Adversarial Attack Prevention in Medical Imaging Deep Learning: A Comprehensive Review

Dr. Dhairya Vyas[1] *, Dr. Sheshang Degadwala*[2]

[1]Post Doctoral Researcher, Lincoln University College, Petaling Jaya, Selangor Darul Ehsan-47301, Malaysia ; [2] Professor & Head, Department of Computer Engineering, Sigma University, Vadodara, Gujarat

[1]pdf.dhairyavyas@lincoln.edu.my, [2]sheshang13@gmail.com

Abstract: Deep learning has also enhanced the predictability of medical imaging systems, allowing them to reliably diagnose using such modalities as X-ray, MRI, and CT scans. Nevertheless, all these models are extremely susceptible to adversarial attacks, where imperceptible modifications may result in critical misdiagnoses, casting serious doubts on the security and reliability of AI in healthcare. The paper is a review of the existing literature on preventive frameworks that can be used to secure medical imaging deep learning models against adversarial threats. It critically reviews the current defense mechanisms such as adversarial training, input preprocessing, robust architecture design, explainable AI integration, and hybrid security mechanisms. Comparative analysis shows the effectiveness of such methods in terms of performance (accuracy, precision, recall, and F1-score) and limitations of such methods in terms of robustness and interpretability. The results indicate that, despite the great strides made, the current methods do not have the generalized defense capabilities and cannot withstand the adaptive attacks. The paper highlights the necessity of comprehensive, extensible, and understandable security systems. These lessons are vital in creating credible AI systems that can be used in the real-world clinical decision support, medical diagnostics, and healthcare security infrastructures.

Keywords: Medical Imaging; Deep Learning; Adversarial Attacks; Model Security; Robust AI

Introduction

Medical imaging has become an essential part of contemporary healthcare, facilitating early diagnosis, monitoring and treatment planning of diseases via non-invasive procedures. The most common imaging modalities include X-ray, Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) which offer detailed information on anatomical structures and pathological conditions. Over the last few years, the combination of deep learning methods has dramatically changed the way medical images are analyzed and provided automated, accurate, and scalable solutions to medical imaging analysis problems, such as classification, segmentation, and anomaly detection. Deep learning models, especially Convolutional Neural Networks (CNNs) and Transformer-based architectures have shown impressive performance in the detection of diseases such as cancer, neurological diseases, and cardiovascular abnormalities [1,2]. The use of artificial intelligence (AI) in healthcare has brought many positive outcomes, such as the increased accuracy of the diagnostic process, the decreased number of human errors, and the increased efficiency of the clinical process as shown in **Figure 1**.

The purpose of this review is to give a thorough description of preventive frameworks that can be used to ensure that deep learning models used in medical imaging are resistant to adversarial attacks. It examines the present situation in research, considers the different defense measures, and identifies some of the major challenges and gaps in research. This work will help to create more robust, secure,

and trustworthy AI systems to be used in medical imaging practices. Finally, to achieve the safety of these systems, it is necessary to secure patient outcomes and the future of intelligent healthcare.

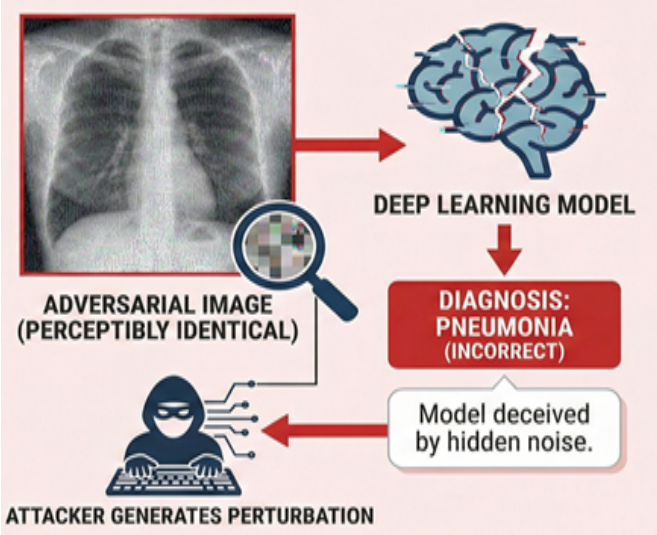

Figure 1. Process of Adversarial attack

Background Study

In **Table 1**, a detailed overview of current methods of medical imaging deep learning and adversarial defense, including the techniques used, their applications, and the security measures applied in the related research [1]-[20]. It demonstrates that an extensive variety of methods, such as federated learning, explainable AI, encryption, and hybrid deep learning models, have been discussed to promote robustness.

Table 1. Comparative Study of previous research by other researchers

Ref.	Method	Technique Used	Adversarial Defense Strategy	Key Contribution
[1]	Federated Generative Models	Federated Learning + Generative Models	Privacy-preserving training	Enhances secure distributed learning
[2]	Transfer Learning for Synthetic Detection	Transfer Learning (CNN)	Synthetic image detection	Prevents fraud using generated images
[3]	Robust XAI Evaluation	Explainable AI (XAI)	Robustness analysis	Evaluates model reliability under attacks
[4]	Image Security Scheme	Encryption + Watermarking	Data-level security	Secures image transmission
[5]	Adversarial Defense Review	Survey of defenses	Multiple defenses	Comprehensive overview of techniques
[6]	Q-VFL Framework	Quantum Federated Learning	Secure training	Combines quantum + federated learning
[7]	Anomaly Detection in FL	Federated Learning	Attack detection	Detects malicious updates
[8]	XAI for Security	Explainable AI	Attack detection	Improves interpretability and trust

[9]	CoDeSwiT Architecture	Hybrid CNN-Transformer	Feature fusion robustness	Enhances global-local features
[10]	Blockchain-based Security	Blockchain	Tamper-proof system	Secures training datasets
[11]	Hybrid Attention Model Attack Study	Transfer Learning + Attention	Attack analysis	Shows vulnerability of hybrid models
[12]	Kernel Density Detection	Feature extraction + KDE	Adversarial detection	Identifies abnormal inputs
[13]	Trusted Framework	Deep Learning Framework	Adversarial robustness	Improves reliability
[14]	Adversarial Defense Model	CNN-based defense	Defense techniques	Enhances robustness
[15]	Hybrid CNN-Transformer + Chaos	Encryption + DL	Data protection	Combines DL and chaos theory
[16]	Physical Attack Study	Camera-based attacks	Attack demonstration	Highlights real-world risks
[17]	Siamese Denoising Autoencoder	Autoencoder	Noise removal	Improves adversarial robustness
[18]	Adversarial Feature Fusion	Ensemble Learning	Attack detection	Detects manipulated images
[19]	Two-Phase Framework	Deep Learning	Robust classification	Improves adversarial resistance
[20]	Noise Removal Technique	Total Variation + Patch Reg.	Noise reduction	Removes adversarial perturbations

Research Gaps

Although there has been significant progress in ensuring that the medical imaging deep learning models are secured, there are still several critical gaps in research that limit the efficacy and practical implementation of adversarial defense models.

1. Deficiency Of Generalized Defense Mechanisms.

The majority of the currently in use defense strategies have been developed to respond to a particular form of adversarial attack, like FGSM or PGD. Nevertheless, such approaches usually do not work when subjected to invisible or adaptive attacks [5].

2. Trade-off Between Accuracy and Robustness

Numerous defense methods, such as adversarial training and input preprocessing, enhance robustness but tend to reduce model accuracy on clean data [14], [19]. This trade-off is a major challenge in medical applications where high accuracy and reliability are needed.

3. Limited Interpretability and Trustworthiness

Despite the introduction of explainable AI (XAI) methods, their combination with adversarial defense mechanisms is not yet widespread. The existing models tend to act as black boxes and therefore the clinicians cannot trust their predictions in case of adversarial conditions [3], [8].

4. Inability to use Standardized Evaluation Metrics.

No standardized standard to assess adversarial robustness in medical imaging exists. Various measures and experimental designs are used in different studies, and this inconsistency in measures and experimental designs can be noted [5], [23].

5. Issues of Computational complexity and scalability.

Most of the defense models (hybrid and ensemble, in particular) are computationally expensive, which restricts the applicability of these methods in real-time clinical settings [18], [19].

Conclusions

This review reflects on the urgent necessity of ensuring deep learning-based medical imaging systems are resistant to adversarial attacks since minor perturbations can cause severe misdiagnoses and pose a threat to patient safety when applied in AI-driven clinical decision-making. The recent studies were analyzed thoroughly, and the techniques included CNNs, Transformers, and hybrid models, adversarial training, autoencoder-based defenses, explainable AI, federated learning, and blockchain-based security. Comparative analysis in terms of such metrics as accuracy, precision, recall, and F1-score demonstrates that despite high diagnostic performance of such models, they are susceptible to adversarial threats. Relatively better robustness is provided by hybrid and ensemble methods, whereas techniques like denoising and adversarial training can only help to increase resilience but do not guarantee ultimate protection. Explainable AI enhances interpretability, and new technologies can support privacy and data integrity, but the robustness against adaptive and real-life attacks is limited. This paper is limited by depending on the existing literature, which is not usually evaluated and validated in practice. Further research ought to be done in designing scalable, generalized protection frameworks, integrating robustness and interpretability, enhancing efficiency, and testing models in real clinical settings.

References

[1] H. Mahmood, Z. Alamgir, S. T. Javed, S. Karim, and M. Awais, “Federated Generative Models in Medical Imaging: Current Advances, Challenges, and Future Directions,” *IEEE Access*, vol. 14, no. December 2025, pp. 5197–5217, 2026, doi: 10.1109/access.2026.3650810.

[2] O. Tariq, M. A. Arshed, M. Kabir, K. Ijaz, Ş. C. Gherghina, and H. B. Batool, “A Transfer-Learning Approach for Detection of Multiclass Synthetic Skin Cancer Images Generated by Deep Generative Models to Prevent Medical Insurance Fraud,” *Mathematical and Computational Applications*, vol. 31, no. 1, pp. 1–17, 2026, doi: 10.3390/mca31010031.

[3] S. Repetto, I. Maljkovic, M. Lotto, A. E. Cinà, S. Vascon, and F. Roli, “Evaluating the robustness of explainable AI in medical image recognition under natural and adversarial data corruption,” *Machine Learning*, vol. 115, no. 1, pp. 1–21, 2026, doi: 10.1007/s10994-025-06919-6.

[4] C. Ye, L. Shi, S. Tan, J. Wang, Q. Zuo, and W. Feng, “A controllable medical image security scheme using selective encryption and watermarking for bit-planes in the TSH domain,” *Cybersecurity*, vol. 9, no. 1, 2026, doi: 10.1186/s42400-025-00535-6.

[5] R. Ma, Y. Zhang, J. Wang, W. Lu, and X. Luo, “Advancements in adversarial example defense for deep learning models: a review,” *Cybersecurity*, vol. 9, no. 1, 2026, doi: 10.1186/s42400-025-00546-3.

[6] A. Kottahachchi Kankanamge Don and I. Khalil, “Q-VFL: quantum-enhanced vertical federated learning with contrastive encoding for privacy-preserving medical AI,” *Quantum Machine Intelligence*, vol. 8, no. 1, 2026, doi: 10.1007/s42484-026-00385-6.

[7] M. Lengl *et al.*, “From Injection to Detection: Analyzing the Impact of Anomalous Data in Federated Learning,” *SN Computer Science*, vol. 7, no. 2, 2026, doi: 10.1007/s42979-026-04775-2.

[8] A. A. ForouzeshNejad, F. Arabikhan, and R. Taheri, “The Role of Explainable AI (XAI) in Enhancing the Security of Machine Learning Systems Against Adversarial Attacks,” *Adversarial Example Detection and Mitigation Using Machine Learning*, pp. 153–170, 2026, doi: 10.1007/978-3-031-99447-0_11.

[9] C. I. Sci, “Article in Press CoDeSwiT : an efficient multi-stream hybrid architecture for Global – Local feature fusion via dynamic gating in medical ımage analysis IN AR IN,” *Journal of King Saud University Computer and Information Sciences*, 2026, doi: 10.1007/s4444 3-026-00726-2.

[10] R. Shinde, S. Patil, K. Kotecha, and S. Mishra, “Ensuring the integrity of AI models: a blockchain-based approach for protecting medical imaging training data,” *Scientific Reports*, vol. 16, no. 1, p. 13989, Mar. 2026, doi: 10.1038/s41598-026-44040-3.

[11] O. I. Obaid and A. Alazzawi, “Adversarial Attacks on Hybrid Attention Integrated Transfer Learning for Lung Cancer CT Classification,” *Mesopotamian Journal of CyberSecurity*, vol. 5, no. 2, pp. 863–885, 2025, doi: 10.58496/MJCS/2025/049.

[12] S. Das, P. K. Keserwani, and M. C. Govil, *Detecting Adversarial Samples using Kernel Density Feature Extractor in Medical Image*, vol. 1, no. 1. Association for Computing Machinery, 2025. doi: 10.1145/3700838.3703675.

[13] T. J. Mohammed, C. Xinying, A. S. Albahri, A. Alnoor, and K. Khai Wah, “A Novel Trusted Framework for Orthopedic Disease Detection With Reliability Against Adversarial Attacks,” *IEEE Access*, vol. 13, no. August, pp. 160193–160220, 2025, doi: 10.1109/ACCESS.2025.3609330.

[14] M. J. Tsai, Y. C. Lee, H. Y. Lien, and C. C. Liang, “Adversarial Defense for Medical Images,” *Electronics (Switzerland)*, vol. 14, no. 22, 2025, doi: 10.3390/electronics14224384.

[15] R. S. Ali, R. K. Hassoun, A. Al-Adhami, S. A. Jabber, and S. H. Hashem, “Proposal Medical Image Protection System Based on Hybrid CNN-Transformer Model and Chaotic Maps,” *IEEE Access*, vol. 13, no. July, pp. 123793–123807, 2025, doi: 10.1109/ACCESS.2025.3586562.

[16] J. Oda and K. Takemoto, “Mobile applications for skin cancer detection are vulnerable to physical camera-based adversarial attacks,” *Scientific Reports*, vol. 15, no. 1, pp. 1–12, 2025, doi: 10.1038/s41598-025-03546-y.

[17] J. Shim and K. Jo, “Siamese Denoising Autoencoders for Enhancing Adversarial Robustness in Medical Image Analysis,” *IEEE Access*, vol. 13, no. April, pp. 86333–86343, 2025, doi: 10.1109/ACCESS.2025.3570898.

[18] A. Ali, H. A. Basha, K. Thanuja, Puneet, S. K. Gupta, and S. Kim, “Enhancing tumor deepfake detection in MRI scans using adversarial feature fusion ensembles,” *Scientific Reports*, vol. 16, no. 1, p. 1667, Dec. 2025, doi: 10.1038/s41598-025-31231-7.

[19] S. B. ul haque and A. Zafar, “Robust Medical Diagnosis: A Novel Two-Phase Deep Learning Framework for Adversarial Proof Disease Detection in Radiology Images,” *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, pp. 308–338, 2024, doi: 10.1007/s10278-023-00916-8.

[20] B. U. H. Sheikh and A. Zafar, “Removing Adversarial Noise in X-ray Images via Total Variation Minimization and Patch-Based Regularization for Robust Deep Learning-based Diagnosis,” *Journal of Imaging Informatics in Medicine*, vol. 37, no. 6, pp. 3282–3303, 2024, doi: 10.1007/s10278-023-00919-5.