Explainable Adversarial Attack Detection in Medical Images Using Layer-wise Progressive Activation Maps (PLAN)

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: Medical image analysis is a key application area for deep learning models, and these models are susceptible to adversarial attacks and can lead to unexpected changes in the results, thus compromising the clinical validity. This paper introduces an explainable adversarial attack detection framework that examines the activation patterns of several convolutional layers rather than just a last layer feature representation, called Layer-wise Progressive Activation Maps (PLAN). PLAN produces progressive activation maps, fuses them into a single explainability map and recovers activation properties such as entropy, activation area ratio, maximum activation, and concentration score. All of these features are aggregated into a PLAN Attack Score: images with scores higher than 1.35 are considered to be adversarial, and lower scores are considered to be clean images. PLAN also includes a quantitative score as well as suspicious regions to visually interpret. The experimental results show the effectiveness of detecting adversarial images and the improvement in transparency, robustness, and trustworthiness in the medical diagnosis process of AI algorithms.

Keywords: Medical Image Analysis; Adversarial Attack Detection; Explainable Artificial Intelligence (XAI); Layer-wise Progressive Activation Maps (PLAN); Deep Learning Security

Introduction

In the field of medical image analysis, deep learning has emerged as a crucial technique with significant success in disease classification, lesion detection, and clinical decision support. Yet, deep neural networks continue to be very susceptible to adversarial attacks, which can introduce small perturbations in the data that strongly affect model predictions and affect the reliability of diagnosis [1,2,3,4]. The rise in the sophistication of the adversarial attacks in computer vision applications has spurred a lot of research into attack detection, defenses, and making the systems more robust to attacks [5,6,7,8,10].

In recent studies, a number of adversarial detection methods have been proposed such as explainability using Grad-CAM [9,11,12,13,14,17,20], clustering methods, deep feature analysis, reconstruction-based learning, optimized adversarial networks, attack-as-detection paradigms, and self-supervised representation learning. However, in the medical imaging field, there have been a few studies on adversarial robustness using federated learning, robust diagnostic models, and defense strategies for pretrained models, along with several reviews of the numerous challenges and opportunities for robust medical image analysis systems [15,16,18,19]. While these methods are effective at making the system more robust, many detection methods rely on feature representations from the last convolutional layer, which makes it difficult to observe the hierarchical evolution of the activations of features throughout the network.

To overcome this drawback, this paper introduces a novel, explainable method of adversarial attack detection in medical images, called Layer-wise Progressive Activation Maps (PLAN). PLAN's hierarchical activation information, combined with a unified explainability framework, offers interpretable visual evidence, and improves the detection of adversarial attacks, paving the way to more trustworthy and transparent AI-based medical imaging systems.

Background Study

Table 1 summarizes representative algorithms, the adversarial attacks they address, along with their major advantages and limitations.

Table 1. Comparison of Existing Adversarial Attack Detection Methods

Algorithms	Target Attacks	Pros	Cons
Hybrid Reverse Knowledge Distillation [1]	FGSM, PGD, CW	High detection accuracy and strong feature discrimination	Requires teacher–student training and high computational cost
Federated Learning [2,11]	Distributed adversarial attacks	Preserves data privacy and improves collaborative robustness	Communication overhead and limited explainability
Grad-CAM + Clustering [9]	FGSM, PGD	Explainable visualization of suspicious regions	Uses only final-layer Grad-CAM and misses hierarchical information
Optimized Weighted Adversarial Network [13]	FGSM, BIM, PGD	Robust adversarial detection	Complex optimization and longer training time
Reconstruction-Based Detection [17]	FGSM, PGD	Captures reconstruction inconsistencies	Reconstruction network increases computational complexity

Methodology

In this proposed work, Layer-wise Progressive Activation Maps (PLAN) are used to detect adversarial attacks by observing the progression of activation maps at different convolutional layers of a pretrained Convolutional Neural Network (CNN). The input medical image is first preprocessed by resizing and normalizing it before being passed on to the CNN. PLAN does not only obtain an activation map from the last convolutional layer, but also from several intermediate layers of the convolutional layers, so that hierarchical features at different semantic levels can be represented. The activation maps are resized back to the original image resolution and then rescaled to a range of 0-1 for easy comparison between the layers. The Structural Similarity Index (SSIM) is computed between two successive activation maps in order to assess the consistency of the feature evolution,

$$S_i = \mathrm{SSIM}(A_i, A_{i+1}), \tag{1}$$

where A_i and A_{i+1} represent the activation maps for i^{th} and $(i+1)^{th}$ convolutional layers, respectively. Typical medical images are clean, with smooth and uniform transitions in activation, while adversarial images have abrupt transitions between layers. Hence, the intensity of the activation transition is given by,

$$T_i = \frac{1}{HW} \sum_{x=1}^{H} \sum_{y=1}^{W} | A_{i+1}(\mathrm{x},\mathrm{y}) - A_i(\mathrm{x},\mathrm{y}) |, \tag{2}$$

Where, the height H and width W of the activation maps are shown.

The layer weights are adaptively set to highlight more informative feature representations, with the layer weight defined as the normalized mean activation of the layer,

$$w_i = \frac{\mu(A_i)}{\sum_{k=1}^{L} \mu(A_k)}, \tag{3}$$

where $\mu(\cdot)$ is the mean activation value and L is the number of all the convolutional layers selected. Lastly, the proposed PLAN adversarial score combines the cross-layer consistency and the intensity of activation transition as,

$$\text{PLAN}_{score} = \sum_{i=1}^{L-1} w_i[\alpha(1 - S_i) + \beta T_i], \tag{4}$$

where α and β are weighting coefficients such that $\alpha + \beta = 1$. The higher the PLAN score, the greater the differences in the hierarchical evolution of features and evidence of adversarial manipulation. The final decision is arrived at with the help of,

$$\text{Image} = \begin{cases} \text{Adversarial}, & \text{PLAN}_{score} \geq \tau, \\ \text{Clean}, & \text{PLAN}_{score} < \tau, \end{cases} \tag{5}$$

Here, τ is the decision threshold that is determined using the validation dataset.

Results

The experiments were conducted using the publicly available COVID-Pneumonia-Normal Chest X-ray Images dataset (sachinkumar413/covid-pneumonia-normal-chest-xray-images) from Kaggle, which contains chest X-ray images of COVID-19, Pneumonia, Normal, and other respiratory conditions. All images were resized and normalized before analysis, and the proposed Layer-wise Progressive Activation Maps (PLAN) framework was implemented and evaluated in the Google Colab environment using Python with GPU acceleration, leveraging the PyTorch, OpenCV, NumPy, and Matplotlib libraries.

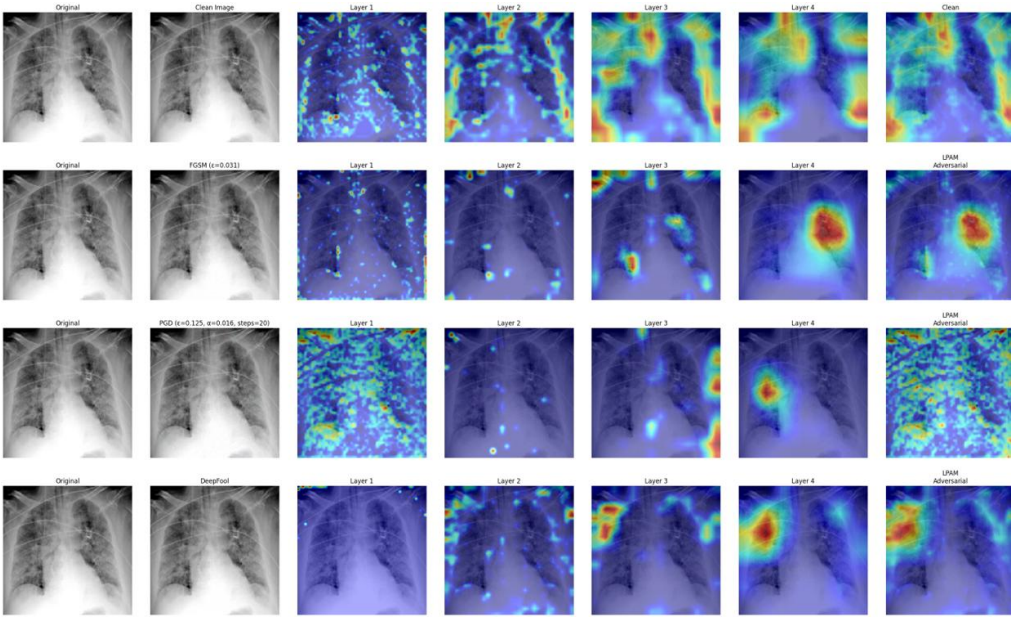

Figure 1. Layer-wise Progressive Activation Maps (PLAN) for Clean and Adversarial Medical Images

Table 2. Quantitative Comparison of PLAN Metrics for Clean and Adversarial Images

Metric	Clean Image	FGSM	PGD	DeepFool
Entropy	10.632690	10.392539	10.677216	10.306990
Activation Area	3432	2128	1475	2496
Area Ratio	0.068399	0.042411	0.029397	0.049745

Maximum Activation	1.000000	1.000000	1.000000	1.000000
Concentration Score	3.407415	5.869023	3.431035	6.232594
Attack Score	1.339325	2.086473	1.357955	2.193663

Proposed Layer-wise Progressive Activation Maps (PLAN) metrics for clean and adversarial chest X-ray images generated using FGSM, PGD and DeepFool attacks are quantitatively compared in **Table 2**. A higher score for an attack means higher activation inconsistency across the layers, due to adversarial perturbations.

Conclusions

This research has presented a solution to the vulnerability of deep learning-based medical image classification systems to adversarial attacks, as well as the fact that the traditional attack detection methods are not very interpretable. To address these challenges, the proposed Layer-wise Progressive Activation Maps (PLAN) is based on progressive changes of activation patterns within different convolutional layers, as well as structural similarity, activation transition intensity, and adaptive layer weighting across layers to provide an explainable score for adversarial detection. The analysis shows that whilst clean medical images have relatively stable hierarchical activation patterns, adversarial examples generated by FGSM, PGD and DeepFool have noticeable alterations in cross-layer activation patterns that PLAN is able to identify and quantify. The current work, however, has its drawbacks as it has been evaluated in a specific medical image dataset, chosen attack strategies, CNN architecture, and a threshold-based detection mechanism. Future research should thus validate PLAN across various medical imaging modalities and architectures, include other white-box and black-box attacks, explore data-driven optimization of the threshold, and demonstrate the generalizability and clinical interpretability of PLAN in larger-scale experiments.

References

[1] H. Kwon, J. B. Maeng, and D.-J. Kim, “Hybrid reverse knowledge distillation for adversarial example detection,” *Scientific Reports*, vol. 16, no. 1, p. 20384, May 2026, doi: 10.1038/s41598-026-50137-6.

[2] M. Mohil, A. Mohil, and A. S. Mohil, “Federated Learning for Enhancing Reliability and Security in Medical Image Analysis Against Adversarial Threats,” *Journal of Engineering and Technology for Industrial Applications*, vol. 12, no. 57, pp. 1045–1053, 2026, doi: 10.5935/jetia.v12i57.3209.

[3] M. H. Kamil, E. P. T. Farma, and S. Basuki, “Robustness Under Attack: Assessing Adversarial Fragility in Deep Learning Models for COVID-19 Radiography Prediction,” *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 8, no. 2, pp. 769–788, 2026, doi: 10.35882/jeeemi.v8i2.1506.

[4] C. Daniel Ponnan and S. Dibbo, “Deep Feature-Based Detection of Adversarial Attacks in Medical Image Classification,” *SSRN*, 2026, doi: 10.2139/ssrn.6598778.

[5] L. Wang, X. Lei, H. He, L. Wang, J. Shi, and Z. Wu, “Over-the-Air Adversarial Attack Detection: from Datasets to Defenses,” *arXiv*, vol. 14, no. 8, pp. 1–13, Sep. 2025, [Online]. Available: http://arxiv.org/abs/2509.09296

[6] Y. Li, P. Angelov, and N. Suri, “Self-supervised Representation Learning for Adversarial Attack Detection,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* , vol. 15118 LNCS, pp. 236–252, 2025, doi: 10.1007/978-3-031-73027-6_14.

[7] D. Pal, R. Sony, and A. Ross, "A Parametric Approach to Adversarial Augmentation for Cross-Domain Iris Presentation Attack Detection," *Proceedings - 2025 IEEE Winter Conference on Applications of Computer Vision, WACV 2025*, pp. 5719–5729, 2025, doi: 10.1109/WACV61041.2025.00558.

[8] K. N. T. Nguyen *et al.*, "A Survey and Evaluation of Adversarial Attacks in Object Detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 9, pp. 15706–15722, 2025, doi: 10.1109/TNNLS.2025.3561225.

[9] J. H. Sim and H. M. Song, "A Generalized Framework for Adversarial Attack Detection and Prevention Using Grad-CAM and Clustering Techniques," *Systems*, vol. 13, no. 2, 2025, doi: 10.3390/systems13020088.

[10] N. Al Roken, H. Hacid, A. Bouridane, and A. Hussain, "On adversarial attack detection in the artificial intelligence era: Fundamentals, a taxonomy, and a review," *Intelligent Systems with Applications*, vol. 27, no. July, p. 200554, 2025, doi: 10.1016/j.iswa.2025.200554.

[11] E. Darzi, F. Dubost, N. M. Sijtsema, and P. M. A. Van Ooijen, "Exploring Adversarial Attacks in Federated Learning for Medical Imaging," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 12, pp. 13591–13599, 2024, doi: 10.1109/TII.2024.3423457.

[12] S. Pal, S. Rahman, M. Beheshti, A. Habib, Z. Jadidi, and C. Karmakar, "The Impact of Simultaneous Adversarial Attacks on Robustness of Medical Image Analysis," *IEEE Access*, vol. 12, no. March, pp. 66478–66494, 2024, doi: 10.1109/ACCESS.2024.3396566.

[13] K. Barik, S. Misra, and L. Fernandez-Sanz, "Adversarial attack detection framework based on optimized weighted conditional stepwise adversarial network," *International Journal of Information Security*, vol. 23, no. 3, pp. 2353–2376, 2024, doi: 10.1007/s10207-024-00844-w.

[14] Z. Zhao *et al.*, "Attack as Detection: Using Adversarial Attack Methods to Detect Abnormal Examples," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 3, 2024, doi: 10.1145/3631977.

[15] M. J. Tsai, P. Y. Lin, and M. E. Lee, "Adversarial Attacks on Medical Image Classification," *Cancers*, vol. 15, no. 17, 2023, doi: 10.3390/cancers15174228.

[16] G. W. Muoka *et al.*, "A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense," *Mathematics*, vol. 11, no. 20, 2023, doi: 10.3390/math11204272.

[17] M. Hussain and J. E. Hong, "Reconstruction-Based Adversarial Attack Detection in Vision-Based Autonomous Driving Systems," *Machine Learning and Knowledge Extraction*, vol. 5, no. 4, pp. 1589–1611, 2023, doi: 10.3390/make5040080.

[18] Revers Leigh, "Enhancement Robustness of Breast Cancer Models Against Adversarial Attacks," *University of Kut Journa*, vol. 00, no. May, pp. 2–3, 2023.

[19] M. Xu, T. Zhang, and D. Zhang, "MedRDF: A Robust and Retrain-Less Diagnostic Framework for Medical Pretrained Models Against Adversarial Attack," *IEEE Transactions on Medical Imaging*, vol. 41, no. 8, pp. 2130–2143, 2022, doi: 10.1109/TMI.2022.3156268.

[20] E. Tcydenova, T. W. Kim, C. Lee, and J. H. Park, "Detection of Adversarial Attacks in AI-Based Intrusion Detection Systems Using Explainable AI," *Human-centric Computing and Information Sciences*, vol. 11, 2021, doi: 10.22967/HCIS.2021.11.035.