Discover Applied Sciences

Research

An explainable AI-driven transformer model for spoofing attack detection in Internet of Medical Things (IoMT) networks

Mohammad A. Alsharaiah[1] · Mohammed Amin Almaiah[1] · Rami Shehab[2] · Mansour Obeidat[3] · Fuad Ali El-Qirem[4] · Theyazn Aldhyani[3]

Received: 13 March 2025 / Accepted: 30 April 2025
Published online: 13 May 2025

Abstract
The increasing sophistication of cyber threats necessitates the development of advanced security mechanisms to protect modern networks. Among these threats, spoofing attacks pose a significant risk by enabling malicious actors to impersonate legitimate entities. To address this challenge, we propose a novel Transformer-based deep learning framework designed for the effective detection of spoofing attacks. The core of our novel model is a Transformer neural network, enhanced with a custom attention mechanism to improve feature extraction and classification accuracy. To enhance model interpretability and foster trust in AI-driven security systems, we integrate Explainable AI (XAI) techniques, specifically SHAP analysis, allowing for a deeper understanding of feature contributions in decision-making. The proposed model utilized the CIC IoMT2024 dataset, a benchmark with limited prior research on spoofing attack detection. Further, our approach incorporates comprehensive data preprocessing techniques and employs over-sampling using the synthetic minority oversampling technique (smote) and cleaning using (tomek) these techniques are integrated into links smotetomek to mitigate class imbalance, ensuring a more representative training dataset. The proposed framework is evaluated using benchmark dataset datasets, demonstrating high binary classification performance in spoofing attacks through key metrics such as accuracy, confusion matrix analysis, and other classification benchmarks. The proposed model archived an exact result with Accuracy 99.71%. The findings highlight the potential of Transformer-based architectures in cybersecurity applications, paving the way for real-time threat detection and adaptive defense mechanisms.

Keywords Explainable AI · Transformer model · Spoofing · Classification · Attack

1 Introduction

The integration of intelligent medical devices into the Internet of Medical Things (IoMT) network has transformed healthcare services with enhanced patient monitoring abilities and quicker medical responses with accurate diagnostic devices [1]. The widespread use of IoT technology has resulted in significant security risks, including spoofing attacks, which allow hackers to imitate real devices to alter data or disrupt system functionality [2]. Conventional security measures cannot protect against such threats because medical devices lack processing capability and IoMT environments are too complex.

✉ Mohammed Amin Almaiah, m_almaiah@ju.edu.jo; ✉ Rami Shehab, Rtshehab@kfu.edu.sa; Mohammad A. Alsharaiah, M.ALsharaiah@ju.edu.jo; Mansour Obeidat, mobaydat@kfu.edu.sa; Fuad Ali El-Qirem, beautywings-1994@hotmail.com; Theyazn Aldhyani, taldhyani@kfu.edu.sa | [1]King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan. [2]Vice-Presidency for Postgraduate Studies and Scientific Research, King Faisal University, 31982 Al-Ahsa, Saudi Arabia. [3]Applied College, King Faisal University, Al-Ahsa, Saudi Arabia. [4]Faculty of Architecture and Design, Al-Zaytoonah University of Jordan, Amman 11733, Jordan.

Intrusion detection systems (IDS) based on deep learning (DL) have become effective instruments for identifying and stopping spoofing attacks in Internet of Medical Things (IoMT) networks [3].The Internet of Medical Things (IoMT) has improved healthcare services but also exposed systems to spoofing attacks, where malicious actors mimic legitimate devices. These attacks can compromise patient safety and data integrity.

Many existing spoofing detection models rely on outdated datasets and struggle to capture complex traffic behaviors. Furthermore, they often lack interpretability, which is critical in medical environments. Limited research has combined Transformer models with explainable AI for spoofing detection in IoMT. The absence of SHAP-based insights and use of the CICIoMT2024 dataset represents a significant gap.

This study introduces a Transformer-based model with SHAP explanations to detect spoofing attacks. It employs self-attention for deep pattern recognition and SMOTETomek preprocessing to handle class imbalance.

Much research into sophisticated intrusion detection methods to counter online threats has been spurred by the growing need for better cybersecurity solutions [4]. Cybercriminals can use spoofing attacks to imitate genuine medical devices due to flaws in authentication systems [5]. These security flaws jeopardize patient safety by compromising the integrity of patient data and interfering with necessary medical procedures. Cybercriminals frequently utilize IP address and MAC address spoofing in conjunction with biometric data fraud to carry out their assaults [6]. Deep learning is the best approach for detecting these attacks since it requires sophisticated analytical models that can manage large amounts of real-time data flow [7].

Since these systems can improve cybersecurity and prevent unwanted access, the need for NIDs has grown over time, with an emphasis on research and development [8]. Furthermore, because NIDS can perform predictive analytics, they are very useful for maintaining cyber systems. Additionally, artificial intelligence [9], a technology that enables machines to mimic human brain functions, has improved cybersecurity and intrusion detection. More powerful machine learning and deep learning methods make it possible to process enormous amounts of network traffic in search of concealed attacks and are far more efficient than humans, boosting the efficiency of NIDS systems [10]. Deep learning models inspired by biological neural networks utilize multiple layers of computation to obtain precise meaningful features and improve target systems' cybersecurity [11]. In addition, such models aid extract attack patterns, reshaping evolutionary NIDS frameworks, and learning from observed patterns [10].

One of the main problems with current Network Intrusion Detection Systems (NIDS) is that they are trained and evaluated on old datasets that do not adequately reflect the traffic patterns of modern IoT networks (e.g., see ref. [12]). This restriction reduces the models' efficacy in real-world applications by impeding their capacity to generalize. Furthermore, poor feature selection, ineffective feature reduction methods, and ineffective sampling strategies can result in biases, a higher chance of overfitting, and the loss of important data [13]. As a result, detection accuracy is generally lowered since uncommon attack episodes are frequently difficult for classification algorithms to identify [14].

To improve classification performance and threat detection accuracy, this study proposes a novel network intrusion detection system (NIDS) based on an Explainable AI-Driven Transformer Model for Spoofing Attacks. The suggested model effectively distinguishes between malicious and benign network traffic. The CICIoMT2024 dataset [15], which was selected for its wide range of applications, varied network activity encompassing both normal and attack conditions, and evenly distributed class distribution, is used to train the framework to optimize efficiency. Additionally, the suggested method is contrasted with current cutting-edge detection models, including Deep Neural Networks (DNN). Experiments show that the model outperforms current frameworks in terms of detection capabilities, accuracy, and F-score, proving its capacity to bolster cybersecurity defenses. Further, Explainable AI (XAI) enhances interpretability in medical cyber-physical security systems by providing transparent decision-making, boosting trust among healthcare professionals. It helps ensure compliance, improves security threat detection, and enables more informed medical decisions, ultimately enhancing patient care. XAI facilitates collaboration between AI and medical experts, improving system safety and performance.

Primarily, this paper introduces a new approach that enhances spoofing attack detection in IoMT environments using An Explainable AI-Driven Transformer Model for Spoofing Attacks The CIC-IoMT2024 dataset is utilized to specifically design and fine-tune the proposed architecture for this task. The fact that this type of proposed deep learning method has never been used on this dataset before is noteworthy and represents a significant advancement in the field. Furthermore, the model achieves extremely accurate classification results, outperforming current approaches. In IoMT networks, the binary classification framework efficiently separates and groups different kinds of data.

However, this paper is organized as follows: In Sect. 2, relevant research is reviewed, and current deep learning techniques for network intrusion detection systems (NIDS) are covered. The suggested model is explained in detail in Sect. 3. In Sect. 4, the effectiveness of the model is evaluated, the experimental methodology is described, and the findings are evaluated.

Section 5 explores the broader implications and potential constraints of the approach. The study ends with a summary of the main conclusions and recommendations for further research.

1.1 Related work

Through the integration of medical devices into networked systems, the Internet of Medical Things (IoMT) has revolutionized healthcare by enhancing data accessibility and patient care [16]. However, there are risks associated with more connectivity as well, chief among them being spoofing attacks, in which malicious actors impersonate reliable devices to gain unauthorized access or disrupt services [17]. To get over these issues, researchers have investigated deep learning models for spoofing attack detection in IoMT networks [18]. One significant contribution to this field is the CICIoMT2024 dataset from the Canadian Institute for Cybersecurity [19]. Because it contains network traffic data from 40 IoMT devices, including 18 distinct cyberattack scenarios, this dataset offers a comprehensive foundation for developing and evaluating intrusion detection systems (IDS).

Using this dataset, Dadkhah et al. assessed a number of machine learning models for binary and multi-class classification tasks, including Random Forest, AdaBoost, Deep Neural Networks (DNN), and Logistic Regression [20]. Their findings demonstrated nearly perfect binary classification accuracy, but the accuracy fell to 73.3% after splitting into 19 classes, highlighting the challenge of identifying multi-class spoofing assaults. The limitation lies in the accuracy drop from near-perfect binary classification to 73.3% in multi-class classification, highlighting challenges in distinguishing between multiple spoofing attack types. This may be due to factors like class imbalance. Sánchez et al. advanced detection methods by investigating the application of optimal Transformer designs using the CICIoMT2024 dataset [21]. Their approach demonstrated promising performance improvements over traditional machine learning techniques, suggesting that Transformer-based models might effectively enhance anomaly identification in IoMT environments. To help these efforts, a study by Maroof et al. identified IoT event spoofing attacks using time-series classification techniques [22]. Their comprehensive analysis of a publicly available real-world dataset demonstrated that temporal feature learning was a viable solution for IoMT security because it was able to identify spoofing attempts using significantly smaller training samples. However, their performance in real-world, large-scale IoMT environments is still uncertain. Additionally, Maroof et al.'s approach relies on smaller training samples, which may not generalize well to larger, more diverse datasets or complex attack scenario [23].

In conclusion, deep learning models—especially DNNs and Transformer-based architectures—have shown great promise in detecting spoofing attacks in IoMT networks when combined with large datasets like CICIoMT2024. These strategies support the development of robust and adaptable security solutions by tackling the unique challenges posed by the varied and dynamic nature of IoMT settings.

Recent studies have looked into the application of transformer-based models in IoMT intrusion detection [24]. For instance, a hybrid model that included transformers and convolutional neural networks (CNNs) had an overall accuracy of 99.49% in identifying IoT dangers using the CICIoT2023 dataset [25]. On the CICIoMT2024 dataset, a different method combined transformer-based neural networks with data augmentation, feature selection, and ensemble learning to achieve ideal detection rates of 99% [26].

By incorporating XAI techniques into these models, the critical need for transparency and dependability in AI-driven security solutions is satisfied [27]. By providing interpretable insights into model decisions, XAI increases the reliability of intrusion detection systems and enables stakeholders to understand and trust the procedures that underlie threat identification [28]. This is particularly crucial in medical cyber-physical systems where patient data confidentiality and privacy are vital.

In conclusion, the combination of XAI and transformer models, backed by huge datasets such as CICIoMT2024, is revolutionary for spoofing attack detection in IoMT networks. In addition to improving detection accuracy, this union guarantees AI systems' transparency, which boosts confidence in automated security measures.

2 The main methodology

This section covers several informative subjects, such as the characteristics of the dataset and the methods for preprocessing and preparation before the training phase. It also examines the criteria used for performance evaluation and the techniques used to develop the proposed model. The general characteristic framework of the proposed model is also described in this section.

2.1 Dataset for ICIoMT2024

Spoofing attacks have become more and more dependent on the combination of Explainable Artificial Intelligence (XAI) and sophisticated deep learning models, like transformers. In this context, the Canadian Institute for Cybersecurity's CICIoMT2024 dataset is a helpful resource [29]. Ransomware, malware injections, Distributed Denial of Service (DDoS) assaults, and man-in-the-middle attacks are among the 18 different cyberthreats that are recorded in the dataset. It includes malicious activity captured from 40 IoMT devices as well as normal network data. Its thoroughness makes it possible to develop and assess machine learning algorithms especially intended to enhance IoMT security.

An excellent benchmark for cybersecurity practices in the healthcare sector is the publicly accessible ICIoMT2024 dataset, which includes roughly 46 attributes and documents 16,047 network activity occurrences covering a range of cyber breaches, including ARP Spoofing. The dataset is important in comparing the effectiveness of intrusion detection systems based on machine learning and deep learning against spoofing threats. It was created to enable the development and assessment of security frameworks for IoMT ecosystems. It supports detailed study and verification of detection models through a mixture of simulated and actual network data, contributing to IoMT threat mitigation strategies improvement [29].

2.2 Data preprocessing

Several cutting-edge techniques are combined in the preprocessing pipeline utilized in this work to enhance feature representation, enhance data quality, and address problems such as class imbalance as shown in Fig. 1. The methodology enhances generality and robustness in spoofing attack detection by guaranteeing that the model receives input data that is properly structured and normalized. Starts with raw data loading and consolidation. Errors, inconsistencies, and duplicate records are then removed during the cleaning process. In this procedure, redundant, inconsistent, and unnecessary data are filtered out; null and infinite values are discarded; duplicate entries are found and fixed; and data integrity issues are addressed.

Preprocessing and data consumption begin when the training and testing datasets are imported. Training and testing datasets are consumed as part of the process's data intake and preparation processes. The SMOTETomek hybrid

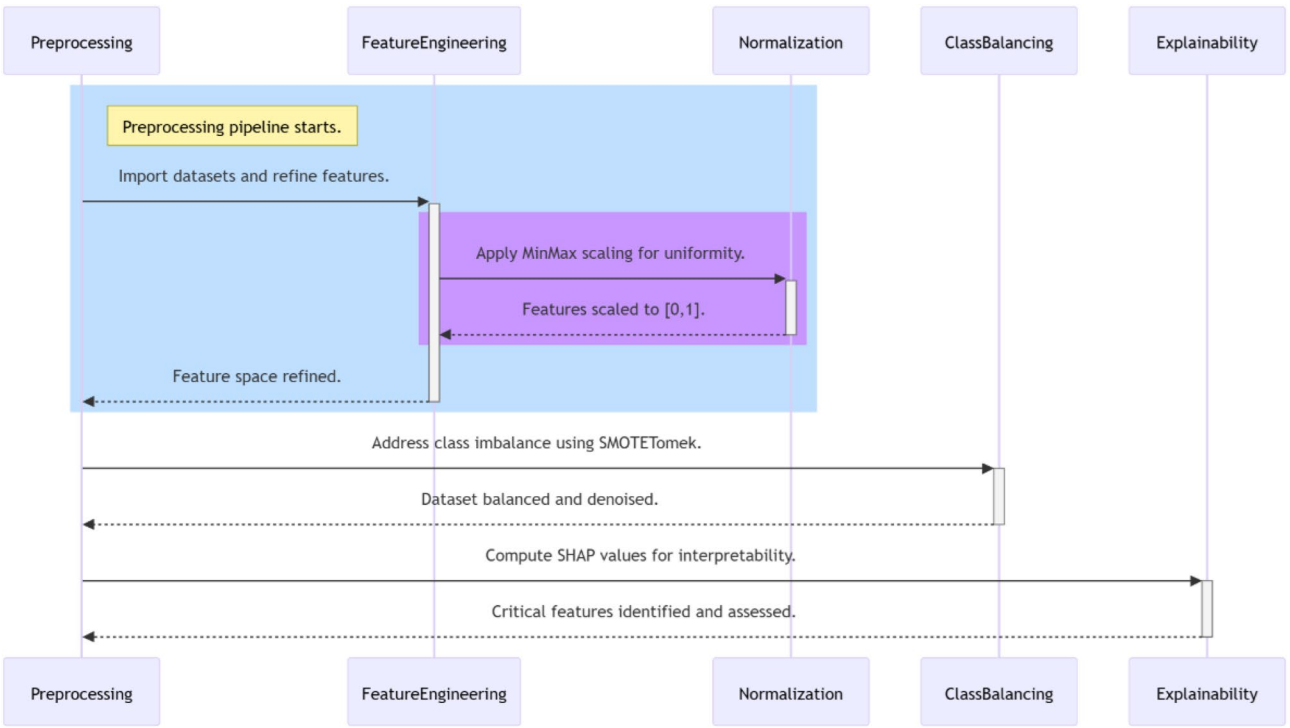


Fig. 1 Data preprocessing pipeline

resampling technique is used to correct this [30]. This method combines the Synthetic Minority Over-Sampling Technique (SMOTE), which artificially creates more instances of the minority class, with Tomek connections, which eliminate ambiguous and borderline material. SMOTETomek is used to balance the dataset, reducing the likelihood of misclassifications and improving the model's capacity to differentiate between legitimate and fraudulent traffic.

There is a notable disparity in class distribution, with only 428 instances assigned to class 1, whereas the remaining 16,047 belong to class 0. This imbalance can create difficulties in training the model, potentially resulting in skewed predictions that favor the dominant class. To mitigate this issue, various strategies such as resampling, adjusting class weights, or employing specialized algorithms can be utilized to promote balanced learning and enhance predictive accuracy.

To tackle the class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) has been applied. This method is effective in managing imbalanced datasets by increasing the representation of the minority class without introducing overfitting. It achieves this by generating synthetic samples that resemble existing minority class instances, thereby reducing the scarcity of data points and ensuring a more evenly distributed feature space.

Explainability techniques like SHapley Additive exPlanations (SHAP) significantly boost the preprocessing process' efficacy [31]. To ensure interpretability and openness in decision-making, SHAP values are calculated after training to assess the contribution of specific qualities to model predictions. The study determines major prediction factors and evaluates whether specific qualities have a substantial impact on the detection process using SHAP. These data pretreatment methods work together to provide a balanced, well-organized, and optimized dataset that raises the deep learning model's detection accuracy. Class balancing, explainability measurements, correlation analysis, and feature scaling is used to guarantee that the proposed model is efficient, understandable, and relevant to real-world spoofing attempts.

The extraction of pertinent features and the conversion of the target variable into a binary classification format based on its median value are two crucial feature engineering procedures that are used to enhance the feature space. Figure 2 illustrates how [32], which finds interdependencies across attributes to assist comprehend feature linkages and possible redundancies, also makes use of a heatmap. MinMax is used in scaling (1) to normalize feature distributions and lessen the effect of different numerical scales [33]. By mapping qualities to a range of [0,1], this normalization strategy ensures uniformity among input variables and keeps some attributes from having an excessive impact on the model.

$$X\prime = \left(X - X_{min}\right)/\left(X_{max} - X_{min}\right) \tag{1}$$

Unlike standardization techniques that assume a Gaussian distribution, MinMax scaling preserves the data's inherent distribution, making it perfect for deep learning models. One major issue with intrusion detection datasets is the presence of class imbalance, where malicious activities, such as spoofing assaults, are underrepresented.

A correlation heatmap, which is frequently used in data analysis to show the correlations between numerical variables, is shown in Fig. 2. It displays a correlation heatmap, which is commonly used in data analysis to display the relationships between numerical variables. The heatmap employs a color gradient, with red hues signifying positive correlations and blue hues indicating negative correlations. The color bar on the right shows how the strength of the association is represented by the color's intensity. relationships, such as protocol-based or statistical similarities. With this image, researchers can identify variables that are redundant or highly linked that may have an impact on how well machine learning models perform. The values inside the cells correspond to Pearson correlation coefficients, which range from -1 to 1. Pearson correlation coefficients in (2), which vary from -1 to 1, are represented by the values inside the cells [34]. Potentially taken from a cybersecurity dataset, variables along both axes seem to represent network traffic attributes.

$$r = \frac{\Sigma\left(x - \bar{x}\right)\left(y - \bar{y}\right)}{\sqrt{\sum\left(x - \bar{x}\right)^2 \sum\langle y - \bar{y}\rangle 2}} \tag{2}$$

In tasks involving feature selection, anomaly detection, or classification, strong correlations—especially clusters of red and blue—indicate dependencies between features. Potential links, such as statistical or protocol-based commonalities, are shown by the hierarchical grouping of features. Researchers can find redundant or strongly correlated features that could affect the performance of machine learning models with the help of this visualization.

The Synthetic Minority Over-sampling Technique (SMOTE) is used to create artificial examples of underrepresented classes to address the inherent class imbalance in intrusion detection datasets. This method lessens bias against the dominant class and improves the model's capacity to identify infrequent spoofing attempts.

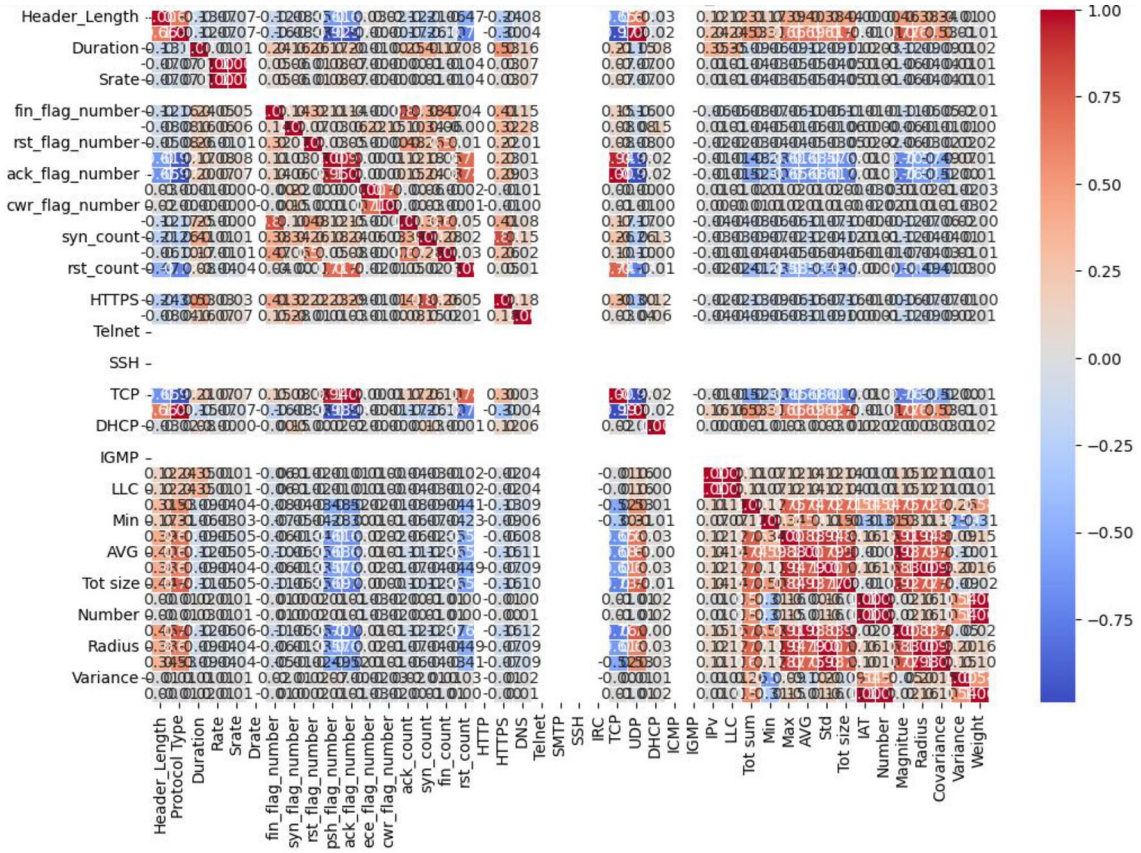


Fig. 2 Features correlations

2.3 The proposed model

The suggested study technique, which includes data preprocessing, model construction, training, evaluation, and interpretability, adheres to a methodically organized framework as shown in Fig. 3. The first step of the procedure is data acquisition and preparation, which involves importing the training and testing datasets and putting several crucial feature engineering strategies into practice as presented in the previous section. For instance, these procedures include converting the target variable into a binary representation, encoding category variables, and extracting relevant properties. MinMax normalization is used to standardize feature distributions after feature correlation analysis is performed to maintain data consistency and improve model efficiency. By the construction and definition of transformer-based model tailored for use in cybersecurity, the Custom Model stage is an important component of the pipeline. To enhance the classification performance, the stage begins by defining a custom attention mechanism that utilizes the self-attention component to dynamically assign importance to different input data. Transformer blocks that implement feed-forward networks and multi-head self-attention layers for efficient processing are employed to build the model. After the attention mechanism is established, the transformer-based model architecture is organized with the required elements like positional encoding to introduce sequential context, an embedding layer (if required) for the conversion of categorical data, and stacked transformer blocks consisting of self-attention, layer normalization, activation function (ReLU) in (3), and dropout layers to avoid overfitting.

$$\text{Relu} = \max(0, x) \tag{3}$$

After the architectural design is finished, the model is built using the Adam optimizer, which is well-known for its effective gradient updates and variable learning rate [35]. During compilation, an appropriate loss function (such as binary or categorical cross-entropy) is selected based on the classification goal, and evaluation metrics like accuracy, precision, recall,

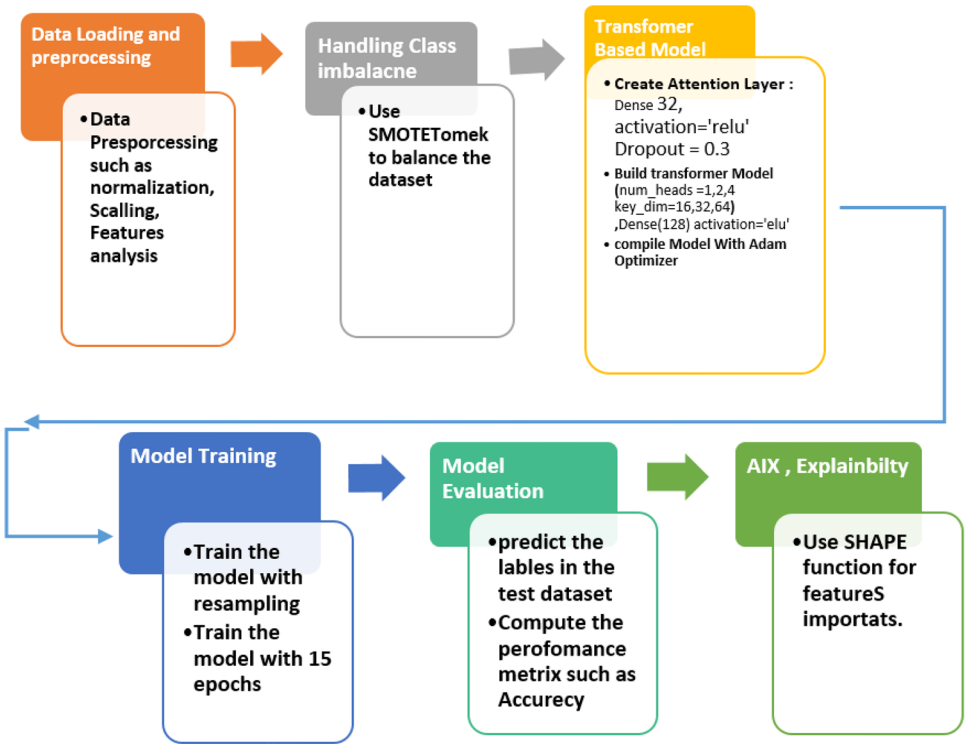


Fig. 3 The framework for the proposed model

and F1-score are used to monitor performance. Convergence during training can also be enhanced through learning rate scheduling. This Custom Model phase is essential for developing a deep learning model that employs transformers to enhance cybersecurity threat detection and ensure dependable performance in identifying cyberattacks within IoMT networks.

During the training phase, the model undergoes mini-batch supervised learning, which iterates through the dataset across multiple epochs to enhance its predictive abilities while lowering the risk of overfitting and underfitting. Following training, the model uses the test dataset to make predictions during the assessment phase. To assess how well it distinguishes between malicious and legitimate network activity, key performance metrics like accuracy and a comprehensive classification report are computed. The model's prediction behavior is further examined using performance visualization techniques. The learning dynamics and convergence features of the model can be better understood by charting training loss and accuracy trends over time. Additionally, by producing feature importance plots and SHAP summary visualizations that clarify the contributions of different input features to the model's decision-making process, SHapley Additive exPlanations (SHAP) is used to improve model interpretability.

Mainly, as shown in Fig. 3 the model incorporating self-attention, designed for binary classification of spoofing attacks. It starts with an input layer followed by a dense layer with 32, units, using ReLU activation. Layer normalization is applied, followed by a custom self-attention layer using Keras's MultiHeadAttention with 1, 2, or 4 heads and key dimensions of 16, 32, or 64. The attention mechanism treats the input features as a sequence to learn inter-feature dependencies. The attention output is normalized again, passed through a second dense layer with half the units of the first (128), and regularized with dropout at rates of 0.3. Finally, a dense output layer with a sigmoid activation provides the binary classification result.

Finally, additional evaluation metrics like precision, recall, and F1-score are computed to provide a comprehensive assessment of categorization performance. This comprehensive, multi-perspective assessment enhances the developed model's suitability for real-world use in spoofing attack detection in Internet of Medical Things (IoMT) systems by guaranteeing that it exhibits both robustness and generalizability.

2.4 Evaluation metrics for detecting spoofing attacks

2.4.1 Confusion matrix

Key metrics from the Confusion Matrix are used to evaluate model performance in machine learning-based [36] spoofing attack detection as shown in Fig. 4. These metrics include: True Positives (TP): Spoofing assaults that were correctly

detected. True Negatives (TN): Accurately identifying legitimate activity as non-attacks. False Positives (FP): Attacks that are legitimate activities that are mistakenly reported as such. False Negatives (FN): Missed attacks, or spoofing attempts that are mistakenly categorized as legitimate activity.

2.4.2 Precision, recall and F1-score

This part represents several performance metrics [37]. For instance, Precision termed positive predictive value is the fraction of relevant instances among the retrieved instances. Mainly, Precision evaluates how many of the predicted spoofing attacks were attacks, a high Precision guarantees that identified spoofing attempts are commonly real attacks, minimizing false alarms (4).

$$\text{Precision} = \text{TP}/\ \text{TP} + \text{FP} \tag{4}$$

Recall indicates the model's ability to correctly detect spoofing attacks, Recall (5) ensures that most spoofing attacks are detected, reducing undetected security breaches.

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{5}$$

The F1 Score balances the two measurements by taking the harmonic mean of Precision and Recall, F1 Score is essential for reliable detection without overwhelming security teams with false positives (6).

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}/\text{Precision} + \text{Recall}) \tag{6}$$

3 Results analysis

The suggested model is evaluated in this section using several performance indicators, including the previously mentioned accuracy, precision, recall, and F1-score. Additionally, it effectively illustrates the model's concert by presenting the notion of the model's correctness, model loss, and confusion matrix. Additionally, it provides the results of our tests, offering a valuable comprehension of the identification and elucidation. The study uses a binary classification framework to differentiate between spoofing assaults and typical activity.

4 Experimental analysis

According to Table 1, numerous tests have been conducted using different system characteristics (such as the number of attention heads, key dimensions, dropout rates, and number of neurons in thick layers).

The model is trained and evaluated for each trial, as shown in Fig. 5, and the results are displayed and compared to determine the best configuration. This line graph illustrates the model's performance in three experiments by calculating its F1 Score, Accuracy, Precision, and Recall. The accuracy is consistently high in all cases. The trends of precision and recall are inverse; in the first trial, precision increases while recall decreases, but eventually converges. This implies a typical categorization task trade-off between precision and recall. An intermediate pattern is followed by the F1 Score, which balances precision and recall and gets somewhat better throughout the studies. These differences suggest that the model tuning has been adjusted, maybe to optimize for various categorization performance factors.

The accuracy of the final configuration is illustrated in the Fig. 6. The blue and orange curves in the picture reflect the training and validation accuracy, respectively, and show the accuracy trends of a spoofing detection model during 15 training epochs. Both measures show a sharp rise in the first few epochs, with accuracy hitting almost 99% in the first five

Fig. 4 Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Table 1 Multiple experiments have been done with varying system parameters

Number of attention heads	Key dimensions	Dense_units	Dropout_rate
1	16	64	0.3
2	32	128	0.2
4	64	256	0.1

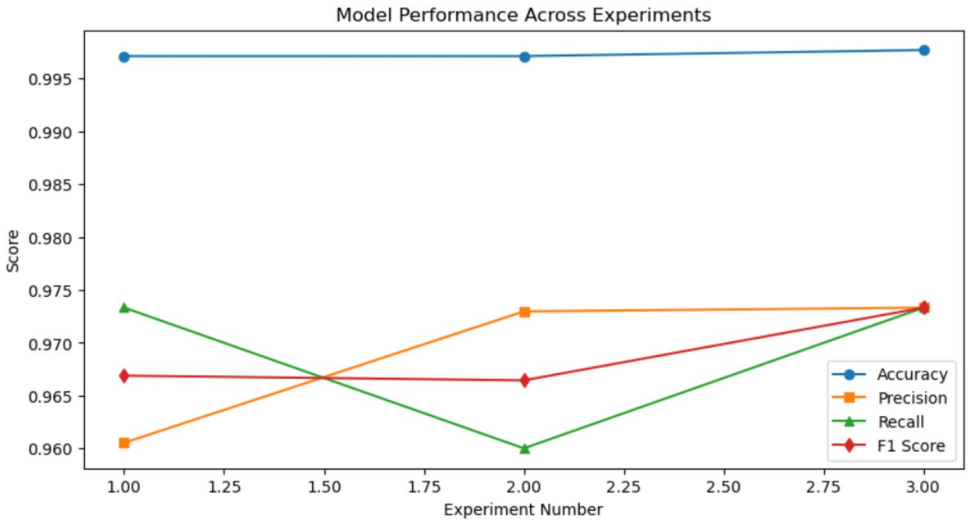


Fig. 5 The proposed model performance across experiments

epochs. This implies that the model successfully picks up patterns in the dataset, most likely as a result of a well-tuned hyperparameter and an optimal deep-learning architecture. Following this first stage, accuracy levels suggesting that the model can distinguish between real and fake traffic. Since the training and validation curves stay tightly matched throughout the procedure, little changes indicate the lack of considerable variance or overfitting. The model appears to generalize well rather than memorize training data, as evidenced by the absence of a performance gap between these metrics. Even though the almost flawless accuracy is encouraging, it is imperative to evaluate the model's resistance to hidden real-world spoofing attempts. Insufficient variety in the training sample could make it difficult for the model to identify more complex attack patterns found in real-world IoMT applications. Consequently, additional assessments using a variety of datasets and adversarial testing are required to guarantee the model's resilience and dependability in practical situations.

Furthermore, the loss values of the training and validation datasets during several epochs during a machine learning model's training process are depicted in Fig. 7. The number of epochs is indicated by the X-axis, while the loss is represented by the Y-axis. In the early epochs, both the training loss (blue line) and validation loss (orange line) show a sharp decline, suggesting that the model is rapidly learning and fine-tuning its parameters. Both loss curves gradually stabilize close to zero as training goes on, indicating that the model is effectively convergent.

Since there isn't any discernible divergence that would point to overfitting, the training and validation loss curves' tight alignment indicates that the model generalizes well to unseen data. The model has also reached a high degree of forecast accuracy, as seen by the near-zero loss values. After a few epochs, the validation loss may fluctuate slightly, which could be explained by model regularization effects or small changes in the dataset. With a balanced fit that reduces the hazards of both underfitting and overfitting, this loss curve shows that the model has been trained successfully overall.

In addition, the confusion matrix offers a performance assessment of the binary spoofing classification model as shown in Fig. 8. Class "0" in this context denotes genuine (non-spoof) instances, but class "1" denotes malicious (spoof) instances. There are four essential values in the confusion matrix. True Negatives (TN = 1667): 1667 valid cases were accurately identified by the model as non-spoof. False Positives (FP = 2): Two valid cases were incorrectly identified as spoofs by the model. False Negatives (FN = 3): Three spoof cases were mistakenly identified as genuine by the model. True Positives (TP = 72): Seventy-two spoofs were accurately detected by the model. However, with a low false positive rate, the model exhibits excellent accuracy and precision in spoofing attack detection. However, increasing recall (lowering false negatives) might increase the efficacy of spoofing detection even more. A greater balance between precision and recall could be achieved by modifying the decision threshold or using additional feature engineering strategies.

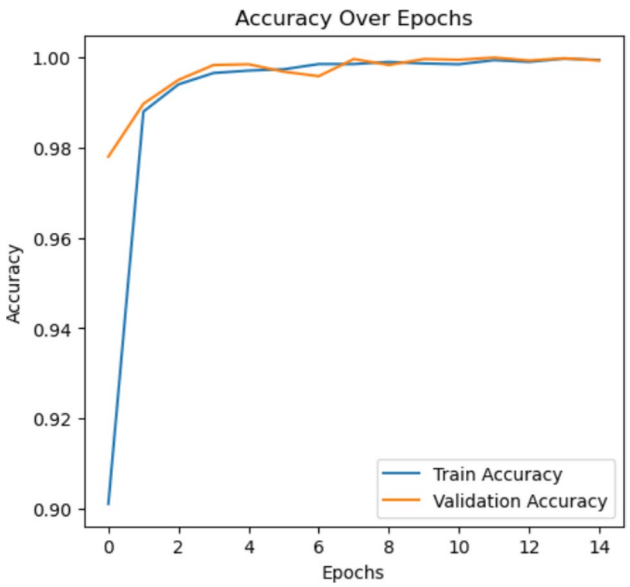


Fig. 6 The proposed model's accuracy over the number of epochs

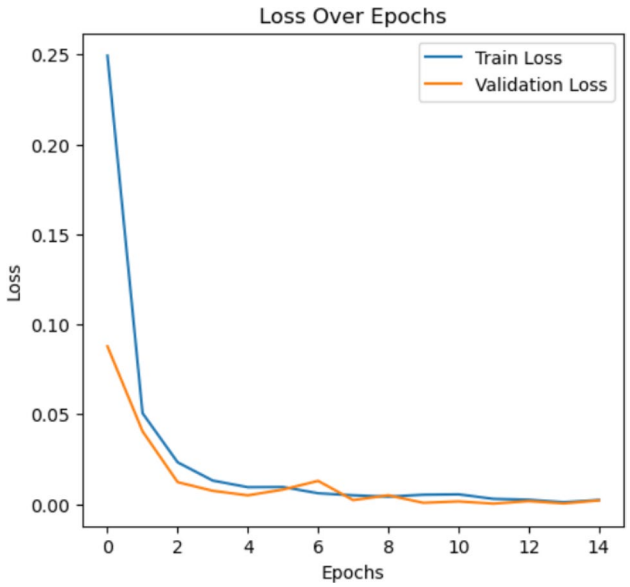


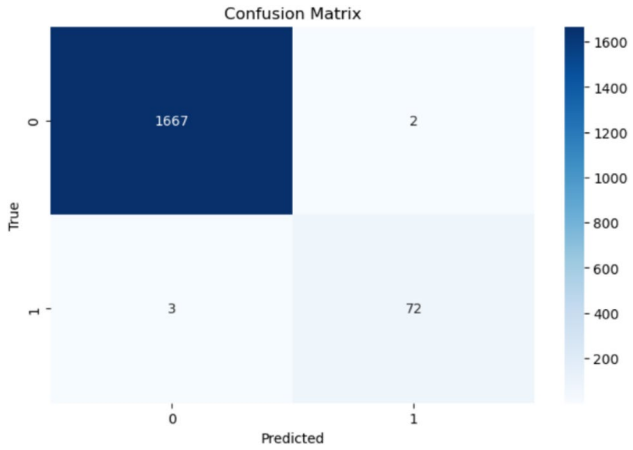
Fig. 7 The proposed model's Loss over the number of epochs

The Fig. 9 presents the SHAP summary plot and offers a comprehensible examination of feature significance and how it affects the model's predictions. The features are listed on the Y-axis in order of significance, with the most important features having the most influence on the model's decision-making. The SHAP values, which show the direction and strength of a feature's impact on the model's output, are represented by the X-axis. Predictions are pushed toward the negative class (such as typical conduct) by negative SHAP values, and toward the positive class (such as an attack or abnormality) by positive values.

Generally, this SHAP analysis improves the interpretability of the model by highlighting critical features and their role in classification decisions, thus increasing transparency and trust in the predictive outcomes. The color gradient, which ranges from blue (low feature values) to red (high feature values), provides insight into how different feature values affect predictions. Notably, features like UDP, Protocol Type, and ack_flag_number exhibit strong influences, with high values generally increasing the likelihood of a positive classification; on the other hand, features like IAT and LLC show mixed contributions depending on their values.

Using the CICIoMT2024 dataset, several models have been assessed for spoofing attack detection in Internet of Medical Things (IoMT) networks. The performance metrics of these models are summarized in Table 2.

Fig. 8 Confusion matrix



With an accuracy of 99.71% and an F1 Score of 96.64%, the Transformer model performs better in spoofing attack detection. For instance, The Random Forest model also yields outstanding results with an accuracy of 95.10%. Notably, the Random Forest model showed almost flawless accuracy in binary classification tasks, An LSTM-based model demonstrated a 97.6% accuracy rate across multiple classes, suggesting its potential utility in complex IoMT scenarios; and this justification can be noted in the rest of the models in Table 2 and Fig. 10. Even if more traditional models perform

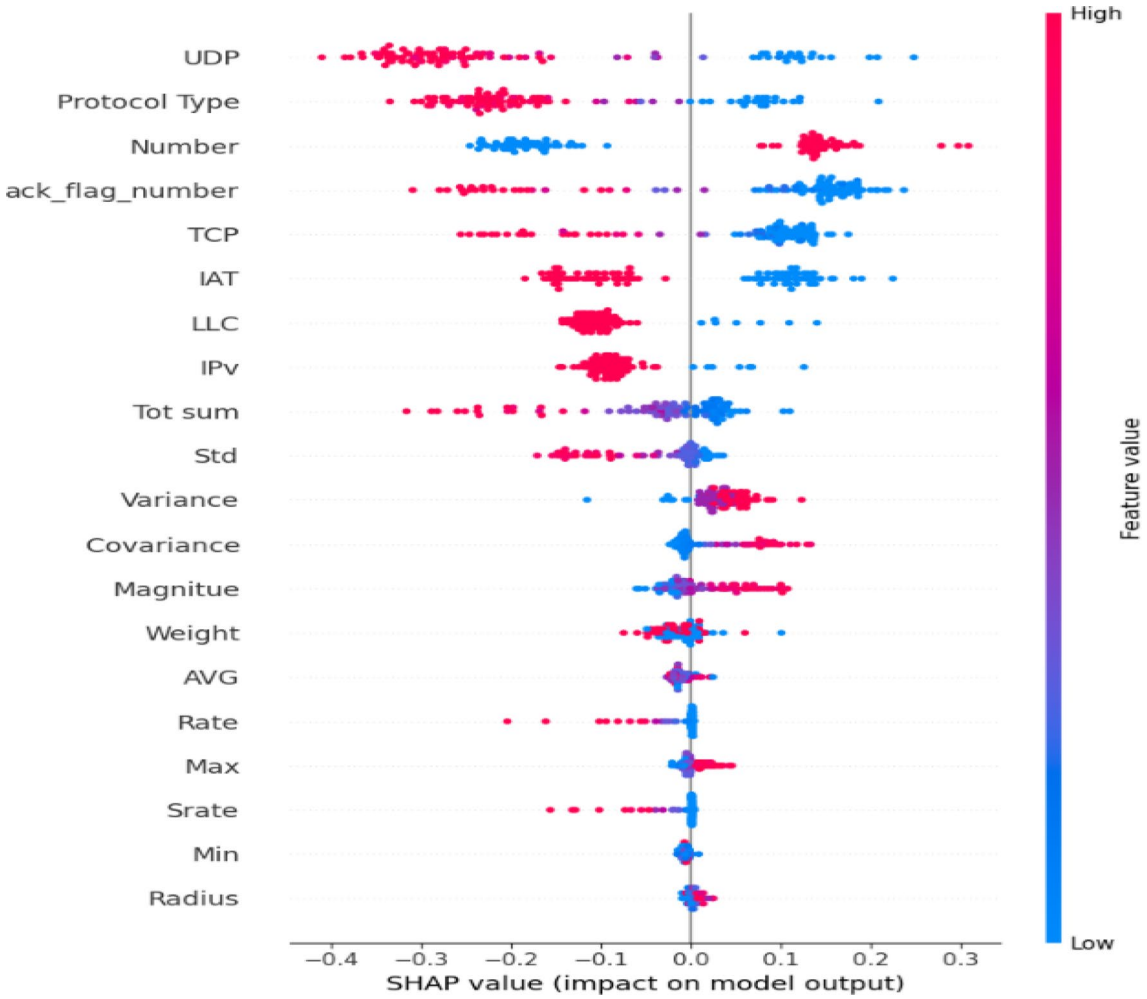


Fig. 9 Features Impact on model output based on SHAP function

Table 2 Comparative analysis between the proposed model and literature models

Model	Accuracy	Precision	Recall	F1 Score	Reference
Random Forest	95.10	95.10	95.10	95.10	[19]
Gradient Boosting	96.5	96	96	95	[38]
CNN, RNN, LSTM and fully Connected Networks (FCNs)	98.5	97.8	98.2	98	[39]
Long Short-Term Memory (LSTM)	97.06	97.06	97.06	96.60	[40]
Hybridization (CNN-LSTM)	97.63	98.47	97	97.73	[39]
The proposed Transformer Model	99.71	97.30	96.00	96.64	

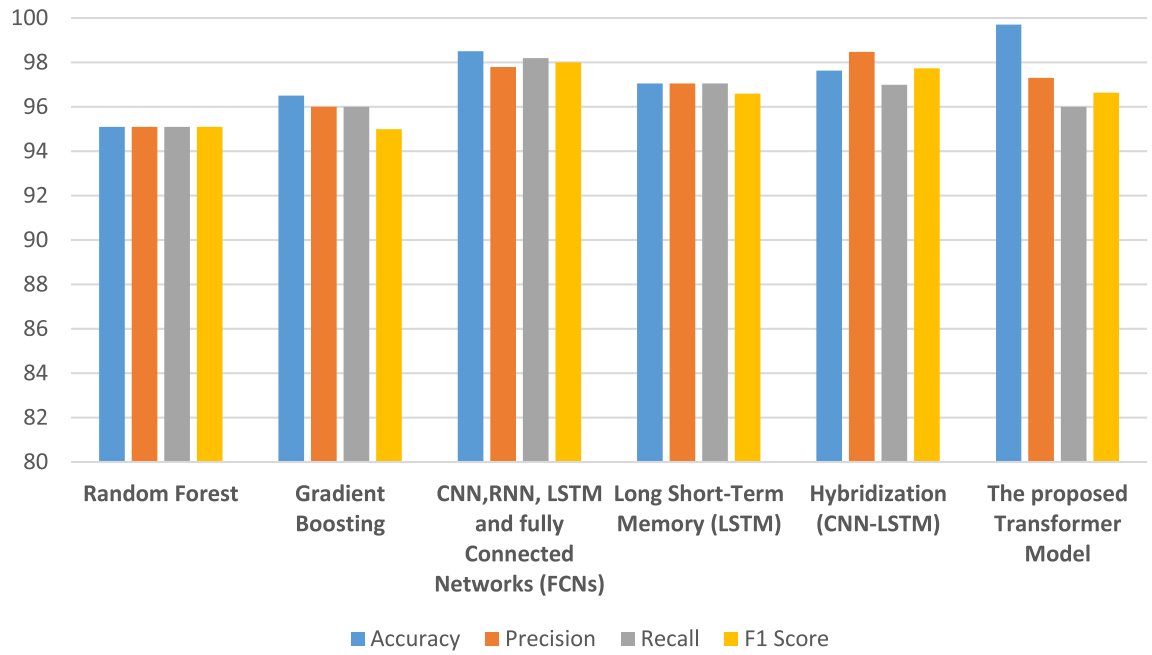


Fig. 10 A comparative study through the proposed model and works of literature models

well, complicated models like Transformers offer better accuracy and robustness overall, particularly in multi-class classification jobs within IoMT networks.

5 Conclusion

The employed Transformer-based deep learning technique demonstrates impressive competence in binary spoofing attack detection, effectively tackling critical network security challenges. The model implementation involved an intention layer to increase the performance level. In addition, this research utilized the application of SMOTETomek aids in balancing the class imbalance issue in the CIC IoMT2024 dataset, leading to balanced and accurate predictions. Furthermore, the employment of Explainable AI (XAI) methods, namely SHAP analysis, enables the proposed model to be more interpretable by highlighting the most significant features underlying its predictions. This level of transparency is a prerequisite for cybersecurity, building trust, facilitating model debugging, and ensuring adherence to security policies. As a consequence, the explainable AI-driven transformer model for spoofing attack detection provides more precise outcomes compared with the published models in the literature. Measures, including confusion matrix analysis and key classification metrics, validate the model's strength, with minimal overfitting and high generalization to unseen data. The findings point to the promise of Transformer-based models in real-time anomaly detection and proactive threat prevention. However, the model's performance may be sensitive to hyperparameter choices such as number of heads, key dimensions, dropout rate, and it might struggle to generalize in more complex real-world scenarios without further tuning or incorporation of temporal-aware architectures like Transformers for time-series. Future research can explore

hyperparameter tuning, integration of ensemble learning techniques, and computational efficiencies to enhance scalability and deployment on large cybersecurity networks.

Author contributions Conceptualization, M.A.; methodology; M.A.A; formal analysis, R.S.; investigation, M.O.; resources, F.A.E. and T.A; writing original draft preparation, M.A; writing—review and editing, M.A.A; supervision, A.A. and M.A.A.; project administration, M.A.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU251782).

Data availability The data can be accessed via the following link: http://205.174.165.80/IOTDataset/CICIoMT2024/Dataset/WiFI_and_MQTT/attacks/CSV/.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. El-Saleh AA, Sheikh AM, Albreem MA, Honnurvali MS. The Internet of Medical Things (IOMT): opportunities and challenges. Wirel Netw. 2024;31:1–18.
2. Shafiq M, Gu Z, Cheikhrouhou O, Alhakami W, Hamam H. The rise of "Internet of Things": review and open research issues related to detection and prevention of IoT-based security attacks. Wirel Commun Mob Comput. 2022;1:8669348.
3. Si-Ahmed A, Al-Garadi MA, Boustia N. Survey of machine learning based intrusion detection methods for Internet of Medical Things. Appl Soft Comput. 2023;140:110227.
4. Moustafa N, Koroniotis N, Keshk M, Zomaya AY, Tari Z. Explainable intrusion detection for cyber defences in the internet of things: opportunities and solutions. IEEE Commun Surv Tutor. 2023;25(3):1775–807.
5. George AS, George AH. The emergence of cybersecurity medicine: protecting implanted devices from cyber threats. Partn Univers Innov Res Publ. 2023;1(2):93–111.
6. Naeem MM, Hussain I, Missen MMS. A technique for safeguarding legitimate users from media access control (MAC) spoofing attacks. J Comput Biomed Inform. 2024;6(02):160–71.
7. Kochhar SK, Bhatia A, Tomer N. Using deep learning and big data analytics for managing cyber-attacks. In: Karthikeyan P, Katina PF, Anandaraj SP, editors. New approaches to data analytics and internet of things through digital twin. Hershey: IGI Global; 2023. p. 146–78.
8. Martins I, Resende JS, Sousa PR, Silva S, Antunes L, Gama J. Host-based IDS: a review and open issues of an anomaly detection system in IoT. Future Gener Computer Syst. 2022;133:95–113.
9. Alohali MA, Al-Wesabi FN, Hilal AM, Goel S, Gupta D, Khanna A. Artificial intelligence enabled intrusion detection systems for cognitive cyber-physical systems in industry 4.0 environment. Cognit Neurodyn. 2022;16(5):1045–110.
10. He K, Kim DD, Asghar MR. Adversarial machine learning for network intrusion detection systems: a comprehensive survey. IEEE Commun Surv Tutor. 2023;25(1):538–66.
11. Bakhshi T, Zafar S. Hybrid deep learning techniques for securing bioluminescent interfaces in internet of bio nano things. Sensors. 2023;23(21):8972.
12. Meliboev A, Alikhanov J, Kim W. Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets. Electronics. 2022;11(4):515.
13. Montesinos López OA, Montesinos López A, Crossa J. Overfitting, model tuning, and evaluation of prediction performance. In: López OAM, López AM, Crossa J, editors. Multivariate statistical machine learning methods for genomic prediction. Cham: Springer; 2022. p. 109–39.
14. Tunde-Onadele O, Lin Y, Gu X, He J, Latapie H. Self-supervised machine learning framework for online container security attack detection. ACM Trans Auton Adapt Syst. 2024;19(3):1–28.

15. Sohail F, Bhatti MAM, Awais M, Iqtidar A. Explainable boosting ensemble methods for intrusion detection in Internet of Medical Things (IoMT) applications. In: 24 4th International Conference on Digital Futures and Transforma. 2024.
16. Gobinath A, Rajeswari P, Suresh Kumar N, Anandan M. Internet of Medical Things (IoMT). In: Kumar A, Gupta M, Sharma S, Sharma EH, Aurangzeb K, editors. Smart hospitals: 5g, 6g and moving beyond connectivity. Hoboken: Wiley; 2024. p. 91–105.
17. Aslan Ö, Aktuğ SS, Ozkan-Okay M, Yilmaz AA, Akin E. A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. Electronics. 2023;12(6):1333.
18. Jayaraj IA, Shanmugam B, Azam S, Thennadil S. Detecting and localizing wireless spoofing attacks on the Internet of Medical Things. J Sens Actuat Netw. 2024;13(6):72.
19. The Canadian Institute for Cybersecurity (CIC). The Canadian Institute for Cybersecurity (CIC). 2024. https://www.unb.ca/cic/datasets/iomt-dataset-2024.html. Accessed 2 Feb 2025.
20. Çetin A, Öztürk S. Comprehensive exploration of ensemble machine learning techniques for IoT cybersecurity across multi-class and binary classification tasks. J Future Artif Intell Technol. 2025;1(4):371–84.
21. Akar G, Sahmoud S, Onat M, Cavusoglu Ü, Malondo E. L2D2: a novel LSTM model for multi-class intrusion detection systems in the era of IoMT. IEEE Access. 2025. https://doi.org/10.1109/ACCESS.2025.3526883.
22. Maroof U, Batista G, Shaghaghi A, Jha S. Detecting Iot event spoofing attacks using time-series classification. Available at SSRN 4922101.
23. Dadkhah S, Neto EC. CICIoMT2024: a benchmark dataset for multi-protocol security assessment in IoMT. Internet Things. 2024;28:10135.
24. Sánchez N, Calvo A, Escuder S, Escrig J, Domenech J, Ortiz N, Mhiri S. Towards enhanced IoT security: advanced anomaly detection using transformer models. In Proc. KDD 4th Workshop Artif Intell Enabled Cybersecurity Anal.
25. Tareq Al-Halboosi I, Elbagoury BM, El-Regaily SA, El-Horbaty ESM. A hybrid-transformer-based cyber-attack detection in IoT networks. Int J Interact Mob Technol. 2024;18(14):90.
26. Naeem H, Alsirhani A, Alserhani FM, Ullah F, Krejcar O. Augmenting Internet of Medical Things Security: deep ensemble integration and methodological fusion. Computer Model Eng Sci. 2024;141(3):2185–223.
27. Chamola V, Hassija V, Sulthana AR, Ghosh D, Dhingra D, Sikdar B. A review of trustworthy and explainable artificial intelligence (XAI). IEEE Access. 2023;11:78994–9015.
28. Neupane S, Ables J, Anderson W, Mittal S, Rahimi S, Banicescu I, Seale M. Explainable intrusion detection systems (x-ids): a survey of current methods, challenges, and opportunities. IEEE Access. 2022;10:112392–415.
29. Ullah F, Mostarda L, Cacciagrano D, Alenazi MJ, Chen CM, Kumari S. Federated edge intelligence for enhanced security in consumer intermittent healthcare devices using adversarial examples. IEEE Trans Consum Electron. 2024. https://doi.org/10.1109/TCE.2024.3511615.
30. Gurcan F, Soylu A. Learning from imbalanced data: integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. Cancers. 2024;16(19):3417.
31. Ekanayake IU, Meddage DPP, Rathnayake U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). Case Stud Constr Mat. 2022. https://doi.org/10.1016/j.cscm.2022.e01059.
32. Haldorai A, Ramu A. Canonical correlation analysis based hyper basis feedforward neural network classification for urban sustainability. Neural Process Lett. 2021;53(4):2385–401.
33. Ahsan MM, Mahmud MP, Saha PK, Gupta KD, Siddique Z. Effect of data scaling methods on machine learning algorithms and model performance. Technologies. 2021;9(3):52.
34. Li G, Zhang A, Zhang Q, Wu D, Zhan C. Pearson correlation coefficient-based performance enhancement of broad learning system for stock price prediction. IEEE Trans Circuits Syst II Express Br. 2022;69(5):2413–7.
35. Reyad M, Sarhan AM, Arafa M. A modified Adam algorithm for deep neural network optimization. Neural Comput Appl. 2023;35(23):17095–112.
36. Sathyanarayanan S, Tantri BR. Confusion matrix-based performance evaluation metrics. Afr J Biomed Res. 2024;27:4023–31.
37. Cahyani DE, Putra AW. Relevance classification of trending topic and twitter content using support vector machine. In: 2021 International Seminar on Application for Technology of Information and Communication (iSemantic). 2021. p. 87.
38. Dahouda MK, Joe I. A deep-learned embedding technique for categorical features encoding. IEEE Access. 2021;9:114381–91.
39. Alsubaei FS, Almazroi AA, Ayub N. Enhancing phishing detection: a novel hybrid deep learning framework for cybercrime forensics. IEEE Access. 2024. https://doi.org/10.1109/ACCESS.2024.3351946.
40. Chen P, Liu H, Xin R, Carval T, Zhao J, Xia Y, Zhao Z. Effectively detecting operational anomalies in large-scale IoT data infrastructures by using a GAN-based predictive model. Computer J. 2022;65(11):2909–25.