SG-VAE: Explainable Semi-Generative Model for Attack Detection in IoT Healthcare Systems

Komal Ghafoor*, Furquan Rustam†, Anca Delia Jurcut†, Devis Bianchiniè*
*Department Information Engineering, Università degli Studi di Brescia (Italy),
†School of Computer Science, University College Dublin (Ireland)
{komal.ghafoor, devis.bianchini}@unibs.it, furqan.rustam@ucdconnect.ie, anca.jurcut@ucd.ie

***Abstract*—The integration of Internet of Things (IoT) technologies in healthcare has revolutionized patient monitoring and medical data collection, but it has also introduced new vulnerabilities to cyberattacks. Ensuring the security of these connected medical devices is essential to maintain patient safety and system reliability. In this study, we propose a lightweight semi-generative deep learning framework based on a Variational Autoencoder (SG-VAE) for detecting attacks in IoT-based healthcare systems. The model compresses high-dimensional network traffic into a latent representation using the VAE encoder and employs a classification head to detect potential cyber threats directly from the latent space. Unlike conventional generative models focused on input reconstruction, our SG-VAE leverages the generative structure purely for efficient and effective classification, resulting in a high-throughput, real-time system in resource-constrained environments. To enhance transparency and interpretability, we integrate two eXplainable AI (XAI) techniques, SHAP and LIME, to provide insights into model decisions, helping stakeholders understand which features influence predictions most. Experimental results on a publicly available IoT-based ICU healthcare security dataset demonstrate that our model achieves near-perfect accuracy, 1.00, and high throughput, 54,286 samples/second, making it both highly effective and practical for deployment in real-world healthcare systems. All code required to reproduce the experiments is publicly available on Dropbox[1].**

***Index Terms*—IoT, Healthcare, Explainable AI, Machine Learning, Network Security, Artificial Neural Network**

I. Introduction

The Internet of Things (IoT) connects physical devices like sensors and actuators to the internet, enabling automation and data exchange [1]. This connectivity enables automation, real-time monitoring, and intelligent decision-making. IoT has been widely adopted across various sectors, improving operational efficiency and enabling data-driven services [2]. Especially in the healthcare domain, the Internet of Medical Things (IoMT) is playing a critical role in transforming medical systems [3]. IoMT technologies have enhanced patient monitoring, enabled real-time data collection, and facilitated automated control of essential medical equipment such as IV bags, ventilators, and vital sign monitors [4]. However, this increased connectivity introduces significant cybersecurity challenges [5]. Medical IoT devices often operate with limited resources in life-critical settings, making them prime targets for cyberattacks [6]. A compromised IV bag sensor, for instance, can result in the delivery of expired or incorrect medication, posing serious risks to patient safety [7]. According to a 2025 report by Sean Blanton[2], cyberattacks on IoMT surged by 123%, with breach costs reaching up to $10 million per incident.

IoMT devices are vulnerable to cyberattacks such as malware, data tampering, and man-in-the-middle attacks, which can disrupt patient care and compromise sensitive data [4]. For example, ransomware can disable critical systems like IV pumps, leading to delays or unsafe treatments [8]. To mitigate these risks, researchers have proposed lightweight AI-based intrusion detection systems (IDS) that can be deployed at the edge [9]. However, traditional IDS still struggles to handle the high volume and speed of IoMT traffic in resource-constrained environments, often resulting in delayed or missed detections [10]. Additionally, the lack of explainability in traditional IDS limits trust and interpretability [11]. Therefore, there is a need for an approach that offers high accuracy, high throughput, low computational cost, and robust explainability to ensure a trustworthy and reliable healthcare security system.

To address the challenges of achieving high accuracy, operating in resource-constrained environments, and ensuring model trustworthiness, this study proposes a lightweight semi-generative model that offers high accuracy, low computational cost, and efficient resource usage, along with explainability features. Specifically, we deploy a semi-generative approach with an integrated classification layer. For explainability, we incorporate eXplainable AI (XAI) techniques such as SHAP and LIME to understand which features most influence attack predictions and to interpret the model's decisions. The key contributions of this study are as follows:

- We propose a lightweight semi-generative Variational Autoencoder (SG-VAE) for attack detection in IoMT. SG-VAE eliminates the need for a decoder during inference, making it suitable for deployment in resource-constrained networks like IoMT.
- We integrate XAI techniques (SHAP[3], LIME[4]) to enhance transparency and interpretability of the model. These techniques show that features such as checksum,

The part of the study carried out by the University of Brescia members has been funded within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004).

[1]https://rb.gy/5y3kuh

[2]https://jumpcloud.com/blog/iot-security-risks-stats-and-trends-to-know-in-2025

[3]https://shap.readthedocs.io/en/latest/

[4]https://christophm.github.io/interpretable-ml-book/lime.html

979-8-3503-9281-4/25/$31.00 ©2025 IEEE

MQTT header length, and topic length contribute positively to detecting attacks, increasing user trust.
- We evaluate the SG-VAE in terms of performance, achieving approximately 1.00 mean accuracy with a standard deviation of ± 0.00, using only 17.39% memory, and delivering a high throughput of 54,286 samples/second.

The structure of this paper is organized as follows: Section II reviews related work, Section III outlines the methodology, Section IV describes the proposed SG-VAE framework, and Section V describes baseline methods. Section VI presents results and discussion, and Section VII concludes the study.

II. Related Work

In this section, we examine recent work on IoMT security and AI-based IDS for IoMT. T. Ganai et al. [12] highlight key cybersecurity challenges in healthcare IoT, including device heterogeneity and resource limits, while N. Sharma and N. Jindal [13] explore the integration of AI, IoT, and cybersecurity, supporting deep learning applications in clinical systems.

To mitigate IoMT security challenges, researchers have proposed several machine learning and deep learning frameworks. Hussain et al. [14] employed classical classifiers such as Random Forest, XGBoost, and ANN to detect malicious traffic in smart healthcare systems, achieving high accuracy, but offering limited interpretability or generative insights. Khan et al. investigated anomaly detection in healthcare IoT using ensemble models, though their framework lacked support for explainable outputs or lightweight adaptability [15].

Several recent studies have explored deep learning and metaheuristic approaches for securing IoT-based healthcare systems. Kacem et al. [16] proposed a GRU-based detection framework using the IoT-Flock dataset to classify cyberattacks in healthcare IoT, showing superior performance over RNN, CNN, and LSTM alternatives, achieving high accuracy and AUC-ROC scores. Goswami et al. [17] developed a metaheuristic-driven IDS, the Lion-Salp-Swarm-Optimization Algorithm (LSSOA), also using IoT-Flock data. By combining multiple optimization techniques, their system effectively reduced false positives and improved detection performance in IoMT environments. Similarly, the study [18] applied Deep Gated Recurrent Unit (D-GRU) on the IoT health dataset to detect anomalies, emphasizing the efficacy of GRU architectures in modeling healthcare traffic patterns and improving attack detection rates. Chakraborty et al. [19] proposed a multi-source transfer learning approach to healthcare cybersecurity, showing promising detection capabilities across distributed data environments. However, the study fell short in leveraging interpretable frameworks or generative paradigms. In contrast, Algethami and Alshamrani [20] developed a hybrid deep learning model combining ANN, GRU, and BiLSTM, achieving near-perfect performance and incorporating explainability, making it one of the few approaches aligned with clinical trustworthiness standards. Similarly, Nasayreh et al. [21] demonstrated the effectiveness of intelligent feature extraction pipelines in identifying healthcare cyberattacks using advanced deep learning techniques.

Among recent advances, deep generative models, particularly Variational Autoencoders (VAEs), have gained attraction for anomaly detection J. Rheey and H. Park [22] applied a spatial-temporal VAE to healthcare IoT data, achieving strong detection performance on ICU sensor streams; however, their approach remains limited due to its reliance on reconstruction error and lack of integration with real-time classification pipelines. In contrast, Manoharan and Thathan [23] introduced the GTPDA model, which combines Group Teaching Optimization with a Conditional Probabilistic Deep Autoencoder to improve intrusion detection in IoMT. Their model achieved 99% accuracy across multiple benchmark datasets, showcasing the effectiveness of optimized feature selection and probabilistic deep learning for IoMT security.

Interpretability remains a key requirement in healthcare applications. L. Antwarg et al. [24] pioneered the use of Kernel SHAP for interpreting autoencoder-based anomaly scores, while B. Sharma et al. [25] later incorporated LIME and SHAP into IDSs to enhance trust and transparency. Despite these advances, few studies have embedded explainability directly into generative modeling for healthcare-specific use cases. Recent efforts have shifted toward integrating deep learning and XAI for IoMT security. Kalakoti et al. [26] proposed a Transformer-based IDS enhanced with XAI techniques like LIME and SHAP, demonstrating high accuracy on the CICIoMT2024 dataset and evaluating explanation quality using faithfulness, sensitivity, and complexity metrics. Similarly, Altrad et al. [27] presented a lightweight ML-based framework utilizing the same dataset, focusing on real-time deployment with LIME-based interpretability and comparative performance across multiple classifiers. Alsharaiah et al. [28] addressed the critical issue of spoofing attacks in IoMT by designing a custom attention-based Transformer model, combined with SMOTE-Tomek preprocessing and SHAP, achieving 99.71% accuracy. Table I shows the summary of related work.

TABLE I
Comprehensive Related Work Summary

Year	Ref.	Model	Accuracy	Limitation	IoT	Med	XAI	Gen.
2021	[14]	RF, NB, XGB, ANN	99.5%	No XAI, no Gen. model	✓	✗	✗	✗
2022	[29]	LR, KNN, SVM, NB, RF, DT, ANN	99%	IIoT testbed, lacks IoT traffic	✓	✗	✗	✗
2023	[19]	Multi-source TL	∼96%	No XAI/GenAI	✓	✗	✗	✗
2024	[16]	GRU	99%	Limited interpretability, No interpretability	✓	✓	✗	✗
2024	[17]	LSSOA	99.59%	Complex optimization	✓	✓	✗	✗
2024	[15]	RF, AdaBoost, LR, Perceptron, DNN	∼96%	Single CIC IoT dataset	✓	✓	✗	✗
2024	[30]	MIC-XGBoost	95.01%	Non-IoT dataset, no XAI/GenAI	✗	✗	✗	✗
2024	[23]	GTPDA	99.100%	No XAI, no interpretability	✓	✓	✗	✗
2024	[26]	Transformer + SHAP, LIME	99%+	No generative model	✓	✓	✓	✗
2025	[18]	D-GRU	96.33	No interpretability, limited generalization	✓	✓	✗	✗
2025	[27]	RF, GB, LR, SVM + LIME	97% (RF/GB)	Simple XAI, no deep models	✓	✓	✓	✗
2025	[28]	Transformer + SHAP	99.71%	No global XAI	✓	✓	✓	✗
2025	Our	SG-VAE, SHAP, LIME	$\sim$1.00 $\pm$0.00	No Real-Time Testing	✓	✓	✓	✓

Limitations & gaps. Despite notable advancements, the current landscape of IoT healthcare security remains fragmented. Few existing approaches effectively integrate deep generative modeling, real-time detection, and XAI. This gap underscores the need for unified solutions like the proposed SG-VAE framework, which combines the compactness and

robustness of variational encodings with enhanced decision interpretability via SHAP and LIME. Moreover, SG-VAE is trained and validated on a realistic ICU dataset, ensuring high predictive accuracy while maintaining transparency and efficiency, key requirements for deployment in resource-constrained, safety-critical medical environments.

III. Proposed Methodology

This study proposes a lightweight SG-VAE approach to protect IoT-based healthcare systems from cyberattacks using an AI-based framework, as illustrated in Figure 1. The methodology consists of several steps: dataset acquisition, data preprocessing, data scaling, model training, evaluation, and explainability.

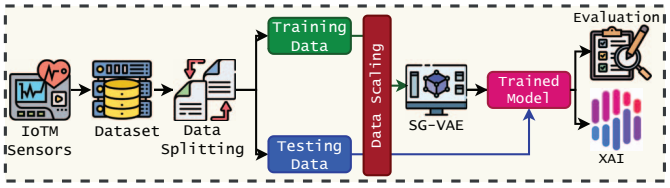

Fig. 1. Overview of the proposed methodology

Dataset description. We utilized the IoT Healthcare Security Dataset (IoT-Flock)[5], which simulates a smart ICU scenario with two beds, each containing nine patient-monitoring sensors and one Bedx-Control-Unit [14]. These devices were generated using the IoT-Flock tool. The dataset contains both normal and malicious network traffic, organized across three files: `attack.csv` (malicious), `patientMonitoring.csv`, and `environmentMonitoring.csv` (both normal). Dataset samples count are shown in Table II

TABLE II
Dataset Distribution by Traffic Type

Traffic Type	Sample Count
Attack Traffic	80,126
Patient Monitoring (Normal)	76,810
Environment Monitoring (Normal)	31,758
Total Samples	188,694

Data Splitting. After preprocessing, the dataset was divided into training and testing sets using an 80:20 ratio. This ensures the model is trained on a majority of the data and tested on unseen samples. Additionally, 5-fold cross-validation was applied to ensure robustness and generalization. Performance was evaluated using mean accuracy and standard deviation (STD). Table III shows the class distribution of the data.

Preprocessing. We applied data normalization using `StandardScaler` to scale numerical features and improve model convergence and performance. Categorical features were transformed using `LabelEncoder` to convert them into a numeric form suitable for model input. This step ensures that the input features are on the same scale, which is crucial for efficient model training.

[5]https://www.kaggle.com/datasets/faisalmalik/iot-healthcare-security-dataset

TABLE III
Class Distribution in Training and Testing Sets

Class Label	Training Samples	Testing Samples
Normal	87,036	21,532
Attack	63,919	16,207
Total	150,955	37,739

Input Features. Each traffic record in the IoT-Flock dataset is represented by 34 features derived from packet headers and MQTT metadata. These include generic network-layer attributes (e.g., `ip.ttl`, `ip.proto`, `tcp.srcport`) and IoT-specific fields (e.g., `mqtt.msgtype`, `mqtt.hdrflags`, `mqtt.topic_len`). Together, these features capture both low-level traffic behavior and application-layer semantics, enabling the model to distinguish normal and malicious activities effectively.

Model Training & Evaluation. The proposed SG-VAE model was trained on the training set and evaluated on the test set. The SG-VAE leverages the encoder of a variational autoencoder to learn compact and informative latent representations of the data (see Section IV), making it ideal for resource-constrained IoT environments. Additionally, 5-fold cross-validation was conducted to reduce overfitting and assess model generalizability. The performance metrics include mean accuracy, STD, CPU usage, memory consumption, training time, and throughput to evaluate system efficiency.

Explainability. To enhance model transparency and build trust, we employed SHAP and LIME explainability methods. SHAP (SHapley Additive exPlanations) provides a game-theoretic approach to explain individual predictions, while LIME (Local Interpretable Model-agnostic Explanations) offers local approximations of the model's decision boundaries. These tools helped identify key features that contribute most significantly to the classification decision, making the model more interpretable and suitable for deployment in critical healthcare environments.

IV. Semi-Generative VAE (SG-VAE)

To enhance model transparency and build trust, we employed two complementary eXplainable AI (XAI) methods: SHAP and LIME. SHAP (SHapley Additive exPlanations) assigns each feature a Shapley value, derived from cooperative game theory, to quantify its contribution to the model output [31]. In our experiments, SHAP highlighted features such as `checksum`, `MQTT header length`, and `topic length` as strong indicators of malicious traffic. LIME (Local Interpretable Model-agnostic Explanations), on the other hand, builds a simple local surrogate model around a single prediction to approximate the decision boundary [32]. For example, LIME identified `tcp.srcport` and `ip.proto` as key drivers of benign classifications in specific instances. Together, SHAP provides a global, consistent view of feature importance across the dataset, while LIME offers instance-level inter-

pretability. This combination strengthens the trustworthiness of SG-VAE in safety-critical IoT healthcare environments.

Encoder. The encoder maps the input $\mathbf{x}$ into a latent space by producing two outputs: the mean $\boldsymbol{\mu}$ and the log-variance $\log \boldsymbol{\sigma}^2$ of a Gaussian distribution. These are computed using two parallel dense layers. Instead of sampling directly from this distribution, which would break backpropagation, we use the reparameterization trick. This trick expresses the latent variable as $\mathbf{z}$ of the input data. $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. This allows the model to learn a compressed, but informative, representation

Latent sampling. Instead of creating a fixed latent vector, SG-VAE learns a distribution and samples from it during training. This approach helps the model generalize better by introducing controlled randomness. We use the standard VAE trick to sample from the distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$.

Classifier and loss function. The latent vector $\mathbf{z}$ is passed through a small neural network that outputs a probability score using a sigmoid function. This score indicates the likelihood that the input represents an attack or not, which is formatted as: $\hat{y} = \sigma(h(\mathbf{z}))$, where $\mathbf{z}$ is the latent representation, σ is the sigmoid activation function defined as $\sigma(z) = \frac{1}{1+e^{-z}}$, and $\hat{y}$ is the predicted probability of the input being an attack. The model is trained using a combination of three loss components. First, the classification loss is computed using binary cross-entropy between the true label y and the predicted label $\hat{y}$. Second, the Kullback-Leibler (KL) divergence encourages the latent space to remain close to a standard normal distribution. Third, the reconstruction loss (used only during training, not inference) measures how well the model can reconstruct the input from the latent vector. The total loss is a weighted sum of these components as $\mathcal{L}_{total} = \mathcal{L}_{cls} + \beta_1 \cdot \mathcal{L}_{KL} + \beta_2 \cdot \mathcal{L}_{recon}$.

During deployment, the decoder is skipped to reduce inference computational overhead. The model focuses solely on leveraging the latent features for classification, making it efficient and suitable for real-time inference on IoMT devices. Figure 2 presents the SG-VAE architecture, which includes an encoder, latent space, and classifier.

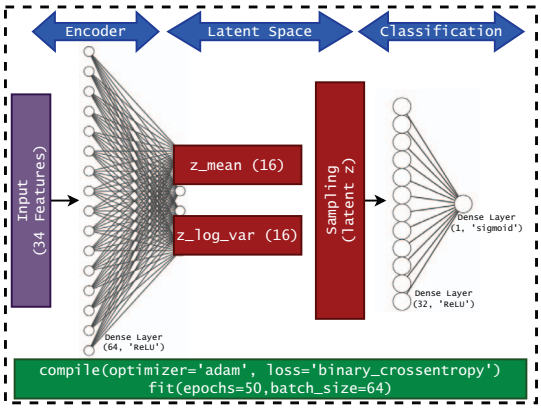

Fig. 2. Proposed SG-VAE Architecture

The encoder consists of a dense layer with 64 neurons and a ReLU activation function. We use ReLU activation due to its computational simplicity and ability to avoid vanishing gradients. Single-layer encoder followed by a 16-dimensional latent space, balancing expressiveness and efficiency. The classifier uses a 32-neuron hidden layer and an output layer with 1 neuron and a sigmoid activation function to produce a binary prediction. Using 64 and 32 neurons keeps the architecture lightweight and suitable for resource-constrained IoMT environments. The model is trained using the Adam optimizer, chosen for its adaptive learning rate and fast convergence, with binary cross-entropy loss, over 50 epochs and a batch size of 64.

V. Baseline Models

To benchmark the performance of the proposed SG-VAE model, we evaluated a comprehensive set of traditional machine learning and deep learning models in terms of accuracy, throughput, and resource consumption.

Traditional Machine Learning Models. We selected a diverse range of classical classifiers known for their applicability in tabular classification tasks [33]: (*i*) *Logistic Regression (LR)*, a simple yet effective linear model, suitable for binary classification; (*ii*) *K-Nearest Neighbors (KNN)*, a non-parametric model using 7 neighbors to classify based on distance; (*iii*) *Support Vector Machine (SVM)*, that utilizes an RBF kernel for capturing non-linear relationships; (*iv*) *Naive Bayes (NB),* a probabilistic classifier based on Gaussian likelihood assumptions.

Deep Learning Baselines. To provide a deep learning comparison, we implemented two widely used architectures: (*i*) *LSTM*, a stacked Long Short-Term Memory network composed of two LSTM layers (64 and 32 units), followed by a dense layer with 64 neurons and ReLU activation, and a sigmoid output layer; this model is suitable for temporal dependencies in sequence-like tabular data; (*ii*) *CNN*, a 1D convolutional neural network composed of two Conv1D layers (32 and 64 filters), a MaxPooling layer, Global Max Pooling, and two dense layers (64 ReLU + 1 Sigmoid). This model is useful for learning spatial hierarchies of features. All deep learning models were trained for 50 epochs with a batch size of 64 using the Adam optimizer and binary cross-entropy loss.

VI. Results & Discussion

Table IV presents the performance comparison between SG-VAE and several baseline models across multiple metrics. The proposed SG-VAE model delivers perfect accuracy with zero standard deviation, demonstrating robust classification performance. It also achieves the highest throughput at 54,286 samples per second while consuming only 17.39 MB of memory, making it both accurate and computationally efficient.

In comparison, deep learning models like CNN and LSTM also achieved perfect accuracy but required significantly more computation time and memory, which limits their practicality in real-time, resource-constrained IoMT deployments. CNN and LSTM performed poorly in terms of throughput, achieving only 25,105 and 1,827 samples per second, respectively, far

from ideal for real-time scenarios. On the other hand, traditional machine learning models such as LR and NB offered fast inference and low memory usage, but fell slightly short in terms of accuracy. Compared to SOTA approaches, SG-VAE offers the best balance across all performance metrics. Figure 3 illustrates per-epoch accuracy and loss values for SG-VAE.

TABLE IV
MODELS' PERFORMANCE FOR ATTACK DETECTION

Model	CT	TP	MU	CPU	MA±STD
SG-VAE	570.88	54286	17.39	92.50	1.00± 0.0000
CNN	1222.44	25105	60.71	101.50	1.00± 0.0000
LSTM	8695.17	1827	162.12	135.50	1.00± 0.0000
LR	22.4341	172572	47.29	44.50	0.98± 0.0008
NB	5.01	170257	45.34	2.50	0.99±0.0002
KNN	186.90	725	57.68	22.50	1.00± 0.0000
SVM	2053.46	2849	21.54	12.50	0.99±0.0003

CT – Computation Time (training time in seconds); **TP** – Throughput (number of samples processed per second); **MU** – Memory Usage (in MB); **CPU** – CPU Usage (in %); **MA** – Mean Accuracy (range: 0 to 1); **STD** – ± Standard Deviation.

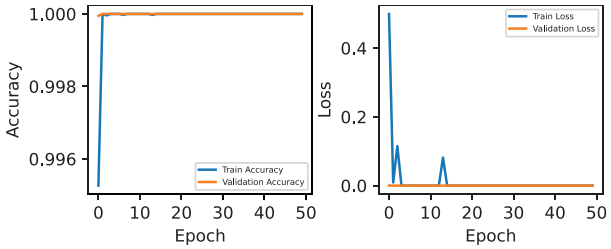

Fig. 3. SG-VAE Per Epoch Performance

A. SG-VAE Explainability

Figure 4 illustrates the SHAP waterfall explanation for a single test sample classified as benign. The base value (model bias) starts at 0.46, and several features push the prediction downward, resulting in a final output of 0.00 (normal traffic). The most influential features that decrease the predicted attack probability include `ip.ttl`, `tcp.srcport`, and `tcp.window_size_value`, with their SHAP values contributing negatively to the decision. Conversely, features such as `tcp.flags.push`, `mqtt.msgtype`, and `mqtt.hdrflags` slightly increase the attack probability, but not sufficiently to cross the classification threshold. This fine-grained interpretation highlights the transparency of our SG-VAE model and how individual feature contributions align with domain knowledge, aiding clinical and network experts in validating model trustworthiness.

Furthermore, figure 5 shows the LIME explanation for a single prediction classified as normal traffic with 100% confidence. The most influential features contributing to this classification include `f5-tcp.srcport`, `f49-ip.proto`, and `f19-tcp.flags.syn`, all having negative values, which pull the decision toward the benign class. LIME provides an interpretable linear approximation of the local decision boundary, enabling domain experts to understand how specific feature values influenced the model's output. Unlike SHAP, LIME does not guarantee global consistency, but it offers

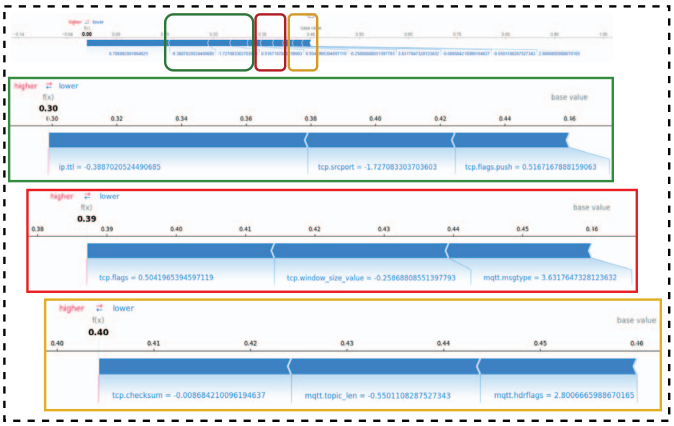

Fig. 4. SHAP Force Plot Illustrating Feature Contributions

intuitive insight at the instance level, complementing the SHAP-based interpretation.

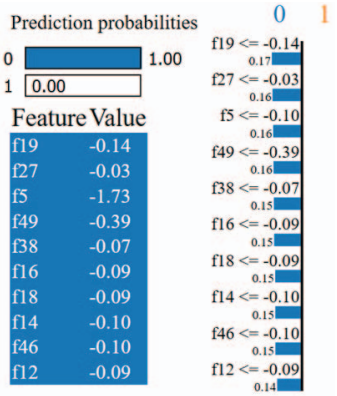

Fig. 5. LIME Explanation for an Instance

B. Comparison With Other Studies

Table V shows the performance of the proposed SG-VAE model in comparison with existing studies using the IoT-Flock dataset. While prior methods such as RF, LSSOA, GRU, and D-GRU show strong accuracy, ranging from 0.9633 to 0.9995, they lack support for XAI, which is crucial in healthcare applications. In comparison, SG-VAE not only achieves near-perfect accuracy 1.00±0.00, but also incorporates XAI for transparent decision-making. Furthermore, it demonstrates high throughput, processing 54,286 samples per second, making it suitable for real-time deployment. Note that the results in Table V are indicative, as they may vary with preprocessing, dataset splits, or hardware environments.

TABLE V
COMPARISON WITH OTHER STUDIES

Ref.	Year	Method	Dataset	Accuracy	XAI	TP
[14]	2021	RF	IoT-Flock	0.9951	×	-
[17]	2023	LSSOA	IoT-Flock	0.9959	×	-
[16]	2024	GRU	IoT-Flock	0.9995	×	-
[18]	2025	D-GRU	IoT-Flock	0.9633	×	-
Our	2025	SG-VAE	IoT-Flock	∼1.00 ±0.00	✓	54286

VII. Conclusion & Future Directions

To tackle IoMT security challenges, this study proposes an AI-based approach that is both highly accurate and lightweight, making it suitable for resource-constrained networks. With integrated explainability features, the framework is designed to be trustworthy and transparent. Through extensive experimentation, the proposed SG-VAE model, using a decoder-less architecture during inference, achieved near-perfect accuracy ($\sim$1.00). SHAP and LIME were employed to interpret the model's decision-making process and to highlight key protocol features relevant to IoMT attack detection. The lightweight architecture also enables high throughput, processing 54,286 samples per second, while maintaining minimal RAM and CPU usage. Despite its strong contributions, the study lacks real-time deployment, leaving accuracy and throughput untested in live settings. Future work will address this by deploying the model on a real-time testbed and incorporating continual learning to enhance adaptability.

References

[1] K. Elgazzar, H. Khalil, T. Alghamdi, A. Badr, G. Abdelkader, A. Elewah, and R. Buyya, "Revisiting the internet of things: New trends, opportunities and grand challenges," 2022.

[2] S. Ahmetoglu, Z. Che Cob, and N. Ali, "A systematic review of internet of things adoption in organizations: Taxonomy, benefits, challenges and critical factors," *applied sciences*, vol. 12, no. 9, p. 4117, 2022.

[3] P. Matthew, S. Mchale, X. Deng, G. Nakhla, M. Trovati, N. Nnamoko, E. Pereira, H. Zhang, and M. Raza, "A review of the state of the art for the internet of medical things," *Sci*, vol. 7, no. 2, 2025.

[4] S. E. El-deep, A. A. Abohany, K. M. Sallam, and A. A. A. El-Mageed, "A comprehensive survey on impact of applying various technologies on the internet of medical things," *Artificial Intelligence Review*, vol. 58, no. 3, p. 86, 2025.

[5] F. Rustam, W. Aljedaani, M. S. Elsayed, and A. D. Jurcut, "Famtds: A novel mfo-based fully automated malicious traffic detection system for multi-environment networks," *Computer Networks*, vol. 251, p. 110603, 2024.

[6] L. Dzamesi and N. Elsayed, "A review on the security vulnerabilities of the iomt against malware attacks and ddos," *arXiv preprint arXiv:2501.07703*, 2025.

[7] SC Media, "Iv pumps riskiest healthcare iot while 50% of medical devices hold critical flaws." https://shorturl.at/UrZ25, 2024. Accessed: July 11, 2025.

[8] S. Madanian, T. Chinbat, M. Subasinghage, D. Airehrour, F. Hassandoust, and S. Yongchareon, "Health iot threats: Survey of risks and vulnerabilities," *Future Internet*, vol. 16, no. 11, 2024.

[9] M. Jouhari and M. Guizani, "Lightweight cnn-bilstm based intrusion detection systems for resource-constrained iot devices," in *2024 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 1558–1563, IEEE, 2024.

[10] T. Shang, H. Lu, P. Wu, and X. Lu, "Method of setting exit advance guide signs in highway tunnels based on the driver's eye movement with markov chains," *IEEE Access*, vol. 9, pp. 24079–24092, 2021.

[11] F. Rustam, R. Shafique, S. K. Posa, and A. D. Jurcut, "Malicious traffic detection in multi-environment network using dual-data trained lightgbm approach," in *2024 IEEE 21st International Conference on Mobile Ad-Hoc and Smart Systems (MASS)*, pp. 598–603, IEEE, 2024.

[12] P. T. Ganai, A. Bag, A. Sable, K. H. Abdullah, S. Bhatia, and B. Pant, "A detailed investigation of implementation of internet of things (iot) in cyber security in healthcare sector," in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 1571–1575, 2022.

[13] N. Sharma and N. Jindal, "Emerging artificial intelligence applications: metaverse, iot, cybersecurity, healthcare-an overview," *Multimedia Tools and Applications*, vol. 83, no. 19, pp. 57317–57345, 2024.

[14] F. Hussain, S. G. Abbas, G. A. Shah, I. M. Pires, U. U. Fayyaz, F. Shahzad, N. M. Garcia, and E. Zdravevski, "A framework for malicious traffic detection in iot healthcare environment," *Sensors*, vol. 21, no. 9, 2021.

[15] M. M. Khan and M. Alkhathami, "Anomaly detection in iot-based healthcare: machine learning for enhanced security," *Scientific Reports*, vol. 14, no. 1, p. 5872, 2024.

[16] T. Kacem, S. Tossou, and A. Muir, "Detecting cyber attacks in healthcare iot systems," in *2024 International Conference on AI x Data and Knowledge Engineering (AIxDKE)*, pp. 80–85, 2024.

[17] N. Goswami, S. Raj, D. Thakral, J. L. Arias-Gonzáles, J. Flores-Albornoz, E. Asnate-Salazar, D. Kapila, S. Yadav, and S. Kumar, "Preserving security in internet-of-things healthcare system with metaheuristic-driven intrusion detection," *Engineered Science*, vol. 25, no. 3, p. 933, 2023.

[18] P. M. Kumar, B. P. Kavin, A. Jagathpally, and T. Shahwar, "Transforming the cybersecurity space of healthcare iot devices using deep learning," in *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*, pp. 1–6, 2025.

[19] C. Chakraborty, S. M. Nagarajan, G. G. Devarajan, T. V. Ramana, and R. Mohanty, "Intelligent ai-based healthcare cyber security system using multi-source transfer learning method," *ACM Trans. Sen. Netw.*, May 2023. Just Accepted.

[20] S. A. Algethami and S. S. Alshamrani, "A deep learning-based framework for strengthening cybersecurity in internet of health things (ioht) environments," *Applied Sciences*, vol. 14, no. 11, 2024.

[21] A. Nasayreh, H. M. Khalid, H. K. Alkhateeb, J. Al-Manaseer, A. Ismail, and H. Gharaibeh, "Automated detection of cyber attacks in healthcare systems: A novel scheme with advanced feature extraction and classification," *Computers & Security*, vol. 150, p. 104288, 2025.

[22] J. Rheey and H. Park, "Robust hierarchical anomaly detection using feature impact in iot networks," *ICT Express*, vol. 11, no. 2, pp. 358–363, 2025.

[23] A. Manoharan and M. Thathan, "Enhanced iomt security framework using group teaching optimized auto-encoder for intrusion detection," *Scientific Reports*, vol. 14, no. 1, p. 30360, 2024.

[24] L. Antwarg, R. M. Miller, B. Shapira, and L. Rokach, "Explaining anomalies detected by autoencoders using shapley additive explanations," *Expert systems with applications*, vol. 186, p. 115736, 2021.

[25] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Explainable artificial intelligence for intrusion detection in iot networks: A deep learning based approach," *Expert Systems with Applications*, vol. 238, p. 121751, 2024.

[26] R. Kalakoti, S. Nõmm, and H. Bahsi, "Explainable transformer-based intrusion detection in internet of medical things (iomt) networks," in *2024 International Conference on Machine Learning and Applications (ICMLA)*, pp. 1163–1170, IEEE, 2024.

[27] A. Altrad, "Iot medical network security system based explainable ai model," in *2025 12th International Conference on Information Technology (ICIT)*, pp. 403–407, IEEE, 2025.

[28] M. A. Alsharaiah, M. A. Almaiah, R. Shehab, M. Obeidat, F. A. El-Qirem, and T. Aldhyani, "An explainable ai-driven transformer model for spoofing attack detection in internet of medical things (iomt) networks," *Discover Applied Sciences*, vol. 7, no. 488, 2025.

[29] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-iiotset: A new comprehensive realistic cyber security dataset of iot and iiot applications for centralized and federated learning," *IEEe Access*, vol. 10, pp. 40281–40306, 2022.

[30] Y. Zhang, D. Zhu, M. Wang, J. Li, and J. Zhang, "A comparative study of cyber security intrusion detection in healthcare systems," *International Journal of Critical Infrastructure Protection*, vol. 44, p. 100658, 2024.

[31] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[33] Monica and P. Agrawal, "A survey on hyperparameter optimization of machine learning models," in *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pp. 11–15, 2024.