Santos, L. & Costa, R. L. C. IoMT-TrafficData: Dataset and tools for benchmarking intrusion detection in internet of medical things. *IEEE Access* **12**, 115370–115385 (2024).
41. Çavuşoğlu, Ü. A new hybrid approach for intrusion detection using machine learning methods. *Appl. Intell.* **49**, 2735–2761 (2019).
42. Arshad, A. et al. A novel ensemble method for enhancing Internet of Things device security against botnet attacks. *Decis. Anal. J.* **8**, 100307 (2023).
43. Ahmed, M. A. O., AbdelSatar, Y., Alotaibi, R. & Reyad, O. Enhancing Internet of Things security using performance gradient boosting for network intrusion detection systems. *Alexandria Eng. J.* **116**, 472–482 (2025).
44. Abdulganiyu, O. H., Ait Tchakoucht, T., Alaoui, A. E. H. & Saheed, Y. K. Attention-driven multi-model architecture for unbalanced network traffic intrusion detection via extreme gradient boosting. *Intell. Syst. with Appl.* **26**, 200519 (2025).
45. Saheed, Y. K. & Chukwuere, J. E. CPS-IIoT-P2Attention: Explainable privacy-preserving with scaled dot-product attention in cyber physical system-industrial IoT network. *IEEE Access* https://doi.org/10.1109/ACCESS.2025.3566980 (2025).
46. Lundberg, S. M. & Lee, S.-I. Explainable AI over the Internet of Things (IoT): Overview, state-of-the-art and future directions. *Adv. Neural Inf. Process. Syst.* **30**, 2106 (2017).
47. Jagatheesaperumal, S. K. et al. Explainable AI over the Internet of Things (IoT): Overview, state-of-the-art and future directions. *IEEE Open J. Commun. Soc.* **3**, 2106–2136 (2022).
48. Saheed, Y. K., Abdulganiyu, O. H. & Ait Tchakoucht, T. Modified genetic algorithm and fine-tuned long short-term memory network for intrusion detection in the internet of things networks with edge capabilities. *Appl. Soft Comput.* **155**, 111434 (2024).
49. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* vol. 14, 1137–1145 (Montreal, Canada, 1995).
51. Banerjee, A. & Ghosh, J. Scalable clustering algorithms with balancing constraints. *Data Min. Knowl. Discov.* **13**, 365–395 (2006).
52. Saheed, Y. K., Omole, A. I. & Sabit, M. O. GA-mADAM-IIoT: A new lightweight threats detection in the industrial IoT via genetic algorithm with attention mechanism and LSTM on multivariate time series sensor data. *Sensors Int.* **6**, 100297 (2025).
53. Movassagh, A. A. et al. Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model. *J. Ambient Intell. Humaniz. Comput.* **14**, 1–9 (2023).

Author contributions

Yousif Hosain conceptualized the research framework, conducted the experiments, implemented the proposed IDS model, and drafted the manuscript. Muhammet Çakmak provided critical supervision, contributed to the methodological design, and revised the manuscript for intellectual content. Both authors were actively involved in analyzing the results and gave their approval to the final version of the manuscript.

Funding

Not applicable.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Y.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.