Received: March 20, 2025. Revised: April 26, 2025.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Identify Vulnerability of Adversarial Attack on Chest X-Ray Image Using Hybrid Refinement - Generative Adversarial Network with Convolutional Block Attention Module

Amudha Gopalakrishnan[1]* **Nalini Joseph[1]** **Umarani Srikanth[2]**

[1]*Department of Computer Science and Engineering, Bharath Institute of Science and Technology, Chennai, India*
[2]*Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India*
* Corresponding author's Email: amudhag.cse@gmail.com

Abstract: Chest X-ray (CXR) imaging is the most widely applied diagnostic tool in healthcare, and plays a vital role in radiological evaluations. However, machine learning techniques such as CXR analysis applied to medical imaging face significant security threats from adversarial attacks. These attacks exploit system vulnerabilities, compromising diagnostic accuracy and reliability by altering or misleading diagnostic images. The challenges in analyzing adversarial attacks effects on CXR images include struggles in distinguishing diagnostic features from unwanted noises. The proposed CXR specific Hybrid Refinement – Generative Adversarial Network with Convolutional Block Attention Module (CXR HRGAN-CBAM) efficiently detects adversarial attacks, refines affected image areas, and optimally enhances both channel and spatial features. Initially, raw data obtained from the NIH CXR dataset, after which pre-processing is carried out with Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve image quality and identify the visibility of small regions in CXR images. The proposed method achieves an accuracy of 99.12% on the NIH CXR dataset, outperforming existing methods such as Self-Attention Generative Adversarial Capsule Network optimized with Sun Flower Optimization algorithm (GACaps-SFO).

Keywords: Chest X-ray, Convolutional block attention module, Contrast limited adaptive histogram equalization, Hybrid refinement-generative adversarial network, Self-attention generative adversarial capsule network.

1. Introduction

The vulnerability of medical image analysis to adversarial attacks, particularly in Chest X-ray (CXR) imaging, is a significant concern as lung diseases continue to be the leading cause of death worldwide [1]. Studies have extensively explored the use of CXR for diagnosis and treatment, emphasizing the critical role of pathological analysis. However, adversarial attacks introduce subtle malicious modifications to input data, posing a significant challenge and compromising model's performance [2]. While these approaches are innovative in advancing medical imaging techniques, they also introduce complex challenges in diagnosis. Deep Learning (DL) methods, mostly employed in medical imaging, are inherently vulnerable to adversarial attacks. These attacks are specifically designed to deceive DL models, undermining their reliability and accuracy in crucial applications [3-4]. DL approaches are hence modelled using realistic chest X-ray datasets of good quality, so as to accelerate methodological advancements in medicine [5]. During adversarial training, Generative Adversarial Network (GAN) is considered one of the most efficient defence techniques to analyse medical images affected by generative adversarial attacks [6]. Challenges in chest X-ray analysis primarily arise due to poor image quality and unwanted noise, minimizing diagnostic accuracy. In order to address these issues, pre-processing is used to enhance image quality before directly handling adversarial attacks [7].

The GANs have significant image-generation capabilities and are mostly applied in medical image-specific tasks with augment detector training sets for

medical image processing [8]. After investigation, the model is trained to utilise adversarial neighbors by incorporating structured signals alongside feature inputs. This signal structure is implicitly generated through adversarial perturbations, involving small data to design modifications [9-10]. However, due to the unavailability of quality data, manipulation techniques and augmentations are employed to expand the sample size [11]. CXR images leverage the GAN model due to its ability to capture more distinctive features from the image [12]. Two types of adversarial attacks, Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), are selected to demonstrate strong and simultaneous attack scenarios [13-14]. Traditional techniques, in contrast to existing approaches, do not efficiently analyse adversarial attacks, leaving medical images vulnerable to potential threats [15]. The key challenge in analysing adversarial attacks in CXR images lies in accurately identifying diagnostic features while mitigating the introduction of unwanted noise, which further degrades image quality. This research proposes CXR-specific Hybrid Refinement- Generative Adversarial Network with Convolutional Block Attention Module (CXR HRGAN-CBAM) to efficiently identify adversarial attacks, refine distinctive image regions, and ensure channel and spatial features are optimally improved. The key contributions of this research are as trails:

- Pre-processing with contrast enhancement improves image quality, while histogram equalization focuses on enhancing the visibility of smaller regions of CXR images, rather than the entire image.
- The proposed CXR HRGAN-CBAM efficiently identifies adversarial attacks, refines the distinctive image regions, and ensures channel and spatial features are optimally improved.
- The hybrid mechanism effectively reconstructs regions affected by adversarial attacks through two refined stages that efficiently segment the affected portions and integrate an attention mechanism to focus on the critical areas of CXR using channel and spatial attention.

The research paper is further organized as follows: Section 2 presents a literature survey of existing methods introduced for CXR image classification, and Section 3 offers a detailed explanation of the proposed methodology. Section 4 determines the experimental results, and Section 5 summarizes the overall conclusion of the paper.

2. Literature survey

Kumar [16] presented a lung disease detection method using Self-Attention Generative Adversarial Capsule Network Optimized with Sun Flower Optimization algorithm (GACaps-SFO) using the NIH CXR dataset. The model captured long-range dependencies and relationships between diverse image regions and enhanced its ability to distinguish between similar CXR images. However, the SFO was trapped into the local optima, rendering CXR image classification challenging.

Iqbal [17] suggested a VDV model designed for complex models of data level ensembles to using three methods of Convolutional Neural Network (CNN) namely, VGG 16, VGG 19 and DenseNet 121 using the CXR dataset. These models were combined to extract features efficiently, reduce overfitting to ensure image quality, and identify non-uniform resolutions of the input image. However, the model struggled to analyse similar tumor regions in CXR images, and lacked effective fine-tuning, resulting in reduced performance and classification accuracy.

Annamalai [18] introduced a CNN model with Auction-based Optimization Algorithm (ABOA) and Depthwise Separable Convolution (DSC) process. The CNN ignored a main feature from the X-ray image while employing an extraction procedure by using the CXR dataset. The CNN automatically learned hierarchical features, minimizing the need for manual feature extraction, while the ABOA selected the relevant features during training, reducing redundancy and focusing on the meaningful patterns in X-ray images. However, this CNN method failed to model dominant features under adversarial influence, impacting overall quality. The ABOA faced difficulties in selecting optimal features, often getting trapped into the local minima.

Ma and Lv [19] developed a Swin Transformer that used a hierarchical feature that captured both local and global features efficiently on the CXR dataset. The swin transformer was applied to a variety of vision tasks and pre-trained weights on large data, further boosting performance in downstream tasks. Nonetheless, the model effectively handled adversarial attacks, limiting its robustness to CXR image classification.

Qi [20] presented a Supervised algorithm using Multi-Instance Learning (MIL) and Class Activation Maps (CAM) to capture attacks using Graph Regularized Embedding Network (GREN). This model exploited intra and inter-image data to locate diseases. This approach captured spatial relationships within an image, reduced noise and ensured robust disease localization while defending against

adversarial attacks in CXR images. Nonetheless, the model faced challenges with GREN construction and regularization which added overhead to training and inferences, leading to the processing of attack images as well, reducing the model's overall efficacy.

Samarla and Maragathavalli [21] designed the CNN approach using three maximum pooling layers (3M-CNN) along with early fusion for the categorization of lung abnormalities. The ensemble model combined the results of the trained sub-models, offering a robust method for the sub-categorization of lung abnormalities. However, the greater reliance on pooling resulted in a loss of complex spatial details in the lung regions. Additionally, the ensemble model increased computational complexity and inference time.

Upadhya [22] introduced ResNet50 for image classification and BERT for sequence analysis. Their approach explored binary multi-modal architectures, including both early and late fusion. The early fusion approach combined image and text data at the input level, performing cross-attention to deepen relational understanding. Meanwhile, the late fusion approach synergized extracted text and image features to improve performance. However, early fusion was complex due to misaligned or imbalanced image-text inputs, while late fusion added model complexity and resulted in overfitting on limited datasets.

From the overall analysis, it can be noted that the existing models face challenges with the analysis of adversarial attacks in CXR images, and struggle to identify and eliminate diagnostic and unwanted noise in CXR images. The proposed CXR HRGAN-CBAM efficiently identifies adversarial attacks, refines diverse image regions, and ensures the channel and spatial features are optimally improved.

3. Proposed methodology

The proposed CXR HRGAN-CBAM efficiently identifies adversarial attacks, refines diverse image regions, and ensures the channel and spatial features are optimally improved. Initially, data is obtained from the NIH-CXR dataset, while pre-processing with contrast enhancement is deployed to improve image quality, and histogram equalization focuses on improving the visibility of smaller and relevant regions, rather than the entire CXR image. The hybrid mechanism effectively reconstructs regions affected by adversarial attacks using two stages of refinement, efficiently segmenting the affected portions. These are then integrated with an attention mechanism focused on the critical areas of the CXR, applying both channel and spatial attention mechanisms.

3.1 Data collection

3.1.1. NIX CXR dataset

The NIH chest X-ray (NIH-CXR) [23] dataset consists of 14 diseases categories, with a total of 91,324 frontal view CXR images of 32,717 patients. It includes X-rays of patients with and without diseases. The NIH dataset is analyzed with samples divided for training, testing, and validation purposes in the ratios of 70%, 15%, and 15%, respectively. A sample of the CXR dataset is represented in Table 1.

3.1.2. CheXpert dataset

The CheXpert dataset consists of 224,316 chest X-ray images from 65,240 patients [24]. Each image is annotated with labels indicating the presence of 14 pathologies, categorized as positive, negative, or indeterminate.

Table 1. Sample CRX dataset

Diseases	No. of Cases	Training Sample	Testing Sample	Validation Sample
Infiltration	9,547	6,683	1,909	955
Atelectasis	4,215	2,950	843	422
Nodule	2,705	1,893	541	271
Pneumothorax	2,194	1,497	427	215
Infiltration	9,547	6,683	1,909	955
Mass	2,139	1,497	427	215
Hernia	105	73	21	10
Consolidation	272	190	54	28
Pneumonia	140	98	28	14
Emphysema	142	99	29	14
Edema	241	169	48	24
Fibrosis	138	99	28	13
Pleural Thickening	165	115	33	16
Cardiomegaly	1010	707	202	101

3.1.3. MIMIC-CXR dataset

The MIMIC Chest X-ray (MIMIC-CXR) dataset [25] is a substantial and publicly available medical dataset that integrates multiple data modalities making it a valuable resource for advancing research in medical imaging and natural language processing (NLP).

The design of MIMIC-CXR required handling various large data sources, including electronic health record data, chest radiographs, and natural language (free-text reports). These diverse data types were collected independently and then integrated to create a comprehensive database. The dataset comprises 377,110 chest radiograph images, each associated with corresponding radiology reports, and is derived from 227,835 radiographic studies conducted between 2011 and 2016 at Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA.

3.2 Pre-processing

After data acquisition, pre-processing using Contrast Limited Adaptive Histogram Equalization (CLAHE) efficiently handles noise in the input image and enhances image quality, particularly where the domain of interest and the background have similar contrast values [26]. CLAHE processes smaller regions within an image, known as tiles, rather than the entire image. Its ability to filter specific portions of the CXR images and eliminate unwanted noise is mathematically expressed in Eqs. [1-2].

$$CLAHE_{pre-processing} = \frac{Processed_{image}}{Original_{image}} \quad (1)$$

$$contrast = \frac{gray\ leve_{values\ of\ image}}{backgroud_{values\ of\ image}} \quad (2)$$

Where, $Processed_{image}$ & $Original_{image}$ denote the contrast values of the image before and after pre-processing, used to remove noise from CXR images. The contrast enhancement improves image quality and removes unwanted noise. The pre-processed image is then used to identify vulnerabilities to attacks and is further processed by the improved GAN method.

3.3 Generative adversarial Attack

In this section, generative adversarial attacks leverage GANs to craft adversarial data instances specifically designed to deceive neural network models. Fast Gradient Sign Method (FGSM) attacks are a type of white-box attack where the attacker has knowledge of the model parameters. Perturbations are added to the input image in the FGSM attack, which impacts the system and degrades its ability to analyze malicious attempts in chest X-ray images. The evaluated CXR image is then modified by adding a small multiple of the sign of the gradient to the input image. The portion of the attack within the neural network is represented in Eq. (3).

$$\begin{gathered} Med_{adv}=Med_{raw}+\epsilon\times loss_{fun} \\ loss_{fun}=sign\left(\nabla^{k}_{Med_{raw}}(\sigma,Med_{raw},Med_{tar})\right) \end{gathered} \quad (3)$$

Where, $loss_{fun}$ denotes the loss function gradient k in terms of Med_{raw}, and sign refers to the sign function. The adversarial image is denoted as Med_{adv}, representing the result of the FGSM attack, while σ denotes the model parameters, and ϵ denotes the perturbation factor.

A Projected Gradient Descent (PGD) attack is similar to an FGSM attack in the context of white-box attacks. However, PGD processes iteratively, using the gradient of the most recent outcomes and assuming non-linearity in the model, as expressed in Eq. (4).

$$\begin{gathered} Med_{adv}(I+1)=clip_{(-\epsilon,\epsilon)}\left(I\left(Med_{raw}+\gamma\times loss_{fun}\right)\right) \\ loss_{fun}=sign\left(\nabla^{k}_{Med_{raw}}(\sigma,Med_{raw},Med_{tar})\right) \end{gathered} \quad (4)$$

Here, the pixels with a perturbation size larger than ϵ are clipped to the parameter ϵ. Med_{raw} denotes the medical raw image, γ denotes the perturbation factor, $Med_{adv}(I+1)$ refers to the adversarial medical image after iteration $I+1$, and Med_{tar} refers to the target image. Both FGSM and PGD attacks are implemented and explored on CXR images using the improved GAN method. This approach signifies a patch attack on a persistent partial position, affecting either FGSM or PGD. It identifies vulnerabilities in the face of concurrent attacks and demonstrates the diversity of improved models to classify efficiently.

3.4 Generative adversarial network

In this section, CXR analysis using GAN involves two deep neural networks: the discriminator and the generator. These networks are trained against each other to optimize performance using a min-max function, as expressed by Eqs. (5) and (6).

$$\begin{gathered} \min_{G}\max_{D}\mathbb{E}_{x\sim P_{data}}(x)[logD(x)]+ \\ \mathbb{E}_{z\sim P_{z}}(z)\left[log\left(1-D\left(G(z)\right)\right)\right], \end{gathered} \quad (5)$$

$$\min_{G}\max_{D}\mathbb{E}_{x\sim P_{data}}(x)[logD(x|c)] + \mathbb{E}_{z\sim P_z}(z)\left[log\left(1-D\big(G(c)\big)\right)\right] \quad (6)$$

Where, G and D denote the generator and discriminator, x denotes the training data and z denotes the arbitrary sample for the noise vector from a predefined distribution of P_z. The Eq. (6) is updated based on the label image condition considering the discriminator and generator using GAN. The label classes are denoted as c and synthetic medical data using GAN condition and the simple condition of GAN does not handle high data dimensionality. GANs identify perturbation regions in CXR images, but an improved GAN enhances identification accuracy and efficiently segments the affected regions. The novelty of integrating HRGAN with CBAM lies in improving adversarial defense capabilities through adaptive feature refinement. Unlike conventional GAN-based approaches, which produce rough reconstructions, the proposed method utilizes CBAM to selectively focus on informative spatial and channel features, thereby enhancing the model's ability to suppress adversarial perturbations. This targeted attention mechanism allows the HRGAN-CBAM to preserve large semantic details while filtering out adversarial noise, offering superior robustness compared to existing GAN-based defenses.

3.4.1. Manipulation of GAN for CXR image

In this section, the GAN process considers the latent space as encoding highly rich semantic data.

The main goal is to manipulate the input image by altering latent codes in specific directions. The latent space is processed at a low rate, which limits the range of reconstruction performance, then focuses on additional high-rate information during the generation process. This approach is referred to as refinement method. The equation $X = G(\omega)$ denotes the original image processed by the generator G, which represents the latent space. Weight modulation is performed using θ, and by minimizing the reconstruction error, θ* is obtained. The input processing is then represented as $X = G(\omega; \theta^*)$. The refinement process encodes latent codes obtained from off-the-shelf encoders, while the weight modulation method predicts L as the image distance, optimizing the refinement process. This is expressed by Eqs. (7) and (8).

$$\theta^* = \arg\min_{\theta} L(x, G(\omega; \theta)) \quad (7)$$

$$f^* = \arg\min_{f} L(x, G(\omega; f)) \quad (8)$$

Where, x denotes the original input image, ω refers to the latent code, θ denotes the set of learnable modulation parameters, θ^* refers to the optimal set of parameters, f denotes the refinement function and f^* refers to the optimal refinement function.

Weight and feature modulation impact image manipulation in various aspects. The feature modulation mechanism stabilizes intermediate feature distributions at a specific layer, ensuring that the vector applied to a previous layer does not alter the features of a later layer. Meanwhile, the weight modulation mechanism adjusts the generator manifold, sharpening the sample point. Therefore, the highly semantic characteristics of the pre-trained generator are disrupted, which decreases its overall capability.

3.4.2. Hybrid refinement in CXR image

In this section, weight modulation achieves better results compared to simple weight modulation mechanisms, while feature modulation offers enhanced reconstruction ability. The section explores the CXR image into in-CXR and out-CXR regions, both of which are enhanced by the GAN. The term "hybrid" refers to the segmentation and identification of the CXR image using domain-specific techniques. Initially, the process involves both in-CXR and out-CXR analysis. In-CXR refers to areas with a distribution similar to the generator's output space, making them easier to invert. Out-CXR, on the other hand, misaligns with the output space, making inversion and instance identification in the medical domain more challenging. In-CXR includes regions like tumors and lungs, while out-CXR encompasses occlusions, backgrounds, and artifacts. Therefore, the hybrid refinement method segments the image into in-CXR and out-CXR regions, applying weight and feature modulation accordingly.

The image embedding module aims to embed the input image X into latent codes, which is achieved by using an off-the-shelf encoder predicts W^+, the latent space of codes. This is represented as $w = E(X)$ where E is the encoder. The binary mask for CXR segmentation prediction, which indicates the in-CXR and out-CXR areas, as expressed by Eq. (9).

$$m = S(X) \quad (9)$$

Where, $m \in \{0,1\}^{H\times W}$, segments the image into two features used for refinement. The hybrid modulation refinement applies weight modulation to

refine the in-CXR area. The reconstruction discrepancy in in-CXR is minimal, with low weight deviation, ensuring the preservation of important features. The out-CXR part undergoes feature modulation to refine spatial information and stabilize feature distribution. The segmentation outcomes m modulate the weight θ and feature f by minimizing the reconstruction error in both in-CXR and out-CXR regions, as expressed by Eq. (10).

$$\theta^*, f^* = \arg\min_{\theta} L(X, G(\omega, f, m; \theta)) \quad (10)$$

The hybrid approach ensures that the generator manifold undergoes minimal changes, effectively maintaining the enhanced CXR image. The two refinements occur simultaneously, and visualizing $G(w, f^*, \theta^*)$ and $G(w; \theta^*)$ separately demonstrates the effect of the method. Fig. 1 illustrates the overall process of the HR GAN technique.

3.4.3. CXR-specific segmentation

In this phase, the CXR-specific segmentation module is used to segment images into two domains as in-CXR and out-CXR images. The traditional segmentation approach effectively learns the CXR image, but requires large annotated data, which is costly. This challenge is overcome by the proposed method where CXR images are segmented without requiring data annotation and designing. The automatic segmentation is carried out in two phases: partition and binaries, which, combined in the model, utilize super pixel algorithm to partition the input image into multiple areas. The partition of each image considers $\{m_s^i\}_{i=1}^S$. The categorizing of every partition into in-CXR and out-CXR without annotation is a challenge. The process begins with latent codes initialized to mean values, which are then optimized over a few steps. In the in-CXR region, it is easier to invert the area outcomes X_{coarse} reconstructed in-CXR areas effectively. The perceptual loss $L(X, X_{coarse})$ between coarse reconstruction images is evaluated. The white areas of the reconstruction indicate a higher loss value, while the black areas indicate a lower loss value. It is observed that the loss values for occluded areas are significantly higher than those for the medical image. The average loss for each partition is calculated as shown in Eq. (11).

$$v_i = \frac{L \odot m_s^i}{\|m_s^i\|} \quad (11)$$

It binarizes the segments by threshold τ to attain m_s. To compensate for missed segmentation in small areas, a superpixel approach is utilized in conjunction with a parsing model to enhance accuracy. The medical image-specific segmentation outcomes are combined by parsing outcome with the superpixel outcomes: $m = m_p \times m_s$. The CXR segmentation module achieves fine segmentation without requiring data annotation and is effective for medical images.

3.4.4. Convolutional block attention module

In this phase, the segmented image is enhanced by incorporating an attention mechanism designed to improve the CBAM for CXR images. This mechanism consists a simple structure consisting of both a channel and spatial attention module. It directly generates the output through a second matrix multiplication, followed by addition and concatenation via a convolutional layer. This process enables non-linear feature fusion, improving information flow and enhancing the effectiveness of the network.

3.4.5. Hybrid modulation refinement

The enhanced image quality and segmentation of the CXR image preserve the capability of the pre-trained GAN and recover image details. This module includes two refinement aspects: weight and feature modulation. The primary goal of weight modulation is to minimize the in-CXR reconstruction error by tuning the generator parameters, while feature modulation is applied to the out-CXR area using the intermediate features. The hybrid refined image is expressed by Eqs. (12) to (15).

$$f' = fl \odot m + F \odot (1 - m) \quad (12)$$

$$L = L_2 + \lambda L_{lpips} \quad (13)$$

$$\nabla f = \frac{\partial}{\partial f}\left(\frac{L \odot (1-m)}{\|1-m\|}\right) \quad (14)$$

$$\nabla \theta = \frac{\partial}{\partial \theta}\left(\frac{L \odot m}{\|m\|}\right) \quad (15)$$

Where, f' denotes the final blended or refined feature representation, fl denotes the locally refined feature, F denotes the original or global feature representation, m denotes a soft mask indicating the region of interest, $\odot$ denotes element-wise multiplication, L denotes the total loss used for optimization, λ denotes weighting factor controlling the influence of perceptual loss relative to denotes pixel-wise loss, ∇f denotes the loss gradient based on the feature f.

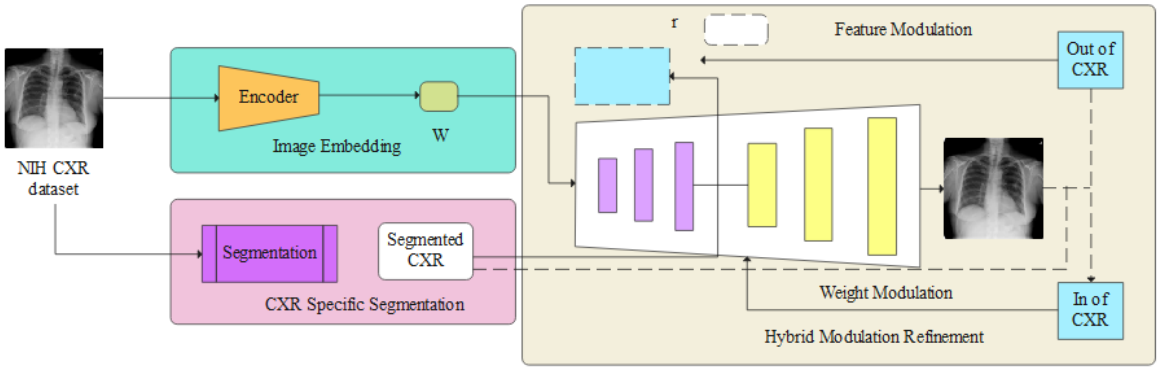


Figure. 1 Overall process of the HR GAN technique

The backward process improves the weight and feature in parallel, allowing them to focus on the corresponding medical image. The mean square error L_2 and perceptual loss L_{lpips} are utilized and the backward error loss is evaluated with the segmentation outcome m.

4. Experimental analysis

The CXR HRGAN-CBAM technique is simulated in Python version 3.8 software tool, with system specifics being 16GB RAM, intel i7 processor, Windows 10 operation system, 16GB GPU and 1TB SSD. The performance of the proposed method is validated based on several metrics: accuracy, precision, specificity, f1-score, recall and AUC, which are defined by Eqs. (16) to (20).

$$Accuracy = \frac{TP+TN}{TN+TP+FN+FP} \tag{16}$$

$$F1-Measure = \frac{2\times TP}{2\times TP+FP+FN} \tag{17}$$

$$Precision = \frac{TP}{TP+FP} \tag{18}$$

$$Recall = \frac{TP}{TP+FN} \tag{19}$$

$$Sensitivity = \frac{TP}{(TP+FN)} \tag{20}$$

Where, FP denotes false positive, TP denotes true positive, FN denotes a false negative, and TN denotes true negative.

4.1 Performance analysis

In this section, a quantitative analysis of the CXR HRGAN-CBAM model is presented, focusing on precision, sensitivity, F1-score, accuracy, and recall. These metrics are presented in Table 2, which highlights the performance of classifiers on the Chest X-ray dataset. The performance metrics for ANN, RNN, GRU, CNN, Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), DeepFool, Ensemble Adversarial Training (EAT), and Feature Squeezing are evaluated and compared with the implemented GAN model. The attained results demonstrate that the implemented GAN approach outperforms others, with performance metrics, such as F1-measure, accuracy, precision, sensitivity, and recall, showing values of 99.12%, 98.53%, 98.25%, 98.52%, 98.42%, and 96.36%, respectively, when compared to the other classifiers. Table 3 provides the performance analysis across various classes evaluated for the proposed method.

K-fold Validation partitions the dataset into k subsets to estimate predictive performance. This approach is trained and tested k times, with each fold serving as a validation set. Performance metrics from each fold are collected to estimate the model's generalization performance. In this study, the dataset is divided into five folds, and the training and testing processes are performed. k=5 yields the highest model performance. Table 4 illustrates the performance evaluation of the Chest X-ray dataset using k-fold validation.

Table 5 demonstrates the statistical analysis and computational complexity of the proposed method compared to existing methods across three datasets. Lower p-values correspond to higher t-test values, indicating stronger statistical significance for the model's performance. The proposed HRGAN-CBAM consistently shows strong results across datasets, with t-values indicating significant performance improvements. Metrics such as computational cost and inference time show minimal results compared to the other existing methods. Fig. 2 demonstrates the confusion matrices of the proposed CXR HRGAN-CBAM method on three datasets: NIH CXR, CheXpert, and MIMIC CXR.

Table 2. Various classifier performances for the Chest X-ray dataset

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Measure (%)	Sensitivity (%)	AUC (%)
ANN	95.23	94.36	94.78	94.16	94.32	92.12
CNN	96.85	95.82	95.43	95.86	95.98	93.03
RNN	97.45	96.45	96.15	96.74	96.54	94.84
GRU	98.23	97.15	97.03	97.63	97.32	95.13
FGSM	97.46	95.37	94.54	93.12	90.12	94.23
PGD	93.48	97.12	90.32	95.39	92.43	95.64
DeepFool	95.39	94.39	92.43	94.24	94.39	94.32
EAT	94.92	96.33	94.33	95.98	93.42	92.47
Feature Squeezing	97.48	97.45	96.56	97.67	95.58	95.46
HRGAN-CBAM	99.12	98.53	98.25	98.52	98.42	96.36

Table 3. Performance analysis of the various classes evaluated for proposed method

Thoracic Diseases	Accuracy (%)	Precision (%)	Recall (%)	F1-Measure (%)	Sensitivity (%)	AUC (%)
Ateleciasis	95.59	93.85	95.63	93.09	95.89	95.63
Mass	96.96	94.23	96.36	94.23	96.18	96.36
Consolidation	96.23	96.45	96.85	96.23	96.45	96.85
Pneumonia	98.68	96.52	98.36	96.15	98.13	98.36
Hervina	97.85	95.14	97.86	95.75	97.83	97.86
Cardiomegaly	95.59	93.85	95.63	93.09	95.89	95.63
Nodule	96.96	94.23	96.36	94.23	96.18	96.36
Edema	96.23	96.45	96.85	96.23	96.45	96.85
Pleural Thickening	94.23	94.96	94.32	94.86	94.26	94.32
Infiltration	95.59	93.85	95.63	93.09	95.89	95.63
Pueumothorax	96.23	96.45	96.85	96.23	96.45	96.85
Emphysema	97.32	97.23	97.63	97.15	97.96	97.63
Fibrosis	98.23	98.12	98.42	98.74	98.23	98.42
Average	99.86	97.87	97.5	97.58	99.85	97.5

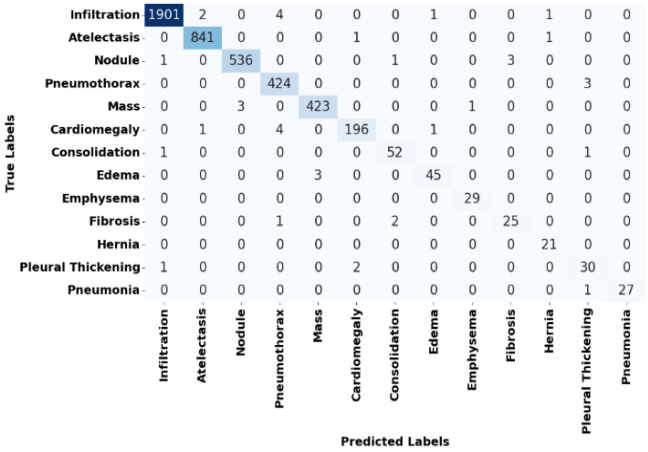
Table 4. Performance evaluation of the Chest X-ray dataset's k-fold validation

K-fold values	Accuracy (%)	Precision (%)	Recall (%)	F1-Measure (%)	Sensitivity (%)	AUC (%)
k=3	98.42	99.40	98.10	99.51	99.42	95.03
k=5	99.12	98.53	98.25	98.52	98.42	96.36
k=8	99.20	98.25	98.64	98.06	98.60	98.12
k=10	98.90	98.33	99.18	99.18	99.35	97.03

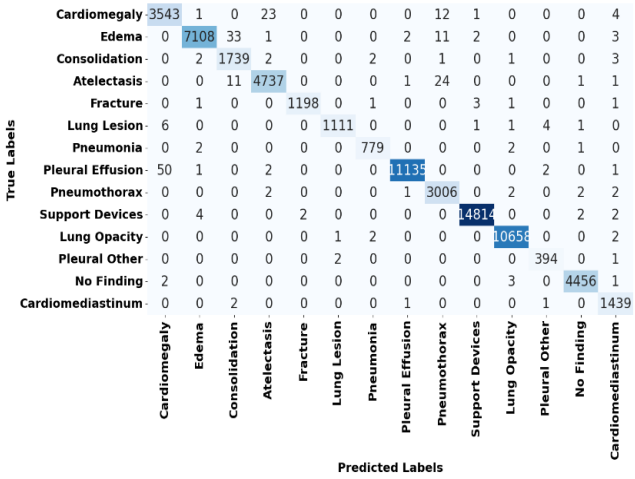
Table 5. Performance estimation of statistical analysis and computational of proposed method with existing methods

Dataset	Method	P-test	t-test	Computational cost (FLOP)	Inference time (s)
NIX CXR	ANN	4.42E-17	13.79	15.38	0.56
	CNN	3.74E-15	8.57	14.47	0.32
	RNN	6.56E-15	9.29	13.48	0.53
	GRU	5.47E-17	10.33	9.37	0.43
	HRGAN-CBAM	3.38E-13	8.94	6.38	0.23
ChesXpert	ANN	6.57E-18	11.37	14.33	0.78
	CNN	6.87E-13	9.22	11.56	0.65
	RNN	4.56E-14	9.31	9.67	0.54
	GRU	3.55E-12	9.18	7.55	0.43
	HRGAN-CBAM	2.01E-19	10.56	6.42	0.34
MIMIC CXR	ANN	7.53E-18	9.52	14.27	0.76
	CNN	6.67E-16	8.55	12.35	0.53
	RNN	6.12E-19	10.11	11.54	0.45
	GRU	8.56E-15	9.07	10.42	0.33
	HRGAN-CBAM	7.87E-14	8.83	9.78	0.24

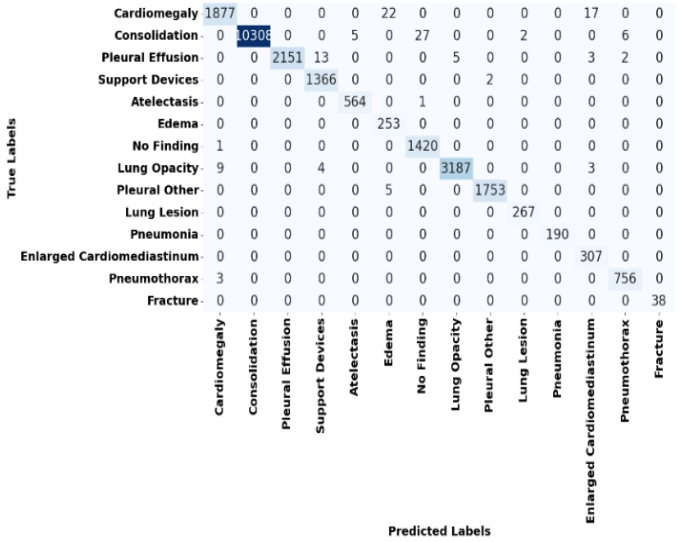
A confusion matrix is validated by comparing the true labels with the predicted classes. Fig. 3 demonstrates the ROC curve of the proposed CXR HRGAN-CBAM method based on three datasets: NIH CXR, CheXpert, and MIMIC CXR. The ROC curve is estimated in terms of the True Positive Rate (TPR) and True Negative Rate (TNR). Estimating the ROC curve through TPR and TNR helps to better evaluate the trade-off between sensitivity and specificity.



(a)



(b)



(c)

Figure. 2 Confusion Matrix: (a) NIX CXR dataset, (b) ChesXpert dataset, and (c) MIMIC CXR dataset

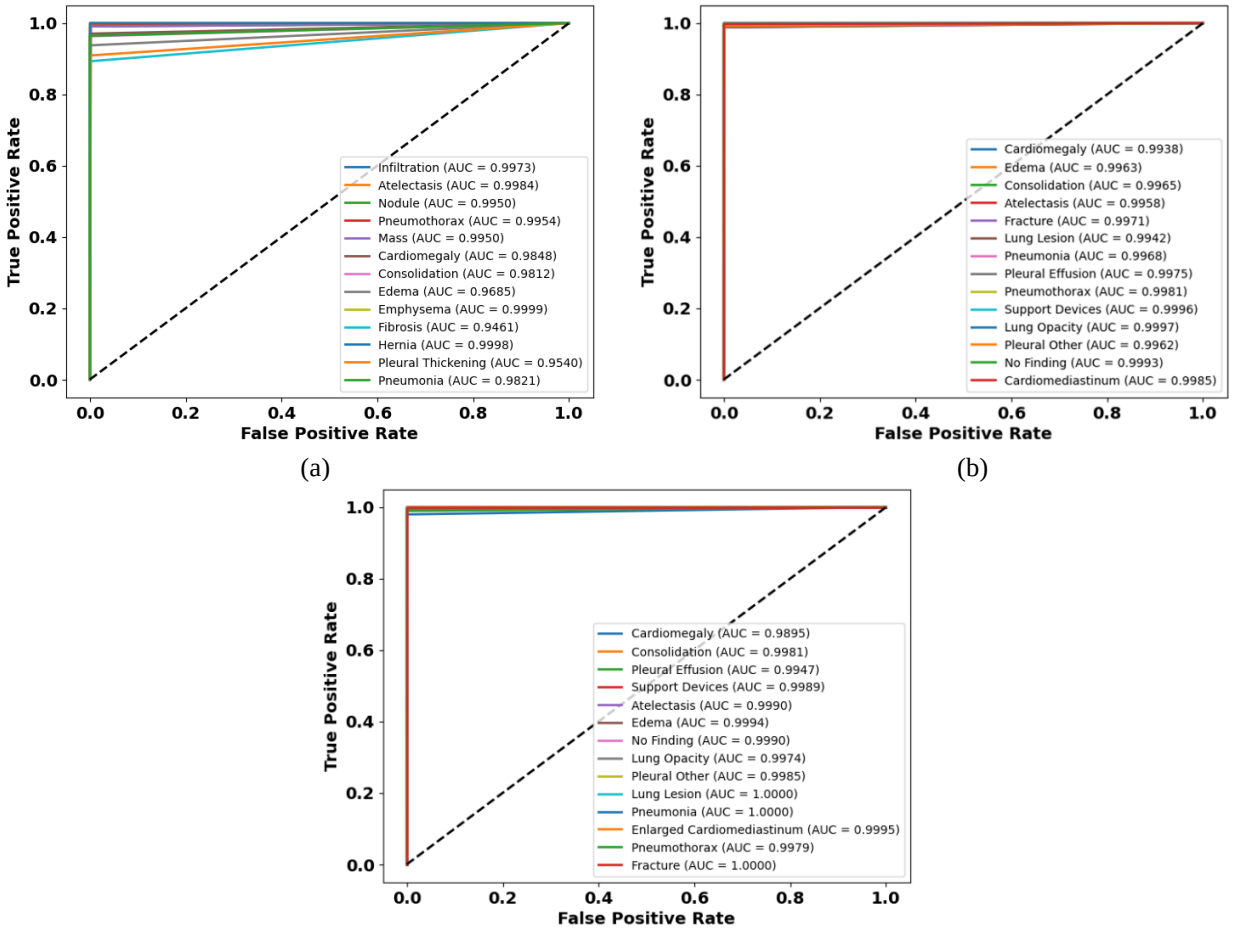


Figure. 3 ROC curve: (a) NIX CXR dataset, (b) ChesXpert dataset and (c) MIMIC CXR dataset

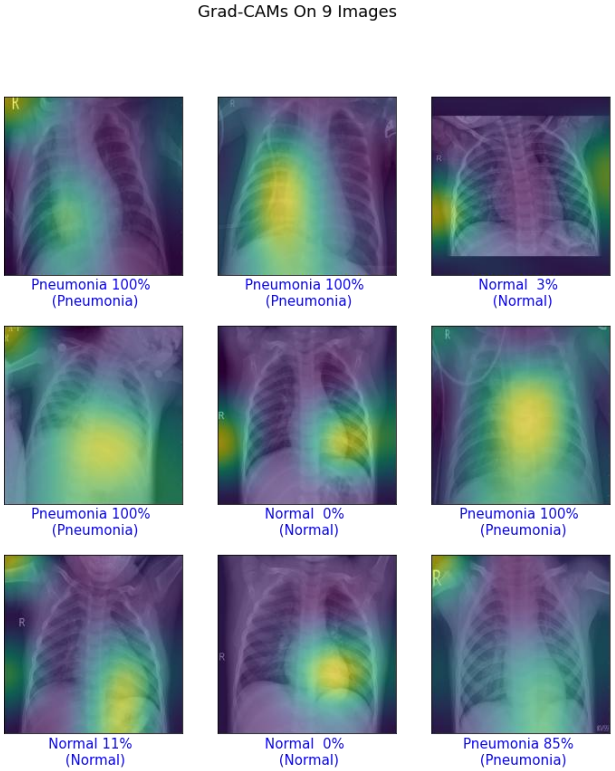


Figure. 4 Grad-CAM map of the NIH CXR dataset

Table 6 demonstrates the ablation study of with and without each component.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Measure (%)	Sensitivity (%)	AUC (%)
Without CBAM						
ANN	95.23	94.36	94.78	94.16	94.32	92.12
CNN	96.85	95.82	95.43	95.86	95.98	93.03
RNN	97.45	96.45	96.15	96.74	96.54	94.84
GRU	95.74	96.38	95.82	94.53	92.37	91.21
HRGAN	98.23	97.15	97.03	97.63	97.32	95.13
With CBAM						
ANN	96.46	95.58	91.26	90.45	91.23	89.46
CNN	97.32	96.32	93.43	92.43	93.21	90.12
RNN	98.63	96.48	95.43	94.43	95.65	92.45
GRU	98.87	97.74	96.56	96.54	97.67	94.47
HRGAN	99.12	98.53	98.25	98.52	98.42	96.36

Table 7. Comparative analysis of proposed method

Dataset	Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Specificity (%)	AUC (%)
NIX CXR	SACaps GAN SDOA-LDC [16]	98.96	97.89	97.68	97.02	97.13	NA
	VDV [17]	NA	NA	90.9	NA	NA	95.0
	ABOA-CNN [18]	96.5	NA	97.3	97.1	96.6	NA
	Swin Transformer [19]	87.3	NA	NA	NA	NA	NA
	GREN [20]	80.74	NA	NA	NA	NA	NA
	Proposed CXR HRGAN-CBAM method	99.12	98.53	98.25	98.52	98.42	96.36
ChesXpert	Early fusion [21]	91.4	NA	NA	NA	NA	NA
	Proposed CXR HRGAN-CBAM method	98.57	96.83	95.45	95.48	95.39	93.35
MIMIC CXR	3M-CNN [22]	98.79	90	89.72	95	NA	NA
	Proposed CXR HRGAN-CBAM method	99.56	97.48	96.49	97.68	98.57	94.38

Fig. 4 demonstrates the Grad-CAM map of the NIH CXR dataset. CBAM's attention mechanism refines feature maps by emphasizing informative spatial and channel-wise cues. This results in robust, high-confidence predictions with enhanced interpretability and classification accuracy.

Table 6 presents the ablation study comparing the performance with and without each component. The HRGAN approach with CBAM achieves better results compared to the HRGAN approach without CBAM. These results demonstrate that the HRGAN with CBAM outperforms other approaches.

4.2 Comparative analysis

This section presents the comparative analysis of the proposed method using performance metrics such as precision, accuracy, F1-score, sensitivity, and recall, as illustrated in Table 7. Existing methods, including SACaps GAN SDOA-LDC [16], VDV [17], ABOA-CNN [18], Swin Transformer [19], GREN [20], Early Fusion Model [21], and 3M-CNN [22], are utilized to assess the classifier's performance. For the GREN approach, the threshold is set to 0.7. The implemented method achieves an accuracy of 99.12% on the NIH CXR images.

4.3 Discussion

The advantages of the proposed hybrid mechanism effectively reconstruct regions affected by adversarial attacks and utilize two refined stages to efficiently segment the portions. These stages are then connected with an attention mechanism that focuses on critical areas of the CXR image by applying both channel and spatial attention. A drawback of existing methods, such as GAN [16], is the difficulty in training due to mode collapse and instability in the discriminator's dynamics, while SFO falls into local optima, making CXR image classification challenging. VDV [17] struggles to analyze similar tumors in CXR images and does not efficiently fine-tune the technique, negatively impacting classification accuracy. CNN [18] struggles to capture the dominant features fully because of adversarial attacks that degrade image quality. ABOA faces challenges in selecting relevant features due to falling into local optima, which in turn affects classification accuracy. The Swin

Transformer [19] struggles with a fixed window size and fails to effectively capture adversarial attacks in CXR images. It also requires parameter tuning, such as window size and shift strategies, for optimal performance. The challenges in constructing and regularizing GREN [20] add overhead to training and inference, leading to difficulty in processing adversarially attacked images.

5. Conclusion

The proposed CXR-specific Hybrid Refinement-Generative Adversarial Network with Convolutional Block Attention Module (CXR HRGAN-CBAM) efficiently detects adversarial attacks, refines different regions of the image, and ensures channel and spatial features are optimally improved. The initial raw data is obtained from the NIH CXR dataset, and pre-processing is performed using the CLAHE technique to improve image quality and identify small regions in CXR images. The hybrid mechanism effectively reconstructs regions affected by adversarial attacks, with two refined stages that efficiently segment the relevant portions. These stages are then integrated with an attention mechanism that focuses on critical areas of the CXR by applying both channel and spatial attention mechanisms. The proposed method achieves an accuracy of 99.12% on the NIH CXR dataset, outperforming existing methods such as the GACaps-SFO algorithm. Future work can focus on addressing various adversarial attacks in medical images using an improved DL approach.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, AG and NJ; methodology, US; software, AG; validation, NJ, US, and AG; formal analysis, NJ; investigation, US; resources, AG; data curation, NJ; writing—original draft preparation, US; writing—review and editing, NJ; visualization, AG; supervision, AG; project administration, AG; funding acquisition, US.

Notation Table

Notations	Descriptions
$Processed_{image}$ & $Original_{image}$	Contrast values image that consider the pre-processed and original image for remove in noise from CXR image
$loss_{fun}$	Gradient of loss function k with respect to Med_{raw}
Med_{adv}	Adversarial image in FGSM attack
σ	Model parameter
ϵ	Strength of FGSM
Med_{raw}	Medical raw image
γ	Perturbation factor
$Med_{adv}(I+1)$	Adversarial medical image after iteration $I+1$
Med_{tar}	Target image
G and D	Generator and discriminator
x	Training data
z	Arbitrary sample for noise vector
P_z	Predefined distribution
x	Original input image
θ^*	Optimal set of parameters
ω	Latent code
θ	Set of learnable modulation parameters
f	Refinement function
f^*	Optimal refinement function
f'	Final blended or refined feature representation
L	Total loss used for optimization
fl	Locally refined feature
F	Original or global feature representation
m	Soft mask indicating the region of interest
$\odot$	Element-wise multiplication
λ	Weighting factor controlling the influence of perceptual loss relative to pixel-wise loss
∇f	Gradient of the loss based on feature f
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

References

[1] F.J.M. Shamrat, S. Azam, A. Karim, K. Ahmed, F.M. Bui, and F.D. Boer, "High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images", *Computers in Biology and Medicine*, Vol. 155, p. 106646, 2023.

[2] N.G. Laleh, D. Truhn, G.P. Veldhuizen, T. Han, M.V. Treeck, R.D. Buelow, R. Langer, B. Dislich, P. Boor, V. Schulz, and J.N. Kather, "Adversarial attacks and adversarial robustness in computational pathology", *Nature communications*, Vol. 13, No. 1, p. 5711, 2022.

[3] L. Alzubaidi, A.D. Khamael, H.A.H. Obeed, A. Saihood, M.A. Fadhel, S.A. Jebur, Y. Chen, A.S.

Albahri, J. Santamaría, A. Gupta, and Y. Gu, "MEFF–A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging", *Intelligent Systems With Applications*, Vol. 22, p. 200355, 2024.

[4] W.Y.R. Fok, A. Fieselmann, C. Huemmer, R. Biniazan, M. Beister, B. Geiger, S. Kappler, and S. Saalfeld, "Adversarial robustness improvement for X-ray bone segmentation using synthetic data created from computed tomography scans", *Scientific Reports*, Vol. 14, No. 1, p. 25813, 2024.

[5] R. Jin, and X. Li, "Backdoor attack and defense in federated generative adversarial network-based medical image synthesis", *Medical Image Analysis*, Vol. 90, p. 102965, 2023.

[6] D. Rodriguez, T. Nayak, Y. Chen, R. Krishnan, and Y. Huang, "On the role of deep learning model complexity in adversarial robustness for medical images", *BMC Medical Informatics and Decision Making*, Vol. 22, No. 2, p. 160, 2022.

[7] S.B.U. 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.

[8] Y.W. Lee, S.K. Huang, and R.F. Chang, "CheXGAT: A disease correlation-aware network for thorax disease diagnosis from chest X-ray images", *Artificial Intelligence in Medicine*, Vol. 132, p. 102382, 2022.

[9] O. Daanouni, B. Cherradi, and A. Tmiri, "NSL-MHA-CNN: a novel CNN architecture for robust diabetic retinopathy prediction against adversarial attacks", *IEEE Access*, Vol. 10, pp. 103987-103999, 2022.

[10] U. Ahmed, J.C.W. Lin, and G. Srivastava, "Mitigating adversarial evasion attacks by deep active learning for medical image classification", *Multimedia Tools and Applications*, Vol. 81, No. 29, pp. 41899-41910, 2022.

[11] K. Kansal, P.S. Krishna, P.B. Jain, R. Surya, P. Honnavalli, and S. Eswaran, "Defending against adversarial attacks on Covid-19 classifier: A denoiser-based approach", *Heliyon*, Vol. 8, No. 10, p. e11209, 2022.

[12] S. Dey, R. Roychoudhury, S. Malakar, and R. Sarkar, "An optimized fuzzy ensemble of convolutional neural networks for detecting tuberculosis from Chest X-ray images", *Applied Soft Computing*, Vol. 114, p. 108094, 2022.

[13] 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, pp. 66478-66494, 2024.

[14] R. Wajgi, G. Yenurkar, V.O. Nyangaresi, B. Wanjari, S. Verma, A. Deshmukh, and S. Mallewar, "Optimized tuberculosis classification system for chest X-ray images: Fusing hyperparameter tuning with transfer learning approaches", *Engineering Reports*, Vol. 6, No. 11, p. e12906, 2024.

[15] X. Shi, Y. Peng, Q. Chen, T. Keenan, A.T. Thavikulwat, S. Lee, Y. Tang, E.Y. Chew, R.M. Summers, and Z. Lu, "Robust convolutional neural networks against adversarial attacks on medical images", *Pattern Recognition*, Vol. 132, p. 108923, 2022.

[16] N.B.M. Kumar, K. Premalatha, and S. Suvitha, "Lung disease detection using Self-Attention Generative Adversarial Capsule network optimized with sun flower Optimization Algorithm", *Biomedical Signal Processing and Control*, Vol. 79, Part 2, p. 104241, 2023.

[17] T. Iqbal, A. Shaukat, M.U. Akram, A.W. Muzaffar, Z. Mustansar, and Y.C Byun, "A hybrid VDV model for automatic diagnosis of pneumothorax using class-imbalanced chest X-rays dataset", *IEEE Access*, Vol. 10, pp. 27670-27683, 2022.

[18] B. Annamalai, P. Saravanan, and I. Varadharajan, "ABOA-CNN: auction-based optimization algorithm with convolutional neural network for pulmonary disease prediction", *Neural Computing and Applications*, Vol. 35, No. 10, pp. 7463-7474, 2023.

[19] Y. Ma, and W. Lv, "Identification of Pneumonia in Chest X-Ray Image Based on Transformer", *International Journal of Antennas and Propagation*, Vol. 2022, No. 1, p. 5072666, 2022.

[20] B. Qi, G. Zhao, X. Wei, C. Du, C. Pan, Y. Yu, and J. Li, "GREN: graph-regularized embedding network for weakly-supervised disease localization in X-ray images", *IEEE Journal of Biomedical and Health Informatics*, Vol. 26, No. 10, pp. 5142-5153, 2022.

[21] S.K. Samarla, and P. Maragathavalli, "Ensemble fusion model for improved lung abnormality classification: Leveraging pre-trained models", *MethodsX*, Vol. 12, p. 102640, 2024.

[22] J. Upadhya, K. Poudel, and J. Ranganathan, "Advancing medical image diagnostics through multi-modal fusion: Insights from mimic chest x-ray dataset analysis", In: *Proc. of 2024 IEEE*

3rd International Conference on Computing and Machine Intelligence (ICMI), pp. 1-8, 2024.
[23] NIH CXR dataset: https://www.kaggle.com/datasets/nih-chest-xrays/data (Accessed on Nov 2024).
[24] ChesXpert dataset link: https://www.kaggle.com/datasets/ashery/chexpert. (Accessed on April 2025).
[25] MIMIC CXR dataset link: https://www.kaggle.com/datasets/wasifnafee/mimic-cxr. (Accessed on April 2025).
[26] S. Bhoopal, M. Rao, and C.H. Krishnappa, "Enhanced diabetic retinopathy detection and classification using fundus images with ResNet50 and CLAHE-GAN", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 35, No. 1, pp. 366-377, 2024.