Minimal but Lethal: A XAI-Driven Approach for Feature-Level Adversarial Attacks on Healthcare 5.0

Lucas P. Siqueira[1], Pedro H. Lui[1], Juliano F. Kazienko[1], Silvio E. Quincozes[2,3], Vagner E. Quincozes[4], and Daniel Welfer[1]

[1]PPGCC – Universidade Federal de Santa Maria (UFSM)
[2]AI Horizon Labs/PPGES – Universidade Federal do Pampa (UNIPAMPA)
[3]PPGCO – Universidade Federal de Uberlândia (UFU)
[4]IC – Universidade Federal Fluminense (UFF)

lucas.pittella@acad.ufsm.br, {pedro.lui, kazienko}@redes.ufsm.br
vequincozes@id.uff.br, silvioquincozes@unipampa.edu.br
daniel.welfer@ufsm.br

Abstract. *In Healthcare 5.0, the expanded attack surface increases the vulnerability of Intrusion Detection Systems (IDS) to sophisticated threats. Among them, adversarial attacks modify features to evade the detection of malicious samples. XAI-driven methods enable the manipulation of fewer — sometimes just one—features while maximizing impact. To date, no XAI-driven adversarial strategy has been applied to cyber-biomedical features in Healthcare 5.0. In this work, we address this gap by employing XAI-Driven approach to maximize IDS degradation through a feature-level adversarial attacks. Our results reveals that a single feature perturbed can drastically reducing F1-Score from 99% to 0% in data alteration scenarios and from 81% to 12% in spoofing attacks.*

1. Introduction

Healthcare 5.0 marks a deep advancement in the digital transformation of medicine by unifying Artificial Intelligence (AI), Internet of Things (IoT) and human-centric design to deliver precise, personalized treatment, predictive diagnostics and continuous monitoring [Gadekallu et al. 2024]. With the global Digital Health market expected to surpass $258 billion by 2029 [Statista 2025], cybersecurity within the Internet of Medical Things (IoMT) has become a critical concern. As connected medical devices, telehealth platforms, and AI-driven diagnostics become foundational to modern healthcare, their dependence on sensitive patient data and networked environments significantly amplifies exposure to cyber threats. A single security breach can endanger patient safety, disrupt essential clinical services, and erode public trust in digital health infrastructure. Alarmingly, recent findings reveal a 44% year-over-year increase in global cyber-attacks, with healthcare identified as the second most targeted industry — highlighting the sector's growing risk profile [Check Point Software Technologies 2025].

These advancements also create complex attack surfaces that traditional Intrusion Detection System (IDSs) are ill-equipped to defend. To address this, recent research has developed network-biomedical datasets that combine network traffic with physiological signals for enhanced intrusion detection in electronic health monitoring systems [Hady et al. 2020]. In addition, a major challenge faced by modern IDSs involves

adversarial attacks, which have gained attention for their ability to subtly manipulate input data in ways that degrade detection performance without raising suspicion. These attacks exploit the decision boundaries of Machine Learning (ML) models, often requiring only minimal perturbations to cause misclassifications — posing a serious threat to systems that rely on AI for anomaly detection.

Explainable AI (XAI) methods — increasingly integrated into Healthcare 5.0 IDS — generate saliency maps and feature-attribution scores that reveal which network or physiological signals trigger an alert. Yet these explanations can be weaponized: attackers use XAI outputs to pinpoint and minimally perturb critical features, mounting stealthy adversarial attacks [Okada et al. 2025] [Yan et al. 2024]. While traditional adversarial techniques have been explored separately for network and sensor data, and XAI-driven attacks tested on pure IDS datasets, no study has assessed explanation-guided attacks on combined IDS and physiological streams or compared their effectiveness against classic methods within Healthcare 5.0. We address this gap by evaluating IDS vulnerability and robustness under XAI-informed threat models in a converged healthcare environment.

In this work, we (i) adapt the XAI-driven adversarial attack technique from [Okada et al. 2025] to evade machine learning–based intrusion detection systems (IDSs), and (ii) evaluate its effectiveness against a classical adversarial approach on a Healthcare 5.0–aligned dataset. We first apply XAI techniques to rank feature importance on the WUSTL-EHMS-2020 dataset [Hady et al. 2020] and then craft targeted perturbations on the most influential features. Although focused on the healthcare domain, our methodology is domain-agnostic and applicable to other sectors. Importantly, this study evaluates adversarial impact in a controlled setting; the deployment and configuration of real-world IDSs are out of scope. Our results show that perturbing a single feature can reduce the F1-score from 99% to 0% in data alteration scenarios and from 81% to 12% in spoofing attacks. These findings highlight the need for explanation-aware defense strategies.

The paper is structured as follows. Section 2 introduces the fundamental concepts. Section 3 reviews related work on XAI-driven adversarial attacks in medical networks and outlines key research gaps. Section 4 details our methodology, including XAI-based feature ranking and the design of targeted perturbations. Section 5 presents the results and analysis. Finally, Section 6 concludes the paper and suggests future research.

2. Background

Firstly, this section outlines the key foundations for our XAI-driven adversarial framework in Healthcare 5.0. We begin by defining Healthcare 5.0 — its smart, data-driven technologies and the security risks they introduce. Next, we cover XAI methods that render complex models interpretable. Then, we trace the evolution of network Intrusion Detection Systems, emphasizing modern ML- and DL-based approaches in healthcare settings. We follow with a discussion of how XAI outputs can be repurposed for adversarial attacks, and finally review classic white-box and black-box attack strategies, contrasting unconstrained and constrained domains.

2.1. Healthcare 5.0

Healthcare 5.0 marks a transformative phase in medical care, emphasizing personalization, anticipatory services, and patient engagement through the adoption of technologies

such as smart sensors, ML, and the IoMT. Devices like wearable biosensors and fitness monitors support continuous health tracking, enabling both remote diagnostics and improved clinical insights through real-time data processing. The convergence of ML and IoMT facilitates early diagnosis, data-driven decision-making, and precision medicine, thereby enhancing healthcare delivery and patient well-being [Tandel et al. 2024].

At the same time, the widespread deployment of networked medical devices introduces new security challenges. The interconnected nature of digital health infrastructures increases vulnerability to cyberattacks, underscoring the need for effective protective measures. Leveraging AI for cybersecurity enables the detection of abnormal behavior, threat identification, and maintenance of device reliability. These measures are vital to safeguard patient information and uphold trust, ensuring that Healthcare 5.0 achieves its goals without compromising security [Khan et al. 2024].

2.2. Explainable AI

In recent years, deep learning models have achieved remarkable performance but their complex, non-linear structure often renders their decision processes opaque and difficult to interpret. XAI addresses this challenge by offering model-agnostic, post-hoc methods that quantify how input features contribute to individual predictions, without requiring access to internal weights or gradients [Okada et al. 2025]. These techniques range from local explanations — highlighting feature importances for a single instance — to global explanations that summarize patterns across a dataset, supporting tasks such as regulatory auditing, model debugging, and scientific discovery [Baniecki and Biecek 2024]. By illuminating the "black box," XAI methods foster greater transparency, trust, and accountability in AI systems.

2.3. Intrusion Detection Systems

Intrusion Detection Systems (IDS) for networks are deployed at strategic points to inspect all traffic and have progressively shifted from purely signature-based approaches to advanced advanced deep learning methods capable of detecting both known and unknown threats [Okada et al. 2025]. These deep learning–powered NIDS can operate in binary or multiclass modes, learning signatures automatically from mixed benign and malicious datasets, or in anomaly-detection mode by modeling only legitimate traffic to flag deviations. Machine learning frameworks grounded in computational statistics and optimization combine misuse detection via learned signatures with anomaly detection to identify zero-day attacks by learning normal patterns in network and biometric data. In healthcare contexts such as Enhanced Healthcare Monitoring Systems, IDS modules analyze sensor-derived biometric signals and network traffic metrics, employing algorithms like Random Forest, K-Nearest Neighbors, Support Vector Machines, and Artificial Neural Networks to detect packet alteration, spoofing, and other irregularities, thereby safeguarding patient data integrity and system availability [Hady et al. 2020].

2.4. XAI-Driven Adversarial Attacks

XAI techniques, designed to interpret machine learning decisions, can also be misused to guide adversarial behavior. Attackers can analyze outputs such as saliency maps or feature attributions to identify key input features and apply minimal perturbations to craft adversarial examples or extract sensitive information [Okada et al. 2025]. These XAI-driven

attacks apply across white-box and gray-box settings, and have enabled model inversion and membership inference. A typical adversarial example (AE) attack seeks a perturbed input that minimally deviates from the original while inducing misclassification:

$$\begin{aligned} \text{minimize} \quad & \|x' - x\| \\ \text{subject to} \quad & f(x') = l', \\ & f(x) = l, \\ & l \neq l', \\ & x' \in [0,1]^m, \end{aligned} \tag{1}$$

where $x \in [0,1]^m$ is an m-dimensional input to classifier f, l is the true label, and $l' \neq l$ is the target label. The adversarial input $x' = x + r$ is crafted to fool the model while remaining as close as possible to x, ideally with imperceptible changes [Bayer et al. 2024].

Adversarial attacks are typically classified as *white-box* (full access to the model's architecture, parameters, activations, and loss), *black-box* (only query/test-set access and output observations), or *gray-box* (partial knowledge in between) [Vázquez-Hernández et al. 2024].

2.5. Adversarial Attacks

Classic adversarial attacks typically exploit model gradients to craft imperceptible perturbations that lead to misclassification. The Fast Gradient Sign Method (FGSM) perturbs every input dimension by applying a small, uniform step in the direction of the loss gradient sign for each feature, resulting in a dense one-step modification of the entire input with minimal computational overhead [Goodfellow et al. 2014]. In contrast, HopSkipJumpAttack is a decision-based, black-box approach: it begins by finding a point on the model's decision boundary via binary search from a large initial perturbation, then approximates the local gradient direction by sampling random unit vectors and using finite-difference estimates, and finally updates the adversarial example by stepping along that estimated gradient and projecting back onto the boundary [Chen et al. 2020]. Both FGSM and HopSkipJumpAttack demonstrate that, whether in white-box or black-box settings, carefully designed perturbations can induce misclassification.

Adversarial attacks can be distinguished by whether they operate in an unconstrained or a constrained domain [Alhajjar et al. 2021]. In an unconstrained setting — common in fields like image recognition — attackers are assumed to have the freedom to adjust every feature of the input without limitation, allowing them to craft perturbations that subtly alter the entire sample to induce errors. In contrast, constrained domains impose strict rules on which features can be modified (for example, some attributes may be binary or categorical), enforce interdependencies among features, and include elements that cannot be changed at all. Under these conditions, adversaries must carefully select and adjust only those characteristics that influence the model's decision and respect the domain's inherent value ranges, correlations and immutable constraints.

3. Related Works

In recent years, there has been growing scholarly interest in combining Healthcare 5.0 concepts with cybersecurity mechanisms, particularly in the context of XAI and adversarial strategies that leverage XAI techniques. This section surveys the existing literature in

Table 1. Comparison to the Related Works.

Reference	XAI-Driven Attack (A)	Classic Adversarial Attack (B)	Comparison (A) × (B)	IDS	Healthcare 5.0
[Okada et al. 2025]	✓	✗	✗	✓	✗
[Yan et al. 2024]	✓	✗	✗	✗	✗
[Kuppa and Le-Khac 2021]	✓	✗	✗	✗	✗
[Bayer et al. 2024]	✓	✗	✗	✗	✗
[Rosenberg et al. 2020]	✓	✗	✗	✗	✗
[Alhajjar et al. 2021]	✗	✓	✗	✓	✗
[Imam 2024]	✗	✓	✗	✗	✓
[Agrawal et al. 2024]	✗	✓	✗	✗	✓
[Brohi and Mastoi 2025]	✗	✓	✗	✗	✓
[Baniecki and Biecek 2024]	✗	✓	✗	✗	✗
[Asiri et al. 2024]	✗	✓	✗	✗	✗
Our work	✓	✓	✓	✓	✓

this interdisciplinary area. A comparative overview is also provided in Table 1, highlighting the methods and shortcomings of prior studies in contrast with the present work.

Building on prior research that developed XAI-driven white-box adversarial attacks against DL-based NIDS, the research [Okada et al. 2025] advances the approach to a practical black-box scenario. By leveraging XAI to identify critical features without requiring internal model details, the proposed method generates adversarial examples (AEs) that evade multiple NIDS models with high efficacy (95.7–100% evasion rates) across diverse attack scenarios. Experimental validation in real-world networks demonstrates the method's generalizability and practicality, enabling robust assessment of NIDS vulnerabilities while preserving the malicious intent of attack traffic. This contribution underscores the dual role of XAI in enhancing both adversarial robustness evaluation and attack mitigation strategies in cybersecurity.

The work [Yan et al. 2024] proposes MEAttack, a model-agnostic explanation-driven method for query-efficient black-box adversarial attacks, addressing limitations of optimization and transfer-based approaches. By leveraging model explanations to interpret decision boundaries and strategically perturb inputs, MEAttack achieves non-targeted success rates 4.54%–47.42% higher than AutoZOOM while reducing queries by 2.6–4.2×, notably attaining 20.8× higher success rates with 10× fewer queries on MNIST. Despite computational overhead in local model training, MEAttack establishes a query-efficient benchmark for label-only black-box attacks.

The work [Kuppa and Le-Khac 2021] demonstrates that counterfactual XAI methods, despite enhancing transparency, expose systems to attacks like membership inference, model extraction, evasion, poisoning, and backdoors. Their four black-box attacks, leveraging three explanation methods, compromise classifier confidentiality and privacy through anti-virus evasion, sensitive data inference, and model extraction on real-world datasets. The study underscores XAI's security-usability trade-off, cautioning against unintended attack surface expansion.

Recent research has demonstrated how XAI can be leveraged not only to interpret model behavior but also to craft more effective adversarial attacks. One line of work utilizes XAI techniques to identify features that models rely on most, enabling the con-

struction of targeted attacks that exploit these dependencies. For example, one approach enhances adversarial training by identifying falsely learned indicators in incorrectly predicted instances and balancing decision rule complexity to boost robustness and mitigate shortcut learning behaviors in NLP transformers [Bayer et al. 2024]. Similarly, in the cybersecurity domain, XAI has been used to guide feature modifications in malware to evade detection, revealing a security-transparency trade-off by showing how explainability can assist attackers in modifying critical parts of executable files without compromising functionality [Rosenberg et al. 2020].

[Alhajjar et al. 2021] explored adversarial attacks on ML-based Network Intrusion Detection Systems (NIDS), using algorithms (particle swarm optimization, genetic algorithm) and GANs to generate evasive perturbations. Evaluated on NSL-KDD and UNSW-NB15 datasets, their methods induced high misclassification rates in 11 ML models and a voting classifier, surpassing Monte Carlo simulations. Key results highlighted extreme vulnerability of SVM and DT classifiers, transferability of adversarial examples (e.g., PSO-generated perturbations evaded gradient boosting and bagging classifiers at more than 97%), and a functionality-preserving feature modification strategy, diverging from prior approaches. The work underscores ML-based NIDS susceptibility to evolutionary and deep learning-driven adversarial manipulation.

In the healthcare domain, studies emphasize the interplay between adversarial robustness and explainability, recognizing both as essential for trustworthy AI deployment. One body of work investigates how adversarial attacks undermine the reliability of the explanation, even in robustly trained models, and proposes combined detection-classification frameworks using XAI methods such as SHAP and Grad-CAM to localize the influence of the attack and assess the degradation of interpretability [Agrawal et al. 2024]. Another effort explores adversarial threats targeting XAI-enabled digital twin systems for smart healthcare, where label-flipping attacks significantly distort model explanations in clinical settings such as stroke prediction. A resilient digital twin framework is proposed to protect interpretability and preserve the integrity of decision-making-making [Imam 2024]. Complementing these studies, a separate evaluation of an optimized MLP model for breast cancer classification under FGSM attacks shows dramatic accuracy loss, underscoring the necessity for lifecycle-wide integration of adversarial defenses and XAI tools to ensure clinical AI robustness [Brohi and Mastoi 2025].

Studies reveal that XAI methods (e.g., SHAP, Grad-CAM) themselves are susceptible to adversarial attacks, which can skew explanations or fairness evaluations [Baniecki and Biecek 2024]. Key gaps include vulnerabilities in transformer-based explainers and fairness metrics, underscoring the need for standardized benchmarks and robust XAI practices. Empirical work shows that FGSM perturbations degrade Saliency Map accuracy and explanation quality (via PIQE) on image datasets — suggesting explanation breakdowns can signal adversarial activity [Asiri et al. 2024].

4. Material & Methods

In this section, we present the methodology behind our proposed adversarial attack in a Healthcare 5.0 scenario. Also, we present the dataset utilized alongside XAI techniques to systematically perturb the most influential features.

4.1. Dataset

The WUSTL-EHMS-2020 dataset [Hady et al. 2020], adopted in this research, provides a unique resource for evaluating intrusion detection systems in Healthcare 5.0 environments. It was assembled from a controlled testbed that simulated a medical monitoring system where Internet of Medical Things (IoMT) sensors transmitted physiological data, such as heart rate and SpO2, to a server. The transmission path was deliberately exposed to interception by malicious actors, enabling the collection of data under both normal conditions and simulated cyberattacks.

Comprising a total of 16,318 samples, the dataset is notable for its composition of multimodal data: it integrates 35 network-flow features with 8 biometric measurements. Malicious traffic accounts for 12.5% of the samples and is classified into two distinct types: spoofing attacks, designed to compromise data confidentiality through packet manipulation, and data-alteration attacks, which target data integrity by modifying packet contents. By merging these two data streams, the dataset allows for a comprehensive analysis of cyber-physical threats, providing a robust foundation for developing and testing intelligent, context-aware security solutions.

4.2. Attacker Model

We consider a realistic threat scenario in which an internal actor — such as a system administrator, a member of the development team, or a third-party contractor — compromises the IDS by leaking the model and its training data to an external adversary. With privileged access to implementation details and potentially the original data, the attacker gains white-box access, including full knowledge of the model architecture, parameters, and feature preprocessing steps. Leveraging this access, the adversary can apply XAI-driven adversarial attack techniques to identify the most influential features in the IDS's decision-making process. By subtly manipulating only these key features, the attacker can craft adversarial inputs that preserve malicious behavior while evading detection.

Although our work focuses on white-box attacks reflecting insider threats, black-box attacks, where the adversary has limited or no knowledge of model internals, remain highly relevant in real-world scenarios due to hidden datasets and restricted model access. In particular, [Okada et al. 2025] applied a similar XAI-driven adversarial methodology in a black-box setting, demonstrating the potential to generate adversarial inputs without full transparency. Future work may extend our approach to cover such black-box scenarios, enhancing its applicability and robustness.

4.3. Methodology

In this subsection, we outline the structured pipeline (Figure 1) used to implement XAI-driven adversarial attacks on the WUSTL-EHMS-2020 dataset. Our approach unfolds in five key stages: data preprocessing, model development, performance evaluation, explainability analysis, and the generation and assessment of adversarial perturbations.

Dataset preprocessing began with manual cleaning to remove non-contributory features. The `SrcMac` attribute was excluded because the testbed used only one machine to simulate benign traffic and another for malicious traffic, making this feature redundant. The `Dir` and `Flgs` attributes were also removed, as they lacked meaningful information and were not described in the original dataset documentation. Additionally, `Packet_num`, which serves merely as a sequential counter for each sample, was

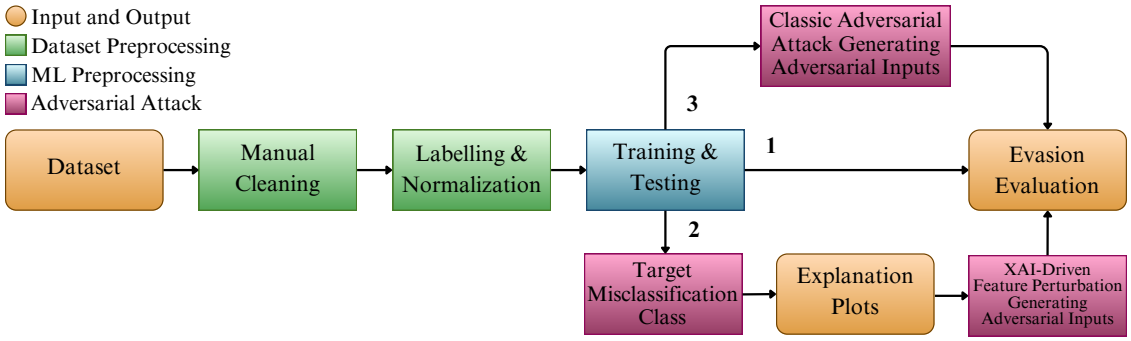


Figure 1. Methodology Flow Chart.

discarded. Three anomalous samples (with indexes `10633`, `7923`, and `8412`) were removed due to the presence of missing (`NaN`) values and categorical characters in a numerical feature. After these steps, the `Attack Category` — comprising the classes Benign, Data Alteration, and Spoofing — was retained as the target variable for classification.

Next, using scikit-learn's Python library[1], categorical variables were converted to numeric values via the `LabelEncoder` function and all features were standardized with `StandardScaler`. The processed data were then split 80%/20% into training and testing sets, and an XGBoost gradient-boosting classifier — known for its learning efficiency and built-in regularization — was trained [Chen and Guestrin 2016]. Figure 1 illustrates three distinct pathways that follow the completion of the Machine Learning Preprocessing phase. In the first pathway, the baseline dataset was utilized to compute precision, recall, and F1-score for each class using scikit-learn's `classification_report` function, along with the confusion matrix, to assess the model's performance prior to exposure to adversarial attacks.

In the second pathway, we interpret and pinpoint the drivers of False Negatives in our target attack class, applying SHAP's `TreeExplainer` [Lundberg and Lee 2017] to the trained XGBoost model. SHAP values were computed for each test instance; we then filtered for samples whose true label matched the target attack but were misclassified by the model. By aggregating the mean absolute SHAP value per feature, we ranked features by their influence on misclassification. The top three most influential features were selected to guide subsequent adversarial perturbations.

Inspired by the approach of [Okada et al. 2025], we visualized True Positive and False Negative samples using a 3D scatter plot based on the top three SHAP-ranked features. We also analyzed the F1-score for each of these features to better understand their individual impact on model performance. To ensure independence between features, we generated a correlation matrix. This combined analysis guided the selection of a single feature for targeted perturbation — one that was both highly influential and relatively uncorrelated with the others.

During the adversarial attack phase, we first compute the standard deviation of the selected feature over the normalized test set using the `std()` function from the `pandas` library. We then define a small perturbation factor (δ), which serves as a scalar to control

[1]Scikit-learn library. Available at: `https://scikit-learn.org/`

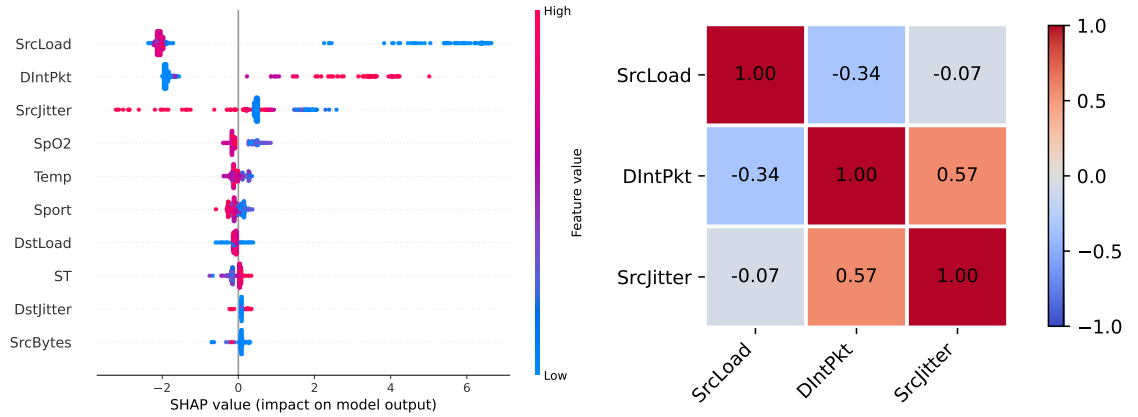
the attack's subtlety. The actual perturbation magnitude (Δ) is computed by multiplying the perturbation factor by the standard deviation of the selected feature, representing the amount by which the adversarial inputs will be altered. Next, we create a copy of the test dataset and apply the perturbation magnitude to the selected feature for all samples belonging to the target class (i.e., malicious instances), thereby generating the adversarial examples. To evaluate the effectiveness of the attack, we employ confusion matrices for all experimental scenarios, the F1-Score as implemented in `scikit-learn`.

In contrast to the XAI-driven attack, the third pathway employs the Adversarial Robustness Toolbox (ART)'s [Nicolae et al. 2018] `XGBoostClassifier` to implement a targeted HopSkipJump attack aimed at misclassifying samples as Benign. Using parameters `max_iter=10`, `max_eval=1000`, and `init_eval=10`, adversarial examples were generated based on the model's outputs. We then evaluated baseline versus adversarial accuracy and targeted success rates on Data Alteration and Spoofing scenarios.

5. Results

The methodology outlined in Section 4.3 is applied here to analyze the two Attack Category classes from the WUSTL-EHMS-2020 dataset: Data Alteration and Spoofing. Using XAI techniques, we systematically identify the most vulnerable features for targeted adversarial perturbations through manual analysis of model interpretation results.

5.1. XAI Analysis



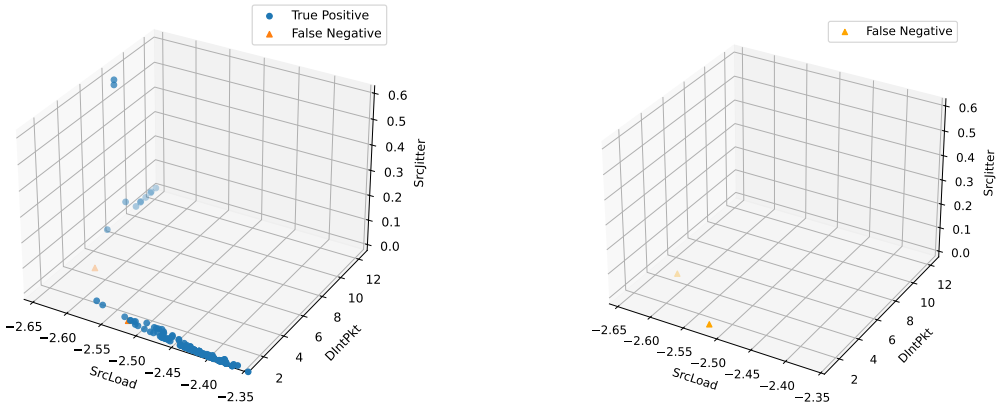
(a) SHAP Plot for Data Alteration.

(b) Correlation Matrix for Data Alteration.

Figure 2. (a) SHAP summary plot under Data Alteration attack, and (b) corresponding correlation matrix of the most impactful features.

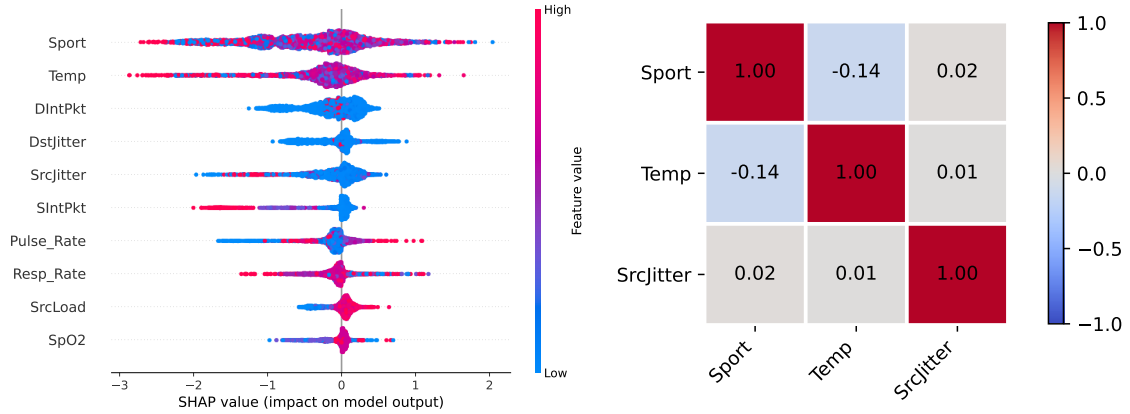
Figure 2 and Figure 3 summarize the explainability evaluation of our XGBoost model, highlighting the three most influential features — `SrcLoad`, `DIntPkt`, and `SrcJitter` — by mean absolute SHAP value. In Figure 2(a), low `SrcLoad` values (blue) and high `DIntPkt` values both drive up SHAP scores, signaling a higher probability of Data Alteration attacks, while `SrcJitter` exerts a consistent, moderate effect indicative of benign traffic. The correlation heatmap in Figure 2(b) confirms minimal interdependence among these features, with only a moderate positive correlation between `DIntPkt` and `SrcJitter`, suggesting that manipulating one feature is unlikely to unduly perturb the others. Finally, the 3D plots of True Positives (Figure 3(a)) and False

Negatives (Figure 3(b)) reveal distinct clusters and boundary-adjacent misclassifications, underscoring `SrcLoad` as the most discriminative — and thus the prime target for adversarial experiments.



(a) TP vs FN in Top-3 Features. (b) FN in Top-3 Features.

Figure 3. 3D scatter plots under Data Alteration attack with (a) TP vs FN samples and (b) only FN samples.

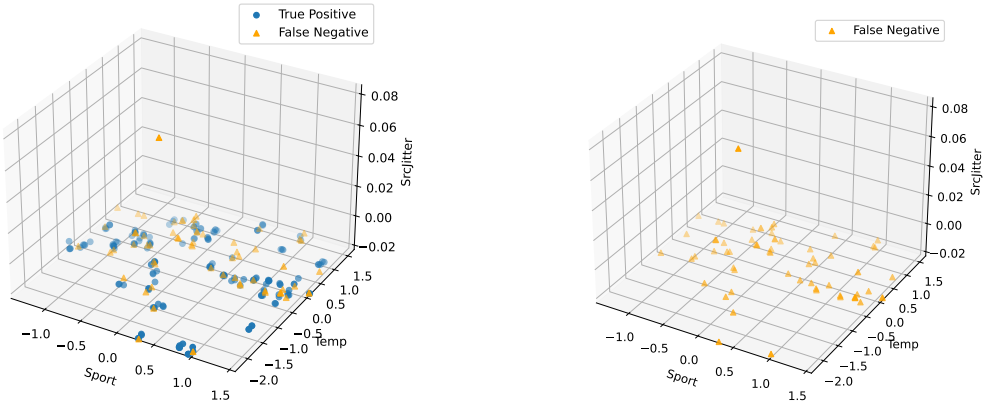


(a) SHAP Plot for Spoofing. (b) Correlation Matrix for Spoofing.

Figure 4. (a) SHAP summary under Spoofing attack, and (b) corresponding correlation matrix of the most impactful features.

The XAI-driven analysis of the Spoofing attack (Figures 4–5) highlights two primary drivers — Source Port (`Sport`) in network traffic and Temperature (`Temp`) in biomedical data — whose SHAP values span –3 to +2, indicating bidirectional influence depending on feature interactions. Lesser contributors such as `DIntPkt`, `DstJitter`, and `SrcJitter` exhibit weaker overall impacts, although low `SrcJitter` correlates with false negatives. A correlation heatmap of the top three false-negative drivers (`Sport`, `Temp`, `SrcJitter`; Figure 4(b)) confirms negligible pairwise dependencies, while the 3D scatter plot (Figure 5) shows true positives and false negatives clustering in regions of sparse `Sport` and `Temp` values with particularly low `SrcJitter`.

We therefore focus our Data Alteration attack on the highly discriminative `SrcLoad` feature and our Spoofing attack on `Sport`, whose pronounced SHAP impacts



(a) TP vs FN in Top-3 Features. (b) FN in Top-3 Features.

Figure 5. 3D scatter plots under Spoofing attack with (a) TP vs FN samples and (b) only FN samples.

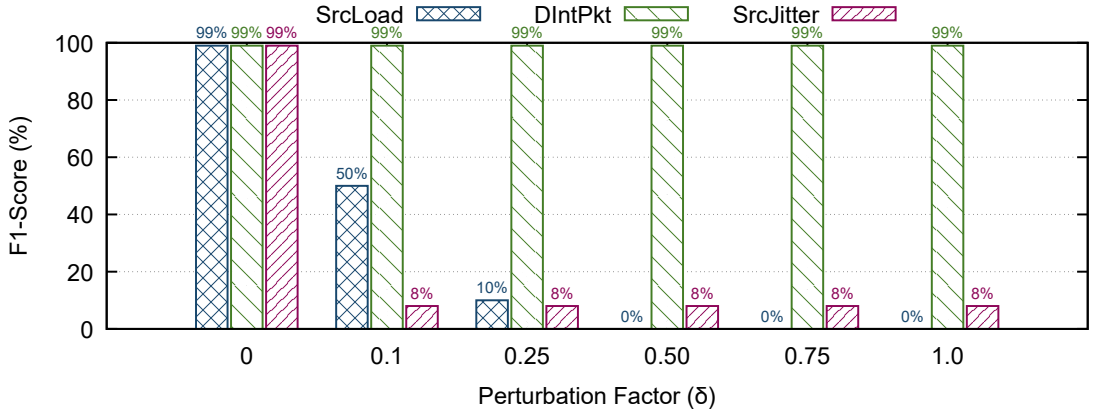
and distinct true-positive and false-negative separation make them ideal for probing the XGBoost model's vulnerabilities.

5.2. Evaluation

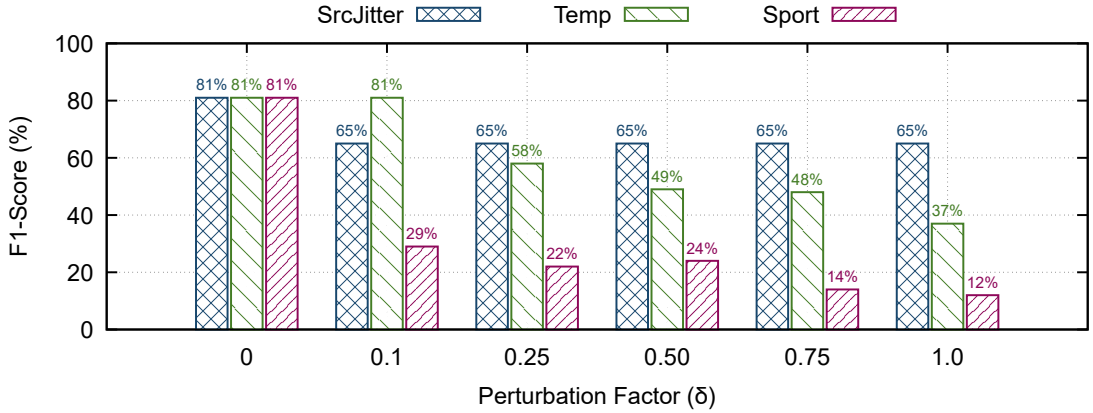
After identifying the key characteristics, we probed their vulnerabilities individually by plotting F1-score degradation curves for the most susceptible candidates on the original, unaltered dataset. We then repeated the same evaluation after applying each adversarial attack, including HopSkipJump, to quantify their impact using correlation matrices to quantify feature inter-dependencies and compare attack success rates.

Figure 6(a) illustrates the impact of the perturbation factor (δ) on the model's performance when applied individually to each of the top three features selected for the Data Alteration attack. As the perturbation factor increases up to 25%, the F1-score for `SrcLoad` and `SrcJitter` drops sharply—from 99% to 10% and 8%, respectively. Notably, `SrcJitter` performance stabilizes beyond this point, while `SrcLoad` continues to decline, reaching an F1-score of zero when the perturbation exceeds 50%. Figure 6(b) shows how targeted perturbations of the top three XAI-selected features affect Spoofing detection. Without any perturbation, all three features achieve an F1-score of 81%. As the perturbation factor grows, `SrcJitter` remains unchanged at 65% across tested magnitudes (10% through 100%), demonstrating robustness. In contrast, the *Temp* feature — representing human sensor data — shows a steady decline in detection performance as perturbation increases, indicating only moderate vulnerability. On the other hand, the *Sport* feature proves significantly more fragile, with detection accuracy dropping sharply even under small perturbations. This suggests that *Sport* is the most effective single-feature vector for evading spoofing detection, making it a key target for adversarial manipulation.

Then we apply our XAI-driven perturbation exclusively to `SrcLoad` on Data Alteration samples. Figure 7(a) illustrates how the F1-score and recall deteriorate as the adversarial perturbation magnitude increases, targeting the single most vulnerable feature for each attack class. The attack achieves 100% evasion — misclassifying all malicious instances as benign — once the perturbation factor reaches 33%. The post-



(a) Data Alteration.



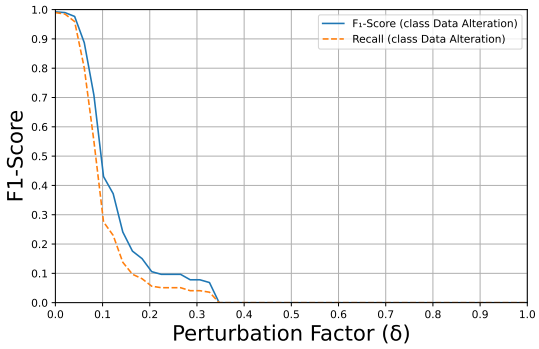
(b) Spoofing.

Figure 6. F1-Score degradation with increasing perturbation size for (a) Data Alteration and (b) Spoofing attacks.

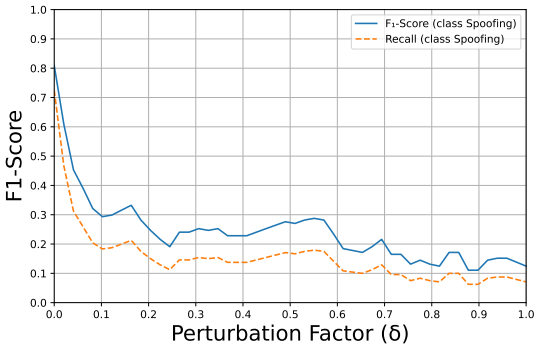
attack confusion matrix (Figure 9(a)) demonstrates that perturbing `SrcLoad` alone, with this level of modification, is sufficient to cause successful evasion.

In the case of the XAI-Driven manipulation on Spoofing attack (Sport feature) 7(b), the performance degradation follows a different trend. The F1-score drops rapidly from 81% to around 30% when the perturbation factor reaches approximately 10%, while recall similarly decreases from about 75% to 2%. However, beyond this point, both metrics show a degree of saturation and oscillate between 1% and 3% as the perturbation factor increases up to 100% leading to F1-score of 12%. This plateau effect suggests a partial robustness to spoofing attacks, where initial perturbations are effective, but further increases yield only marginal gains in evasion.

In addition to our targeted XAI-driven perturbations, we evaluated the resilience of the Healthcare and IDS merged system against HopSkipJump attack. Unlike single-feature manipulations guided by feature-importance explanations, HopSkipJump indiscriminately perturbs all input dimensions, although with heterogeneous magnitudes, across both the data-alteration and spoofing threat models. This complementary analysis not only benchmarks our method against a well-established adversarial strategy but also elucidates how targeted explanation-guided perturbations compare with global, gradient-



(a) F1-Score and Recall vs. Perturbation Factor on SrcLoad Feature.



(b) F1-Score and Recall vs. Perturbation Factor on Sport Feature.

Figure 7. Comparison of F1-Score under increasing adversarial perturbation for (a) Data Alteration (SrcLoad) and (b) Spoofing (Sport) attacks.

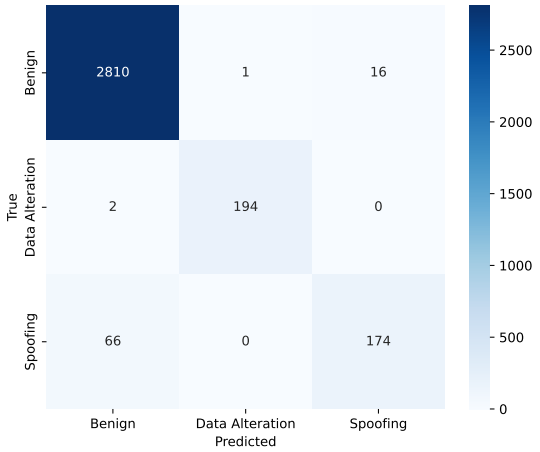
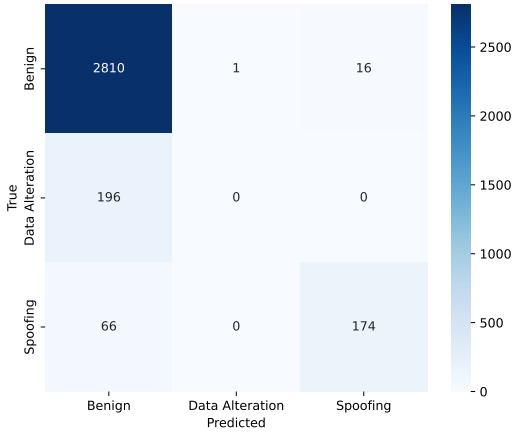


Figure 8. Confusion matrix of baseline.

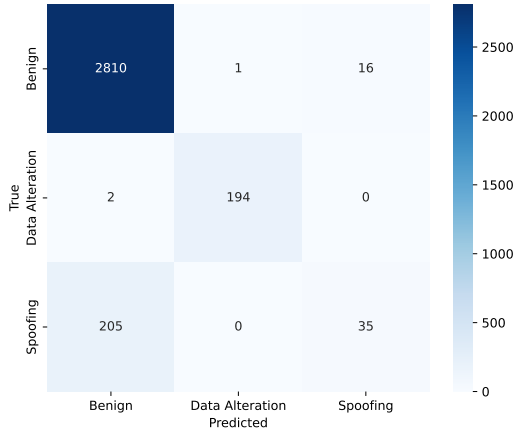
free optimization approaches in a Healthcare 5.0 context. By subjecting both attack paradigms to identical network and data preprocessing pipelines, we ensure a fair comparison that highlights the unique advantages and limitations of each technique.

We begin by presenting the confusion matrix for the baseline input data, where baseline refers to the original, unaltered test sample (Figure 8). Before the generation of adversarial examples, the model demonstrates high classification accuracy. However, after being subjected to an XAI-driven targeted attack on the Data Alteration class, the model misclassifies all test samples (Figure 9(a)). Similarly, under a Spoofing attack, the model achieves an evasion rate of 93% with a perturbation magnitude of approximately 90%. In contrast, our instantiated HopSkipJump attacks achieve 100% evasion when targeting both the Data Alteration class (Figure 9(c)) and the Spoofing class (Figure 9(d)).

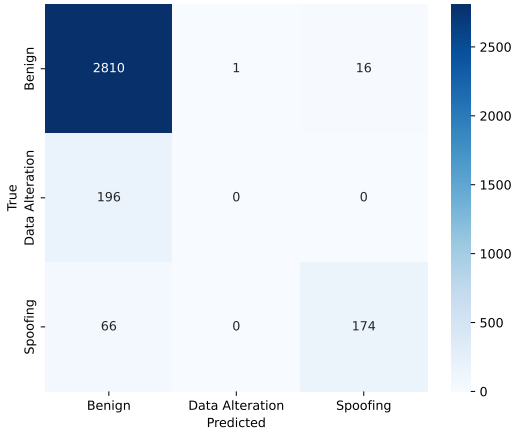
The XAI-guided attack achieved perfect evasion in the data alteration scenario and maintained a high success rate in the spoofing scenario, all while modifying only a single input feature and requiring minimal runtime per sample. In contrast, the HopSkipJump attack achieved 100% evasion across both scenarios; however, it operated by perturbing multiple input features simultaneously, as illustrated by the multidimensional perturbation



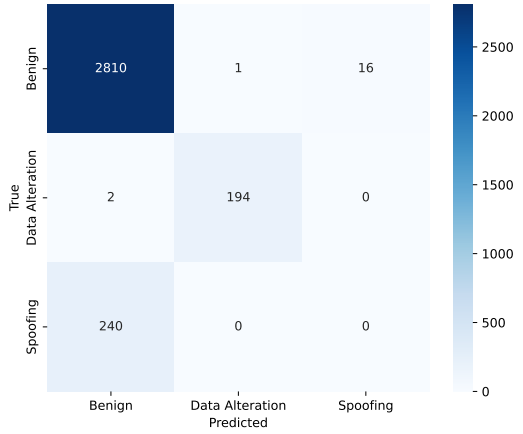
(a) Confusion matrix after XAI-driven perturbation for Data Alteration.



(b) Confusion matrix after XAI-driven perturbation for Spoofing.



(c) Confusion matrix after HopSkipJump Attack on Data Alteration.



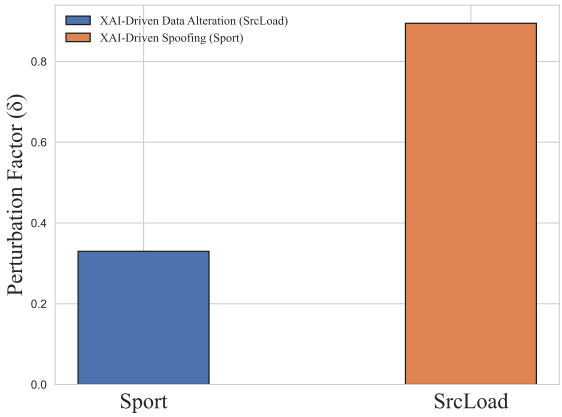
(d) Confusion matrix after HopSkipJump Attack on Spoofing.

Figure 9. Confusion matrices showing misclassification after XAI-driven perturbation for (a) data alteration and (b) spoofing scenarios.

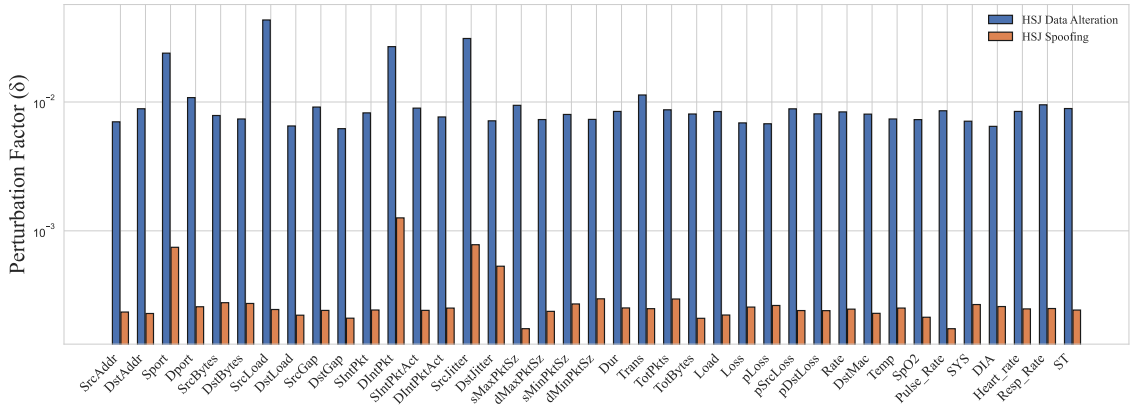
factors in Figure 10. This contrast reveals an important insight for adversarial planning in resource-constrained environments like healthcare: explanation-based attacks can provide near-optimal stealth with low latency, whereas high-dimensional attacks, although more powerful, may be impractical under strict time or compute constraints. Moreover, the HopSkipJump scenario illustrates that crafting multidimensional adversarial examples can be more complex and resource-intensive than single-feature perturbations, depending on the context. These results suggest that XAI-driven adversarial strategies are a viable and efficient alternative for evading advanced IDS, and emphasize the need for defenses that can withstand both broad-spectrum and targeted, explanation-informed attacks.

6. Conclusion and Future Works

In this study, we reveal that the shift toward Healthcare 5.0 — marked by pervasive machine-driven decision systems — significantly amplifies the risk of cyberattacks, as adversaries can leverage XAI tools to probe and exploit critical model features. Although interpretability methods aim to improve transparency, they can be manipulated through



(a) XAI-Driven perturbation factor applied unidimensionally.



(b) HopSkipJump attack perturbation factor applied multidimensionally.

Figure 10. Comparison of average perturbation factor (δ) under different attack strategies. Subfigure (a) shows the perturbation factor when a single, most influential feature is altered (unidimensional XAI-driven attack), while (b) presents the perturbation factors resulting from the multidimensional HopSkipJump attack, which perturbs multiple features simultaneously to achieve successful misclassification.

single-feature perturbations to achieve complete evasion, confirming similar observations in recent work [Okada et al. 2025]. Our experiments uncover a clear trade-off between stealth and complexity of attack: an XAI-guided strategy alters just one pivotal attribute to evade detection entirely and corrupt downstream decisions, whereas the HopSkipJump approach must tamper with every input dimension to match that level of evasion. These outcomes not only spotlight a new class of threats to patient safety and data integrity but also underscore the necessity of extending adversarial assessments to diverse healthcare datasets and realistic testbed scenarios.

These findings present the imperative to extend the study of XAI-driven adversarial attacks across diverse datasets and realistic testbed environments, and to devise resilient countermeasures — such as training regimes that reinforce critical features, dynamic filtering of explanation outputs, and continuous anomaly monitoring — to protect full packet-level integrity under real-world bandwidth, latency, and protocol constraints.

Acknowledgments

This research effort is sponsored in part by resources from "Edital PRPGP/UFSM N.050/2024 - Programa de Fortalecimento e Redução de Assimetrias da Pós-Graduação da UFSM".

References

Agrawal, N., Pendharkar, I., Shroff, J., Raghuvanshi, J., Neogi, A., Patil, S., Walambe, R., and Kotecha, K. (2024). A-xai: adversarial machine learning for trustable explainability. *AI and Ethics*, 4(4):1143–1174.

Alhajjar, E., Maxwell, P., and Bastian, N. (2021). Adversarial machine learning in network intrusion detection systems. *Expert Systems with Applications*, 186:115782.

Asiri, A., Wu, F., Tian, Z., and Yu, S. (2024). From the perspective of ai safety: Analyzing the impact of xai performance on adversarial attack. In *GLOBECOM 2024-2024 IEEE Global Communications Conference*, pages 4982–4987. IEEE.

Baniecki, H. and Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, page 102303.

Bayer, M., Neiczer, M., Samsinger, M., Buchhold, B., and Reuter, C. (2024). Xai-attack: Utilizing explainable ai to find incorrectly learned patterns for black-box adversarial example creation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17725–17738.

Brohi, S. and Mastoi, Q.-u.-a. (2025). From accuracy to vulnerability: Quantifying the impact of adversarial perturbations on healthcare ai models. *Big Data and Cognitive Computing*, 9(5):114.

Check Point Software Technologies (2025). Check Point Software's 2025 Security Report Finds Alarming 44% Increase in Cyber-Attacks Amid Maturing Cyber Threat Ecosystem. Available at: https://www.checkpoint.com/press-releases/check-point-softwares-2025-security-report-finds-alarming-44-increase-in-cyber-attacks-amid-maturing-cyber-threat-ecosystem/. Accessed on February 20, 2025.

Chen, J., Jordan, M. I., and Wainwright, M. J. (2020). Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Gadekallu, T. R., Maddikunta, P. K. R., Boopathy, P., Deepa, N., Chengoden, R., Victor, N., Wang, W., Wang, W., Zhu, Y., and Dev, K. (2024). Xai for industry 5.0-concepts, opportunities, challenges and future directions. *IEEE Open Journal of the Communications Society*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hady, A. A., Ghubaish, A., Salman, T., Unal, D., and Jain, R. (2020). Intrusion Detection System for Healthcare Systems Using Medical and Network Data: A Comparison Study. *IEEE Access*, 8:106576–106584.

Imam, N. H. (2024). Adversarial examples on xai-enabled dt for smart healthcare systems. *Sensors*, 24(21):6891.

Khan, N., Ahmad, K., Tamimi, A. A., Alani, M. M., Bermak, A., and Khalil, I. (2024). Explainable AI-based Intrusion Detection System for Industry 5.0: An Overview of the Literature, associated Challenges, the existing Solutions, and Potential Research Directions. *arXiv preprint arXiv:2408.03335*.

Kuppa, A. and Le-Khac, N.-A. (2021). Adversarial xai methods in cybersecurity. *IEEE transactions on information forensics and security*, 16:4924–4938.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I., and Edwards, B. (2018). Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069.

Okada, S., Jmila, H., Akashi, K., Mitsunaga, T., Sekiya, Y., Takase, H., Blanc, G., and Nakamura, H. (2025). Xai-driven black-box adversarial attacks on network intrusion detectors. *International Journal of Information Security*, 24(3):1–15.

Rosenberg, I., Meir, S., Berrebi, J., Gordon, I., Sicard, G., and David, E. O. (2020). Generating end-to-end adversarial examples for malware classifiers using explainability. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–10. IEEE.

Statista (2025). Digital Health - Worldwide. Available in: `https://www.statista.com/outlook/hmo/digital-health/worldwide`. Accessed on February 20, 2025.

Tandel, V., Kumari, A., Tanwar, S., Singh, A., Sharma, R., and Yamsani, N. (2024). Intelligent wearable-assisted digital healthcare industry 5.0. *Artificial Intelligence in Medicine*, 157:103000.

Vázquez-Hernández, M., Morales-Rosales, L. A., Algredo-Badillo, I., Fernández-Gregorio, S. I., Rodríguez-Rangel, H., and Córdoba-Tlaxcalteco, M.-L. (2024). A survey of adversarial attacks: An open issue for deep learning sentiment analysis models. *Applied Sciences*, 14(11):4614.

Yan, A., Liu, X., Li, W., Ye, H., and Li, L. (2024). Explanation-guided adversarial example attacks. *Big Data Research*, 36:100451.