future internet

IMPACT FACTOR 3.6

CITESCORE 8.3

Article

Explainable AI Assisted IoMT Security in Future 6G Networks

Navneet Kaur and Lav Gupta

Special Issue
Toward 6G Networks: Challenges and Technologies

Edited by
Prof. Dr. Jianwu Zhang and Dr. Léo Mendiboure

MDPI

https://doi.org/10.3390/fi17050226

future internet

MDPI

Article

Explainable AI Assisted IoMT Security in Future 6G Networks

Navneet Kaur and Lav Gupta *

Department of Computer Science, University of Missouri, St. Louis, MO 63121, USA; nk62v@umsystem.edu
* Correspondence: lavgupta@umsl.edu

Abstract: The rapid integration of the Internet of Medical Things (IoMT) is transforming healthcare through real-time monitoring, AI-driven diagnostics, and remote treatment. However, the growing reliance on IoMT devices, such as robotic surgical systems, life-support equipment, and wearable health monitors, has expanded the attack surface, exposing healthcare systems to cybersecurity risks like data breaches, device manipulation, and potentially life-threatening disruptions. While 6G networks offer significant benefits for healthcare, such as ultra-low latency, extensive connectivity, and AI-native capabilities, as highlighted in the ITU 6G (IMT-2030) framework, they are expected to introduce new and potentially more severe security challenges. These advancements put critical medical systems at greater risk, highlighting the need for more robust security measures. This study leverages AI techniques to systematically identify security vulnerabilities within 6G-enabled healthcare environments. Additionally, the proposed approach strengthens AI-driven security through use of multiple XAI techniques cross-validated against each other. Drawing on the insights provided by XAI, we tailor our mitigation strategies to the ITU-defined 6G usage scenarios, with a focus on their applicability to medical IoT networks. We propose that these strategies will effectively address potential vulnerabilities and enhance the security of medical systems leveraging IoT and 6G networks.

Keywords: explainable AI; artificial intelligence; SHAP; LIME; DiCE; counterfactual explanations; 6G networks; healthcare security; IoMT; 6G; 6G security; XAI; 6G usage scenarios

check for updates

Academic Editors: Jianwu Zhang and Léo Mendiboure

Received: 13 April 2025
Revised: 4 May 2025
Accepted: 8 May 2025
Published: 20 May 2025

Citation: Kaur, N.; Gupta, L. Explainable AI Assisted IoMT Security in Future 6G Networks. *Future Internet* **2025**, *17*, 226. https://doi.org/10.3390/fi17050226

1. Introduction

Advancements in wireless communication have revolutionized healthcare, enabling innovations like surgical robotics, real-time remote monitoring, and XR (extended reality)-assisted treatments, benefiting rural, remote, and emergency care [1,2]. The rise of the Internet of Medical Things (IoMT) has further improved patient care through continuous data access, supporting timely interventions and personalized treatments [3]. Future 6G networks are expected to accelerate these advancements with ultra-low latency, massive device connectivity, and seamless AI and edge cloud integration. According to the ITU-R M.2160 framework, 6G will enable network-wide AI integration, multi-access edge computing (MEC), enhanced XR, and integrated sensing and communication for more efficient, intelligent, and responsive real-time patient care [4,5].

These advancements brought by the integration of 6G networks in healthcare are anticipated to introduce substantial security challenges. The incorporation of sensing, edge computing, and cloud services expands the attack surface, putting critical medical devices like cardiac monitors, infusion pumps, and respirators at risk from both internal and external threats [6]. Introducing AI further exposes healthcare systems to adversarial and model poisoning attacks, endangering patient safety [7]. Additionally, immersive technologies, such as AR-assisted surgeries, introduce new vulnerabilities, offering potential entry points for unauthorized access [4,8]. The fragmented management of edge and cloud infrastructures increases system vulnerabilities, making it more challenging to maintain data integrity [9]. As 6G networks integrate with public land mobile networks (PLMNs), traditional threats like DDoS, MiTM attacks, and SYN floods will intensify, alongside more sophisticated attacks [10]. The rapid growth of IoMT devices, edge nodes, and real-time data streams adds risks, such as disruptions in telemedicine or remote surgeries due to DDoS-induced outages or network congestion, with indicators like high CPU usage and memory depletion [11,12]. Other concerns, like eavesdropping, session hijacking, and spoofing, can be detected through latency spikes or unusual traffic patterns, threatening patient safety and network integrity [12,13].

As 6G networks rely more heavily on edge computing and cloud infrastructure, securing these decentralized environments will become significantly more complex than current network configurations [13]. Edge nodes processing real-time healthcare data are prime targets for attacks, with CPU disruptions and network anomalies serving as early indicators of potential intrusions. The challenge is further compounded by the lack of centralized control, as managing these nodes across multiple operators complicates real-time attack detection and mitigation [13]. Unauthorized access to critical systems could allow attackers to manipulate patient records, inject malicious data, or disable monitoring systems, resulting in incorrect diagnoses, delayed treatment, or the failure of life-saving procedures [14]. As 6G networks evolve and integrate with existing networks, strengthening the security frameworks for these decentralized systems is crucial for ensuring safe and resilient healthcare services. Explainable AI (XAI) can help address these challenges by providing transparent, actionable insights into AI-driven decision-making, enabling healthcare providers to effectively identify and respond to security threats [15,16].

This paper explores the critical role of XAI in healthcare applications within IoMT environments, utilizing methods such as Shapley Additive Explanations (SHAP) [17], Local Interpretable Model-agnostic Explanations (LIME) [18], and Diverse Counterfactual Explanations (DiCE) [19], supported by an advanced dataset derived from an IoT wireless medical testbed [5]. To enhance our research outcomes, we cross-validated feature importance using XAI methods to identify consistent indicators of attack and normal traffic. This revealed key features and their values (high or low) that impacted security outcomes, improving threat mapping for various 6G scenarios. Additionally, we analyzed the correlations between these features, which allowed us to differentiate between malicious and benign activities, offering a clearer understanding of potential security threats. We then mapped these insights to identify the relevant security threats and performance issues that each 6G scenario may face. Based on these findings, we proposed targeted mitigation strategies that address the unique vulnerabilities and requirements of each usage scenario. Furthermore, we offer security enhancements derived from XAI analysis, enabling security administrators to gain actionable insights and implement targeted security measures tailored to the specific needs of each 6G usage scenario.

By making AI-driven decisions more transparent and understandable, our work fosters greater trust in AI-powered security solutions. This not only strengthens the security of healthcare systems, but also supports better-informed decision-making in critical healthcare environments, ensuring the reliability and protection of medical data and devices. The key contributions of this paper include the following:

- Comprehensively mapping the emerging vulnerabilities of 6G usage scenarios within the healthcare domain, using the authoritative ITU-R IMT-2030 framework as a foundation.

- Leveraging and synthesizing SHAP, LIME, and DiCE to identify and interpret model behavior—uncovering the hidden patterns and feature sensitivities that expose critical security flaws in 6G-enabled IoMT environments.
- Cross-validating multiple XAI methods to assess the consistency and reliability of feature importance in model predictions.
- Aligning XAI insights with the specific needs and challenges of each 6G use case to identify relevant threats and performance issues.
- Proposing targeted mitigation strategies to address the unique vulnerabilities and requirements of each 6G usage scenario's.
- Providing XAI-driven security enhancements to empower administrators with actionable insights and targeted security measures tailored to each 6G use case.

The remainder of the paper is organized as follows: Section 2 reviews the relevant literature on 6G security in healthcare. Section 3 discusses the security challenges associated with 6G IoMT applications. Section 4 outlines the materials and methods used in this research, including the application of XAI techniques. Section 5 presents the experiments and results, focusing on the identification of key features and vulnerabilities. Section 6 discusses the cross-validation of XAI methods and the consistency of feature importance. Section 7 proposes targeted mitigation strategies tailored to the unique requirements of each 6G usage scenario. Finally, Section 8 concludes the paper.

2. Review of Existing Research and Novel Contributions

2.1. Existing Research

Recent studies have increasingly focused on the integration of XAI with healthcare, with the aim of improving data privacy, interpretability, and clinical decision-making. For instance, the authors in [6] propose a framework that combines XAI with mass surveillance to improve epidemic monitoring. This framework uses deep learning and edge computing to improve data privacy through blockchain integration. In [7], the authors apply XAI to heart disease diagnosis and emphasize the need for transparent AI models to foster trust among medical professionals. Techniques such as LIME and SHAP are used to illustrate how specific clinical features influence predictions. Similarly, paper [8] presents a framework that combines federated learning (FL) with XAI to boost both privacy and interpretability in next-generation healthcare networks, resulting in enhanced prediction accuracy and user trust.

Researchers have employed XAI to explain clinical diagnoses. In paper [9], authors use SHAP to support clinicians in understanding the predictive factors in unplanned hospital readmissions of elderly patients. In [20], the authors use XAI to address safety, trust, and reliability concerns regarding using AI in brain tumor diagnosis. Similarly, Ref. [21] explores the integration of XAI with neuromorphic computing to develop consumer-friendly healthcare applications, using techniques such as SHAP, LIME, and ELI5 to overcome transparency limitations. Several studies focus on XAI in security-sensitive healthcare settings. For instance, in [22], authors propose an explainable malicious traffic detection system using LIME, SHAP, ELI5, and Integrated Gradients (IGs) to ensure interpretability in models monitoring intensive care patient datasets. Similarly, Refs. [23,24] present intrusion detection systems in IoMT environments, both using the WUSTL-EHMS-2020 dataset and SHAP to explain deep learning model decisions. Finally, Ref. [25] introduces a privacy-preserving ECG classification framework based on federated learning, incorporating XAI to assist healthcare professionals in validating and interpreting model predictions.

While many studies emphasize the role of XAI in healthcare, they often overlook the deeper integration of wireless networks and the emerging security challenges associated with the impending transition from 5G to 6G. Furthermore, due to the developmental stage of 6G, contemporary research to fully investigate how XAI can be incorporated into 6G-enabled healthcare environments—where dynamic system parameters and AI-driven security features are critical—is largely missing. Additionally, to the best of our knowledge, no prior work has leveraged multiple XAI techniques in combination for cross-validation, which is a critical step in ensuring the reliability and consistency of AI-generated explanations. Moreover, the current research lacks a strategic framework for translating insights gained from XAI into actionable risk mitigation strategies that are specifically tailored to 6G usage scenarios in the healthcare domain.

Our research addresses these critical gaps by carrying out the following steps: comprehensively mapping 6G healthcare use cases using the ITU-R IMT-2030 framework; applying and cross-validating multiple XAI methods (SHAP, LIME, and DiCE) to uncover consistent and trustworthy feature attributions; linking these insights to security threats and performance bottlenecks within each 6G usage scenario; and proposing XAI-driven mitigation strategies and enhancements that empower administrators to take targeted, scenario-specific security action in 6G medical IoT environments.

2.2. Novel Contributions and the Proposd Approach

We align our study with the ITU-R IMT-2030 framework [4] to systematically assess security risks in 6G networks. This ensures that our analysis is grounded in globally accepted standards and delivers relevant insights into securing next-generation networks. A significant and distinctive contribution of this work is its focus on network security rather than traditional textual data analysis. While network features and their roles in security have been studied in prior research, our approach takes a significant step forward by uniquely extracting knowledge from neural network models and leveraging XAI techniques to extract deeper, actionable insights. Unlike in some existing work, our focus is not just on identifying which features need monitoring, but also on exploring the intricate relationships between multiple network features to show how their interplay can predict or trigger security threats. What further sets our work apart is the mapping of these XAI-derived insights to specific 6G usage scenarios, particularly within the medical IoT domain. This enables not only the identification of potential vulnerabilities, but also the formulation of scenario-specific mitigation strategies. By demonstrating how XAI can be used beyond model interpretability—to proactively inform defense planning and guide strategic decision-making—our work highlights its broader applicability. Just as we have shown in the context of 6G-enabled healthcare systems, the same methodology can be extended to other domains such as autonomous vehicles, smart manufacturing, and critical infrastructure protection, making XAI a powerful tool for enhancing security preparedness across next-generation technologies.

Figure 1 presents the proposed explainable security framework, showcasing our key contributions and the role of XAI in strengthening system resilience. The collected dataset [5] was processed using the standard Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, ensuring better generalizability to real-world scenarios [26]. We employed Logistic Regression, K-Nearest Neighbor (KNN), Random Forest, and a Convolutional Neural Network (CNN) for model evaluation, while SHAP, LIME, and DiCE were applied for explainability. Cross-validation was performed to ensure consistent and reliable insights. These insights provided a deeper understanding of the security risks faced by each 6G use case, leading to the proposal of tailored mitigation strategies for each 6G IoMT scenario, enhancing security and facilitating informed decision-making. Finally, to maximize the usability and impact of our findings, we present the extracted insights in a clear and interpretable format, making them accessible to both technical and non-technical stakeholders. This approach promotes transparency, fosters trust in AI-driven security mechanisms, and facilitates informed decision-making within 6G-enabled healthcare systems.

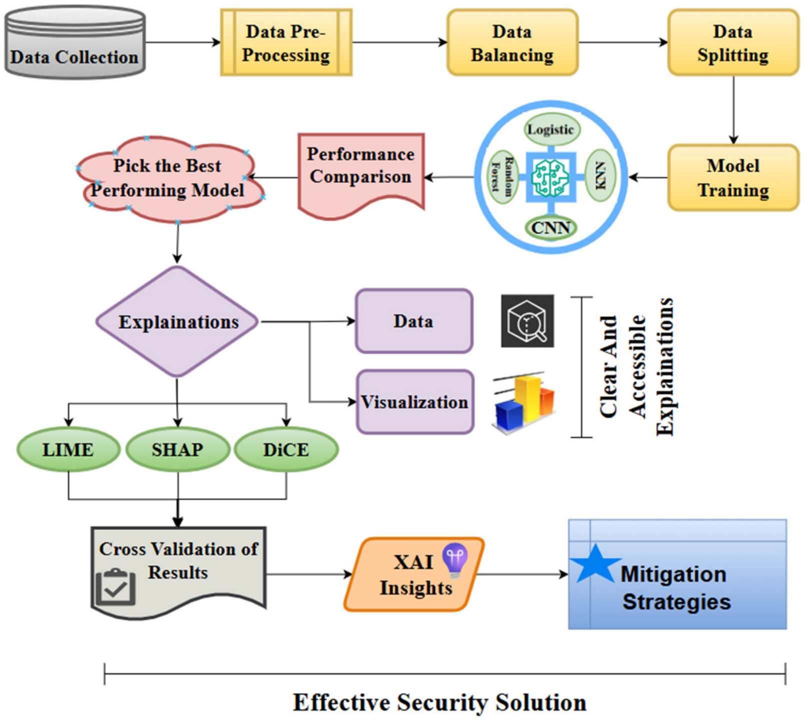


Figure 1. The proposed explainable security framework.

3. Emerging Security Vulnerabilities in IoMT-Driven 6G Healthcare Scenarios

With the rapid evolution toward 6G, healthcare applications powered by the IoMT are set to become more intelligent, interconnected, and latency sensitive. As illustrated in Figure 2, some of these 6G usage scenarios are extensions of the 5G framework, while others, such as ubiquitous connectivity, native AI, and integrated sensing and communication, are entirely new scenarios. For further technical details, readers may refer to the ITU-R IMT-2030 framework [4].

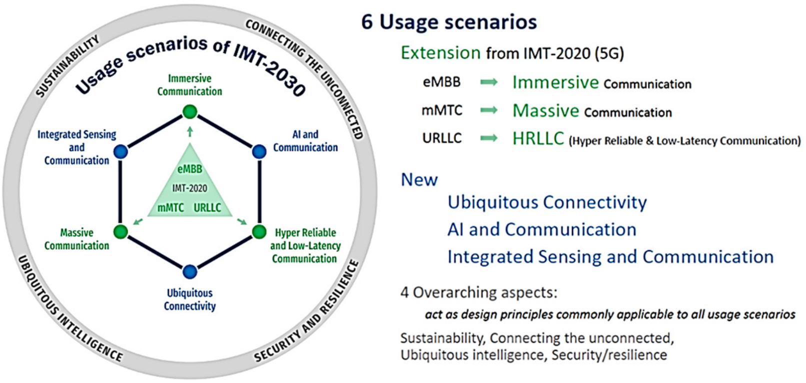
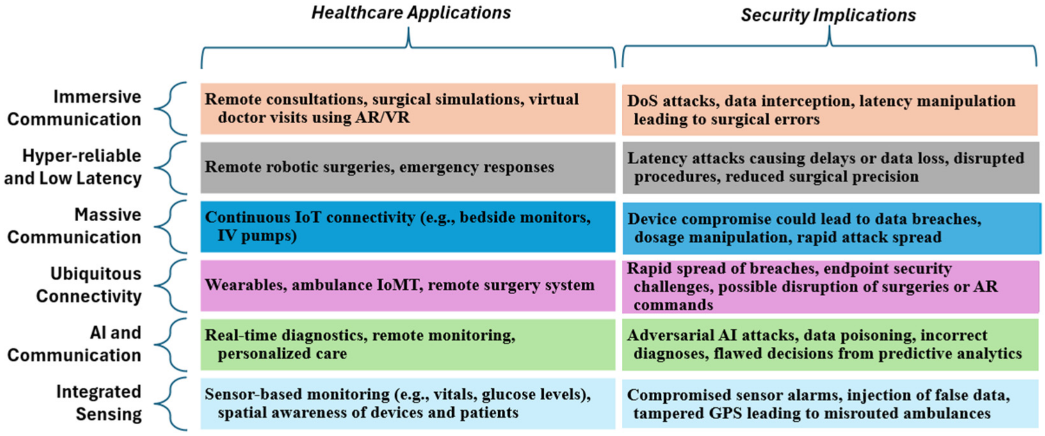

Usage scenarios of IMT-2030
SUSTAINABILITY
CONNECTING THE UNCONNECTED
UBIQUITOUS INTELLIGENCE
SECURITY AND RESILIENCE
Immersive Communication
AI and Communication
Hyper Reliable and Low-Latency Communication
Ubiquitous Connectivity
Massive Communication
Integrated Sensing and Communication
eMBB
IMT-2020
mMTC URLLC
6 Usage scenarios
Extension from IMT-2020 (5G)
eMBB ➡ Immersive Communication
mMTC ➡ Massive Communication
URLLC ➡ HRLLC (Hyper Reliable & Low-Latency Communication)
New
Ubiquitous Connectivity
AI and Communication
Integrated Sensing and Communication
4 Overarching aspects:
act as design principles commonly applicable to all usage scenarios
Sustainability, Connecting the unconnected,
Ubiquitous intelligence, Security/resilience

Figure 2. The six usage scenarios for 6G (source: ITU-R M.2160).

While 6G advancements offer great potential for healthcare, they also create significant security challenges. Each usage scenario involves distinct operational dynamics that expose unique vulnerabilities. Figure 3 highlights these scenarios alongside associated threats such as data tampering, denial-of-service (DoS) attacks, latency manipulation, and AI poisoning [9,16,27,28]. These risks are particularly critical in IoMT-driven environments, where security breaches can quickly propagate and impact patient safety and care delivery.



	Healthcare Applications	*Security Implications*
Immersive Communication	Remote consultations, surgical simulations, virtual doctor visits using AR/VR	DoS attacks, data interception, latency manipulation leading to surgical errors
Hyper-reliable and Low Latency	Remote robotic surgeries, emergency responses	Latency attacks causing delays or data loss, disrupted procedures, reduced surgical precision
Massive Communication	Continuous IoT connectivity (e.g., bedside monitors, IV pumps)	Device compromise could lead to data breaches, dosage manipulation, rapid attack spread
Ubiquitous Connectivity	Wearables, ambulance IoMT, remote surgery system	Rapid spread of breaches, endpoint security challenges, possible disruption of surgeries or AR commands
AI and Communication	Real-time diagnostics, remote monitoring, personalized care	Adversarial AI attacks, data poisoning, incorrect diagnoses, flawed decisions from predictive analytics
Integrated Sensing	Sensor-based monitoring (e.g., vitals, glucose levels), spatial awareness of devices and patients	Compromised sensor alarms, injection of false data, tampered GPS leading to misrouted ambulances

Figure 3. Healthcare applications and security implications of 6G usage scenarios in IoMT environments.

As healthcare systems transition to fully connected, intelligent 6G environments, the increasing reliance on AI can hinder threat detection and strategic decision-making for risk management [11,29]. Integrating XAI is essential for ensuring transparency, building trust, and maintaining secure, resilient infrastructures [30]. In the following sections, we will discuss the methods, materials, experiments, and results, demonstrating how XAI contributes to developing effective mitigation strategies for 6G usage scenarios, ultimately enhancing security and system resilience.

4. Materials and Methods

The experiments were conducted using Python 3.10 in a Google Colab environment with GPU acceleration to enhance computational efficiency. TensorFlow (v2.17.1) and Keras (v3.5.0) were used for model development and training, while Matplotlib (v3.10.0) and Seaborn (v0.13.2) supported result visualization. This setup ensured efficient processing, robust model development, and clear visual insights—key for building trust and security in intelligent 6G networks.

For analysis, we used a comprehensive, publicly available dataset designed for security research in medical networks [15]. The dataset comprises 77 features across 145K samples—132K normal and 13K attack instances—capturing both network and host-based attributes. The testbed includes a core network, local network, multi-access edge computing (MEC) servers, user equipment (UE), and a routing system. To simulate realistic threat scenarios, tools like Simu5G and Stateful IP/ICMP Translation (SIIT) were integrated. The dataset highlights four major threats, each representing a critical security challenge in medical network environments, as shown in Figure 4:

- Man-in-the-Middle (MiTM): This attack involves intercepting and potentially modifying communications between two parties, thereby enabling unauthorized access to sensitive information.
- Distributed Denial-of-Service (DDoS): In this type of attack, adversaries inundate systems with excessive traffic, depleting resources and resulting in service disruptions or complete outages.
- Ransomware: A form of malicious software that encrypts critical organizational data and demands a ransom for its decryption, leading to significant disruption of routine operations.

- Buffer overflow: This attack exploits vulnerabilities in memory management to inject and execute malicious code, which can compromise the integrity and security of the system.

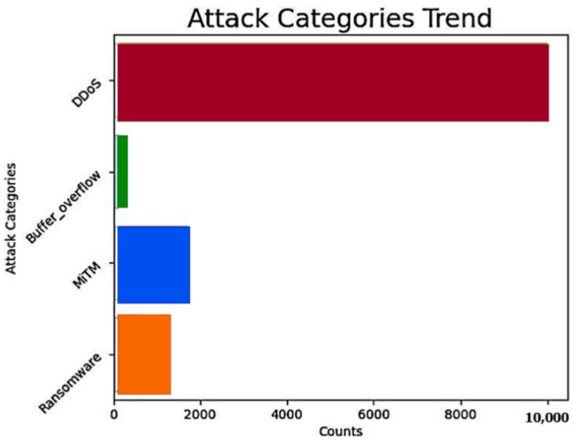
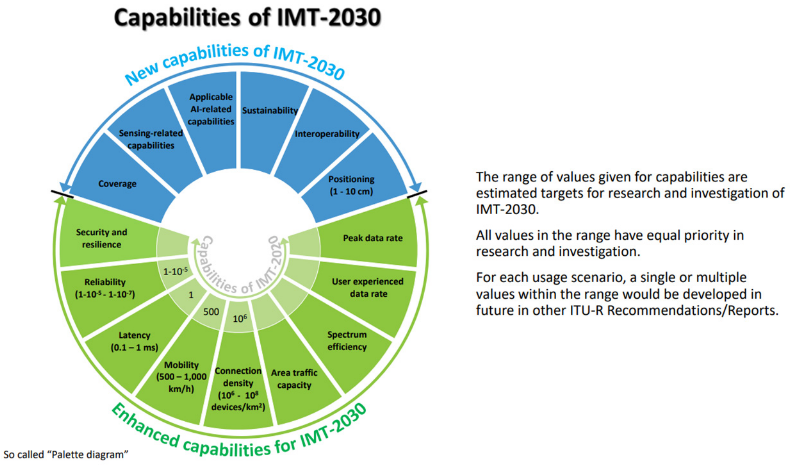


Figure 4. Visualization plot of different class types and their counts.

Although this dataset was originally designed for 5G, its structure and complexity make it highly suitable for evaluating AI-driven security frameworks in 6G environments. It aligns with the ITU-R IMT-2030 framework (M.2160) depicted in Figure 5, which envisions 6G as an evolution of 5G, incorporating both enhanced capabilities and entirely new features. Building on the foundation established by 5G, 6G introduces innovations such as native AI integration, sub-terahertz (THz) communication, reconfigurable intelligent surfaces (RISs), and integrated sensing and communication (ISAC), while also improving areas like spectrum efficiency, energy efficiency, and network resilience. This dataset is particularly relevant for evaluating the resilience and adaptability of security frameworks in the 6G context, as it includes critical network components such as routers, servers, and MEC nodes.


Capabilities of IMT-2030
New capabilities of IMT-2030
Applicable AI-related capabilities
Sustainability
Sensing-related capabilities
Interoperability
Coverage
Positioning (1 - 10 cm)
Security and resilience
Peak data rate
1-10⁻⁵
Capabilities of IMT-2020
Reliability (1-10⁻⁵ - 1-10⁻⁷)
User experienced data rate
1
500
10⁶
Latency (0.1 – 1 ms)
Spectrum efficiency
Mobility (500 – 1,000 km/h)
Connection density (10⁶ - 10⁸ devices/km²)
Area traffic capacity
Enhanced capabilities for IMT-2030

The range of values given for capabilities are estimated targets for research and investigation of IMT-2030.

All values in the range have equal priority in research and investigation.

For each usage scenario, a single or multiple values within the range would be developed in future in other ITU-R Recommendations/Reports.

So called "Palette diagram"

Figure 5. The capabilities of IMT—2030 (source: ITU—R M.2160).

Moreover, the dataset simulates network conditions like high bandwidth, ultra-low latency, and diverse traffic scenarios, reflecting the core characteristics of 6G, particularly relevant for mission-critical healthcare applications such as robotic surgeries and AI-assisted diagnostics. It also incorporates security-specific data, such as traffic logs and simulated attack patterns, making it indispensable for 6G healthcare threat modeling. By addressing IPv6/IPv4 compatibility and including a range of cyber threats, including IoT-based attacks and network anomalies, it provides a solid foundation for developing and testing advanced security solutions for future wireless networks. Also, given the unavailability of 6G network data, this dataset remains the most reliable resource for evaluating potential security threats and developing mitigations, particularly in the context of healthcare applications in the forthcoming 6G era.

After data collection, we addressed class imbalances, especially in attack subclasses, using standard SMOTE [26]. SMOTE generates new minority instances using nearest neighbors, preserving key majority-class information without duplication. This approach generates synthetic instances by interpolating between a given minority class instance, x_i, and one of its k nearest neighbors, x_{nm}. The new synthetic instance $\mathrm{x}_{\mathrm{new}}$ is computed as follows:

$$x_{new} = x_i + \lambda.(x_{nm} - x_i) \tag{1}$$

In Equation (1), λ is a randomly selected value from a uniform distribution in the range [0, 1]. This technique ensures that the synthetic data points are systematically positioned along the line segment connecting x_i and x_{nm}, effectively enhancing the diversity of the minority class without simply duplicating the existing data.

The dataset was pre-processed for binary classification (attack vs. normal) and split into 80–20 ratio for training and testing purposes. Various machine learning models, including Logistic Regression [31], Random Forest [32], KNN [33], and a CNN [34], were evaluated to classify the network traffic. These models were selected for their distinct strengths: Logistic Regression for its simplicity and interpretability, Random Forest for its robustness to overfitting and ability to handle complex interactions, KNN for its effectiveness in non-linear data classification, and the CNN for its ability to automatically learn hierarchical features from raw data. Each model was chosen to provide a balanced comparison of performance.

To ensure model interpretability, XAI techniques were used to interpret the model's decisions and to identify key factors that are influencing model predictions. These techniques are essential for building trust and accountability in cybersecurity applications. Specifically, SHAP quantifies the contribution of each feature to the model's prediction, assigning specific values to each feature [17]. The value for feature i in prediction x is computed using the following formula:

$$\varnothing_i(x) = \sum_{S \sqsubseteq \{1,2,\ldots p\} \setminus \{i\}} \frac{|S|!(p - |S| - 1)!}{p!} [f(S \cup \{i\}) - f(S)] \tag{2}$$

In Equation (2), $\varnothing_i(x)$ denotes the SHAP value for feature i at instance x, S is the subset of the features excluding i, $|S|$ denotes the number of features in the subset S, $|S|!(p - |S| - 1)!$ is the number of ways to select subset S, p represents the total number of features, and f is the model's prediction function. This method provides a comprehensive explanation of how each feature affects the model's decision.

The other XAI method, LIME, approximates the model's prediction function f locally around an instance x by minimizing a loss function [18]. This method enables the interpretation of complex models in the vicinity of a specific instance. The optimization problem is expressed as follows:

$$\hat{g} = argmin_g \; L(f, g, \; \pi_x) + \Omega(g) \tag{3}$$

In Equation (3), $\hat{g}$ denotes the interpretable model, π_x represents the proximity measure around x, L is denoted as a loss function, and Ω(g) is a penalty term ensuring the complexity of g.

Also, the DiCE method generates multiple counterfactuals, or alternative explanations, for a model's prediction, helping us to understand what changes could lead to a different outcome [19]. This process aids in uncovering the sensitivity of the model's predictions to variations in the feature values. The objective of optimization for DiCE is to identify counterfactuals that are close to the original input, providing insights into the sensitivity of the model's decisions. The optimization objective is as follows:

$$\min_{x'}||x'-x||^2+\lambda.\mathcal{L}(f(x',y_{target})+\beta.Diversity(x',\{x_i\})) \tag{4}$$

In Equation (4), x is the original instance, x' is the counterfactual instance, y_{target} is the desired output, and $\mathcal{L}$ is a loss function that ensures alignment with the target. The term $||x'-x||^2$ ensures that the counterfactual instance is close to the original, while $Diversity(x',\{x_i\})$ promotes variability among counterfactuals.

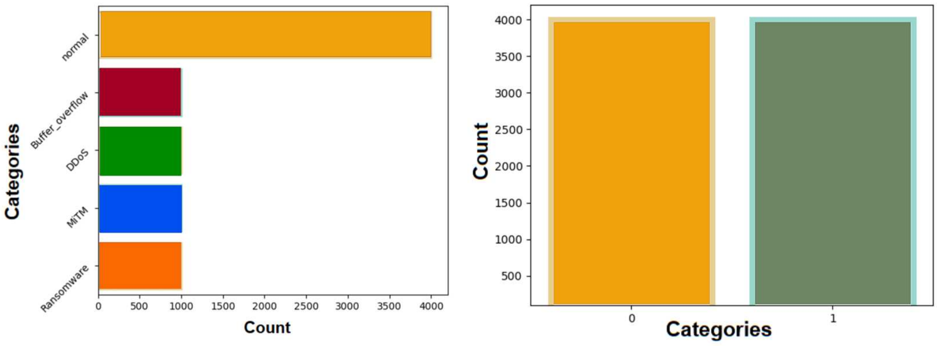
5. Experiments and Results

This section outlines the experimental setup and presents the performance results of the proposed models, including insights from the XAI techniques (SHAP, LIME, and DiCE) used to interpret the models' decision-making processes.

5.1. Data PreProcessing and Splitting

The collected dataset [5] was preprocessed into Pandas DataFrame [35], addressing infinite and missing values, removing duplicates, converting categorical labels to numerical formats, and normalizing features. Features with zero variance were excluded, reducing the attribute count from 77 to 52, ensuring a suitable dataset for model training.

To address class imbalances and reduce biases, we applied the standard SMOTE, standardizing each attack subclass to match the sample size of the normal class, as shown in Figure 6a. This balancing improved the model's ability to generalize and enhanced the classification accuracy. The problem was reformulated as a binary classification task to distinguish between malicious (attack) and non-malicious (normal) network traffic, aiming to identify attacks accurately. Figure 6b shows the balanced distribution of the two classes—normal (0) and attack (1)—improving the model's reliability in real-world network security applications.



(**a**) Category balancing. (**b**) Sub-category balancing.

Figure 6. Visualization of class categories and their counts after using SMOTE.

Given the size of the dataset, we randomly selected 4000 records from each class (normal and attack traffic) for model evaluation. This subset enables us to effectively apply and evaluate XAI techniques while maintaining computational efficiency, aligning with our primary objective of developing interpretable and transparent models to enhance network security, rather than focusing solely on maximizing model performance across the full dataset. Table 1 presents the sample distribution before and after applying SMOTE, which was used to balance the dataset. Achieving this balance is essential for accurate threat detection and consistent model performance in realistic 6G network scenarios.

Table 1. Dataset categorization.

Class Types	Total Count (Before SMOTE)	Class Instance (Before SMOTE)	Total Count (After SMOTE)	Class Instance (After SMOTE)
Normal	132,884	132,884	4000	4000
DOS	9971	12,339	1000	4000
MiTM	1672		1000	
Ransomware	528		1000	
Buffer_Overflow	68		1000	

5.2. Model Training

The models—Logistic Regression, Random Forest, KNN, and a CNN—were trained and evaluated on the same post-SMOTE dataset using key metrics, with the results summarized in Table 2 for reproducibility:

- **Logistic Regression**: The model is configured with random_state = 7 to ensure consistent results, max_iter = 5000 to allow sufficient iterations for convergence, solver = ‘lbfgs’ for efficient optimization using limited-memory BFGS, and class_weight = ‘balanced’ to automatically adjust for class imbalances in the dataset.
- **CNN**: The architecture consists of dense layers with 120, 80, 40, and 20 neurons, followed by a single-neuron output layer. It utilizes a batch size of 10 and the Adam optimizer for adaptive learning rate adjustments and is trained for 100 epochs to ensure convergence and minimize loss.
- **Random Forest**: The model uses n_estimators = 100 decision trees, with random_state = 42 sets to ensure the reproducibility of results across different runs by maintaining the same random state in the training process.
- **KNN**: The model is initialized with default hyperparameters, and random_state = 42 ensures consistent results in the training and evaluation process by controlling the randomization.

Table 2. Performance evaluation metrics.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	98.37%	0.984	0.983	0.983
K-Nearest Neighbor	98.87%	0.988	0.988	0.988
Random Forest	99.85%	0.998	0.998	0.998
CNN	75.25%	0.752	0.752	0.752

While further tuning could enhance accuracy, the focus remains on interpretability through XAI techniques—essential for informed security decisions, detecting vulnerabilities, and ensuring robust protection in future 6G networks.

5.3. Model Selection for XAI Technique Application

Given its strong performance, the Random Forest model was prioritized, and XAI techniques—SHAP, LIME, and DiCE—were used to interpret its decision-making. While Random Forest provides some level of interpretability through built-in feature importance, it still behaves like black box models in high-dimensional scenarios. XAI techniques help to uncover complex feature interactions and offer instance-level explanations, thereby improving transparency in healthcare security. Figure 7 presents the feature importance plot generated by the Random Forest model using the scikit-learn Python library [36]. Among the features, ‘scputimes_idle’ holds the highest importance, followed by ‘scputimes_user’ and 'scpustats_interrupts'. Features that were assigned to zero importance have been excluded from the plot, as they do not contribute to the model's predictive performance.

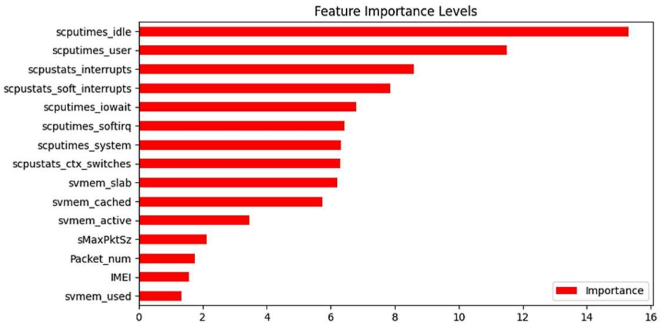


Figure 7. Feature importance plot for Random Forest.

5.4. Implementation of Explainable AI Techniques: SHAP, LIME, and DiCE

To illustrate the decision-making process of the selected model, we analyzed a random test sample from the post-SMOTE dataset to assess how specific features impact the final classification ('Normal' or 'Attack'). This analysis highlights the key factors influencing the model's decisions, improving transparency and supporting ongoing refinement to ensure adaptability to the evolving security challenges in 6G network environments.

5.4.1. SHAP–Global Behavior Analysis

The SHAP global plot identifies the features with the greatest influence on classifying an instance as either an attack or normal, providing insights into the model's overall behavior. Figure 8 visualizes the impact of each feature on the model's predictions across all instances, with blue representing Class 1 (attack) and red representing Class 0 (normal). The balanced distribution of blue and red lines for each feature indicates that the model considers both attack and normal instances, showing that its decision-making process is influenced by patterns in both classes.

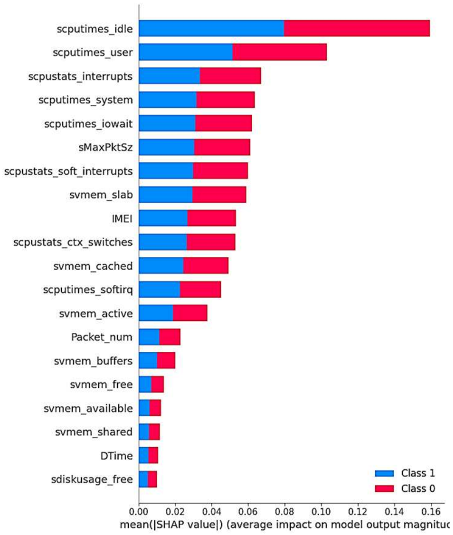


Figure 8. SHAP force plot.

5.4.2. SHAP—Local Behavior Analysis

The SHAP global summary plot shows overall feature importance but doesn't explain individual predictions. This highlights the need for local interpretability which is provided by the SHAP force plot. This plot visually provides instance-specific insights, enhancing interpretability and trust in AI-driven security and healthcare applications. The plot employs the following:

- Red indicates positive contributions (pushing prediction higher), increasing the likelihood of the predicted class.
- Blue represents negative contributions (pulling prediction lower), usually decreasing the likelihood of the predicted class, favoring the opposite prediction.
- The width of each bar reflects the magnitude of the feature's influence, with wider bars indicating a stronger effect.
- It starts with the base value (average model output) and accumulates feature contributions to display the final prediction at the end.

1. SHAP Force Plot for Test Sample Record 1—'Normal' Traffic Prediction

Figure 9 shows the SHAP force plot for sample record 1, highlighting how individual features influenced the model's prediction. The actual target is labeled as "Normal", meaning that in the dataset, this instance is truly a non-attack instance. The model's prediction output (f(x) = 1.00) suggests that the model classified this instance as not an attack.

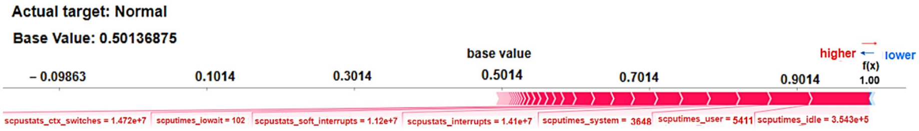


Figure 9. SHAP force plot for record sample 1 from test dataset (local explanation)—predicting Class 0—"Normal" traffic.

Table 3 lists the key features from the SHAP force plot that contribute to the "Normal" traffic prediction. It includes feature descriptions, their roles in the model's decisions, and how specific values support the classification as non-attack traffic.

Table 3. SHAP force plot for normal traffic classification.

Features	Full Form	Interpretation	Reason for 'Normal' Prediction
scputimes_ idle	CPU Idle Time	It indicates high idle time where resources are not being excessively consumed by malicious processes or attack-related activities.	Positive contribution indicates that the system is not stressed and can handle additional tasks smoothly. No indication of failure or abnormal behavior.
scputimes_ user	CPU Time in User Mode	Suggests that the system's CPU is being used by typical user-level applications or processes.	Positive contribution reflects standard user activity, without being overloaded by the malicious or abnormal processes.
stats_soft_ interrupts	CPU Software Interrupts Count	In a normal scenario, software interrupts are expected to be at a reasonable level, as they reflect routine system activities.	Positive contribution indicates that the system is processing normal, expected software interrupts without any significant anomaly or malicious activity.

Table 3. *Cont.*

Features	Full Form	Interpretation	Reason for 'Normal' Prediction
scputimes_ iowait	CPU Time Waiting for I/O Operations	I/O wait time is typical, reflecting normal operations waiting for data.	Positive contribution suggests expected I/O wait time, indicating typical system conditions without anomalies.
scputimes_ system	CPU Time in System Mode	Represents system CPU time, a normal indicator of workload. Higher values suggest expected system usage.	Positive contribution aligns with normal system activity and suggests that CPU is actively managing core tasks, typical in a normal state handling routine processing, resource management, and device I/O.
scpustats_ interrupts	CPU Hardware Interrupts Count	The number of interrupts suggests normal system activity with adequate handling of processes.	Positive contribution indicates the system is operating under expected conditions, with no significant interruptions or failures.
scpustats_ ctx_ switches	CPU Context Switches Count	High number of context switches indicates normal multitasking or process handling.	Positive contribution indicates regular system activity and resource allocation, reinforcing a stable system state.

2. SHAP Force Plot for Test Sample Record 2—'Attack' Traffic Prediction

The SHAP force plot for sample record 2 from the testing dataset, with a base value of 0.50 and a final prediction of 'Attack', is shown in Figure 10. The model's final prediction of 0.00, which is notably lower than that of the base value, indicates that the feature values substantially decreased the model's confidence in classifying the instance as an attack. The primary features (depicted in blue) are contributing to this reduction in the prediction value, thereby decreasing the likelihood of the event being classified as an attack. The following observations can be made:

- The actual target is labeled as "Attack", meaning that in the dataset, this instance is truly an attack.
- The model's prediction output (f(x) = 0.00) suggests that the model classified this instance as not an attack (likely normal or benign traffic).
- The cumulative effect of these features outweighs the benign contributions of others, leading the model to classify the prediction as an attack. For instance, high system resource usage (user, I/O wait time, and system-level processes) and interrupts (hardware and software) indicate a pattern of abnormal activity that is typical of an attack (e.g., DDoS, resource exhaustion, or other malicious behaviors).

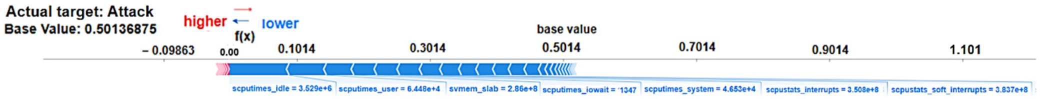


Figure 10. SHAP force plot for record sample 2 from test dataset (local explanation)—predicting Class 1–"Attack" traffic.

Table 4 lists the features, their interpretations, and the contributions to the model's prediction, explaining why the prediction is classified as an attack. While some features suggest normal behavior, their individual contributions are insufficient to change the decision. However, the combined influence of certain features, in context, leads to the classification of an attack, highlighting the complex interplay of feature contributions in the decision-making process.

Table 4. Feature contributions and interpretations in SHAP force plot for network traffic classification.

Features	Full Form	Interpretation	Reason for 'Attack' Prediction
scputimes_ idle	CPU Idle Time	The idle CPU time typ-ically suggests benign behavior, but its impact is not enough to over-ride the attack predic-tion.	The idle CPU time alone is not sus-picious, but the combination with high scpustats_interrupts (sug-gesting system disruption) and scputimes_user (indicating us-er-level activity) points to a possible attack.
scputimes_ user	CPU Time in User Mode	High CPU time spent on user processes suggests active user behavior but is not dominant enough to change the classification.	scputimes_user suggests active user processes, which, along with scputimes_iowait (I/O wait times, indicating delays) and svmem_slab (high memory usage), points to abnormal system behavior typical of attacks.
svmem_ slab	Kernel Slab Memory Usage	High memory usage by kernel objects can signal system stress, possibly due to attack activity.	High svmem_slab indicates memory manipulation, which, combined with high scpustats_interrupts (potential overload), suggests that the system is under attack, trying to overwhelm the resources.
scputimes_ iowait	CPU Time Waiting for I/O Operations	High I/O wait time suggests delays, potentially from attack-related activities.	Elevated scputimes_iowait indicates delayed I/O operations, possibly due to attack-induced resource contention. Combined with high scpustats_interrupts (system disruptions), this signals a denial-of-service or resource exhaustion attack.

Table 4. *Cont.*

Features	Full Form	Interpretation	Reason for 'Attack' Prediction
scputimes_ system	CPU Time in System Mode	High CPU time for system tasks could signal attack-related resource manipulation.	scputimes_system represents background system tasks, and in combination with abnormal interrupts and elevated system-level resource consumption, suggests a coordinated attack manipulating system resources.
scpustats_ interrupts	CPU Hardware Interrupts Count	High interrupt counts can indicate heavy traffic or attack-induced interruptions.	High interrupt counts, combined with increased svmem_slab (high memory usage) and CPU processes like scputimes_user, point to network or system disruptions, often associated with DDoS or other attack strategies.
scpustats_ soft_ interrupts	CPU Software Interrupts Count	High number of software interrupts suggest network activity or potential attack.	High scpustats_soft_interrupts with high number of hardware interrupts (scpustats_interrupts) and high system process usage (scputimes_system) suggest an ongoing attack, such as DDoS or resource manipulation.

5.4.3. LIME—Local Behavior Analysis

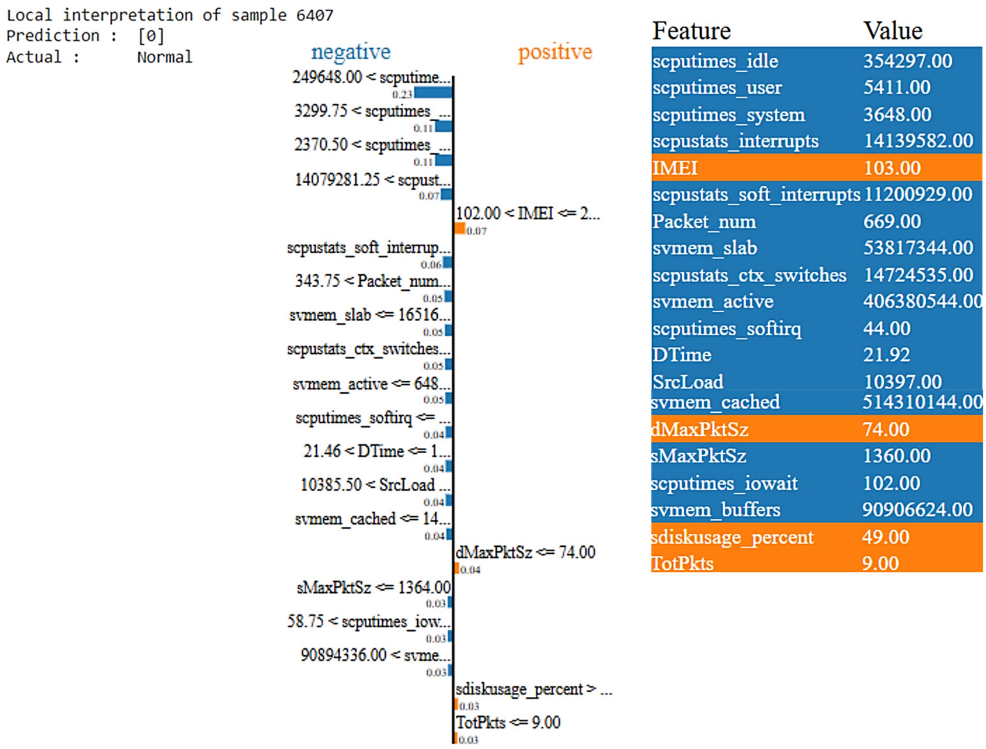
LIME are divided into three sections with a consistent color scheme:

- The left shows the model's predicted probability for the instance.
- The middle section shows key features: blue for benign (0), orange for attack (1). Bar length reflects feature impact, with longer bars indicating greater influence.
- The right section shows actual feature values. The features highlighted in blue contribute negatively to the prediction and the features highlighted in orange contribute positively to the prediction.

1. LIME Plot for Test Sample Record 1—'Normal' Traffic Prediction

Figure 11 illustrates a LIME for the "Normal" class instance. The following conclusions can be drawn:

- Although several features contribute negatively (blue) in the LIME plot, the prediction is still classified as "Normal" because the model evaluates the overall interactions of all features rather than assessing them individually.
- The positively contributing features (orange), such as dMaxPktSz, IMEI, and TotPkts, help to counterbalance the negative influences, leading to the final classification.



```
Local interpretation of sample 6407
Prediction :  [0]
Actual :      Normal
```

Feature	Value
scputimes_idle	354297.00
scputimes_user	5411.00
scputimes_system	3648.00
scpustats_interrupts	14139582.00
IMEI	103.00
scpustats_soft_interrupts	11200929.00
Packet_num	669.00
svmem_slab	53817344.00
scpustats_ctx_switches	14724535.00
svmem_active	406380544.00
scputimes_softirq	44.00
DTime	21.92
SrcLoad	10397.00
svmem_cached	514310144.00
dMaxPktSz	74.00
sMaxPktSz	1360.00
scputimes_iowait	102.00
svmem_buffers	90906624.00
sdiskusage_percent	49.00
TotPkts	9.00

Figure 11. LIME plot for record sample 1 from test dataset (local explanation)—predicting Class 0—"Normal" traffic.

Table 5 highlights the features and their contributions to the classification of network traffic as "Normal", based on the LIME plot, with interpretations of the feature values and the reasons for their positive or negative influence on the final prediction.

Table 5. LIME-based feature contribution analysis in normal prediction.

Features	Full Form	Interpretation	Reason for 'Normal' Prediction
IMEI	International Mobile Equipment Identity	A valid IMEI typically suggests legitimate device activity. No signs of spoofed or unauthorized device access.	Combined with normal packet sizes and disk usage, it reinforces legitimate network behavior.
dMaxPktSz	Maximum Packet Size in Flow	Network packet sizes are within the standard limits (74.00).	Along with low total packets and stable CPU usage, it confirms an absence of data tampering, data exfiltration, and packet flooding.
sdiskusage_percent	Percentage of Disk Space Used	Normal disk usage, with no indications of disk exhaustion or malicious activities like disk space manipulation or attacks.	The disk usage at 49% is moderate, indicating no signs of resource exhaustion, which is consistent with normal system operation.

Table 5. *Cont.*

Features	Full Form	Interpretation	Reason for 'Normal' Prediction
TotPkts	Total Packets Transmitted	Regular packet flow and no signs of a Distributed Denial-of-Service (DDoS) or heavy packet flooding attack.	With only nine packets transmitted, this indicates minimal network traffic, reinforcing the normal classification.
scputimes_ idle	CPU Idle Time	High idle time (354297.00) suggests that the system may be underutilized or have background processes like cryptojacking.	The features such as low packet traffic (TotPkts = 9) and moderate disk usage (e.g., sdiskusage_percent = 49%) dominate the normal classification.
scputimes_ user	CPU Time in User Mode	CPU time spent on user-level processes can signal abnormal activities or background attacks.	While the user CPU time is slightly higher (5411.00), it does not overwhelm the system, and the overall network behavior remains consistent with normal operations, with no abnormal spikes in traffic or disk usage.
scputimes_ system	CPU Time in System Mode	Increased system CPU time could point to resource manipulation during an attack.	The increase in system-level tasks is moderate (3648.00), and combined with stable traffic and disk usage, it does not indicate a full-fledged attack, supporting the normal classification.
scpustats_ interrupts	CPU Hardware Interrupts Count	A high number of interrupts (14139582.00) might indicate network stress or DDoS activity.	Although a high number of interrupts suggests some system load, the low packet transmission (nine total packets) confirms no DDoS activity or overwhelming network traffic, supporting a normal prediction.
scpustats_ softinterrupts	CPU Software Interrupts Count	A high number of software interrupts could indicate attack-induced disruptions like DDoS.	Despite the number of software interrupts being high (Value = 11200929.00), the network's low packet count (TotPkts = 9) and stable disk usage suggest that the system is not under attack, thus maintaining the normal prediction.

Table 5. *Cont.*

Features	Full Form	Interpretation	Reason for 'Normal' Prediction
Packet_num	Total Number of Packets in Flow	Elevated packet count might suggest malicious activity like cryptojacking.	The 669 packets could indicate some background activity, but the system does not show signs of attack. The normal disk usage and low packet size (dMaxPktSz = 74) reinforce normal network behavior.
svmem_slab	Kernel Slab Memory Usage	High kernel slab memory usage (53817344.00) could indicate memory leaks or excessive resource consumption, triggering significant memory anomalies like buffer overflows.	Despite the higher memory usage, the overall system behavior remains normal due to low traffic volume (TotPkts = 9) and moderate disk usage (sdiskusage_percent = 49%).
scpustats_ctx_switches	CPU Context Switches Count	High context switching (14724535.00) may indicate excessive task switching due to an attack.	Even with high context switching, other features like normal packet flow and moderate disk usage override the potential attack indicators, leading to the final normal prediction.
svmem_active	Active RAM Usage	High RAM usage could indicate background malicious activities consuming system resources.	Though active RAM usage is high (Value = 406380544.00), the system shows low traffic (TotPkts = 9) and moderate disk usage (sdiskusage_percent = 49%), confirming that the behavior is typical for a regular system not under attack.
scputimes_softirq	CPU Time Spent Handling Software Interrupts	Increased soft IRQ processing may point to a DDoS or other overload-based attack.	While the soft IRQ time is slightly higher (Value = 44.00), the system's low packet count and normal disk usage (sdiskusage_percent = 49%) maintain a normal classification despite the slight indication of an attack.

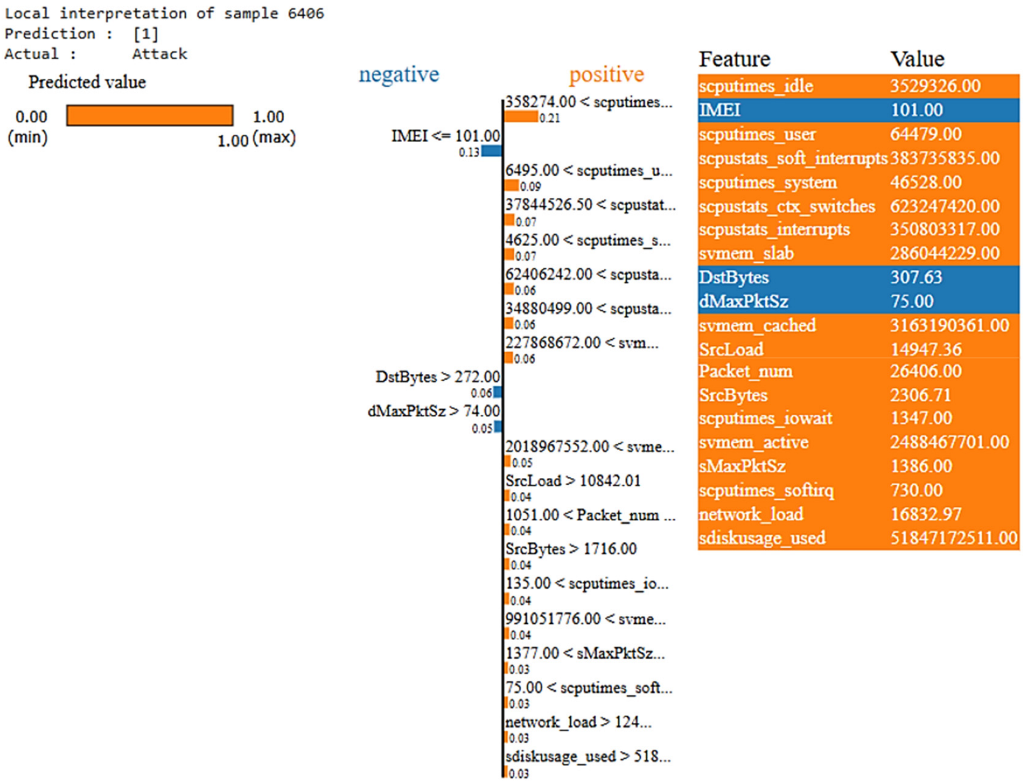
Table 5. *Cont.*

Features	Full Form	Interpretation	Reason for 'Normal' Prediction
DTime	Flow Duration (Seconds)	Extended flow duration could suggest slow exfiltration or DoS activities.	Even though the flow duration is higher (Value = 21.92), the low traffic (TotPkts = 9) and moderate system resource usage confirm that there is no attack, supporting the normal classification.
SrcLoad	Source Device Load	High source device load (Value = 10397.00) indicates the signs of excessive processing or stress typically associated with attack scenarios.	Although the source device load is high, the low packet transmission and normal disk usage reinforce the overall network stability, leading to the normal prediction.
svmem_ cached	Cached Memory Usage	High cached memory usage could indicate hidden attacks like cache poisoning or excessive resource hoarding.	The system remains stable with low packet flow (TotPkts = 9) and moderate disk usage (sdiskusage_percent = 49%), confirming that the system is not under attack.
sMaxPktSz	Maximum Packet Size Sent	Standard packet size suggests typical network traffic without evasion tactics.	The normal packet size (Value = 1360.00) aligns with expected network behavior and does not indicate any attack-related manipulation, confirming the normal classification.
scputimes_ iowait	CPU Time Waiting for I/O Operations	Increased I/O wait time could be indicative of delays due to attacks like DoS.	A low I/O wait time (Value = 102.00) suggests that disk operations are not experiencing delays, reinforcing that the system is not under excessive load from attack-driven I/O processes.
svmem_ buffers	Buffered Memory Usage	High buffered memory usage (Value = 90906624.00) might suggest resource exhaustion or buffer overflow attacks.	Despite higher buffered memory usage, the system's low packet transmission (TotPkts = 9) and moderate disk usage (sdiskusage_percent = 49%) confirm no attack, reinforcing the normal classification.

2. LIME Plot Test Sample Record 2—'Attack' Traffic Prediction

Figure 12 presents the LIME plot for test record 2, with the prediction of "Attack". The plot illustrates how different features influence the model's prediction.

- The blue features (IMEI, DstBytes and dMaxPktSz) contribute slightly towards a benign classification, as they do not exhibit strong attack characteristics. Their presence slightly reduces the likelihood of an attack but does not override the dominant attack-related features.
- The orange features dominate the decision, reflecting the abnormal CPU, memory, network, and disk behaviors commonly associated with attacks.



```
Local interpretation of sample 6406
Prediction :  [1]
Actual :      Attack
```

Feature	Value
scputimes_idle	3529326.00
IMEI	101.00
scputimes_user	64479.00
scpustats_soft_interrupts	383735835.00
scputimes_system	46528.00
scpustats_ctx_switches	623247420.00
scpustats_interrupts	350803317.00
svmem_slab	286044229.00
DstBytes	307.63
dMaxPktSz	75.00
svmem_cached	3163190361.00
SrcLoad	14947.36
Packet_num	26406.00
SrcBytes	2306.71
scputimes_iowait	1347.00
svmem_active	2488467701.00
sMaxPktSz	1386.00
scputimes_softirq	730.00
network_load	16832.97
sdiskusage_used	51847172511.00

Figure 12. LIME plot for record sample 1 from test dataset (local explanation)—predicting Class 1—"Attack" traffic.

Table 6 effectively summarizes how LIME interprets the feature contributions in determining the final prediction.

Table 6. LIME-based feature contribution analysis for attack classification.

Feature	Full Form	Interpretation	Reason for 'Attack' Prediction
IMEI	International Mobile Equipment Identity	A valid IMEI typically suggests legitimate device activity, reinforcing the normal prediction.	Despite the benign suggestion from the IMEI, other features like high CPU interrupts (350803317), high memory usage (2488467701), and high packet number (26406) dominate, pushing the prediction towards attack.
DstBytes	Destination Bytes	A moderate number of destination bytes suggests typical network behavior.	While DstBytes = 307.63 contributes slightly to the benign prediction, a high number of CPU interrupts (350803317) and high memory usage (2488467701) override this, resulting in an attack classification.

Table 6. *Cont.*

Feature	Full Form	Interpretation	Reason for 'Attack' Prediction
dMaxPktSz	Maximum Packet Size in Flow	Packet size within standard limits suggests no data exfiltration or packet flooding, contributing to a benign classification.	Despite dMaxPktSz = 75 indicating normal packet size, other features like a high number of CPU interrupts (350803317) and high system memory usage (2488467701) contribute more heavily, leading to an attack prediction.
scputimes_ idle	CPU Idle Time	High idle time suggests that the system may be underutilized, with potential background processes like cryptojacking.	scputimes_idle = 3529326 indicates idle CPU time, but a high number of CPU interrupts (350803317) and context switches (623247420) and high memory usage (2488467701) suggest attack-related activity, leading to an attack classification.
scputimes_ user	CPU Time in User Mode	Increased CPU time in user mode could indicate background attack activities, such as cryptojacking.	scputimes_user = 64479 suggests background activities but combined with a high number of CPU interrupts (350803317) and high memory usage (2488467701), this leads to an attack classification.
scpustats_ soft_ interrupts	CPU Software Interrupts Count	High number of software interrupts could indicate attack-related disruptions such as DDoS.	scpustats_soft_interrupts = 383735835 suggests attack-related disruptions, reinforcing the attack classification along with a high number of CPU interrupts and high memory usage.
scputimes_ system	CPU Time in System Mode	High CPU usage for system-level tasks might indicate resource manipulation, often seen in attack scenarios.	scputimes_system = 46528 shows moderate CPU usage, but a high number of CPU interrupts (350803317) and memory anomalies push the final prediction towards attack.
scpustats_ ctx_ switches	CPU Context Switches Count	High context switch counts suggest system instability and resource contention, often linked to attack scenarios.	scpustats_ctx_switches = 623247420 shows abnormal context switching, contributing to the attack prediction when combined with a high number of interrupts and high memory usage.
scpustats_ interrupts	CPU Hardware Interrupts Count	A high number of hardware interrupts indicate network stress or DDoS activity, strongly contributing to the attack classification.	scpustats_interrupts = 350803317 indicates a high number of interrupts, which, when combined with other orange features like memory usage (2488467701), strongly signal an attack.
svmem_ slab	Kernel Slab Memory Usage	High kernel slab memory usage could indicate potential memory manipulation during attacks.	svmem_slab = 286044229 indicates high kernel slab memory usage, contributing to attack prediction when combined with other high resource usage features.

Table 6. *Cont.*

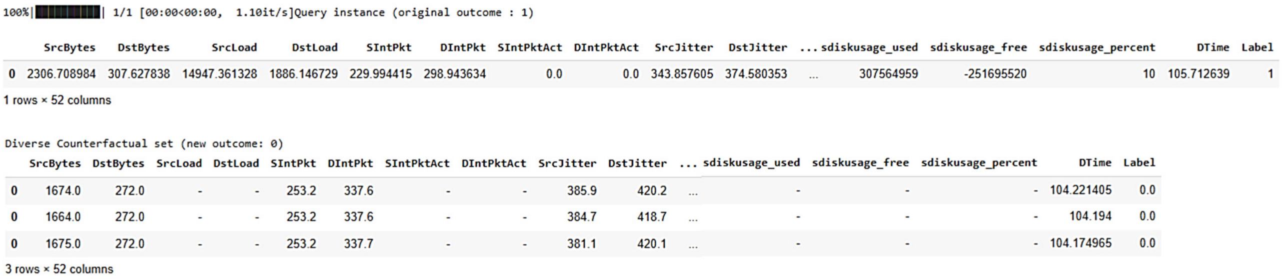
Feature	Full Form	Interpretation	Reason for 'Attack' Prediction
svmem_ cached	Cached Memory Usage	Cached memory represents data temporarily stored for quick access. A decrease in cached memory can indicate abnormal system activity.	svmem_cached = 3163190361 shows that the cached memory is relatively normal, but paired with high memory usage in other areas (svmem_active = 2488467701) and a high number of CPU interrupts (350803317), this indicates a system under heavy stress, possibly from an attack such as DDoS or data exfiltration.
SrcLoad	Source Device Load	High source load indicates that the source device is under heavy resource usage, often tied to attack scenarios.	SrcLoad = 14947.36 indicates that the source device is under heavy load. Combined with a high number of CPU interrupts (350803317) and high memory usage (2488467701), this contributes to the classification of the situation as an attack.
Packet_ num	Total Number of Packets in Flow	Elevated packet count might suggest background malicious activities like cryptojacking.	Packet_num = 26406 suggests higher packet flow, while a high number of CPU interrupts (350803317) and memory anomalies (2488467701) cause the final prediction to be an attack.
SrcBytes	Source Bytes	The number of bytes sent from the source device. An increase suggests high traffic, which could indicate a potential attack.	SrcBytes = 2306.71 suggests a moderate amount of traffic; however, it is overshadowed by other factors like a high number of CPU interrupts (350803317) and memory anomalies (2488467701), suggesting an attack.
scputimes_ iowait	CPU Time Waiting for I/O Operations	High I/O wait time can suggest system delays or bottlenecks, often indicative of attack-related activities.	scputimes_iowait = 1347 shows a moderate I/O wait time, but when combined with a high number of CPU interrupts (350803317) and high system memory usage (2488467701), it reinforces the attack prediction.
svmem_ active	Active RAM Usage	High active memory usage indicates excessive resource consumption, likely due to malicious activity.	svmem_active = 2488467701 shows high memory usage, signaling a potential attack, reinforced by other features like a high number of CPU interrupts (350803317).
sMaxPktSz	Maximum Packet Size Sent	The maximum size of network packets. Large packets might indicate a denial-of-service attack or abnormal network activity.	sMaxPktSz = 1386 indicates large packet sizes, to overwhelm network resources. Combined with a high CPU number of interrupts (350803317) and high memory usage (2488467701), this points to network stress and a potential data flooding attack, reinforcing the classification of an attack.

Table 6. *Cont.*

Feature	Full Form	Interpretation	Reason for 'Attack' Prediction
scputimes_softirq	CPU Time Spent Handling Software Interrupts	High soft IRQ processing could be a sign of a DDoS or other attack-related disruptions.	scputimes_softirq = 730 suggests an elevated level of CPU time spent handling software interrupts. When combined with a high number of CPU interrupts (350803317), high memory usage (2488467701), and network load (16832.97), it strongly suggests that the system is under attack, pushing the final prediction towards attack.
network_load	Network Load	High network load suggests unusual traffic patterns, commonly seen in attacks.	network_load = 16832.97 indicates high network traffic, reinforcing the attack classification when combined with other abnormal features.
sdiskusage_used	Percentage of Disk Space Used	High disk usage may indicate resource exhaustion, often used in attack strategies.	sdiskusage_used = 51847172511 shows abnormal disk usage, which, combined with a high number of CPU interrupts (350803317), points to an attack.

5.4.4. DiCE—Local Behavior Analysis

The DiCE is illustrated in Figure 13, where the top row displays the original sample with a prediction of '1' (indicating "Attack"). The bottom row presents three counterfactual scenarios, each with a prediction of '0' (indicating "Normal").



```
100%|██████████| 1/1 [00:00<00:00,  1.10it/s]Query instance (original outcome : 1)

     SrcBytes    DstBytes       SrcLoad      DstLoad     SIntPkt     DIntPkt  SIntPktAct  DIntPktAct   SrcJitter   DstJitter  ... sdiskusage_used  sdiskusage_free  sdiskusage_percent       DTime  Label
0 2306.708984  307.627838  14947.361328  1886.146729  229.994415  298.943634         0.0         0.0  343.857605  374.580353  ...       307564959       -251695520                  10  105.712639      1
1 rows × 52 columns

Diverse Counterfactual set (new outcome: 0)
  SrcBytes  DstBytes  SrcLoad  DstLoad  SIntPkt  DIntPkt  SIntPktAct  DIntPktAct  SrcJitter  DstJitter  ... sdiskusage_used  sdiskusage_free  sdiskusage_percent       DTime  Label
0   1674.0     272.0        -        -    253.2    337.6           -           -      385.9      420.2  ...               -                -                   -  104.221405    0.0
0   1664.0     272.0        -        -    253.2    337.6           -           -      384.7      418.7  ...               -                -                   -     104.194    0.0
0   1675.0     272.0        -        -    253.2    337.7           -           -      381.1      420.1  ...               -                -                   -  104.174965    0.0
3 rows × 52 columns
```

Figure 13. DiCE visualization of feature values and counterfactual instances.

From this, the following conclusions can be drawn:

- The combination of high traffic volume (SrcBytes, DstBytes), elevated system load (SrcLoad, DstLoad), and network instability (SrcJitter, DstJitter) indicates an anomalous network condition. These feature values typically suggest an ongoing attack, such as a DDoS attack, characterized by excessive data flow and network disruptions.
- The counterfactual examples state "normal" traffic due to significant reductions in traffic volume and jitter. Lower values for SrcBytes and DstBytes indicate typical network behavior with normal data flow, while the reduced SrcJitter and DstJitter suggest stable network conditions, which are characteristics of non-attack traffic patterns.
- Some values are marked with '-', indicating that these features were not altered in the counterfactual examples.

Table 7 shows the features identified via DiCE that contribute to the attack prediction, highlighting their influence on the model's decision. It details how changes in feature values affect the classification, offering insights into the factors driving the shift from normal to attack behavior.

Table 7. List of features identified via DiCE that contribute to attack prediction.

Features	Full Form	Value	Reason for Attack Prediction
SrcBytes	Source Bytes	2306.7	High number, possibly indicating abnormal data sent from the source, suggesting a DoS attack or data exfiltration attempt.
DstBytes	Destination Bytes	307.63	Relatively high value could indicate excessive data received by the destination, which is often linked to malicious activity like data stealing or flooding.
SrcLoad	Source Load	14947.4	High load on the source system, likely a result of malicious processes running, contributing to the attack classification.
DstLoad	Destination Load	1886.1	High load on the destination system, indicating abnormal resource usage typical of an attack scenario.
SrcJitter	Source Jitter	229.99	High jitter suggesting instability in the network, which is common during a network flooding attack.
DstJitter	Destination Jitter	298.94	High jitter, indicative of network instability due to an attack, like a DDoS or data breach.

Table 8 presents the features identified via DiCE and their counterfactual values that shift the prediction from attack to normal. It highlights the key features influencing the model's decision and shows the necessary changes to alter the classification outcome, providing insights into the decision boundaries of the predictive model.

Table 8. List of features identified via DiCE that contribute to normal prediction.

Feature	Full Form	Value 1	Value 2	Value 3	Reason for Normal Prediction
SrcBytes	Source Bytes	1674.0	1664.0	1675.0	Reduced value of SrcBytes indicates normal data sent from the source, reducing the likelihood of data exfiltration or DoS attack.
DstBytes	Destination Bytes	272.0	272.0	272.0	Lower DstBytes suggest that the data received by the destination are within the expected ranges, reflecting normal network behavior.

Table 8. *Cont.*

Feature	Full Form	Value 1	Value 2	Value 3	Reason for Normal Prediction
SrcLoad	Source Load	-	-	-	Unchanged, but its impact is less significant due to other feature changes.
DstLoad	Destination Load	-	-	-	Similar to SrcLoad, unchanged, but now less relevant as other features have been adjusted.
SrcJitter	Source Jitter	253.2	253.2	253.2	Lower jitters indicate stable network conditions, a key indicator of normal traffic.
DstJitter	Destination Jitter	337.6	337.6	337.7	Reduced jitters suggest stable network performance, removing the erratic behavior typical of attacks.

Note: '-' indicates that the values remain unchanged.

By analyzing DiCE features, key indicators such as traffic volume, system load, and network stability effectively distinguish between normal and attack scenarios. In 6G networks, with higher data rates and extensive device connectivity, these factors become critical for identifying security threats. A reduction in abnormal traffic volumes and network instability—indicated by lower SrcBytes and DstBytes and more stable SrcJitter and DstJitter—signals a stable, non-malicious network state, ensuring secure communication in 6G's high-speed environment. This analysis supports proactive risk monitoring, enhances response times, and optimizes resource allocation during security events.

6. Cross-Validation Results of XAI Methods

This section evaluates various XAI techniques, specifically, SHAP, LIME and DiCE, using cross-validation to assess their contributions to securing healthcare application within 6G environments. The analysis emphasizes the consistency and interpretability of feature importance across different XAI methods to improve transparency, trust, and robustness in AI-driven security systems for healthcare.

The interpretability analysis with SHAP, LIME, and DiCE offers critical insights into decision-making, enhancing the model's reliability in distinguishing normal from attack traffic. SHAP associate high system time ('scputimes_system'), user time ('scputimes_user'), and idle time ('scputimes_idle') with normal operations, whereas LIME warn that excessive idle time ('scputimes_idle') may indicate underutilization, potentially masking malicious activity, and high user time ('scputimes_user') could signal cryptojacking or resource-intensive attacks. This suggests that while CPU activity is a strong indicator of normal behavior, extreme deviations can imply anomalies. In network activity, DiCE identify source bytes ('SrcBytes') and destination bytes ('DstBytes') as indicators of normal traffic, with unusually high values signaling potential data exfiltration or flooding. Similarly, low source jitter ('SrcJitter') and destination jitter ('DstJitter') support a normal prediction, while high jitter suggests network instability or an attack, aligning with LIME' identification of unusual packet activity ('SIntPktAct') as a threat. Regarding system resource availability, LIME highlight available virtual memory ('svmem_free') and minimum packet size ('sMinPktSz') as stabilizing factors in normal predictions, complementing SHAP and DiCE, which focus more on CPU and network metrics. This cross-validation confirms that normal traffic is characterized by stable CPU usage, typical network activity, and sufficient resources, while anomalies in idle time ('scputimes_idle'), jitter ('SrcJitter', 'DstJitter'), and unusual packet behavior ('SIntPktAct') indicate potential attacks. This empirical analysis illustrates how a comprehensive, multi-faceted approach strengthens the security framework of 6G-enabled healthcare systems.

This evaluation confirms the reliability of explainability, showing that the identified features genuinely influence model predictions, not just artifacts of a specific method. The consistency across XAI techniques reinforces the trustworthiness of AI-driven decisions in critical applications, such as 6G medical network security, strengthening confidence in the transparency and dependability of the approach.

7. Mitigation Strategies in IoMT-Driven 6G Usage Scenarios

XAI insights play a vital role in enhancing the security of 6G-enabled medical IoT systems by transparently identifying critical features that influence both performance and potential vulnerabilities [11,30]. In Table 9, we analyzed various 6G usage scenarios with the help of XAI methods, which allowed us to identify key features contributing to both attack and non-attack behaviors. XAI techniques revealed which specific features influenced security outcomes—highlighting whether a high or low value of a given feature was associated with malicious or normal activity. By mapping these feature-based insights onto the functional requirements of each 6G usage scenario, we were able to pinpoint potential threats relevant to that scenario. Based on this analysis, we proposed targeted mitigation strategies aligned with the critical system behaviors and security risks of each use case. Furthermore, we presented XAI-driven security enhancements that empower security administrators to extract actionable insights and implement precise, scenario-specific measures—ensuring robust protection tailored to the unique demands of each 6G medical IoT use case.

Table 9. Mitigation strategies and security improvements for 6G usage scenarios based on XAI insights.

Usage Scenario	XAI Insights (SHAP, LIME, DiCE)	Security Risk Identified	Mitigation Strategy	Security Enhancements from XAI Insights
Immersive Communication	SHAP and LIME show high CPU I/O wait and memory usage. DiCE shows abnormal byte transfer and jitter values.	System overload, data exfiltration, latency spikes, and delayed data access.	AI-driven workload balancing, anomaly detection systems, and blockchain logging for data integrity [37,38].	SHAP and LIME insights support dynamic load balancing in real-time applications, reducing the risk of service disruption and improving system resilience. Meanwhile, DiCE identify suspicious byte patterns, which informs the integration of anomaly detection and blockchain-based logging—enhancing secure data transmission and threat response mechanisms.

Table 9. *Cont.*

Usage Scenario	XAI Insights (SHAP, LIME, DiCE)	Security Risk Identified	Mitigation Strategy	Security Enhancements from XAI Insights
High-Reliability Low-Latency Communication (HRLLC)	SHAP show high user CPU time and low idle time. LIME identify large packet sizes. DiCE detects packet anomalies.	DDoS, cryptojacking, malware, and latency disruption.	Behavioral-based detection, packet rate limiting, and secure boot with integrity checks [39–41].	SHAP identify signs of cryptojacking, prompting secure boot mechanisms to block unauthorized mining activities. LIME and DiCE detect abnormal packet sizes and flooding attempts, enabling proactive rate-limiting and behavioral firewalls to maintain system availability under high load conditions.
Massive Machine-Type Communication (mMTC)	LIME and SHAP detect spikes in memory usage (svmem_active, svmem_slab) and packet numbers. DiCE detects traffic anomalies.	Cryptojacking, DDoS botnets, network congestion, and incorrect data.	Memory profiling tools, smart traffic filtering, and task scheduling on MEC [37,39,40].	LIME′ detection of memory spikes facilitates memory leak identification and profiling. SHAP′ analysis of slab usage supports enhanced memory management, while DiCE′ detection of traffic anomalies enables intelligent traffic filtering—together ensuring the stability and efficiency of IoT systems.
Ubiquitous Connectivity	SHAP show I/O waits and memory inefficiencies. LIME indicates excessive context switches and interrupts. DiCE detect jitter.	Resource exhaustion, botnet attacks, and unstable connectivity.	Dynamic resource allocation, optimized routing, and jitter buffers [37,39,40].	SHAP′ identification of I/O delays prompts dynamic resource reallocation to prevent system slowdowns. LIME detect kernel-level stress, leading to targeted system tuning, while DiCE′ jitter detection enables optimized routing and jitter buffering—collectively ensuring stable and reliable connectivity in 6G-enabled IoT environments.
AI and Communication	SHAP and LIME show high CPU/memory use and large destination byte sizes. DiCE highlight packet count and jitter anomalies.	Cryptojacking, DDoS, and degraded AI inference.	AI-based anomaly detection, blockchain logging, and adaptive model offloading/routing [37,38].	SHAP and LIME insights facilitate dynamic load distribution for AI models, preventing system overloads. DiCE′ detection of jitter and traffic spikes drives the implementation of blockchain-backed logging and adaptive routing, ensuring AI service reliability even in unstable network conditions.

Table 9. *Cont.*

Usage Scenario	XAI Insights (SHAP, LIME, DiCE)	Security Risk Identified	Mitigation Strategy	Security Enhancements from XAI Insights
Integrated Sensing and Communication (ISAC)	SHAP identifies high disk and active memory usage. LIME and DiCE show large packet volumes and jitter irregularities.	DDoS, data exfiltration, ransomware, and flooding attacks.	Zero-trust access control, predictive analytics for attack detection, and backup and isolation mechanisms [41,42].	SHAP′ identification of high disk and memory usage triggers backup and isolation to prevent ransomware damage. LIME′ and DiCE′ insights into heavy traffic and jitter patterns enable predictive anomaly detection, while zero-trust access control blocks further unauthorized activity.

Understanding the security implications of these attacks will help to safeguard emerging 6G technologies—ensuring resilient, safe, and intelligent healthcare systems in the future.

8. Conclusions

The convergence of advanced wireless communication and healthcare technologies presents significant risks to the integrity of sensitive patient data and the performance of medical applications. As the transition to 6G accelerates, the increased integration of virtualization, extended reality, intelligent sensing, and AI into communication frameworks expands the attack surface, increasing risks to patient data integrity and the reliability of life-critical medical applications. Left unaddressed, these emerging threats could result in severe service disruptions and potential catastrophic outcomes in clinical environments. While AI is instrumental in detecting and mitigating complex security threats, its inherent black box nature can hinder transparency, making it difficult for security teams to fully understand and trust automated decisions. This limitation highlights the necessity of XAI in the context of 6G-enabled healthcare systems.

Our research highlights the crucial role of XAI in closing the explainability gap in AI-driven security systems. By integrating SHAP, LIME, and DiCE, we propose a comprehensive framework that demystifies model decisions and makes security insights both interpretable and actionable for technical and non-technical audiences alike. This multi-perspective analysis enhances the understanding of model behavior, enabling more adaptive and informed security responses. Our approach not only reinforces the consistency and dependability of XAI-based assessments but also underlines the significance of explainability in fostering trust in AI systems, particularly in sensitive and high-stakes environments.

As healthcare systems advance toward 6G connectivity, the reliability and security of AI-powered medical technologies will be essential for delivering intelligent, timely, and safe patient care. This study supports that goal by contributing to the development of secure, transparent, and adaptive AI models, specifically designed for integration into next-generation healthcare infrastructure. Future work will focus on incorporating datasets specific to 6G, either original or simulated, instead of using 5G-originated datasets, and exploring more advanced models tailored to 6G networks to further enhance the robustness and scalability of AI-based security systems in these evolving environments.

Author Contributions: Conceptualization, L.G. and N.K.; Methodology, L.G. and N.K.; Software, N.K.; Validation, L.G. and N.K.; Formal analysis, N.K.; Investigation, N.K.; Resources, L.G. and N.K.; Data curation, N.K.; Writing—original draft, N.K.; Writing—review & editing, L.G.; Visualization,

N.K.; Supervision, L.G.; Project administration, L.G. and N.K.; Funding acquisition, L.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shen, Y.T.; Chen, L.; Yue, W.W.; Xu, H.X. Digital technology-based telemedicine for the COVID-19 pandemic. *Front. Med.* **2021**, *8*, 646506. [CrossRef] [PubMed]
2. Osama, M. Internet of medical things and healthcare 4.0: Trends, requirements, challenges, and research directions. *Sensors* **2023**, *23*, 7435. [CrossRef] [PubMed]
3. Yaqoob, T.; Abbas, H.; Atiquzzaman, M. Security vulnerabilities, attacks, countermeasures, and regulations of networked medical devices—A review. *IEEE Comst* **2019**, *21*, 3723–3768. [CrossRef]
4. Framework and Overall Objectives of the Future Development of IMT for 2030 and Beyond. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2160-0-202311-I!!PDF-E.pdf (accessed on 11 July 2024).
5. Ghubaish, A.; Yang, Z.; Jain, R. HDRL-IDS: A Hybrid Deep Reinforcement Learning Intrusion Detection System for Enhancing the Security of Medical Applications in 5G Networks. In Proceedings of the 2024 International Conference on Smart Applications, Communications and Networking (SmartNets), Harrisonburg/Washington, DC, USA, 28–30 May 2024; pp. 1–6.
6. Hossain, M.S.; Muhammad, G.; Guizani, N. Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics. *IEEE Netw.* **2020**, *34*, 126–132. [CrossRef]
7. Dave, D.; Naik, H.; Singhal, S.; Patel, P. Explainable ai meets healthcare: A study on heart disease dataset. *arXiv* **2020**, arXiv:2011.03195.
8. Bárcena, J.L.C.; Ducange, P.; Marcelloni, F.; Nardini, G.; Noferi, A.; Renda, A.; Ruffini, F.; Schiavo, A.; Stea, G.; Virdis, A. Enabling federated learning of explainable AI models within beyond-5G/6G networks. *Comput. Commun.* **2023**, *210*, 356–375. [CrossRef]
9. Mohanty, S.D.; Lekan, D.; Jenkins, M.; Manda, P. Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare. *Patterns* **2022**, *3*, 100395. [CrossRef]
10. Nguyen, V.L.; Lin, P.C.; Cheng, B.C.; Hwang, R.H.; Lin, Y.D. Security and privacy for 6G: A survey on prospective technologies and challenges. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 2384–2428. [CrossRef]
11. Kaur, N.; Gupta, L. Securing the 6G–IoT Environment: A Framework for Enhancing Transparency in Artificial Intelligence Decision-Making Through Explainable Artificial Intelligence. *Sensors* **2025**, *25*, 854. [CrossRef]
12. Musa, A. History of security and privacy in wireless communication systems: Open research issues and future directions. In *Security and Privacy Schemes for Dense 6G Wireless Communication Networks*; The Institution of Engineering and Technology: Stevenage, UK, 2023; pp. 31–60. [CrossRef]
13. Xiao, Y.; Jia, Y.; Liu, C.; Cheng, X.; Yu, J.; Lv, W. Edge computing security: State of the art and challenges. *Proc. IEEE* **2019**, *107*, 1608–1631. [CrossRef]
14. Deng, J.; Han, R.; Mishra, S. Decorrelating wireless sensor network traffic to inhibit traffic analysis attacks. *Pervasive Mob. Comput.* **2006**, *2*, 159–186. [CrossRef]
15. Wang, S.; Parsons, M.; Stone-McLean, J.; Rogers, P.; Boyd, S.; Hoover, K.; Meruvia-Pastor, O.; Gong, M.; Smith, A. Augmented reality as a telemedicine platform for remote procedural training. *Sensors* **2017**, *17*, 2294. [CrossRef]
16. Newaz, A.I.; Sikder, A.K.; Rahman, M.A.; Uluagac, A.S. A survey on security and privacy issues in modern healthcare systems: Attacks and defenses. *ACM Trans. Comput. Healthc.* **2021**, *2*, 27. [CrossRef]
17. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Systems* **2017**, *30*.
18. Ribeiro, M.T.; Singh, S.; Guestrin, C. "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, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
19. Mothilal, R.K.; Sharma, A.; Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–29 January 2020; pp. 607–617.
20. Amin, A.; Hasan, K.; Zein-Sabatto, S.; Chimba, D.; Ahmed, I.; Islam, T. An explainable ai framework for artificial intelligence of medical things. In Proceedings of the IEEE Globecom Workshops (GC Wkshps), Kuala Lumpur, Malaysia, 4–8 December 2023; pp. 2097–2102.

21. Sai, S.; Sharma, S.; Chamola, V. Explainable ai-empowered neuromorphic computing framework for consumer healthcare. *IEEE Trans. Consum. Electron.* **2024**. [CrossRef]
22. Gürbüz, E.; Turgut, O.; Kök, I. Explainable ai-based malicious traffic detection and monitoring system in next-gen iot healthcare. In Proceedings of the 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), Istanbul, Türkiye, 25–27 July 2023; pp. 1–6.
23. Alani, M.M.; Mashatan, A.; Miri, A. Explainable Ensemble-Based Detection of Cyber Attacks on Internet of Medical Things. In Proceedings of the 2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Abu Dhabi, United Arab Emirates, 13–17 November 2023; pp. 0609–0614.
24. Alani, M.M.; Mashatan, A.; Miri, A. XMeDNN: An Explainable Deep Neural Network System for Intrusion Detection in Internet of Medical Things. In Proceedings of the 9th International Conference on Information Systems Security and Privacy (ICISSP 2023), Lisbon, Portugal, 22–24 February 2023; pp. 144–151.
25. Raza, A.; Tran, K.P.; Koehl, L.; Li, S. Designing ECG monitoring healthcare system with federated transfer learning and explainable AI. *Knowl.-Based Syst.* **2022**, *236*, 107763. [CrossRef]
26. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
27. Nasralla, M.M.; Khattak, S.B.A.; Rehman, I.U.; Iqbal, M. Exploring the role of 6G technology in enhancing quality of experience for m-health multimedia applications: A comprehensive survey. *Sensors* **2023**, *23*, 5882. [CrossRef]
28. Wood, A.D.; Stankovic, J.A. A taxonomy for denial-of-service attacks in wireless sensor networks. In *Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems*; CRC Press: Boca Raton, FL, USA, 2004; Volume 4, pp. 739–763.
29. Kaur, N.; Gupta, L. An approach to enhance iot security in 6g networks through explainable ai. *arXiv* **2024**, arXiv:2410.05310.
30. Kaur, N.; Gupta, L. Enhancing IoT Security in 6G Environment With Transparent AI: Leveraging XGBoost, SHAP and LIME. In Proceedings of the 2024 IEEE 10th NetSoft, Saint Louis, MO, USA, 24–28 June 2024; pp. 180–184.
31. Cramer, J.S. *The Origins of Logistic Regression*; Social Science Research Network: Rochester, NY, USA, 2002.
32. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [CrossRef]
33. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
34. Wu, J. Introduction to convolutional neural networks. *Natl. Key Lab Nov. Softw. Technol. Nanjing Univ. China* **2017**, *5*, 495.
35. Pandas. Dataframe. Available online: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html (accessed on 2 February 2024).
36. Permutation Importance vs. Random Forest Feature Importance (MDI). Available online: https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-py (accessed on 26 March 2024).
37. Ramamoorthi, V. Exploring AI-Driven Cloud-Edge Orchestration for IoT Applications. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2023**, *9*, 385–393. [CrossRef]
38. Adele, G.; Borah, A.; Paranjothi, A.; Khan, M.S.; Poulkov, V.K. A Comprehensive Systematic Review of Blockchain-based Intrusion Detection Systems. In Proceedings of the 2024 IEEE World AI IoT Congress (AIIoT), Melbourne, Australia, 24–26 July 2024; pp. 605–611.
39. El-Hajj, M. Enhancing Communication Networks in the New Era with Artificial Intelligence: Techniques, Applications, and Future Directions. *Network* **2025**, *5*, 1. [CrossRef]
40. Paulsen, S.; Uhl, T.; Nowicki, K. Influence of the jitter buffer on the quality of service VoIP. In Proceedings of the 2011 3rd International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), Budapest, Hungary, 5–7 October 2011; pp. 1–5.
41. Sharma, H. Behavioral Analytics and Zero Trust. *Int. J. Comput. Eng. Technol.* **2021**, *12*, 63–84.
42. Quattrociocchi, W.; Caldarelli, G.; Scala, A.A. Self-healing networks: Redundancy and structure. *PLoS ONE* **2014**, *9*, e87986. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.