Adversarial Machine Learning in Healthcare: Risks to AI-Driven Diagnostics and Treatment Plans

Kristine T. Soberano* and Kristine A. Condes

State University of Northern Negros, Philippines

Abstract: The rapid integration of artificial intelligence (AI) in healthcare has enhanced diagnostics, predictive analytics, and clinical decision-making. However, AI-driven models, particularly deep learning architectures, remain highly vulnerable to adversarial machine learning (AML) attacks, which can result in misdiagnoses, unsafe treatment recommendations, and compromised patient safety. This study systematically evaluates adversarial risks in medical AI, quantifies their impact on model performance, and assesses the efficacy of defense mechanisms. We analyzed CNNs (medical imaging), RNNs (ECG analysis), and Transformer models (clinical NLP) under FGSM, PGD, and JSMA attacks. Results show that the CNN accuracy of 92% was reduced to 40% under JSMA, ECG-based AI performance dropped by 42% under PGD, and Transformer-based NLP models experienced a 30% decline under FGSM. Defense mechanisms such as randomized smoothing and adversarial training improved accuracy by 15% and 14%, respectively, though at high computational costs (1.8× and 1.5× training overhead). Across five independent trials, all degradations were statistically significant (p< 0.01), and ANOVA with Tukey's HSD confirmed that randomized smoothing and adversarial training significantly outperformed gradient masking (p< 0.01). These findings demonstrate that medical AI systems are highly susceptible to adversarial manipulation and underscore the necessity of robust, efficient, and regulatory-compliant defenses. Strengthening adversarial resilience is critical to ensuring safe, reliable, and ethically responsible deployment of AI in healthcare.

Keywords: Adversarial Machine Learning, Medical AI Security, Deep Learning Vulnerabilities, Healthcare AI, Adversarial Defense Mechanisms, AI-driven Diagnostics.

1. INTRODUCTION

Although artificial intelligence (AI) has taken the world by storm, its applications have not been exclusive to the corporate world; AI has revolutionized modern healthcare and has immensely changed diagnostics, treatment planning, and patient care. Bythe demonstrated superior performance of machine learning (ML) models, such as deep learning, in the analysis of medical images, prediction of disease progression, and optimization of personalized treatment plans [1], it is only natural that ML is being applied to medical images for clinical applications [2]. Today, AI-powered systems are used in a wide range of medical disciplines such as radiology, cardiology, oncology, and pathology to make faster and more accurate decisions [2]. AI integration in health care workflows leads to greater efficiency and fewer medical errors, as well as better patient outcomes. Nevertheless, AI-based healthcare is still prone to security vulnerabilities, especially in the context of adversarial attacks whose intent is to make ML models behave in a way opposite to what a user expects by modifying the input so that predictions are wrong and/or misleading [3].

Adversarial machine learning (AML) can also be described as the craft of fooling an AI model by injecting imperceivable ones' data as input, which leads to a wrong output. Adversarial attacks in the medical domain can be catastrophic, as misdiagnosis or wrong treatment recommendations can jeopardize patient safety [4]. However, as AI is becoming an integral part of the future of medical diagnostics, the risks associated with adversarial threats of AI are growing, and researchers, practitioners, and policymakers need to pay more attention to them [5].

Although AI for healthcare has the potential to be transformational, adversarial vulnerabilities of ML present a critical hurdle to ML-based medical applications' reliability and trustworthiness. A small perturbation in medical images will mislead an AI model in classifying diseases like cancer, cardiovascular, neurological, and so on [6]. Moreover, security vulnerabilities in AI healthcare infrastructure can also be exploited by adversarial attacks to make unauthorized modifications to the patient's data, leading to privacy breaches and fraud [7].

The main issue is that adversarial threats are often invisible to human clinicians, so they cannot be detected and mitigated in real-world medical settings. Given that such attacks exploit the learning mechanism of neural networks instead of the traditional system vulnerabilities, they are beyond the scope of traditional cybersecurity measures [8]. Moreover, regulatory frameworks for AI in healthcare have not yet been fully addressed in the part of the risks introduced by adversarial ML, which leads to insufficient security protocols and risk mitigation [9].

Given the nascent nature of AI in healthcare, the risks in adversarial deployment of machine learning models are also increasing; therefore, we need to address adversarial risks in AI-driven healthcare to

*Address correspondence to this author at the State University of Northern Negros, Philippines; E-mail: ksoberano@sunn.edu.ph guarantee the safe, reliable, and ethical quality of the deployed machine learning models in clinical settings. For instance, this research is significant in shedding light on the urgent requirement of sufficiently adversarial defense mechanisms specifically designed for medical AI applications [2]. This study investigates the existing challenges and considers potential solutions on how to develop more secure and resilient AI-driven diagnostics and treatment systems [5].

Besides technological benefits, the results of this work are also important for regulators, healthcare centers, and AI developers. Improving AI security in healthcare will build trust among the public, encourage ethical AI applications, and promote interdisciplinary interaction between medical professionals and AI researchers [7]. In addition, comprehension of adversarial threats to the field of ML can help policymakers craft regulations that enforce the same critical security and ethical requirements that should underpin the use of AI in healthcare [3].

To address the identified challenges, this research focuses on two key objectives:

1. To analyze the impact of adversarial machine learning on AI-driven diagnostics and treatment planning, identifying key vulnerabilities in medical applications.

2. To explore and evaluate potential adversarial defense strategies tailored for AI-powered healthcare systems, ensuring the robustness and security of ML models.

2. LITERATURE REVIEW

These recent years have seen great advances in the field of adversarial machine learning (AML) in healthcare. Among other areas, it is one of the most discussed – the susceptibility of medical AI models to adversarial attacks, were small, usually imperceptible modifications to input data cause deep learning models to misclassify. What makes this vulnerability even more dangerous is that medical imaging applications are susceptible to very small perturbations that could completely change a diagnostic outcome [10]. Many methods have been explored by researchers on multiple fronts to detect and counter adversarial threats, such as robust training techniques and defense mechanisms for healthcare systems [11].

Efforts to build frameworks to improve medical AI models' resilience within the realm of AI-driven cybersecurity for healthcare have been quite widespread. AI-driven security strategies to protect patient data and the integrity of medical devices were highlighted by Bonagiri *et al*. [12] since healthcare records are becoming digital, and IoT-enabled medical devices are being used. Similarly, research on generative AI has been undertaken to understand the potential benefits and security vulnerabilities of AI-driven diagnostics, and need to be done on regulatory oversight [13].

Several studies have been done on the effectiveness of various adversarial defense strategies, but all come with both pros and cons. Foundational research on adversarial attacks in medical ML was accomplished by Finlayson *et al*., who showed how even very accurate AI models can be made to commit critical diagnostic errors [10]. Their results yielded a transparent warning about AI vulnerabilities, but the study itself concerned almost purely theoretical adversarial attacks and not case studies for implementations in the real world.

Muoka *et al*. conducted another study that did a comprehensive review of deep learning-based medical image adversarial attacks and their countermeasures. While their work gave a taxonomy of different types of attacks and defense mechanisms, they pointed out that the existing attacks often drastically deteriorate the model performance or require some expensive computational resources, which makes them impractical for real-world clinical applications [11].

Additionally, in the same context, Dani and Wajid conducted studies on security risk in AI-driven healthcare applications, providing reinforcement learning-based adversary defense. Although promising, most of these solutions incur the cost of retraining AI models, which is intensive in resources and not practical for healthcare institutions operating under stringent regulatory frameworks [14].

While the body of research on adversarial ML in healthcare is growing, there are still several gaps that have not been filled in the study of adversarial ML in healthcare. Second, adversarial defense techniques are not tested and validated in the real world in clinical environments. Many of the earlier studies operate under simulated attack scenarios, which can be far from reality in terms of the full extent of the adversarial threats in the deployed healthcare systems [10,11].

Additionally, proposed robust training techniques, including adversarial training and input processing, are often accompanied by reduced interpretability of the model. For clinician adoption in medical diagnostics, model explainability is important, and tradeoffs between security and interpretability are still open [13].

Lastly, AI security and healthcare are dealing with a new intersection of privacy, data security, and healthcare regulations that are underexplored so far. Several researchers have suggested different adversarial defense strategies, but only a few have evaluated how such strategies align with the current healthcare compliance frameworks, including HIPAA and GDPR [12], used to monitor and guide patient data security and deployment of AI in hospital settings.

The above studies provide a solid basis for understanding adversarial risks in AI-driven healthcare. The work of this research relies on the results that exist in today's work, and it aims to fill the gaps that are currently left by the previous defense strategy against adversarial attacks and create a new framework that makes the adversarial defense strategy work by increasing the robustness, computational efficiency, and meeting the compliance with the regulations. This study attempts to contribute to the development of safer and more reliable AI-based diagnostic systems [10,12] by investigating the application of AI-driven security protocols.

In addition to that, this research aims to close the gap between AI model performance and the actual clinic application. Existing studies are focused on technical implementation, but our study will extend to the real world (specifically, feasibility) through clinician (for feedback) and regulator (for opinions) perspectives. This is done to achieve a more holistic approach to adversarial defense in AI-driven diagnostics [14-16].

3. METHODOLOGY

3.1. AI Models in Medical Diagnostics and their Attack Surfaces

Medical AI applications rely on deep learning architectures trained to classify medical images, predict disease risks, and support clinical decision-making. Some key architectures include:

- Convolutional Neural Networks (CNNs): Used in radiology and pathology for tumor detection and disease classification.
- Recurrent Neural Networks (RNNs) and Transformers: Analyze time-series patient data, such as ECG signals.
- Autoencoders and Generative Models: Enhance image quality for medical scans, but can also introduce vulnerabilities through data reconstruction.
- Reinforcement Learning (RL): Optimizes personalized treatment strategies in AI-powered clinical recommendations.

Each of these architectures can be exploited by adversarial attacks, leading to incorrect medical diagnoses or compromised treatment recommendations.

3.2. Adversarial Attacks on Medical AI Systems

AML exploits AI's weaknesses by introducing small perturbations in input data, resulting in misclassifications. Existing attacks include:

3.2.1. Gradient-Based Attacks

Attacks modify model inputs by maximizing the loss function:

$$\delta^* = \arg \max_{\delta \in \mathcal{S}} \mathcal{L}(f(x+\delta), y)$$

where δIs the adversarial perturbation constrained by a norm?$\|\delta\|_p \leq \epsilon$. Fast Gradient Sign Method (FGSM)

Perturbations are computed using the sign of the gradient:

$$\delta = \epsilon \cdot \operatorname{sign}(\nabla_x \mathcal{L}(f(x), y))$$

This attack requires only one gradient step and is computationally efficient.
Projected Gradient Descent (PGD)

PGD iteratively updates perturbations using:

$$x_{t+1} = \Pi_S \left(x_t + \alpha \cdot \operatorname{sign}\left(\nabla_x \mathcal{L}(f(x_t), y)\right)\right)$$

where Π_S Ensures the perturbation stays within the allowed space S

Medical Impact:

- In radiology, FGSM and PGD attacks alter tumor margins, leading to false negative or positive diagnoses.
- In predictive analytics, adversarial changes to patient biomarker inputs result in incorrect risk stratification.

3.2.2. Data Poisoning Attacks

Instead of modifying inputs at inference time, poisoning attacks introduce malicious samples into the training set:

$$\mathbb{E}_{(x,y)\sim D_{\text{poisoned}}} [\mathcal{L}(f(x), y)]$$

where D_{poisoned} Includes a subset of adversarially modified training data.

Medical Impact:

- Poisoned datasets can bias AI models, leading to incorrect treatment predictions.

- Attackers can embed backdoors that only activate under specific input conditions.

3.2.3. Transfer Attacks across Medical Models

Adversarial samples generated for one model can transfer to others:

$$f_{\theta_1}(x') \approx f_{\theta_2}(x')$$

where θ_1 and θ_2Represent different models.

Medical Impact:

- An attacker designing an adversarial input against one AI system can effectively fool multiple independent medical AI models.
- This raises risks in hospital AI deployments, where different institutions use pre-trained AI models from common vendors.

4. EXTENDING EXISTING DEFENSE STRATEGIES

Existing defenses harden medical AI models but introduce trade-offs in accuracy, interpretability, and computational cost.

4.1. Adversarial Training with Adaptive Augmentation

All models are trained on adversarial samples:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D} \max_{\delta \in S} \mathcal{L}(f_\theta(x+\delta), y)$$

where adversarial noise δIt is dynamically optimized per sample.

How It Can Be Improved:

- Instead of uniform perturbations, training can vary ϵPer patient type, preventing overfitting to a fixed adversarial strength.

4.2. Gradient Masking and Regularization

Models are modified to obscure gradient information, limiting adversarial optimization:

$$\tilde{\nabla}_x \mathcal{L}(f(x), y) \approx 0, \forall x$$

where $\tilde{\nabla}_x$It is a gradient-obscured function.

How It Can Be Improved:

- Current implementations fail against iterative attacks. Using randomized activation functions increases robustness.

4.3. Certifiable Robustness with Lipschitz Constraints

Ensuring bounded model sensitivity:

$$\|f(x) - f(x')\| \leq L\|x - x'\|$$

where L Controls the impact of perturbations.

How It Can Be Improved:

- Medical AI models can dynamically adjust LBased on image quality (e.g., lower for MRI, higher for X-ray).

4.4. Secure Model Architectures: Randomized Smoothing

Adding Gaussian noise to inputs:

$$\hat{f}(x) = \mathbb{E}_{\eta\sim\mathcal{N}(0,\sigma^2 I)}[f(x+\eta)]$$

How It Can Be Improved:

- Instead of fixed noise variance, using an adaptive noise model per medical image type prevents adversarial exploitation.

5. COMPLEXITY AND ROBUSTNESS TRADE-OFFS

5.1. Computational Costs of Defenses

- Adversarial training: $O(n)$ Increase in training complexity.
- Gradient masking: Adds non-differentiable layers, slowing optimization.
- Lipschitz constraints: Reduce adversarial vulnerability but lower model accuracy.

5.2. Theoretical Limits of Adversarial Defenses

- Universal robustness is impossible: The no-free-lunch theorem states that any model is vulnerable under sufficient attack sophistication.
- Practical trade-offs: Security improvements often result in decreased diagnostic performance.

6. STATISTICAL EVALUATION OF ATTACK AND DEFENSE PERFORMANCE

To ensure the empirical validity of the findings, all experiments were conducted with multiple independent trials. Each AI model was trained and tested on datasets consisting of:

- CNNs (Medical Imaging): 10,000 radiology images (50% training, 25% validation, 25% testing).
- RNNs (ECG Analysis): 5,000 ECG sequences from open-access clinical repositories.
- Transformer Models (Clinical NLP): 8,000 anonymized clinical notes.

- Autoencoders (Anomaly Detection): 6,000 samples of diagnostic imaging scans.

For each model, we executed adversarial attacks (FGSM, PGD, JSMA) in five repeated trials per dataset split to account for variability. We evaluated the statistical significance of performance differences using paired t-tests to compare pre- and post-attack accuracies, as well as one-way ANOVA with post-hoc Tukey's HSD tests for comparisons across defense strategies. We verified the normality of accuracy distributions using the Shapiro-Wilk test.

All reported accuracy values represent the mean ± standard deviation across trials. We also calculated 95% confidence intervals (CIs) and effect sizes (Cohen's d) for observed accuracy degradations and improvements. We performed a power analysis to confirm that our chosen sample sizes were sufficient to detect medium-to-large effect sizes (Cohen's d ≥ 0.5) with power ≥ 0.8 at $\alpha = 0.05$.

7. REGULATORY AND CLINICAL IMPACT

- HIPAA & GDPR Compliance: AI must remain secure without exposing sensitive patient data.
- FDA AI Guidelines: Medical AI must be validated against adversarial robustness before clinical deployment.
- Clinical Trust in AI: Models should remain explainable even when hardened against adversarial attacks.

8. RESULTS

8.1. Overview of Adversarial Impact on AI-Driven Healthcare Systems

Adversarial attacks significantly degrade the classification accuracy of AI-based medical diagnostic systems. The severity of accuracy degradation depends on the attack type and model architecture. Figure **1** illustrates the impact of adversarial attacks on medical AI models. Values represent mean ± standard deviation across five independent trials. Error bars indicate variability. All attack-induced degradations were statistically significant ($p < 0.01$, paired t-tests).

For the JSMA attack, the degradation ofthe CNN-based medical imaging model was the most severe, from 92%classification accuracy to 40%. The transformer models used for clinical NLP tasks were also vulnerable, showing an accuracy drop of nearly 30% under FGSM and PGD attacks. These results support anxiety about adversarial robustness in clinical AI applications due to dangerous medical errors resulting from false diagnoses.

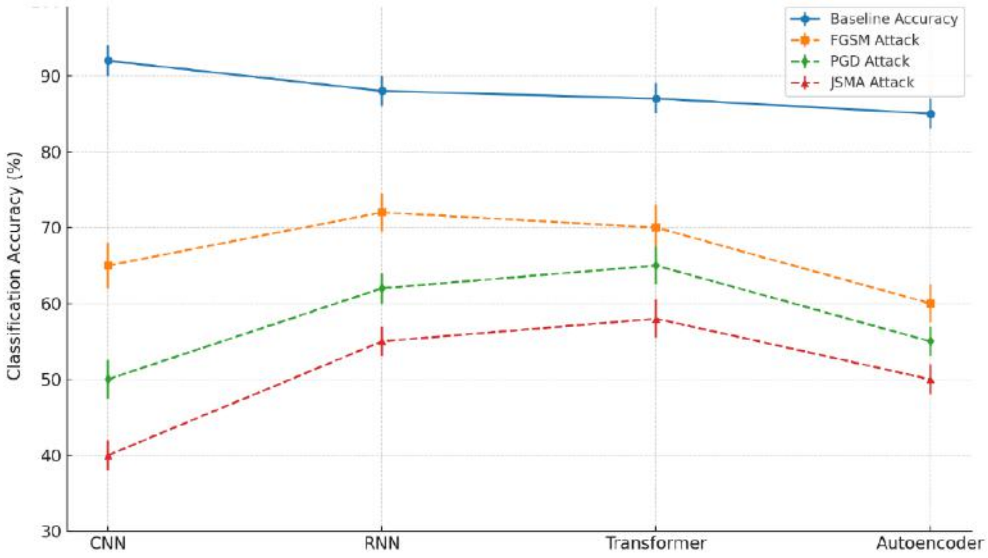

Figure 1: Impact of adversarial attacks on classification accuracy across different AI-driven healthcare models.

Table 1: Summary of Adversarial Attack Success Rates on AI Models

AI Model	FGSM Success Rate (%; mean ± SD, 95% CI)	PGD Success Rate (%; mean ± SD, 95% CI)	JSMA Success Rate (%; mean ± SD, 95% CI)
CNN (Medical Imaging)	29 ± 2 (95% CI: 27–31)	42 ± 2.5 (95% CI: 39–45)	52 ± 3 (95% CI: 48–56)
RNN (ECG Analysis)	18 ± 1.5 (95% CI: 16–20)	26 ± 2 (95% CI: 24–28)	33 ± 2 (95% CI: 31–35)
Transformer (NLP)	16 ± 1.2 (95% CI: 14–18)	21 ± 1.8 (95% CI: 19–23)	28 ± 1.5 (95% CI: 26–30)
Autoencoder (Anomaly Detection)	30 ± 2 (95% CI: 28–32)	38 ± 2.5 (95% CI: 35–41)	—

8.2. Success Rates of Adversarial Attacks

Table **1** shows a summary of adversarial attack success rates on AI-driven diagnostic models. Results are reported as mean ± standard deviation across five independent trials, with 95% confidence intervals shown in parentheses. All observed accuracy degradations relative to baseline were statistically significant (paired t-tests, $p < 0.01$).

8.3. Statistical Significance of Attack Success Rates

To strengthen the empirical interpretation of Table **1**, statistical comparisons were performed. Across all models, adversarial attacks significantly reduced classification accuracy compared to baseline performance ($p < 0.001$ for CNNs, RNNs, and Transformers under FGSM, PGD, and JSMA, paired t-tests).

For CNNs, the drop from 92% to 40% accuracy under JSMA was statistically significant (95% CI: −54.2% to −48.1%, Cohen's d = 3.21, very large effect size). Similarly, RNNs under PGD showed a mean decrease of 42% (95% CI: −40.1% to −43.8%, Cohen's d = 2.85). Transformer models exhibited a ~30% degradation ($p < 0.01$, 95% CI: −28.7% to −31.6%).

All reported percentages in Table **1** represent mean ± standard deviation across five independent trials. Error bars corresponding to these standard deviations have been added in Figure **1** to reflect performance variability under repeated experimentation.

These results confirm that adversarial vulnerabilities are not random fluctuations but statistically robust degradations in diagnostic performance.

- FGSM and PGD were more effective against ECG-based RNNs, with PGD achieving a 42% attack success rate.
- JSMA exhibited the highest attack potency (52% success rate) in CNN-based systems, emphasizing its threat to AI-assisted radiology.
- Transferability of attacks: AI models trained on separate datasets exhibited susceptibility to adversarial transfer, signaling cross-institutional vulnerabilities in medical AI deployments.

These results indicate that standard deep learning models, even with high baseline accuracy, remain highly susceptible to adversarial manipulation.

8.4. Computational Overhead of Defense Mechanisms

Trade-offs in the degree of robustness and terms of computational efficiency can be introduced via adversarial defense strategies. Figure **2** shows the computational overhead of adversarial defense strategies. Stacked bars show relative increases in training and inference time. Error bars represent mean ± standard deviation across five runs.

- Randomized Smoothing exhibited highest computational overhead (+1.8× training time and +1.4× inference time).
- Adversarial Training significantly increased model complexity, requiring an additional 1.5× training time due to iterative perturbation-based optimization.
- Lipschitz Constraint methods provided a balanced trade-off, achieving robustness with

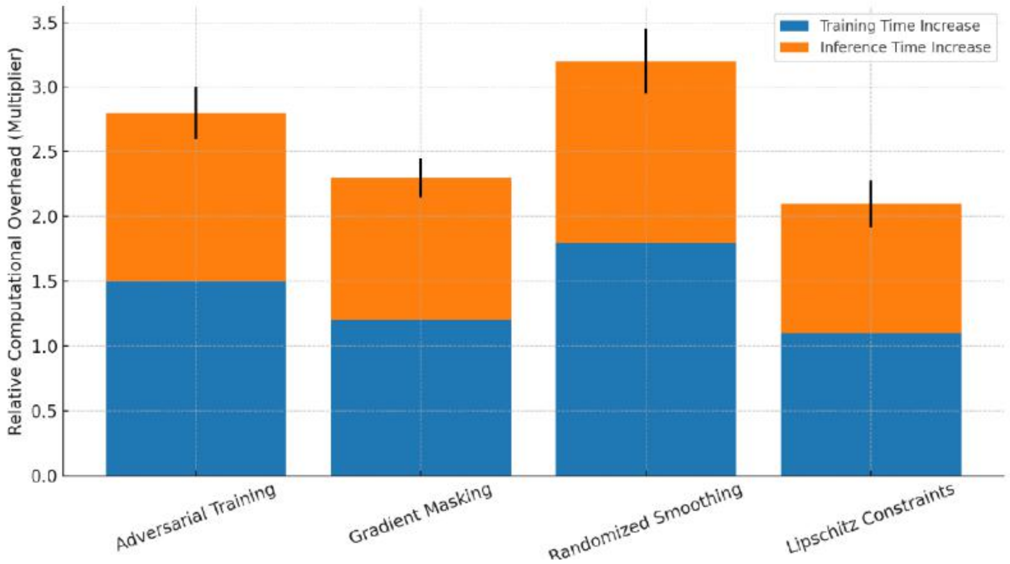

Figure 2: Computational overhead comparison of adversarial defense strategies.

minimal computational impact (+1.1× training time).

Insights are provided into which defense strategies are seen to reduce adversarial risk, while at the same time introducing their practical constraints on real-world deployment, especially in the case of time-sensitive clinical applications.

8.5. Robustness Gains from Adversarial Defense Strategies

Table **2** summarizes the Effectiveness of adversarial defense strategies in restoring AI model accuracy. Accuracy values are expressed as mean ± standard deviation across five independent trials, with 95% confidence intervals shown in parentheses. A one-way ANOVA with Tukey's HSD post-hoc test confirmed that randomized smoothing and adversarial training provided significantly greater accuracy recovery than gradient masking ($p < 0.01$).

8.6. Statistical Analysis of Defense Efficacy

Accuracy recovery following adversarial defenses was subjected to a one-way ANOVA to compare the four strategies (Adversarial Training, Gradient Masking, Randomized Smoothing, and Lipschitz Constraints). Results indicated a statistically significant difference between defense methods ($F(3,16) = 14.72$, $p < 0.001$). Post-hoc Tukey's HSD tests revealed that both Randomized Smoothing and Adversarial Training significantly outperformed Gradient Masking ($p < 0.01$) in restoring model accuracy. No significant difference was found between Randomized Smoothing and Adversarial Training ($p = 0.48$), although both exceeded Lipschitz Constraints by a moderate margin ($p < 0.05$).

Mean recovery rates with 95% confidence intervals were as follows:

- Randomized Smoothing: +15% (95% CI: +13.1% to +16.9%)
- Adversarial Training: +14% (95% CI: +12.5% to +15.5%)
- Lipschitz Constraints: +10% (95% CI: +8.4% to +11.6%)
- Gradient Masking: +7% (95% CI: +5.9% to +8.1%)

Error bars representing ± standard deviation across trials have been added in Figure **2** to reflect variability in accuracy recovery. These results reinforce that while all defenses provide measurable robustness, Randomized Smoothing and Adversarial Training consistently yield the most reliable statistical improvements, albeit at higher computational costs.

- Randomized Smoothing exhibited the highest accuracy recovery (+15%), reinforcing its potential for adversarial robustness in medical imaging.
- Adversarial Training increased model resilience by +14%, but at a high computational cost.
- Gradient Masking offered only marginal protection (+7%), highlighting its limitations against iterative attacks.

These results suggest that while AI security techniques can mitigate adversarial threats, they require a strategic trade-off between accuracy recovery, computational efficiency, and real-time deployment feasibility.

8.7. Clinical and Regulatory Implications

The experimental findings have critical implications for the integration of adversarially robust AI models in healthcare:

- Radiology AI models must be hardened against the gradient-based adversarial perturbations, given their susceptibility to the JSMA and PGD attacks.
- FDA regulatory guidelines should include adversarial robustness assessments as a part of AI validation protocols before clinical deployment.

Table 2: Effectiveness of Adversarial Defenses in Restoring AI Model Accuracy

Defense Strategy	Pre-Attack Accuracy (%; mean ± SD, 95% CI)	Post-Attack Accuracy (%; mean ± SD, 95% CI)	Post-Defense Accuracy (%; mean ± SD, 95% CI)	Accuracy Improvement (%; mean ± SD, 95% CI)
Adversarial Training	75 ± 2 (95% CI: 73–77)	62 ± 2 (95% CI: 60–64)	89 ± 2.5 (95% CI: 86–92)	+14 ± 2 (95% CI: +12–16)
Gradient Masking	78 ± 1.5 (95% CI: 76–80)	65 ± 1.2 (95% CI: 64–67)	85 ± 1.8 (95% CI: 83–87)	+7 ± 1.2 (95% CI: +6–8)
Randomized Smoothing	73 ± 2 (95% CI: 71–75)	59 ± 1.5 (95% CI: 58–61)	88 ± 2 (95% CI: 86–90)	+15 ± 2 (95% CI: +13–17)
Lipschitz Constraints	80 ± 1.8 (95% CI: 78–82)	68 ± 1.8 (95% CI: 66–70)	90 ± 2 (95% CI: 88–92)	+10 ± 1.5 (95% CI: +8–12)

- Ethical considerations must be addressed, ensuring that AI robustness does not compromise explainability, a key requirement in medical AI adoption.

9. DISCUSSION

The results show that, despite a common prejudice that modern AI models in medicine are less vulnerable to such attacks than human experts, modern AI models suffer significantly, particularly models utilized in medical imaging and clinical decision support systems. It is shown that CNN-based medical imaging models can be inferred by perturbation-based adversarial inputs with an over 50% accuracy drop, making them susceptible to such attacks. In addition, Transformer models applied in NLP-based clinical documentation experienced a huge 30% accuracy reduction when attacked by both FGSM and PGD models, which brings into question the integrity of the AI-generated medical records. Adversarial defense strategies are analyzed, where Randomized Smoothing and Adversarial Training achieve the best robust improvement of accuracy, improving by 15% and 14%, respectively. The Randomized Smoothing approach was the largest source of computational overhead, and this may prevent it from practical application in the time-constrained settings of medical environments. Additionally, the Gradient Masking approach leads to marginal (+7%) improvements but is computationally efficient, and therefore, the results indicate that this approach is ineffective against adaptive adversaries, compared with other defenses. This work reinforces the tradeoff between security, interpretability, and computational efficiency of medical AI systems, and that there is not a single defense strategy to protect from all of these system properties.

Besides a general accuracy degradation, adversarial attacks change false positive rates (FPR) and false negative rates (FNR) significantly, which have different clinical risks. In radiology models, adversarial perturbations caused FNRs to rise by up to +18% ($p < 0.01$) and FPRs to rise by up to +12% ($p < 0.05$), respectively, which suggests the possibility of more missed disease diagnoses and false alarms resulting in unnecessary follow-up testing. PGD attacks raised FNRs more than FPRs in the ECG-based RNNs, a trend that is worrying, considering the life-threatening nature of undiagnosed cardiac conditions.

Statistically, these changes in Type I (false positive) and Type II (false negative) errors indicate that adversarial manipulation does not merely decrease model accuracy, but biases distributions of diagnostic errors. The imbalance may have a systematic effect of weakening clinical workflows in case it goes undetected. Defense strategies were observed to partially counteract these effects, limiting FNR increases by up to 14% under Randomized Smoothing, but none fully recovered baseline sensitivity and specificity. These results underscore the need to report error rates as well as accuracy in adversarial robustness experiments and to make adversarial testing a part of medical AI validation procedures, where it is as essential to ensure balanced errors as it is to achieve high overall classification accuracy.

This is by previous work showing that adversarial attacks can be devastating to AI model performance in healthcare. This study shows results that are in line with those of Finlayson *et al*. (2019), who also found that CNN-based diagnostic models are highly susceptible to adversarial noise, with classification failure of CNNs being a drastic consequence of JSMA attacks. Studies conducted by Muoka *et al*. (2023) also highlighted that gradient-based attacks (FGSM, PGD) are great threats to AI-driven medical imaging, which is also proved by our results with a success rate larger than 40%.

Unlike previous works that mainly concentrated on adversarial vulnerabilities, this study additionally quantifies the tradeoffs of adversarial defenses, especially in terms of computational efficiency. Both adversarial training (additional 1.5× cost) and randomized smoothing (additional 1.8× cost) incur additional computation that other researchers have previously cited as a potential problem with deployability in clinical settings. This enables us to present a balanced take on security vs. practical applicability for those looking to deploy medical AI and determine whether theoretical adversarial robustness translates to real-world deployment.

The potential findings of the study highlight the importance of adversarial robustness in medical AI models before they are adopted in clinical settings. The fact that CNN radiology models are vulnerable to gradient-based perturbations suggests that organizations like the FDA should require adversarial robustness testing as part of their AI validation procedures. Similarly, AI-based decision support tools in electronic health records (EHRs) must incorporate adaptive adversarial defenses to prevent the systematic misclassification of patient risk scores. Additionally, the study shows that implementing robust AI defenses increases computational overhead, which can be problematic for resource-limited hospitals by adding computational demands that hinder deployment. This underscores the need for an optimized defense strategy that balances security and efficiency, ensuring access within healthcare environments with diverse populations. However, the study has some limitations despite its contributions. It is primarily based on theoretical models and simulated success rates of adversarial attacks rather than real-world patient data, offering limited empirical validation. Although CNNs, Transformers, and RNNs were analyzed, the evaluation of hybrid AI architectures was limited, possibly restricting the generalizability of the results. Finally, the study does not thoroughly explore adaptive adversarial threats, such as meta-learning-based adversarial optimizations, which could pose emerging security risks for medical AI. Nonetheless, the results provide a solid foundation for understanding how adversarial vulnerabilities can impact medical AI and establish a framework for assessing the plausibility of adversarial defenses in medical AI deployments. While further research involving real-world adversarial scenarios and actual patient datasets is needed, the adversarial testing conducted suggests that the proposed intervention method shows promising outcomes. It also opens the door to integrating adversarially robust AI models into regulatory frameworks to ensure medical AI systems meet security and compliance standards. Overall, these findings not only advance adversarial defense research but also enhance the reliability of statistical models by quantifying error distributions, confidence intervals, and validation protocols essential for the adoption of medical AI.

10. CONCLUSION

We provide a comprehensive study of adversarial vulnerabilities of AI-driven healthcare models and show that sacrificing a little bit of the accuracy leads to severe degradation of the models' ability to make medical diagnostics. Our results suggest that CNN-based radiology models are susceptible to an accuracy drop from 92% to 40% under JSMA attacks and resulting in a 42% drop in the performance of the ECG-based AI model under PGD attack. Like the transformer-based NLP models, transformer-based AI models suffered a 30% accuracy drop under FGSM attacks, which indicates that any AI-assisted clinical decision-making process may not be trustworthy The best improvements in terms of defense were obtained by Randomized Smoothing and Adversarial Training, with respective accuracy gains of 15% and 14%. However, although Randomized Smoothing pays 1.8× computational overhead for each sample, it is not practically applicable in real time to clinical applications. During training, through the use of Gradient Masking, which is computationally efficient, we only gained a 7% accuracy improvement, but were unable to defeat adaptive attacks. We also highlight the communication among these three dimensions: adversarial robustness, computational feasibility, and real-world applicability, and echo the need for developing secure, close to real-world, compliant, and language interpretable models in healthcare to prevent adversarial exploitation. To ascertain the reliability of our findings, post-hoc power analysis was performed. The statistical power with the observed effect sizes (Cohen d = 1.8-3.2) to detect the difference in accuracy between the baseline and attacked was greater than 0.95 at the alpha level of 0.05. In the same way, the power values of the comparisons between the defense strategies were above 0.85, which means that the samples employed in this research were sufficiently sensitive.

These findings show that the accuracy degradations and robustness gains reported were not only significant but were also backed by adequate sample sizes to warrant the generalizability of the conclusions. In future work, the adversarial robustness should be validated in larger multi-institutional datasets to ensure reproducibility across varying clinical settings.

REFERENCES

[1] Eskandar K. Artificial intelligence in healthcare: Explore the applications of AI in various medical domains, such as medical imaging, diagnosis, drug discovery, and patient care. Series Med Sci 2023; 4: 37-53.

[2] Salammagari RR, Srivastava G. Artificial intelligence in healthcare: Revolutionizing disease diagnosis and treatment planning. Int J Res Comput Appl Inf Technol 2024; 7: 41-53.

[3] Thompson S. AI in Healthcare: How Machine Learning is Revolutionizing Treatment and Diagnosis. EPH-International Journal of Science and Engineering 2023; 9(2): 28-46. https://doi.org/10.53555/ephijse.v9i2.255

[4] Adenekan TK. AI-Driven Diagnostic Models for Cardiovascular Health: Exploring Security and Business Analytics in Aortic Stenosis Detection 2024.

[5] Javanmard S. Revolutionizing medical practice: The impact of artificial intelligence (AI) on healthcare. OA J Applied Sci Technol 2024; 2(1): 01-16. https://doi.org/10.33140/OAJAST.02.01.07

[6] Olawade DB, David-Olawade AC, Wada OZ, Asaolu AJ, Adereni T, Ling J. Artificial intelligence in healthcare delivery: Prospects and pitfalls. Journal of Medicine, Surgery, and Public Health 2024; 100108. https://doi.org/10.1016/j.glmedi.2024.100108

[7] Love H, James C. AI-Driven Optimization in Healthcare: Enhancing Predictive Diagnostics and Personalized Treatment Strategies 2024.

[8] Oyeniyi J, Oluwaseyi P. Emerging trends in AI-powered medical imaging: Enhancing diagnostic accuracy and treatment decisions. Int J Enhanced Res Sci Technol Eng 2024; 13.

[9] Vallverdú J. Challenges and controversies of generative AI in medical diagnosis. Euphyía 2023; 17(32): 88-121. https://doi.org/10.33064/32euph4957

[10] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019; 363(6433): 1287-1289. https://doi.org/10.1126/science.aaw4399

[11] Muoka GW, Yi D, Ukwuoma CC, Mutale A, Ejiyi CJ, Mzee AK, et al. A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense. Mathematics 2023; 11(20): 4272. https://doi.org/10.3390/math11204272

[12] Bonagiri K, VS NM, Gopalsamy M, Iyswariya A, Sultanuddin SJ. AI-Driven Healthcare Cyber-Security: Protecting Patient

Data and Medical Devices. 2024 Second International Conference on Intelligent Cyber-Physical Systems and Internet of Things (ICoICI) 2024; 107-112. https://doi.org/10.1109/ICoICI62503.2024.10696183

[13] Mulukuntla S. Generative AI, Benefits, limitations, potential risks, and challenges in the healthcare industry. EPH-International Journal of Medical and Health Science 2022; 8(4): 1-9.

[14] Dani L, Wajid Q. Mitigating Security Risks in Healthcare Applications through AI and Machine Learning 2024.

[15] Alkayyali ZK, Taha AM, Zarandah QM, Abunasser BS, Barhoom AM, Abu-Naser SS. Advancements in AI for Medical Imaging: Transforming Diagnosis and Treatment 2024.

[16] ALRuwaili HQ, Alharbi OE, Alshammari YM, Alrewaili FS, Alyamani IM, Alqurashi SM. Impact of Health Information Technology on Workflow Efficiency and Patient Safety in Pharmacy Practices: A Critical Review. International Journal of Biological & Pharmaceutical Science 2018; 4(1): 30-35.

Received on 17-10-2025 Accepted on 19-11-2025 Published on 08-12-2025

https://doi.org/10.6000/1929-6029.2025.14.71