Journal of Electronics, Electromedical Engineering, and Medical Informatics
Homepage: jeeemi.org; Vol. 8, No. 2, April 2026, pp: 769-788 e-ISSN: 2656-8632

RESEARCH ARTICLE OPEN ACCESS

Robustness Under Attack: Assessing Adversarial Fragility in Deep Learning Models for COVID-19 Radiography Prediction

Muhammad Hisyam Kamil, Elga Putri Tri Farma, and Setio Basuki

Department of Informatics Engineering, Universitas Muhammadiyah Malang, Malang, Indonesia

Corresponding author: Setio Basuki (e-mail: setio_basuki@umm.ac.id), **Author(s) Email**: Muhammad Hisyam Kamil (e-mail: hisyamkamil99@webmail.umm.ac.id, Elga Putri Tri Farma (e-mail: elgafarma@webmail.umm.ac.id)

Abstract **Deep learning, especially Convolutional Neural Network (CNN) architectures, has significantly improved medical image analysis for predicting lung diseases through chest X-ray (CXR) images, including pneumonia and COVID-19. However, despite achieving high diagnostic precision, CNN models remain highly susceptible to adversarial attacks, defined as small, visually imperceptible alterations optimized to exploit non-linear decision boundaries that cause high-confidence mispredictions. This vulnerability presents a critical concern in clinical settings, where deterministic diagnostic errors directly compromise patient safety. This paper systematically implements white-box adversarial attacks to quantify the resilience of CNN models in multi-class CXR image classification. This paper utilizes the COVID-19 Radiography Dataset, comprising four diagnostic categories: COVID-19, Lung Opacity, Normal, and Viral Pneumonia. A DenseNet-121 architecture was employed for feature extraction, and the trained model was subsequently subjected to Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks under varying L∞-bounded epsilon settings. The empirical experiments reveal three critical findings: 1) The implementation of sub-pixel adversarial attacks causes severe performance degradation, where the PGD attack constrained at an epsilon of 0.1/255 reduced the global model accuracy from a baseline of 95.42% to 25.32%; 2) Iterative attacks (PGD) represent the absolute worst-case scenario for model reliability by efficiently discovering high-dimensional manifold gaps, whereas the model demonstrates relative resilience to linear, single-step FGSM perturbations; and 3) Gradient-weighted Class Activation Mapping (Grad-CAM) analysis verifies that this performance collapse is associated with a deterministic semantic shift, displacing the model's spatial attention from clinically relevant pulmonary regions toward spurious background noise. In conclusion, this paper empirically proves that despite exhibiting high accuracy on clean data, unprotected CNNs remain fundamentally unsafe for autonomous clinical deployment due to their acute vulnerability to gradient-based perturbations, necessitating the future integration of robust adversarial training frameworks.**

Keywords **adversarial attacks; chest x-ray; convolutional neural network; COVID-19; Grad-CAM**

I. Introduction

The advancement of Deep Learning (DL) has brought a drastic change in the automated diagnosis of chest X-ray (CXR) images [1]. Convolutional neural networks (CNNs) have demonstrated strong effectiveness in detecting various lung abnormalities, including pneumonia, lung opacity, and COVID-19 [1][2][3]. The model has vast potential to help radiologists and speed up the diagnosis process. Furthermore, CNNs are applicable across multiple medical imaging modalities, not only microscopic images, highlighting the broad utility of deep learning in medical diagnosis [4]. Recent research has further confirmed the combination of ensemble CNNs with Explainable AI (XAI) to improve the diagnostic accuracy of complex medical imaging tasks [5]. Despite their high accuracy, CNN-based models are vulnerable to adversarial attacks. Mathematically, an adversarial attack involves adding a small, optimized perturbation δ to a clean input image x, generating an adversarial example formulated as $x' = x + \delta$. To ensure this modification remains visually imperceptible to clinicians, the magnitude of the perturbation is strictly bounded by the constraint $\| \delta \|_\infty \leq \epsilon$, where ϵ (epsilon) dictates the maximum allowable attack strength. The objective of this bounded noise is to force the model into making a high-confidence misclassification, which leads to substantial performance drops [6][7]. In the medical domain, such

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

vulnerability is of utmost importance as it could lead to potentially life-threatening misclassifications of patients [2][8][9]. Many studies have shown that high-performing deep learning models on common datasets could have disastrous performances when faced with substantially different inputs, thus raising substantial concerns about their clinical safety [7][10][11][12][13]. The vulnerability of AI models to adversarial attacks is not only a concern in the medical image data, but has also been consistently demonstrated in text-based medical data, where generative adversarial attacks can take advantage of vulnerabilities in clinical text processing models [14]. Moreover, recent reviews in the radiology field have emphasized that both white-box and black-box attacks could decrease model performance metrics, such as AUC, to near zero, thus underlining the importance of thorough security analyses [15][16]. This includes not only evasion attacks but also backdoor attacks, which are a rapidly increasing trend in deep learning security threats [17].

Among the different adversarial attack types, Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are recognized as effective gradient-based attack approaches [6][7][18]. FGSM involves adding a one-step perturbation in the direction of the loss gradient, whereas PGD involves iterating to create images that look the same to humans but are different to the model [7]. Recent reviews highlight the importance of recognizing that these attacks target the inherent weaknesses of model training, requiring a comprehensive taxonomy of detection and defense strategies [19][20]. In previous studies, PGD is considered a robust adversary baseline to evaluate the model's robustness against gradient-based attacks [2]. Recent studies also indicate that these weaknesses are not only present in classification models but also in segmentation models involving feature statistics transformation [15] and federated learning frameworks, where privacy-preserving protocols can be attacked [16].

However, current studies on medical imaging attacks rely mostly on quantitative metrics, such as accuracy scores. This approach hides the true cause of the failure. Specifically, there is a distinct absence of research utilizing Grad-CAM to track how the model's attention shifts specifically within the COVID-19 Radiography Dataset during an attack. Previous work does not explain whether the model is merely flipping a label or if it is fundamentally looking at the background instead of the lungs. This paper fills that gap by investigating a phenomenon we formally define as semantic decoupling. We characterize semantic decoupling as the systematic, forced separation between the true pathological features present in the image and the models' learned spatial attention. Under this condition, the model's predictive focus completely diverges from clinically relevant anatomical structures (e.g., pulmonary opacities) and rigidly anchors onto spurious, non-diagnostic background noise. We go beyond simple test scores to measure this topological divergence, explicitly quantifying the shift using spatial similarity metrics across Grad-CAM activation maps to show exactly how these attacks break the link between the disease features and the model's prediction.

Based on these situations, this paper implements FGSM and PGD to simulate the resilience of CNN models in the CXR image classification. To be more specific, this paper aims to: 1) evaluate the magnitude of the effect of attack parameters on CNN performance, and 2) understand the shift in model inference focus post-attack through Grad-CAM visualization [21][22]. The proposed experiments are expected to provide in-depth insight into the vulnerability of medical AI models for the development of future defense mechanisms [23]. This paper delivers several contributions as follows:

- Robustness evaluation of DenseNet-121 against FGSM and PGD adversarial attacks on CXR images.
- The model shows significant performance degradation under iterative PGD attacks ($\epsilon = 0.1/255$), where the accuracy drops to 25.32%.
- PGD outperforms FGSM in exploiting high-dimensional vulnerabilities, especially at low epsilon values.
- Grad-CAM reveals semantic decoupling by showing attention shifts from pathological regions to background noise.

II. Method

This section details the architectural framework used to construct the adversarial attack system proposed in Fig. 1. The primary data source utilized in this paper is the COVID-19 Radiography Dataset, a publicly accessible repository provided by researchers from Qatar University and the University of Dhaka, which can be found at the following link (https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database) [1]. To ensure a reproducible and systematically structured investigation, the proposed methodology follows four distinct operational stages: initial dataset acquisition and strict preprocessing to standardize input vectors, the configuration of the CNN architecture and specific training protocols for optimal feature extraction, the rigorous mathematical formulation of the gradient-based adversarial attack techniques, and finally, the execution of quantitative evaluation metrics alongside Grad-CAM interpretability protocols to accurately measure structural failure.

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

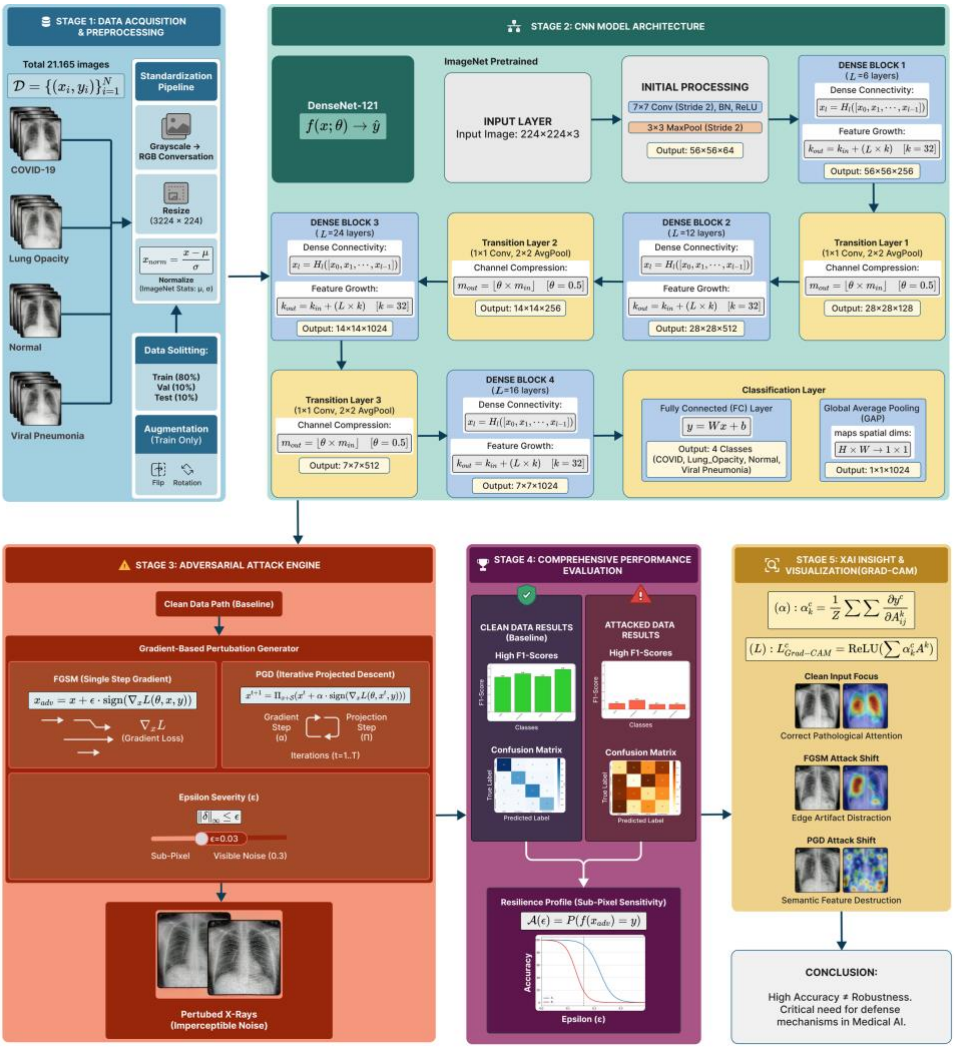


Fig. 1. Diagram Architecture and Adversarial Attacks Pipeline

A. Data Acquisition and Preprocessing

The acquired dataset comprises a total of 21,165 CXR images systematically categorized into four primary diagnostic groups: Normal (10,192 images), Lung Opacity (6,012 images), COVID-19 (3,616 images), and Viral Pneumonia (1,345 images) [1]. Sample images representing each pathological category are presented in Fig. 2. The distribution of the dataset used in this paper is detailed in Table 1.

Table 1. Distribution of COVID-19 Radiography Dataset

Class Category	Number of Images	Percentage (%)
Normal	10,192	48.15%
Lung Opacity	6,012	28.41%
COVID-19	3,616	17.08%
Viral Pneumonia	1,345	6.36%
Total	**21,165**	**100%**

Preprocessing is fundamental. It ensures that all medical image data maintains identical quality and formatting before input into the CNN model. We designed this process to prepare raw data for optimal machine learning performance without stripping critical diagnostic information. Initially, we collected all image paths and labels using Python. We then partitioned the dataset using stratified sampling to strictly maintain class balance: 80% for training (16,932 images), 10% for validation (2,116 images), and 10% for testing (2,117 images). We applied rigorous preprocessing procedures to every image to guarantee consistent input vectors. First, we converted the input format from grayscale to RGB. While CXR images lack inherent color information, this transformation is mandatory to align input channel dimensions with the DenseNet-121 architecture, which uses pretrained weights from the

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

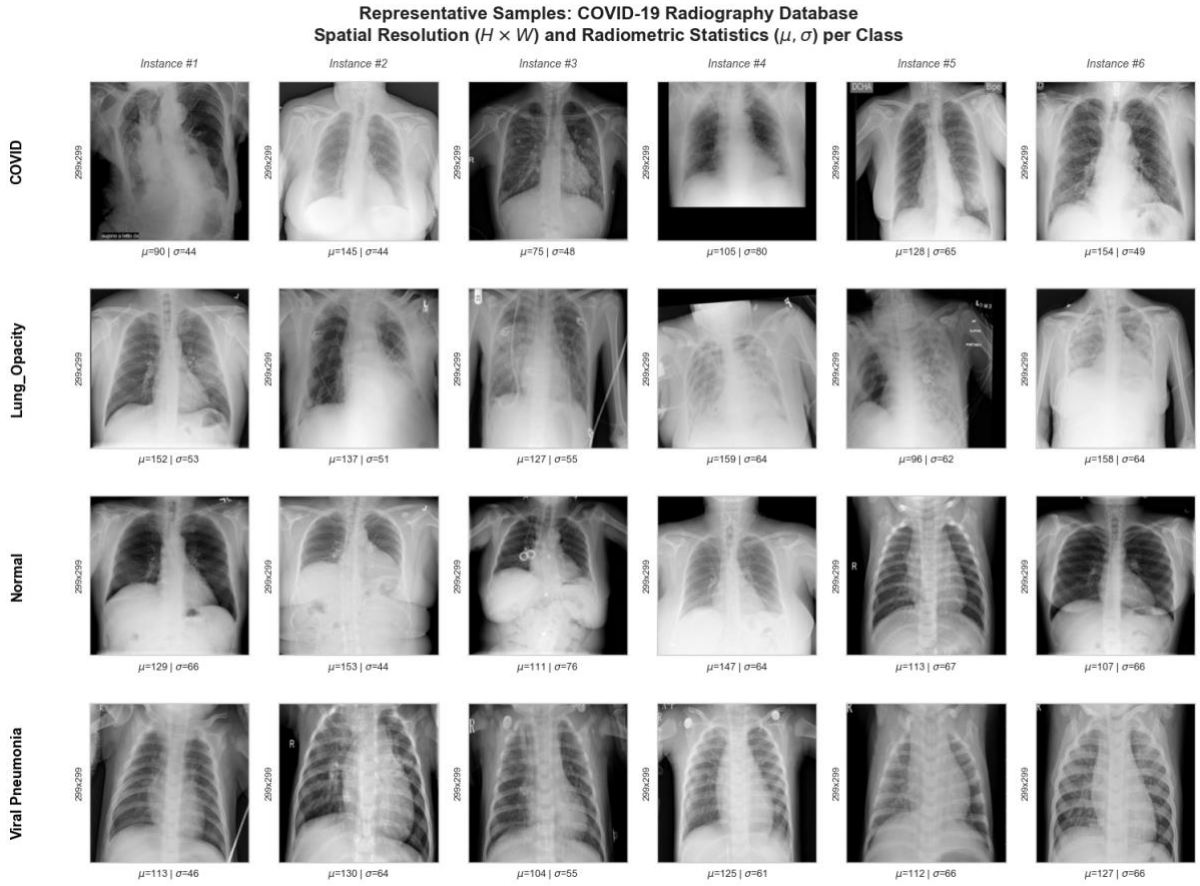


Fig. 2. Chest X-Ray Dataset COVID-19 Radiography Dataset Sample Images

ImageNet dataset [1][21]. Next, we resized all images to the dimensions of 224×224 pixels. Finally, we mathematically formulated the pixel intensity normalization. Let $x^{(c)}$ denote the original pixel intensity of the resized image for the color channel $c \in \{R, G, B\}$. The normalized pixel value $x_{norm}^{(c)}$ is calculated using Z-score standardization as shown in Eq. (1) [1][21] :

$$x_{norm}^{(c)} = \frac{x^{(c)} - \mu_c}{\sigma_c} \quad (1)$$

where μ_c and σ_c represent the ImageNet mean ($[0.485, 0.456, 0.406]$) and standard deviation ($[0.229, 0.224, 0.225]$) for each respective channel. This application of ImageNet-derived statistics normalizes the distribution of novel data and significantly expedites convergence during transfer learning. The subsequent processing flow implements conditional phases based on data subsets. Specifically, for the training data, we employed data augmentation techniques, including random horizontal flips and 10° random rotations. This enhancement augments training diversity, thereby improving model generalizability and mitigating overfitting risks [24]. We strictly excluded augmentation from the validation and testing datasets to maintain consistency and impartiality during performance evaluation.

B. The architecture of CNN

We selected DenseNet-121 as the primary feature extractor due to its dense connectivity pattern, which mitigates the vanishing gradient problem and encourages aggressive feature reuse [10][21]. Unlike traditional architectures that sum feature maps, DenseNet concatenates the output feature maps of all preceding layers. Formally, let x_l be the output of the l-th layer. The feature propagation in a dense block is defined as in Eq. (2) [10][21]:

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \quad (2)$$

where $[\ldots]$ denotes the concatenation operation of feature-maps produced in layers $0, \ldots, l-1$, and $H_l(\cdot)$ represents a composite non-linear transformation comprising Batch Normalization (BN), Rectified Linear Unit (ReLU), and Convolution (3×3). Recent advancements further support this choice, as similar architectures like DenseNet-201 and attention-based variants have demonstrated superior interpretability and robustness in multi-class classification tasks [18][25][26]. Furthermore, extensive assessments in

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

related fields confirm the efficacy of transfer learning with pre-trained architectures for optimal feature extraction [26][27].

We initialized the model with pretrained weights from ImageNet to minimize convergence time. Structurally, we modified the architecture by converting the final classifier into a linear layer mapping to a four-class output dimension. For optimization, we utilized the Adam optimizer with an initial learning rate of 10−4 and a batch size of 32. We executed training for a maximum of 50 epochs, optimizing the network parameters θ using the standard categorical Cross-Entropy loss function, formally defined as shown in Eq. (3) [7][28]:

$$J(\theta, x, y) = \sum_{i=1}^{K} y_i \log(\hat{y}_i) \quad (3)$$

where $K = 4$ represents the number of diagnostic classes, y_i is the binary indicator of the ground-truth class, and $\hat{y}_i$ is the predicted softmax probability. Crucially, in our threat model, both FGSM and PGD adversarial attacks utilize the exact gradients from this identical loss function $J(\theta, x, y)$ to construct perturbations, representing a strict white-box attack scenario rather than relying on a surrogate model approximation. To prevent overfitting, we implemented an early stopping mechanism with a patience parameter of 5.

C. Adversarial Attack Techniques

This section outlines the mathematical formulations and implementation methodologies for the gradient-based attacks used to assess model resilience. We executed adversarial attacks, specifically the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), on the CNN model post-training. This step evaluates resilience against gradient-based perturbations [2][7]. Such assessments are critical. Recent research on pretrained models indicates that ensemble attacks and region-guided attacks often inflict more damage than singular attack vectors [28][29]. Moreover, the threat landscape is evolving; physical attacks on smart systems and vulnerabilities in foundation models highlight the need for rigorous testing [29][30].

1. Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) is a "one-shot" attack algorithm grounded in the linearity hypothesis of high-dimensional neural networks. As formulated by Kansal et al. [7] and analyzed by Rahman et al. [28], FGSM approximates the optimal adversarial perturbation by assuming the model's decision boundary remains linear in the local vicinity of the input.

We compute the adversarial example x_{adv} by taking a single step along the sign of the loss function's gradient $\nabla_x J(\theta, x, y)$ with respect to the input image x [7][28]. Eq. (4) [7][28] defines this calculation:

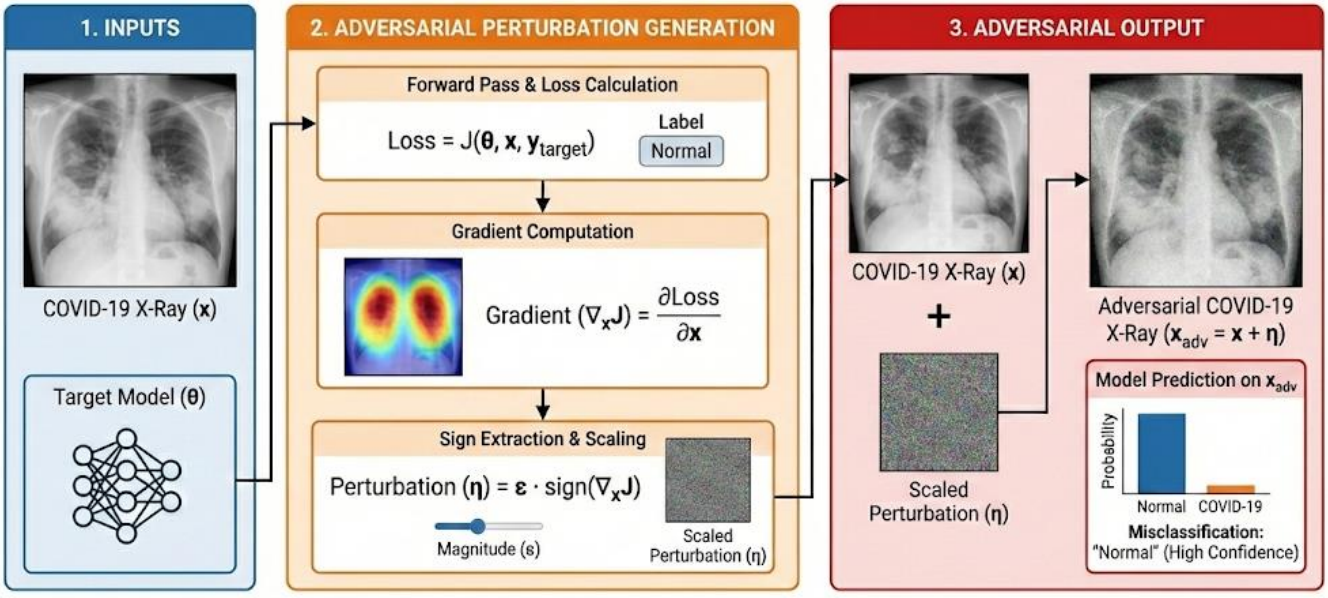


Fig. 3. Illustration of Fast Gradient Sign Method (FGSM)

$$x_{adv} = x + \epsilon \cdot sign(\nabla_x J(\theta, x, y)) \quad (4)$$

Here, y denotes the ground-truth label, θ represents the model parameters, $J(\theta, x, y)$ is the classification loss function (e.g., cross-entropy), and ϵ is a hyperparameter constraining the maximum perturbation strength. To systematically evaluate the model's robustness, we applied FGSM across a spectrum of sub-pixel perturbation bounds $\epsilon \in \{0.01/$

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

$255,0.025/255,0.05/255,0.1/255,0.3/255\}$. Due to its single-step nature, FGSM generates perturbations that correlate highly with the raw gradient of the image structure. This creates noise that does not distribute randomly. Instead, it concentrates on high-frequency spatial features, such as object boundaries. Fig. 3 visually exemplifies this characteristic. The perturbation manifests as structured noise clinging to anatomical edges, such as rib cages and clavicles. These "edge artifacts" distort the structural integrity of the Region of Interest (ROI) by exploiting the model's reliance on shape-based features [6][18].

2. Projected Gradient Descent (PGD)

To overcome the limitations of single-step approximations, we employed Projected Gradient Descent (PGD). This method is widely recognized as the "universal" first-order adversary [2]. Fig. 4 illustrates this phenomenon, visualizing the "salt-and-pepper" style noise characteristic of iterative optimization. PGD formulates the attack as an iterative optimization problem constrained within an ϵ-ball. Unlike FGSM, PGD iteratively ascends the loss landscape. It recalculates the gradient at each step to navigate non-linear decision boundaries effectively [23]. Eq. (5) [2][23] defines the update rule for the t-th iteration:

$$x^{t+1} = \Pi_{x+S}(x^t + \alpha \cdot sign(\nabla_x J(\theta, x^t, y))) \quad (5)$$

Here, x^t is the adversarial image at iteration t, α is the step size, $J(\theta, x^t, y)$ is the loss function, and Π_{x+S} denotes the projection operator. This operator clips the perturbed image back into the feasible set S (the ϵ-neighborhood of x) to ensure the perturbation strictly remains within the defined constraints.

As shown in Fig. 4, the iterative refinement of PGD results in a fundamentally different perturbation profile compared to FGSM. This scattered perturbation disrupts the semantic coherence of the image globally rather than locally, allowing the attack to exploit inherent vulnerabilities in the model's feature extraction mechanism that simple linear probes cannot reach [2][22]. The selection of attack hyperparameters for this iterative process is mathematically ruled by the optimization dynamics required to fully exploit the $L\infty$-constrained space. Consistent with our FGSM evaluation, we tested PGD across the sub-pixel spectrum $\epsilon \in \{0.01/255, 0.025/255, 0.05/255, 0.1/255, 0.3/255\}$ [7][10]. We deterministically set the step size to $\alpha = \epsilon/4$ and the maximum iterations to $N = 10$. This specific configuration provides a rigorous mathematical guarantee of spatial coverage; the total traversable distance during optimization, $N \cdot \alpha = 10 \cdot (\epsilon/4) = 2.5\epsilon$, strictly exceeds the theoretical diameter of the $L\infty$-ball (2ϵ). Consequently, from any random initialization $x^0 \in S$, the gradient ascent trajectory is mathematically guaranteed to span the entire feasible perturbation volume and securely converge to a strong local maximum within the loss landscape [2][31]. Such rigorous parameterization is mandatory to prevent sub-optimal optimization, which often leads to a false sense of robustness (gradient masking) in medical AI evaluations [15], as well as in privacy-preserving protocols [16].

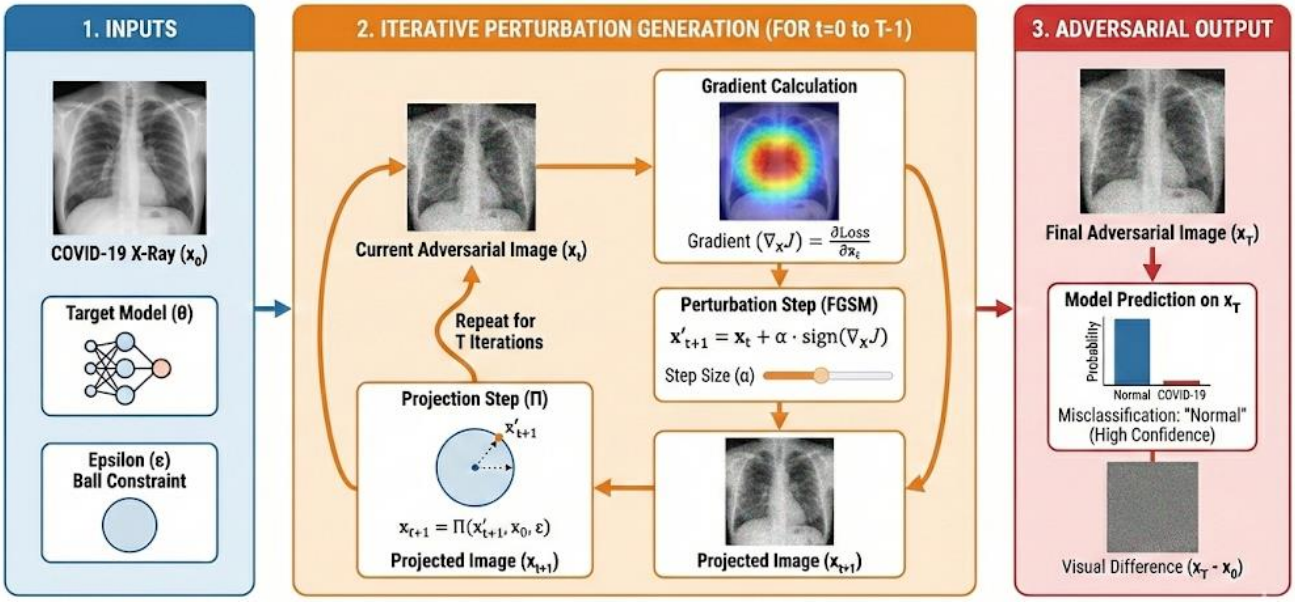


Fig. 4. Illustration of Projected Gradient Descent (PGD)

D. Model Performance Evaluation

We evaluated the model's diagnostic performance and its degradation under adversarial conditions using standard quantitative metrics: Accuracy, Precision,

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

Recall, and F1-score [7][23]. To standardize the interpretation across our multi-class framework, let TP denote True Positives, TN denote True Negatives, FP denote False Positives, and FN denote False Negatives for a given class. The metrics are formally defined as follows at Eq. (6) - (9) [7][23]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (9)$$

We applied these exact formulations to both the original and adversarially perturbed test sets to measure the specific performance weakening caused by the FGSM and PGD attacks. Such rigorous quantitative assessment is essential, as the potential for gradient-based attacks to invisibly compromise model integrity demands precise impact analysis [9][28]. We complemented this quantitative assessment with qualitative interpretability using Gradient-weighted Class Activation Mapping (Grad-CAM) [21][22]. Mathematically, Grad-CAM utilizes the gradient information flowing into the last convolutional layer of the DenseNet-121 architecture to assign importance values to each neuron for a specific class prediction. Let A^k represent the k-th feature map activation of the final convolutional layer. The neuron importance weights α_k^c for a target class c are computed by globally averaging the gradients of the predicted class score y^c with respect to the feature map pixels A_{ij}^k as follows at Eq. (10) [21][22] :

$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (10)$$

where Z is the total number of pixels in the feature map. The final spatial localization map $L_{Grad-CAM}^C$ is subsequently generated by applying a ReLU activation to the linear combination of the forward activation maps as follows at Eq. (11) [21][22]:

$$L_{Grad-CAM}^C = ReLU\left(\sum_k a_k^c A^k\right) \quad (11)$$

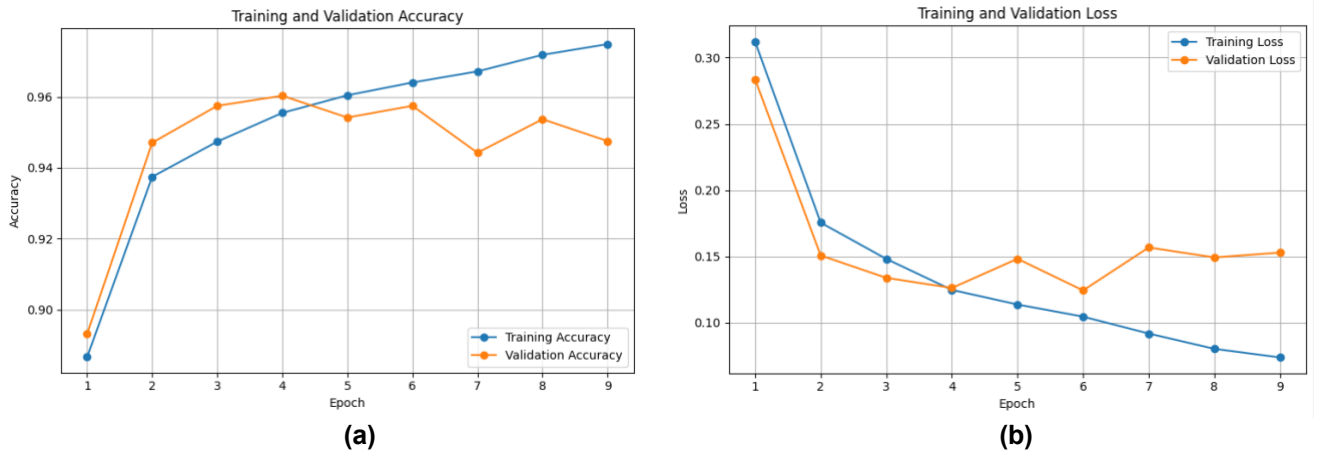
This formulation ensures that only features having a positive influence on the target class are highlighted. This method effectively identifies causal pathological lesions in complex medical tasks, including COVID-19 detection [32][33][34][35]. Recent comparisons highlight the necessity of selecting interpretability tools that are rigorously grounded in model gradients to accurately depict network behavior and build clinical trust [36].

III. Result

This section presents experimental findings. We evaluate model robustness through quantitative performance metrics and qualitative Grad-CAM interpretability maps. Table 2 summarizes the performance comparison between the Original (Clean) data, FGSM-perturbed data, and PGD-perturbed data.

A. Model Performance on Original Data

The DenseNet-121 model achieved an overall accuracy of 95.42% on the original test set. The architecture's dense connectivity allows efficient feature propagation. By reusing features from preceding layers, the network captures fine-grained local textures and high-level semantic patterns essential for identifying lung opacity and consolidation [10][21]. Training dynamics exhibited stability; as shown in Fig. 5, the loss curves converged with minimal oscillation. The close alignment between



(a) (b)

Fig. 5. Training and Validation a) Accuracy, b) Loss.

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

Table 2. Model Performance Comparison: Original vs. Adversarial Attacks ($\epsilon = 0.1/255$)

Category	Metric	Original (Clean)	FGSM Attack	PGD Attack
COVID-19	Precision	0.9914	0.4844	0.154
	Recall	0.9586	0.5138	0.1796
	F1-Score	0.9747	0.4987 (↓ 48.83%)	0.1658 (↓ 82.99%)
Lung Opacity	Precision	0.9566	0.4576	0.096
	Recall	0.9153	0.4568	0.113
	F1-Score	0.9355	0.4572 (↓ 51.13%)	0.1038 (↓ 88.90%)
Normal	Precision	0.9368	0.6989	0.4095
	Recall	0.9745	0.6742	0.3288
	F1-Score	0.9553	0.6863 (↓ 28.16%)	0.3647 (↓ 61.82%)
Viral Pneumonia	Precision	0.9847	0.7181	0.4024
	Recall	0.9627	0.7985	0.5075
	F1-Score	0.9736	0.7562 (↓ 22.33%)	0.4488 (↓ 53.91%)
Overall Accuracy	**Accuracy**	**95.42%**	**59.28%**	**25.32%**

training and validation loss indicates the model generalized to unseen data without overfitting, yielding baseline precision. The "Original (Clean)" column in Table 2 reveals that the COVID-19 and Viral Pneumonia classes achieved F1-scores exceeding 0.97. Fig. 6 displays the confusion matrix, confirming the low misclassification rate on clean data. This indicates the model recognized typical radiological patterns, such as bilateral ground-glass opacities, which are distinct markers of viral infections [1]. This aligns with prior studies confirming that Transfer Learning from ImageNet captures global features relevant to COVID-19 [1][21]. The Lung Opacity class yielded comparatively lower performance in recall, suggesting the model occasionally misclassifies non-specific opacities with other pathologies [24].

B. Model Performance on Perturbed Data

Experiments using FGSM and PGD at epsilon ϵ=0.1/255 resulted in measurable performance degradation.

1. Classification Metric Analysis

To confirm the statistical variance of the observed performance degradation under attacks, we computed the 95% Confidence Intervals (CI) for accuracy based on the exact binomial proportion limits [37][38][39]. On the original clean test set (N=2,117), the model achieved an accuracy of 95.42% (95% CI: 94.53% - 96.31%). Under the FGSM attack (ϵ=0.1/255), the global accuracy decreased to 59.28% (95% CI: 57.19% - 61.37%) [37][38][39][40]. We applied McNemar's test with a continuity correction on the paired nominal predictions (Clean vs. FGSM). The test yielded a

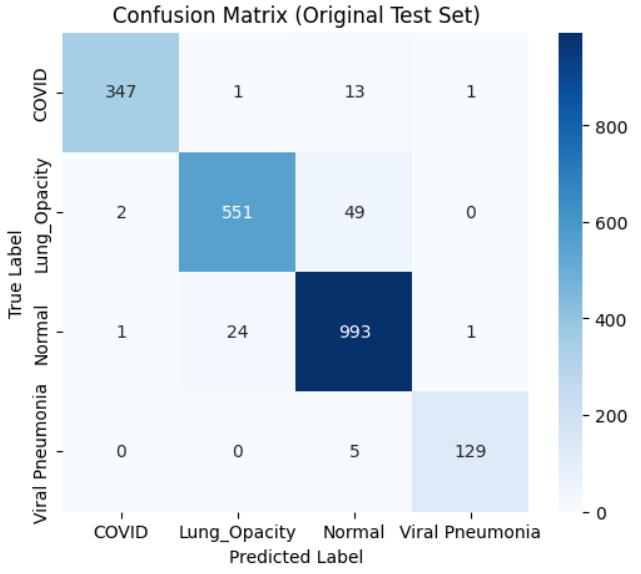


Fig. 6. Confusion Matrix on Original Dataset

statistically significant difference ($\chi 2=763.00, p \ll 0.001$) [37][38][39], proving that the structural alteration under single-step linear perturbations is deterministic. Fig. 7 visualizes this impact in the FGSM confusion matrix.

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

The iterative optimization of the PGD attack caused further degradation. As shown in Table 3, the accuracy dropped to 25.32% (95% CI: 23.42% - 27.12%).

Table 3. Attack Success Rate (ASR) Comparison ($\epsilon = 0.1/255$)

Method	Model Accuracy on Adv (%)	Attack Success Rate (ASR %)
FGSM	59.28	40.72
PGD	25.32	74.68

McNemar's test for Clean vs PGD predictions confirmed statistical significance ($\chi 2=1483.00$, $p \ll 0.001$). The absolute non-overlap between the

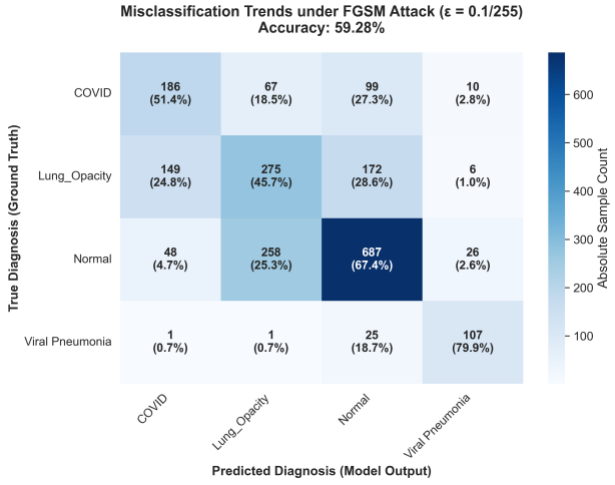


Fig. 7. Confusion matrix detailing misclassification trends under the FGSM attack (ϵ=0.1/255). The matrix illustrates the absolute sample count and class-normalized percentages, demonstrating how single-step linear perturbations predominantly compromise the prediction of Lung Opacity and COVID-19 cases.

lower bound of the clean CI (94.53%) and the upper bound of the PGD CI (27.12%) systematically proves that iterative methods efficiently exploit high-dimensional curvature within the model's decision boundary [2][7][37][38][39][40][41].

The expanded class-wise F1-score comparisons in Table 2 reveal asymmetric feature robustness [42][43]. The explicit calculation of F1-score degradation demonstrates that the Lung Opacity and COVID-19 classes experienced reductions of 88.90% and 82.99% under PGD attacks, respectively. This indicates that the latent features defining these classes (e.g., subtle ground-glass opacities) lie close to the decision boundary, making them susceptible to gradient exploitation [2][42][43]. Conversely, the Viral Pneumonia class exhibits relative resilience, experiencing a significantly lower F1-score drop (53.91% under PGD and 22.33% under FGSM) [38][42][43]. This distinct class-wise variance confirms that classes relying on diffuse, low-contrast textural

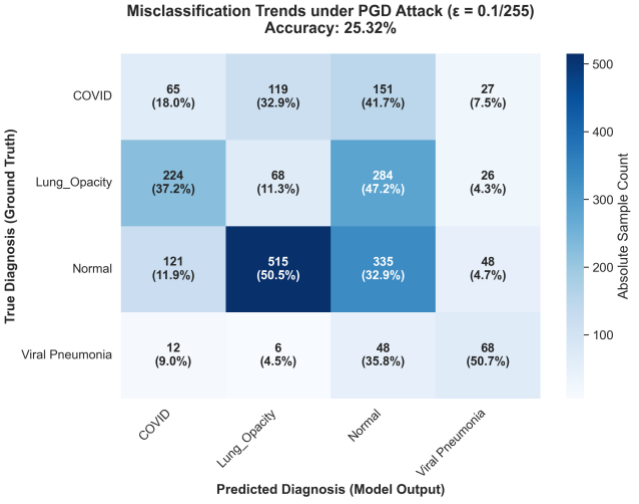


Fig. 8. Confusion matrix detailing misclassification trends under the PGD attack (ϵ=0.1/255). Each cell presents the absolute sample count alongside its class-normalized percentage, explicitly quantifying the asymmetrical vulnerability of the DenseNet-121 architecture, where the Lung Opacity class suffers near-total predictive collapse.

patterns are disproportionately compromised compared to those characterized by macro-structural abnormalities. Fig. 8 illustrates the extensive misclassification under PGD. At a sub-pixel magnitude of ϵ=0.1/255, PGD achieved an ASR of 74.68% [37][38][40]. Escalating the perturbation bound to ϵ=0.3/255 forced the ASR to 99.95%. These findings confirm that baseline high performance on clean data does not correlate with resilience against structurally optimized sub-pixel attacks [9][10][38][40][44].

2. Visual Progression and Grad-CAM Analysis

We utilized Gradient-weighted Class Activation Mapping (Grad-CAM) to interpret the quantitative metrics [22][45][46][47][48]. We formalized the measurement of spatial attention displacement using Cosine Similarity (Sc) between the flattened activation maps of the clean (M_{clean}) and perturbed (M_{adv}) images [45][47][48][49]. The similarity score is defined as shown in Eq. (12) [45][49]:

$$S_c = \frac{M_{clean} \cdot M_{adv}}{\| M_{clean} \| \| M_{adv} \|} \quad (12)$$

This mathematically quantifies the degree of semantic retention [45][47][48][49]. Empirical analysis reveals that

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

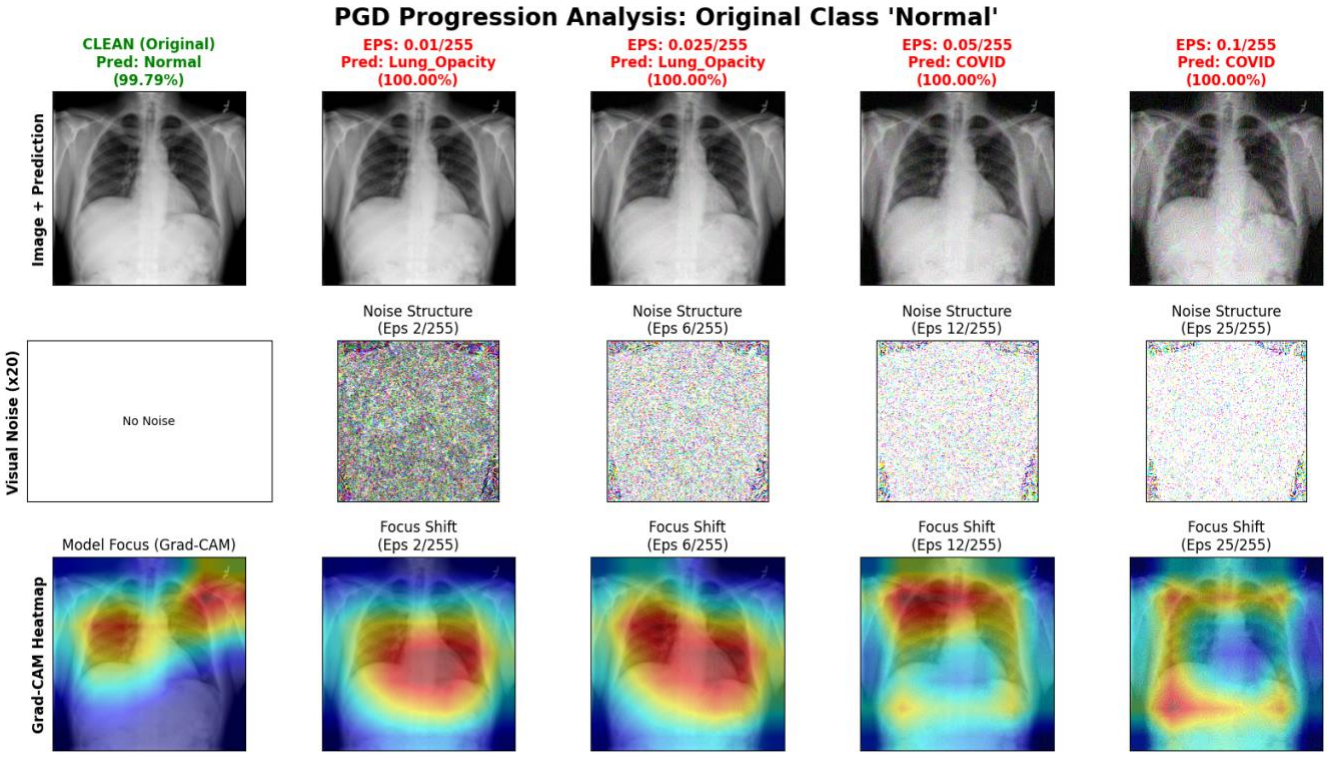


Fig. 9 (a). Visualization of PGD in Normal Class: Prediction shifts to COVID-19

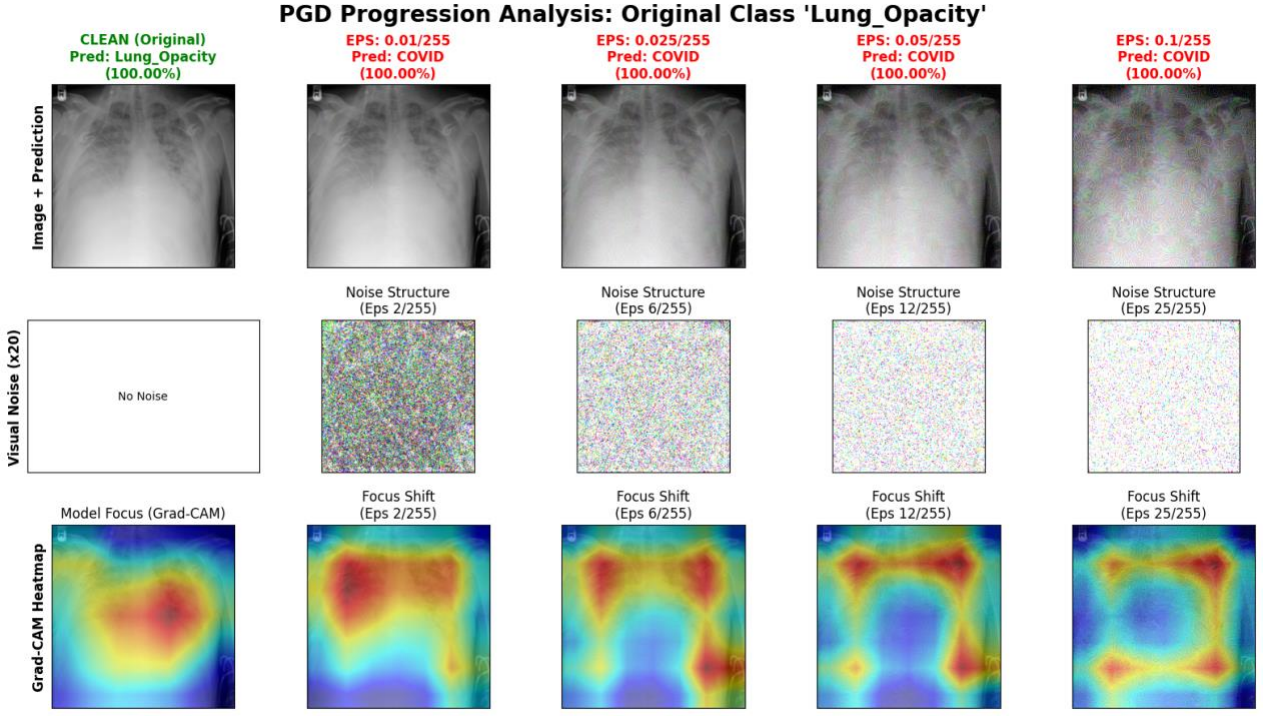


Fig. 9 (b). PGD Visualization on Lung Opacity Class: Model attention is dispersed, causing misclassification to COVID-19

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

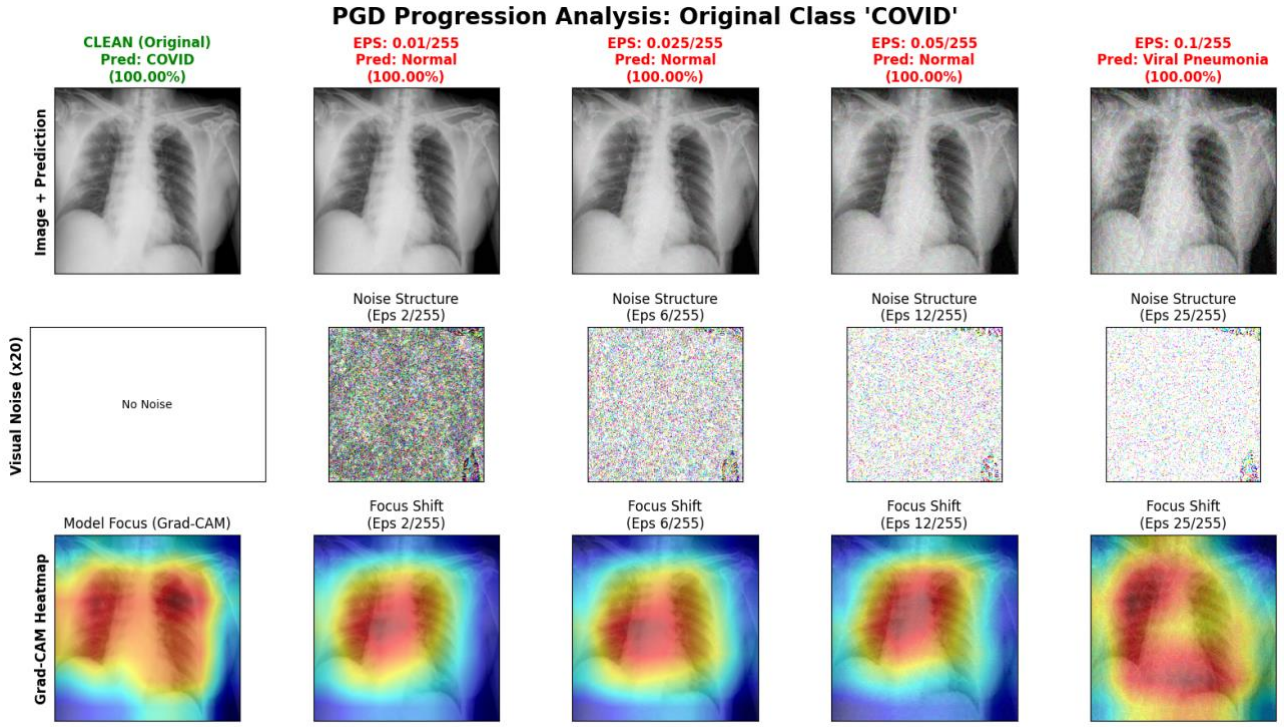


Fig. 9 (c). PGD Visualization on COVID-19 Class: Model attention is dispersed, causing misclassification to Viral Pneumonia

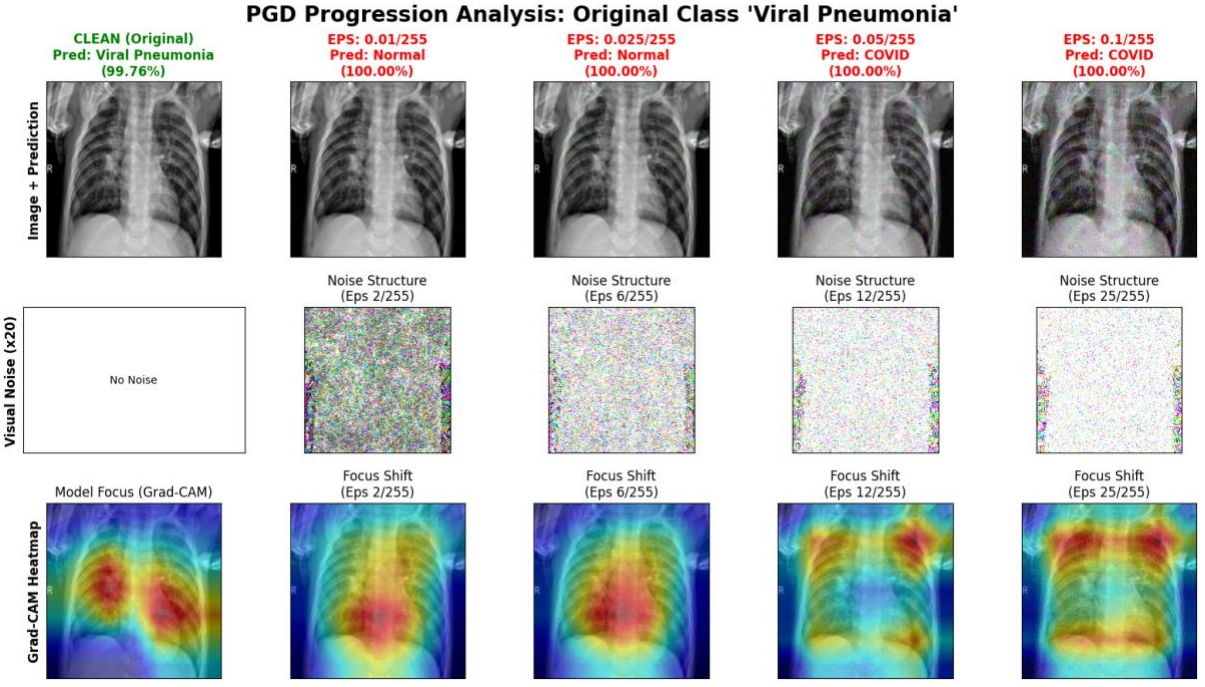


Fig. 9 (d). PGD Visualization in Viral Pneumonia Class: Prediction of turning into COVID-19 due to disturbance in pulmonary features

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

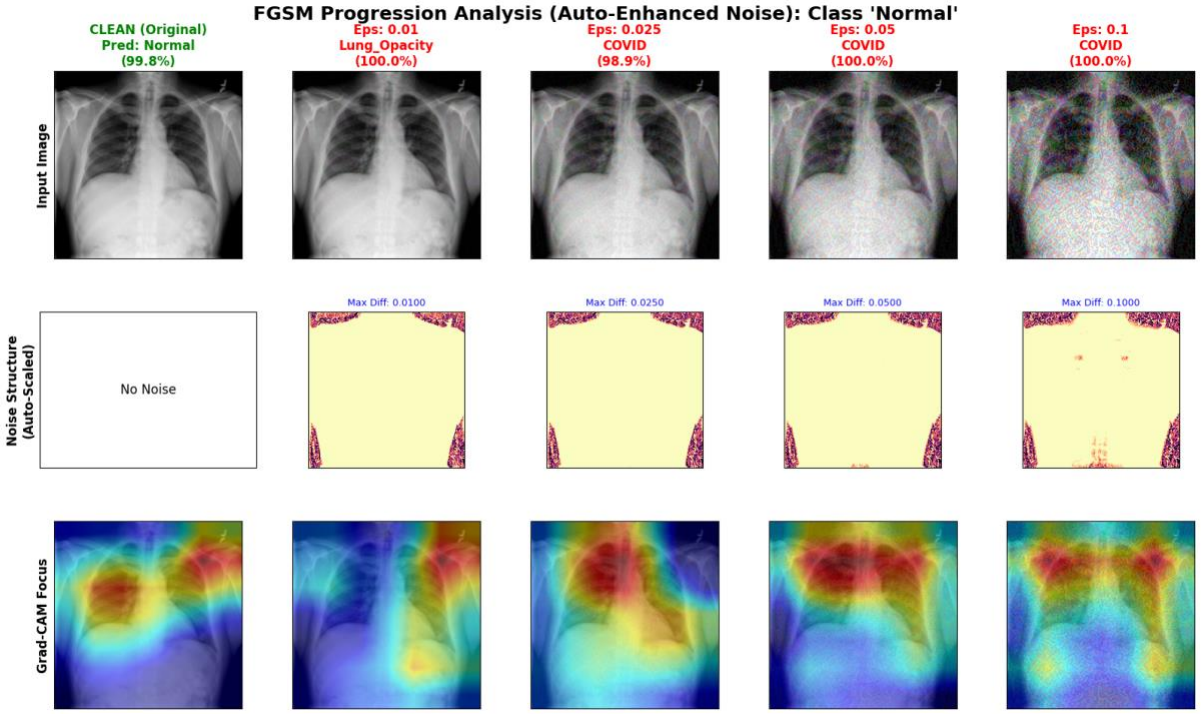


Fig. 10 (a). FGSM Visualization in Normal Class: Disturbances in organ edges trigger COVID-19 prediction

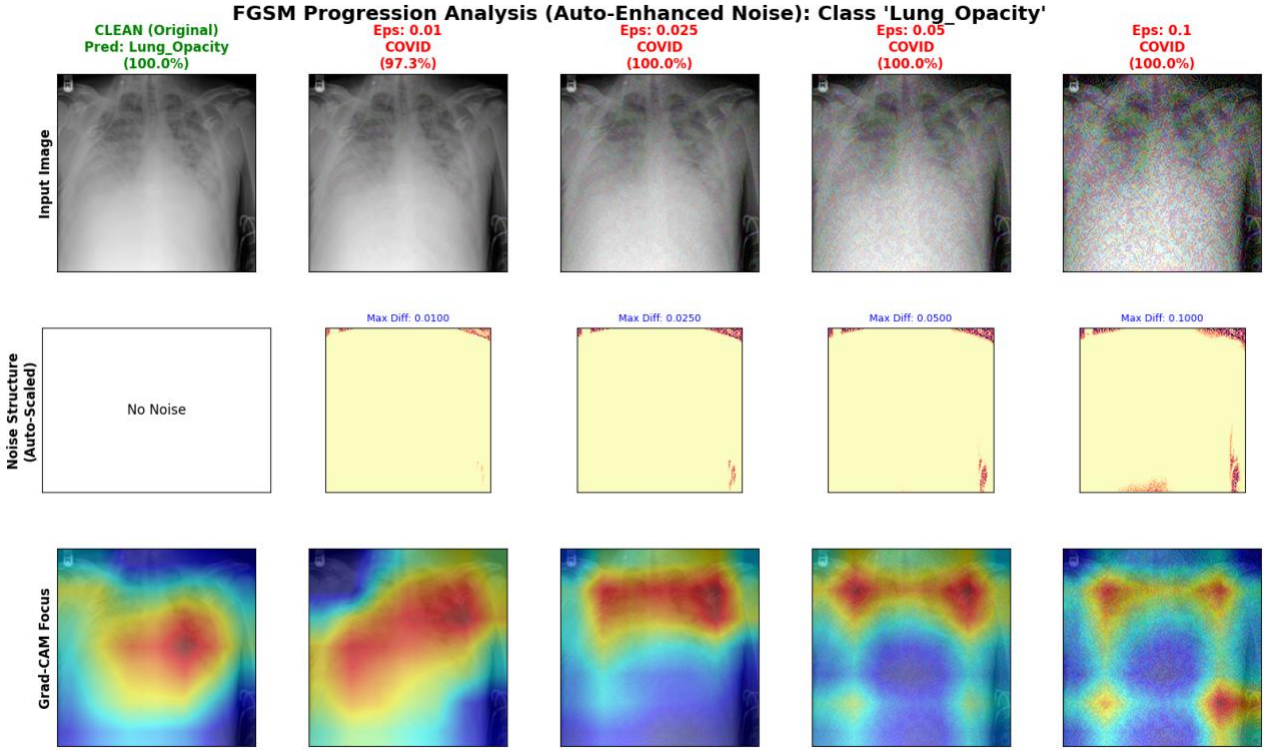


Fig. 10 (b). FGSM Visualization of Lung Opacity Class: Grad-CAM focus shifted extreme to shoulder area

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

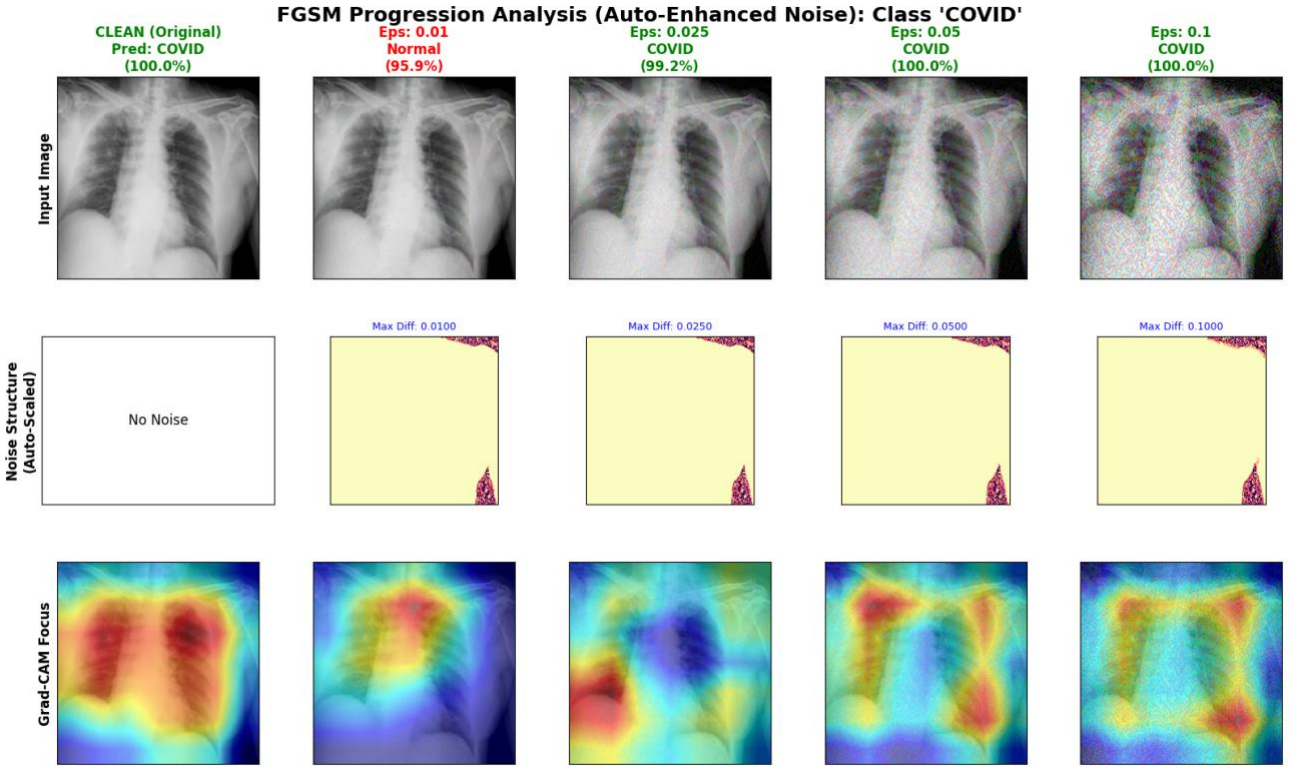


Fig. 10 (c). FGSM Visualization on COVID-19 Class: Model shows high resistance; predictions remain accurate (100%) despite visual noise

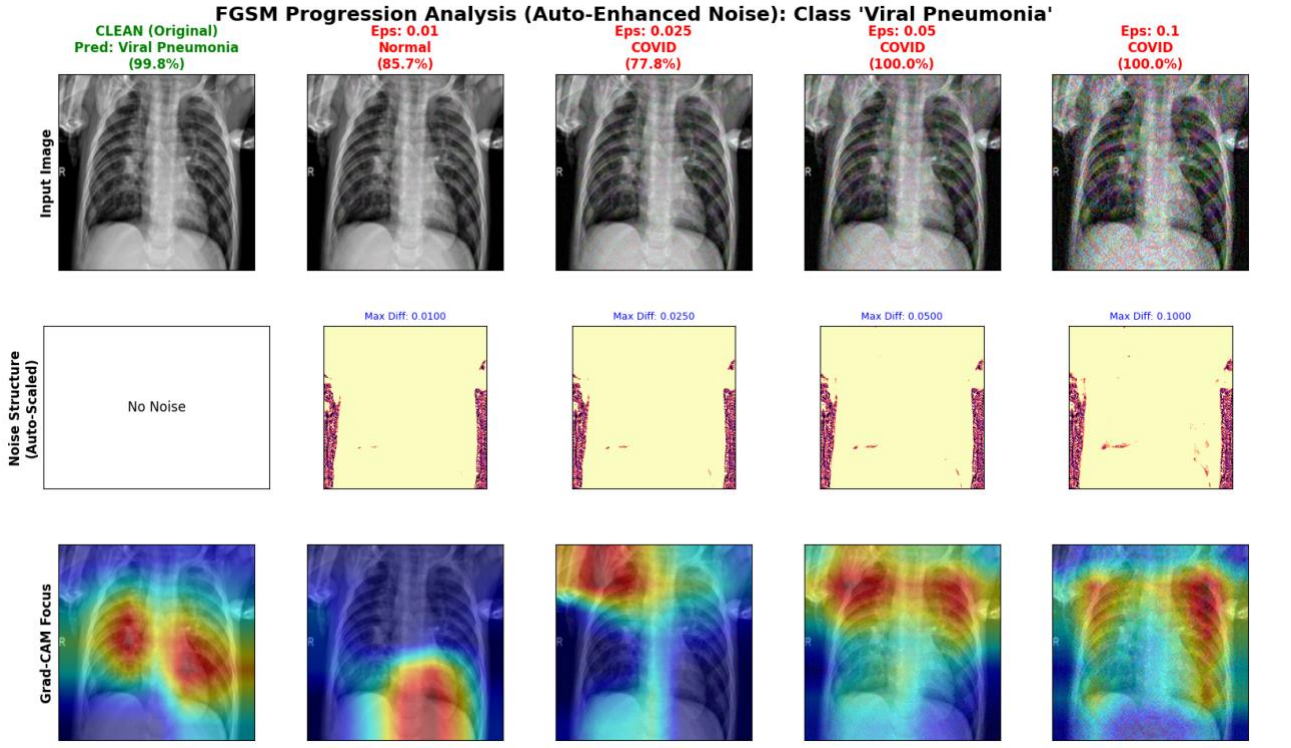


Fig. 10 (d). FGSM Visualization on Viral Pneumonia Class: Gross disturbance pattern leads to prediction change to COVID-19.

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

the PGD attack (ϵ=0.1/255) induces two distinct, class-dependent modes of topological shift. In the COVID-19 category, PGD induced semantic inversion [40][44]. This is evidenced by a high spatial correlation (Sc≈0.79), indicating that the model's spatial attention remained anchored on pulmonary structures. However, the adversarial noise synthesized localized textural features inside these boundaries, forcing the model to misclassify COVID-19 cases as Viral Pneumonia with high confidence [40][44]. Conversely, the Normal and Lung Opacity classes exhibited semantic decoupling [46][47][48][49]. The iterative optimization degraded the spatial coherence of the feature maps, resulting in a mathematical divergence in attention (Sc≪0.50). The model's attention shifted away from the pulmonary regions to background artifacts [21][22][46][47][48][49]. The heatmaps reveal concentrated activation in non-diagnostic edge zones. This explicitly links the concept of semantic decoupling to empirical metric collapse [18][40][44][45][47][49]. Fig. 9 (a-d) depicts these progressive shifts. Against FGSM, the DenseNet-121 model maintained higher prediction probabilities for the COVID-19 class at $\epsilon = 0.1/255$. However, FGSM effectively shifted predictions for Normal and Lung Opacity classes [2]. Fig. 10 (a-d) presents these results.

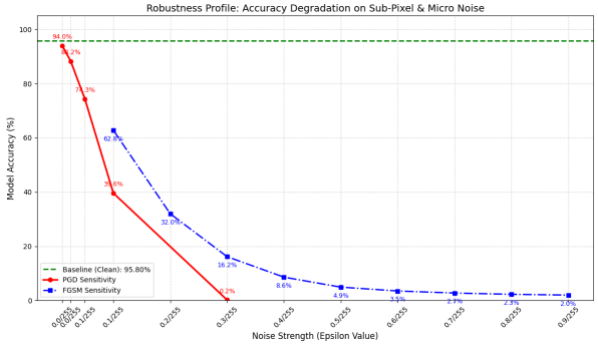


Fig. 11. Robustness Profile: Accuracy Degradation in Sub-Pixel & Micro Noise

C. Profiles between Original and Perturbed Models

1. Original vs. Perturbed Model Performance Comparison

Table 4 compares the performance metrics of the original and perturbed models. The original model achieved a baseline accuracy of 95.42%. Under the PGD attack, the accuracy decreased to 25.32%. This quantitative shift signifies that the CNN's decision boundary is mathematically susceptible to gradient-based perturbations [2][23]. Analysis of the confidence scores indicates a systemic misalignment between prediction accuracy and model certainty under adversarial conditions. In the original data, high confidence was correlated with correct predictions (96.05% average confidence). On the perturbed data, the model maintained an average confidence score of 85.42% for PGD despite an accuracy of 25.32%. In a clinical diagnostic pipeline, this confident misclassification introduces specific operational risks

Table 4. Original vs Perturbed (Epsilon 0.1/255) Comparison

Dataset Type	Accuracy (%)	Attack Success Rate (ASR %)	Avg. Confidence (%)
Original	95.42	-	96.05
FGSM	59.28	40.72	82.16
PGD	25.32	74.68	85.42

[2][7]. Adversarial False Negatives (e.g., misclassifying COVID-19 as Normal with 85% confidence) could lead to the inappropriate discharge of contagious patients. Conversely, Adversarial False Positives (e.g., misclassifying a healthy scan as COVID-19) could trigger unwarranted quarantine protocols and the misallocation of limited medical resources.

2. Resilience Profile: Sub-Pixel Sensitivity Assessment

A fine-grained sensitivity analysis was conducted using sub-pixel epsilon intervals to evaluate the models' performance threshold. Fig. 11 illustrates the degradation curves: FGSM causes a linear decline, whereas PGD triggers an immediate exponential drop.

Table 5 details the models' sensitivity to single-step linear perturbations. At $\epsilon = 0.1/255$, FGSM yields an ASR of 40.72%, after which the degradation curve exhibits a plateau, indicating that single-step attacks reach a computational limit when navigating non-linear boundaries [2]. In contrast, Table 6 presents the performance under the iterative PGD attack. At an epsilon of ϵ=0.1/255, the accuracy is 25.50%. Escalating the perturbation to ϵ=0.3/255 reduces the accuracy to 0.04%. These data empirically confirm that

Table 6. Sub-Pixel PGD Experiment

Label	Epsilon Val	Accuracy (%)	Attack Success Rate (ASR %)
0.1/255	0.000039	92.820028	7.179972
0.025/255	0.000098	86.301370	13.698630
0.05/255	0.000196	66.650921	33.349079
0.1/255	0.000392	25.507794	74.492206
0.3/255	0.001176	0.047237	99.952763

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

Table 5. Sub-Pixel FGSM Experiment

Label	Epsilon Val	Accuracy (%)	Attack Success Rate (ASR %)
0.1/255	0.000392	59.282003	40.717997
0.2/255	0.000784	32.876712	67.123288
0.3/255	0.001176	21.728862	78.271138
0.4/255	0.001569	15.162966	84.837034
0.5/255	0.001961	12.139821	87.860179
0.6/255	0.002353	10.392064	89.607936
0.7/255	0.002745	9.305621	90.694379
0.8/255	0.003137	9.258385	90.741615
0.9/255	0.003529	9.022201	90.977799

the DenseNet-121 architecture is highly susceptible to iterative optimization, with PGD exploiting manifold gaps at sub-pixel perturbation levels well below human visual perception thresholds [7].

IV. Discussion

This section interprets the findings, connecting observed vulnerabilities to safety implications and defense mechanisms. The results demonstrate that DenseNet-121 95.42% performance on clean data does not correspond to robustness under adversarial conditions. Comparing original and perturbed datasets reveals a distinct performance drop [2][6][7]. Research on both medical texts and images confirms that models degrade despite high initial accuracy [6]. Consequently, mitigating these risks requires multi-faceted defense techniques, such as feature squeezing [50]. The performance gap between PGD and FGSM highlights the non-linear nature of high-dimensional decision boundaries in deep CNNs. FGSM operates under a linearity hypothesis, taking a single step along the raw gradient, which frequently underfits the true adversarial subspace. In contrast, PGD utilizes iterative gradient optimization. By executing multiple smaller steps and recalculating the local gradient, PGD effectively navigates the high-dimensional curvature of the loss landscape, exploiting structural fragility and inducing metric decline [2][7][23].

Grad-CAM visualizations indicate that adversarial modifications disrupt internal feature representations. The model shifts its focus from clinically significant lung regions to non-diagnostic areas [21], exploiting its inability to distinguish between causal pathological features and irrelevant noise [21][22]. In robust samples, Grad-CAM reveals that the model maintained focus on pertinent lung regions, indicating attack success depends on modifying the internal feature distribution [2][21]. Similar vulnerabilities in biometrics underscore the transferability of these threats [51][52].

Error analysis mechanistically explains the asymmetric vulnerability observed across diagnostic categories, particularly the relative resilience of the Viral Pneumonia class. Under the PGD attack (ϵ=0.1/255), Viral Pneumonia maintained a higher recall (0.5075) and F1-score (0.4488) compared to COVID-19 (0.1658) and Lung Opacity (0.1038). Pathologically, COVID-19 and Lung Opacity rely on subtle, low-contrast, high-frequency textural anomalies. These localized features lie close to the decision boundary and are disrupted by $L\infty$-constrained high-frequency noise. Viral Pneumonia typically presents with macro-structural infiltrates that dominate lower spatial frequencies, which are less susceptible to restricted gradient-based attacks. This demonstrates that adversarial vulnerability is governed by the spatial frequency of the diagnostic biomarkers.

To mitigate these vulnerabilities, future defense strategies must transition from standard Empirical Risk Minimization (ERM) to robust optimization frameworks. The architecture's susceptibility advocates for the implementation of Adversarial Training (AT) [53]. Mathematically, AT reformulates the training objective as a robust min-max saddle point problem. By continuously generating strong PGD adversaries during the inner maximization loop and subsequently updating the network parameters in the outer minimization loop, the model is mathematically forced to flatten its decision boundaries and resist high-frequency spatial distortions [2][43][41][53].

Furthermore, addressing semantic decoupling suggests that standard AT must be coupled with attention-guided regularization. By penalizing the Cosine Similarity divergence between clean and adversarial activation maps during training, the network is constrained to align its feature extraction strictly with clinically relevant anatomical structures [41][45][46][47][48][49]. Grad-CAM serves as a vital bridge between quantitative performance loss and internal model mechanics [22], providing transparency in medicine [18][21]. Integrating interpretability analysis is mandatory to ensure decisions rely on appropriate clinical information [22].

V. Conclusion

This paper aimed to evaluate the adversarial robustness of a DenseNet-121 model against Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks in classifying COVID-19 chest X-ray images. The experimental results quantitatively demonstrate a critical vulnerability in the unprotected CNN architecture: while baseline accuracy on clean data reached 95.42%, the introduction of $L\infty$-

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

constrained sub-pixel perturbations with epsilon=0.1/255 significantly reduced the overall accuracy to 59.28% under single-step FGSM and 25.32% under iterative PGD attacks. Furthermore, an interpretability analysis utilizing Grad-CAM confirmed that this severe performance degradation is driven by semantic decoupling, where the model's predictive focus deterministically shifts from clinically relevant pulmonary features to spurious background noise. These findings highlight that high baseline accuracy on clinical datasets does not provide an empirical guarantee of model reliability or patient safety under adversarial conditions. Future work will focus on integrating robust optimization frameworks, specifically adversarial training paradigms and attention-guided feature regularization, to strengthen the decision boundaries of medical imaging architectures against high-dimensional gradient-based attacks.

Acknowledgment

The authors express gratitude to the Informatics Department at the Universitas Muhammadiyah Malang for the facilities and computational resources provided to support this research.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The COVID-19 Radiography Dataset used in this paper is publicly available at: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database [1]

Author Contribution

M. H. Kamil conceptualized the study, performed the experiments, and drafted the manuscript. E. P. T. Farma assisted in data preprocessing and visualization analysis. S. Basuki supervised the project, reviewed the methodology, and refined the final manuscript. All authors approved the final version.

Declarations

Ethical Approval

All datasets used are publicly available and have been utilized strictly for academic research purposes, in compliance with their respective terms of use.

Consent for Publication Participants.

Consent for publication was given by all participants

Competing Interests

The authors declare no competing interests.

References

[1] M. E. H. Chowdhury et al., “Can AI Help in Screening Viral and COVID-19 Pneumonia?,” IEEE Access, vol. 8, pp. 132665–132676, 2020, doi: 10.1109/ACCESS.2020.3010287.

[2] Y. Li and S. Liu, “The Threat of Adversarial Attack on a COVID-19 CT Image-Based Deep Learning System,” Bioengineering, vol. 10, no. 2, Feb. 2023, doi: 10.3390/bioengineering10020194.

[3] A. B. Godbin and S. G. Jasmine, “Leveraging Radiomics and Genetic Algorithms to Improve Lung Infection Diagnosis in X-Ray Images Using Machine Learning,” IEEE Access, vol. 12, pp. 47656–47671, 2024, doi: 10.1109/ACCESS.2024.3383781.

[4] A. E. Minarno, T. N. Izzah, Y. Munarko, and S. Basuki, “Classification of Malaria Using Convolutional Neural Network Method on Microscopic Image of Blood Smear,” JOIV: International Journal on Informatics Visualization, vol. 8, no. 3, 2024, doi: 10.62527/joiv.8.3.2154.

[5] R. Rajpoot, M. Gour, S. Jain, and V. B. Semwal, “Integrated ensemble CNN and explainable AI for COVID-19 diagnosis from CT scan and X-ray images,” Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-75915-y.

[6] R. Senthil, L. Ravishankar, S. D. Dunston, and V. Mary Anita Rajam, “Universal Adversarial Perturbation Attack on the Inception-Resnet-v1 model and the Effectiveness of Adversarial Retraining as a Suitable Defense Mechanism,” in 2023 International Conference on Innovative Trends in Information Technology, ICITIIT 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICITIIT57246.2023.10068722.

[7] K. Kansal, P. S. Krishna, P. B. Jain, S. R, P. Honnavalli, and S. Eswaran, “Defending against adversarial attacks on Covid-19 classifier: A denoiser-based approach,” Heliyon, vol. 8, no. 10, Oct. 2022, doi: 10.1016/j.heliyon.2022.e11209.

[8] N. Dietrich, B. Gong, and M. N. Patlas, “Adversarial artificial intelligence in radiology: Attacks, defenses, and future considerations,” Nov. 01, 2025, Elsevier Masson s.r.l. doi: 10.1016/j.diii.2025.05.006.

[9] S. Brohi and Q. U. A. Mastoi, “From Accuracy to Vulnerability: Quantifying the Impact of Adversarial Perturbations on Healthcare AI Models,” Big Data and Cognitive Computing, vol. 9, no. 5, May 2025, doi: 10.3390/bdcc9050114.

[10] W. Tajak, K. Nurzyńska, and A. Piórkowski,

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

"Vulnerability to One-Pixel Attacksof Neural Network Architecturesin Medical Image Classification," Bio-Algorithms and Med-Systems, vol. 21, no. 1, pp. 58–70, Oct. 2025, doi: 10.5604/01.3001.0055.3261.

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

[12] J. Malik, R. Muthalagu, and P. M. Pawar, "A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies," IEEE Access, vol. 12, pp. 99382–99421, 2024, doi: 10.1109/ACCESS.2024.3423323.

[13] V. Sorin, S. Soffer, B. S. Glicksberg, Y. Barash, E. Konen, and E. Klang, "Adversarial attacks in radiology – A systematic review," Oct. 01, 2023, Elsevier Ireland Ltd. doi: 10.1016/j.ejrad.2023.111085.

[14] M. A. S. Maulana, S. Basuki, and A. A. Wardhana, "Exploiting vulnerabilities of machine learning models in medical text via generative adversarial attacks," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, Aug. 2025, doi: 10.22219/kinetik.v10i3.2280.

[15] W. Lee, M. Ju, Y. Sim, Y. K. Jung, T. H. Kim, and Y. Kim, "Adversarial Attacks on Medical Segmentation Model via Transformation of Feature Statistics," Applied Sciences (Switzerland), vol. 14, no. 6, Mar. 2024, doi: 10.3390/app14062576.

[16] E. Darzi, F. Dubost, N. M. Sijtsema, and P. M. A. van Ooijen, "Exploring adversarial attacks in federated learning for medical imaging," arXiv preprint arXiv:2310.06227, 2023, doi: 10.48550/arXiv.2310.06227.

[17] E. Mahamud, N. Fahad, M. Assaduzzaman, S. M. Zain, K. O. M. Goh, and M. K. Morol, "An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning," Decision Analytics Journal, vol. 12, Sep. 2024, doi: 10.1016/j.dajour.2024.100499.

[18] M. Aasem and M. Javed Iqbal, "Toward explainable AI in radiology: Ensemble-CAM for effective thoracic disease localization in chest X-ray images using weak supervised learning," Front Big Data, vol. 7, 2024, doi: 10.3389/fdata.2024.1366415.

[19] A. Askhatuly, D. Berdysheva, A. Berdyshev, A. Adamova, and D. Yedilkhan, "Adversarial Attacks and Defense Mechanisms in Machine Learning: A Structured Review of Methods, Domains, and Open Challenges," 2025, Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/ACCESS.2025.3624409.

[20] 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," Sep. 01, 2025, Elsevier B.V. doi: 10.1016/j.iswa.2025.200554.

[21] J. Zhang, H. Chao, G. Dasegowda, G. Wang, M. K. Kalra, and P. Yan, "Revisiting the Trustworthiness of Saliency Methods in Radiology AI," Radiol Artif Intell, vol. 6, no. 1, 2024, doi: 10.1148/ryai.220221.

[22] 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, Feb. 2025, doi: 10.3390/systems13020088.

[23] 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, Feb. 2024, doi: 10.1007/s10278-023-00916-8.

[24] J. Yao, Z. Guo, X. Zhang, N. Yan, Q. Wang, and W. Yu, "Cross-domain lung opacity detection via adversarial learning and box fusion," Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-82719-7.

[25] S. Alphonse, S. Abinaya, and N. Kumar, "Pain assessment from facial expression images utilizing Statistical Frei-Chen Mask (SFCM)-based features and DenseNet," Journal of Cloud Computing, vol. 13, no. 1, Dec. 2024, doi: 10.1186/s13677-024-00706-9.

[26] E. Hassan, S. A. Ghazalah, N. El-Rashidy, T. A. El-Hafeez, and M. Y. Shams, "DenseNet Model with Attention Mechanisms for Robust Date Fruit Image Classification," International Journal of Computational Intelligence Systems, vol. 18, no. 1, Dec. 2025, doi: 10.1007/s44196-025-00809-4.

[27] Z. Sari and S. Basuki, "Transfer Learning Approaches for Non-Organic Waste Classification: Experiments with MobileNet and VGG-16," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, Oct. 2025, doi: 10.22219/kinetik.v10i4.2319.

[28] M. Rahman, P. Roy, S. S. Frizell, and L. Qian, "Evaluating Pretrained Deep Learning Models for Image Classification Against Individual and Ensemble Adversarial Attacks," IEEE Access, vol. 13, pp. 35230–35242, 2025, doi:

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

10.1109/ACCESS.2025.3544107.

[29] X. Liu, F. Shen, and J. Zhao, "Region-guided attack on the segment anything model," Neural Networks, vol. 193, Jan. 2026, doi: 10.1016/j.neunet.2025.108058.

[30] A. Guesmi, M. A. Hanif, B. Ouni, and M. Shafique, "Physical Adversarial Attacks for Camera-Based Smart Systems: Current Trends, Categorization, Applications, Research Challenges, and Future Outlook," IEEE Access, vol. 11, pp. 109617–109668, 2023, doi: 10.1109/ACCESS.2023.3321118.

[31] J. Sun, H. Yu, and J. Zhao, "An Adversarial Attack via Penalty Method," IEEE Access, vol. 13, pp. 18123–18140, 2025, doi: 10.1109/ACCESS.2025.3529217.

[32] Y. Kumaran S, J. J. Jeya, R. Mahesh T, S. B. Khan, S. Alzahrani, and M. Alojail, "Explainable lung cancer classification with ensemble transfer learning of VGG16, Resnet50 and InceptionV3 using grad-cam," BMC Med Imaging, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12880-024-01345-x.

[33] M. Akçiçek, H. Bingöl, B. Petik, S. Ünlü, and M. Yıldırım, "Detection of osteochondral lesion of talus in ankle magnetic resonance images with GradCAM-based hybrid CNN model," Egyptian Journal of Radiology and Nuclear Medicine, vol. 57, no. 1, Dec. 2026, doi: 10.1186/s43055-025-01661-4.

[34] H. C. Yoon and L. P. Lin, "Brain Tumor Classification in MRI: Insights From LIME and Grad-CAM Explainable AI Techniques," IEEE Access, vol. 13, pp. 154172–154202, 2025, doi: 10.1109/ACCESS.2025.3603272.

[35] M. A. Farhad, A. Razaque, S. B. Mukhanov, D. S. M. Hassan, and H. Mohan Rai, "Enhanced Lesion Localization and Classification in Ocular Tumor Detection Using Grad-CAM and Transfer Learning," IEEE Access, vol. 13, pp. 167762–167777, 2025, doi: 10.1109/ACCESS.2025.3610183.

[36] M. Ennab and H. Mcheick, "Enhancing Pneumonia Diagnosis Through AI Interpretability: Comparative Analysis of Pixel-Level Interpretability and Grad-CAM on X-ray Imaging With VGG19," IEEE Open Journal of the Computer Society, vol. 6, pp. 1155–1165, 2025, doi: 10.1109/OJCS.2025.3582726.

[37] W. Villegas-Ch, A. Jaramillo-Alcázar, and S. Luján-Mora, "Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW," Big Data and Cognitive Computing, vol. 8, no. 1, Jan. 2024, doi: 10.3390/bdcc8010008.

[38] M. Xu, T. Zhang, Z. Li, M. Liu, and D. Zhang, "Towards Evaluating the Robustness of Deep Diagnostic Models by Adversarial Attack," Medical Image Analysis, vol. 69, 2021, doi: 10.1016/j.media.2021.101977.

[39] R. Olivier and B. Raj, "How many perturbations break this model? Evaluating robustness beyond adversarial accuracy," arXiv preprint arXiv:2207.04129, 2023, doi: 10.48550/arXiv.2207.04129.

[40] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," arXiv preprint arXiv:2003.01690, 2020, doi: 10.48550/arXiv.2003.01690.

[41] Y. Dong, Z. Deng, T. Pang, H. Su, and J. Zhu, "Adversarial distributional training for robust deep learning," arXiv preprint arXiv:2002.05999, 2020, doi: 10.48550/arXiv.2002.05999.

[42] Q. Tian, K. Kuang, K. Jiang, F. Wu, and Y. Wang, "Analysis and Applications of Class-wise Robustness in Adversarial Training," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2021, pp. 1561–1570. doi: 10.1145/3447548.3467403.

[43] L. He, Q. Ai, X. Yang, Y. Ren, Q. Wang, and Z. Xu, "Boosting adversarial robustness via self-paced adversarial training," Neural Netw., vol. 167, pp. 706–714, Oct. 2023, doi: 10.1016/j.neunet.2023.08.063.

[44] Z. Li, P. Y. Chen, and T. Y. Ho, "GREAT Score: Global Robustness Evaluation of Adversarial Perturbation using Generative Models," in Advances in Neural Information Processing Systems, Neural information processing systems foundation, 2024. doi: 10.52202/079017-1236.

[45] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, "A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision," Jul. 01, 2025, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/s25134166.

[46] M. Bhandari, T. B. Shahi, B. Siku, and A. Neupane, "Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI," Comput Biol Med, vol. 150, Nov. 2022, doi: 10.1016/j.compbiomed.2022.106156.

[47] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of Explainable AI Techniques in Healthcare," Jan. 01, 2023, MDPI. doi: 10.3390/s23020634.

[48] M. H. Ashraf et al., "HIRD-Net: An Explainable CNN-Based Framework with Attention

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506

Mechanism for Diabetic Retinopathy Diagnosis Using CLAHE-D-DoG Enhanced Fundus Images," Life, vol. 15, no. 9, Sep. 2025, doi: 10.3390/life15091411.

[49] Q. Abbas, W. Jeong, and S. W. Lee, "Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges," Sep. 01, 2025, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/healthcare13172154.

[50] A. Yinusa and M. Faezipour, "A multi-layered defense against adversarial attacks in brain tumor classification using ensemble adversarial training and feature squeezing," Sci Rep, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-00890-x.

[51] S. H. Park, S. H. Lee, M. Y. Lim, P. M. Hong, and Y. K. Lee, "A Comprehensive Risk Analysis Method for Adversarial Attacks on Biometric Authentication Systems," IEEE Access, vol. 12, pp. 116693–116710, 2024, doi: 10.1109/ACCESS.2024.3439741.

[52] R. Chen et al., "Transferable adversarial attacks on human pose estimation: A regularization and pruning framework," Inf Sci (N Y), vol. 723, Jan. 2026, doi: 10.1016/j.ins.2025.122674.

[53] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu, "On the convergence and robustness of adversarial training," arXiv preprint arXiv:2112.08304, 2022, doi: 10.48550/arXiv.2112.08304.

Author Biography

Muhammad Hisyam Kamil is an undergraduate student in the Informatics Study Program at Universitas Muhammadiyah Malang. He also serves as a Laboratory Assistant, contributing to academic support and practical instruction within the department. His research interests lie at the intersection of computer vision and machine learning, with a particular emphasis on improving the robustness of Convolutional Neural Networks (CNNs) against adversarial attacks. He is also engaged in designing and developing agentic AI systems, actively exploring the integration of multi-agent architectures with deep learning techniques for deployment in critical and serious domains. In parallel with his academic pursuits, he is currently participating in the Apple Developer Academy @ UC, Jakarta, where he is further strengthening his expertise in software engineering, human-centered design, and scalable application development within a collaborative, industry-oriented environment.

Elga Putri Tri Farma is currently pursuing a bachelor's degree in the Informatics Study Program at Universitas Muhammadiyah Malang, Indonesia, where she also serves as a Laboratory Assistant. In this role, she supports practical sessions, assists students in understanding technical concepts, and helps maintain a productive learning environment. Her research interests include Data Science and Frontend Engineering, with a strong focus on combining analytical thinking with intuitive visual design. She is actively involved in developing data-driven web applications, aiming to transform complex computational models into responsive, user-friendly interfaces. She also continues to improve her skills in modern web technologies and data processing, while exploring how effective design can enhance user experience and decision-making in real-world applications.

Setio Basuki received the bachelor's degree from STT Telkom (Telkom University) in 2007 and the master's degree from Institut Teknologi Bandung (ITB), Indonesia, in 2015. He obtained the Ph.D. degree in Computer Science and Engineering from Toyohashi University of Technology, Japan, where his doctoral research focused on artificial intelligence, particularly the automation of scientific peer-review processes using large-scale textual data. His work involved extensive computational experiments on a corpus comprising more than 90,000 scholarly articles and millions of sentences, leveraging high-performance computing infrastructure. Since 2009, he has been with the Informatics Study Program, Universitas Muhammadiyah Malang, Indonesia, where he serves as a faculty member and contributes to teaching, research, and academic development. His research interests include machine learning, natural language processing, and data-driven intelligent systems, with an emphasis on scalable models for evaluating and improving the quality of scientific publications.

Manuscript received 27 January 2026; Revised 10 March 2026; Accepted 16 April 2026; Available online 26 April 2026
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v8i2.1506