Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.11 n.56, p. 157-167. November/December, 2025.

DOI: https://doi.org/10.5935/jetia. v11i56.1941

RESEARCH ARTICLE OPEN ACCESS

EFFICIENTNET AND DENSENET ENSEMBLE WITH GAN AUGMENTATION FOR IMBALANCED COVID-19 AND VIRAL PNEUMONIA CLASSIFICATION IN CHEST X-RAYS

Aissa Snani[1], Mohammed Tarek Khadir[2] and Modawy Adam Ali Abdalla[3]

[1, 2]LabGED, Department of Computer Science, University of Badji Mokhtar, PO Box 12, 23000, Annaba, Algeria
[3]Department of Electrical and Electronic Engineering, College of Engineering Science, University of Nyala, Nyala 63311, Sudan.

[1]http://orcid.org/0009-0008-9503-0689, [2]http://orcid.org/0000-0002-1741-0263, [3]http://orcid.org/0000-0002-7227-6941

Email: aissa.snani@univ-annaba.dz, khadir@labged.net, brojacter88@yahoo.com

ARTICLE INFO

Article History
Received: April 27, 2025
Revised: May 20, 2025
Accepted: December 1, 2025
Published: December 31, 2025

Keywords:
EfficientNet–DenseNet Ensemble,
GAN-Based Data Augmentation,
Imbalanced Classification,
Chest X-Ray Classification,
Computer-Aided Diagnosis,

ABSTRACT

Chest X-ray imaging is a critical tool for diagnosing various diseases and plays a significant role in computer-aided diagnosis (CAD) systems. Recent research has leveraged CXR datasets and deep learning models to detect multiple diseases; however, these studies often suffer from dataset imbalance, where one class is overrepresented. This imbalance hampers deep learning model training, leading to models that perform well on the majority class but struggle with underrepresented classes. To address this challenge in CXR datasets comprising three classes COVID-19, Viral pneumonia, and Normal images, we employ a generative adversarial network to generate synthetic CXR images for the minority classes, thereby enhancing model robustness. Additionally, we fine-tune DenseNet201 and EfficientNetV2-B3 and integrate their predictions using a soft voting ensemble. Our method reaches an accuracy of 99.33%, an F1-score of 99.01%, and a Matthews correlation coefficient (MCC) of 98.59%, which helps reduce the problems caused by uneven class distribution. These results show that using GAN-based augmentation with an ensemble of deep learning methods can enhance the classification of different types of CXR, leading to more trustworthy AI-assisted diagnoses.

I. INTRODUCTION

Medical professionals use chest X-rays (CXR) to diagnose a wide range of conditions affecting the heart, lungs, blood vessels, airways, and skeletal structures of the chest and spine [1]. Its widespread use is due to the low cost and widespread availability of equipment, which makes it accessible even in economically challenged areas. Furthermore, the ability to acquire images quickly distinguishes it from more complex methodologies such as computed tomography (CT) and magnetic resonance imaging (MRI), which is critical in emergency medical situations. CXR imaging is a cornerstone in diagnosing respiratory diseases, offering rapid, cost-effective insights into conditions like COVID-19 and viral pneumonia. While manual interpretation remains challenging due to subtle pathological patterns and inter-class similarities, artificial intelligence (AI) has emerged as a transformative tool for automated detection. Deep learning models, particularly convolutional neural networks (CNNs), have successfully classified CXRs.

However, their performance hinges critically on balanced training data; a requirement rarely met in real-world medical datasets. Conventional class imbalance solutions face notable limitations in medical imaging contexts. On the algorithmic side, cost-sensitive learning and focal loss [2], [3] recalibrate loss terms to penalize minority-class misclassification but do not fully capture the hierarchical feature representations of convolutional neural networks. Medical imaging adds further complexity through class ambiguity, up to 22% of COVID-19 CXRs) exhibit visually indistinguishable opacities from early viral pneumonia [4], making loss reweighting difficult to tune correctly. Data-level methods like SMOTE (Synthetic minority over-sampling technique) [5] and ADASYN (Adaptive synthetic sampling) risk generating unrealistic, artifact-laden samples. COVID-19 lesions, for instance, progress via nonlinear expansions from

peripheral to diffuse opacities [6], whereas SMOTE interpolates linearly, creating implausible intermediate images. Undersampling approaches, especially those removing the "easiest" majority samples [7] or entire subgroups via cluster-based selection [8], often discard clinically indispensable negatives or unique phenotypes (e.g., pediatric cases), thereby hampering model generalization. Standard image augmentations such as rotations and flips cannot capture pathology-specific spatial distributions, e.g., COVID-19's bilateral lower-zone predominance [9]. Instead of enhancing the training set, these crude transformations risk adding noise and distorting decision boundaries. Another method for oversampling is generative adversarial networks (GANs) overcome these drawbacks by learning the underlying manifold of complex pathological structures directly from data. Rather than linearly interpolating existing samples or discarding large portions of the majority class, GAN-based approaches synthesize entirely new images that mirror real lesions' subtle variations and spatial progression. This helps preserve morphological details and class-specific features crucial for accurate diagnosis [10]. This paper proposes a dual-strategy framework combining synthetic data generation GANs with ensemble deep learning.

First, a GAN synthesizes high-fidelity CXR images for minority classes, preserving pathological features while balancing class distributions. Second, a two-model EfficientNetV2B3, an improved architecture, and DenseNet-201 architecture leverage complementary feature extraction capabilities through soft voting, enhancing robustness against lesion appearance and positioning variability. Our contributions are fourfold: a GAN-based augmentation pipeline that generates clinically realistic CXRs for imbalanced classes, validated via Fréchet Inception Distance; an ensemble model achieving state-of-the-art performance (99.33% accuracy, 99.01% F1-score) on a multi-class dataset; empirical evidence that GAN-based augmentation outperforms oversampling methods, particularly improving recall and MCC for minority classes; and Grad-CAM visualizations providing feature-level insights into the regions most influential for model decision-making, demonstrating that predictions are based on clinically meaningful anatomical structures. The remainder of this paper is organized as follows: Section II reviews existing approaches to imbalanced CXR classification. Section III presents our proposed GAN architecture and ensemble learning framework. The dataset, preprocessing steps, experimental results, and comparisons with baseline methods are analyzed in Section IV. Finally, conclusions are presented in Section V.

II. RELATED WORKS

The classification of CXR images for detecting COVID-19, viral pneumonia, and Normal CXR has received significant attention. Researchers have explored various CNN architectures to automatically learn and extract discriminative features from CXR images, aiming to accurately identify subtle differences between diseases [11], [12]. Despite the promising results, many studies are limited by imbalanced datasets, which can restrict the generalizability of the models in real-world scenarios. For instance, Ozsoz et al. [13] utilized a pre-trained AlexNet model to classify CXR images, achieving an accuracy of 93.42% on a dataset comprising 371 COVID-19 pneumonia, 4237 non-COVID-19 viral pneumonia, 4078 bacterial pneumonia, and 2882 Normal cases. Similarly, Mangal et al. [14] introduced the CovidAID model, which attained accuracies of 87.2% for four classes and 90.5% for three classes on a dataset with 1341 Normal, 2530 viral pneumonia, 1337 bacterial pneumonia, and 115 COVID-19 samples. In binary classification, Mahin et al. [15] achieved an accuracy of 98% using MobileNetV2 on a dataset of 1,142 COVID-19 and 4,237 pneumonia images. Other studies have leveraged deeper architectures: Asif et al. [16] employed an Inception V3-based DCNN to achieve 93% accuracy, while Tiwar [17] combined a pre-trained VGG-16 network with an inception module to reach 97.67% accuracy on a dataset containing 3616 COVID-19, 6012 lung opacity, 1345 viral pneumonia, and 10192 Normal images. Bashar et al. [18] and Nillmani et al. [19] further explored VGG16, reporting accuracies of 95.63% and 96.63%, respectively, across multi-class datasets.

Hybrid approaches have also shown impressive results. Jin et al. [20] integrated a pre-trained AlexNet with the ReliefF algorithm and a support vector machine (SVM) to achieve an overall accuracy of 98.64%. Gupta et al. [21] demonstrated that AlexNet could deliver an accuracy of 97.6% on a dataset including 219 COVID-19, 1341 healthy, and 1345 viral pneumonia cases. KC et al. [22] evaluated eight pre-trained models, with DenseNet121 reaching a test accuracy of 98.69% and an F1-score of 99% for four-class classification. Apostolopoulos and Mpesiana [23] compared several models (e.g., VGG19, MobileNet, Inception, Xception, and Inception ResNet v2) on two datasets, reporting best-case accuracy, sensitivity, and specificity values of 96.78%, 98.66%, and 96.46%, respectively, using a CADx system based on VGG16. Other notable contributions include Khan et al. [24] with the CoroNet model based on Xception (accuracy: 89.6%), and Ozturk [25], who proposed the DarkCovidNet model (accuracy: 87.02%) using a darknet-based architecture [26]. Alharbi et al. [27] achieved both accuracy and F1 score of 99% with a six-layer CNN on a large dataset of 16490 COVID-19, 15000 Normal, and 5856 pneumonia images. Although many of these models report high accuracy, their reliance on imbalanced datasets poses a significant challenge.

Imbalanced class distributions can limit a model’s ability to generalize effectively. Several studies have proposed strategies to balance the data to mitigate this issue. Win et al. [28] combined 11 deep learning models with techniques such as weighted loss, image augmentation, undersampling, oversampling, and hybrid resampling, achieving average accuracies of 99.23%, sensitivity of 99.27%, and an F1-measure of 98.3% across COVID-19, viral pneumonia, and Normal CXR images. Naseem et al. [29] constructed two balanced datasets by merging four sources, ensuring equal representation for each class. Their transfer learning approach with various pre-trained CNNs yielded maximum accuracies of 99.83%, 98.11%, and 97.00% for two, three, and four-class classifiers, respectively. In a similar vein, Chamseddine et al. [30] addressed class imbalance using SMOTE combined with class weights across three datasets, achieving the highest performance with DenseNet201 and VGG19 with 98.87% accuracy, 98.21% F1-score, 98.86% sensitivity, 99.43% specificity, and 100% precision. In summary, while CNN-based methods have significantly advanced the automated classification of COVID-19 and pneumonia from CXR images, addressing data imbalance remains a critical challenge to enhancing model generalizability and clinical applicability.

III. METHODOLOGY

III.1 GENERATIVE ADVERSARIAL NETWORKS AND PROPOSED ARCHITECTURE

GANs, introduced by Ian Goodfellow et al. [31], are deep learning models designed to generate synthetic data that closely resemble real-world distributions. A GAN consists of two neural networks: the generator, which learns the underlying data distribution, and the discriminator (or critic), which differentiates between real and generated samples. This adversarial setup establishes a min-max optimization framework, where the generator improves to produce realistic data while the discriminator enhances its ability to distinguish real from synthetic samples. Figure 1 illustrates the architecture of the proposed GAN model. The generator begins with a 128-dimensional random noise vector, *z*. This vector is first processed through a fully connected layer containing 4 × 4 × 256 units, reshaped into a (4, *4, 256*) tensor, forming the foundation for spatial upscaling. A sequence of five upsampling blocks progressively increases the spatial resolution from 4×4 to 128×128. Each block employs nearest-neighbor upsampling via an UpSampling2D layer, followed by a Conv2D layer. The number of filters in these convolutional layers gradually decreases in the sequence (128, *64, 32, 16, 8)*, with Batch Normalization and ReLU activation applied at each step. The ReLU function is defined as:

$$ReLU(x) = \max(0, x) \tag{1}$$

After the upsampling blocks, the generator includes 16 residual blocks that use skip connections to help gradients move easily and make training more stable. This architectural design addresses the performance degradation issue in deep networks, enabling the effective training of deeper models. The following formulation forms the core of the residual block:

$$y = F(x; W_i) + W_s x \tag{2}$$

Where $F(x; W_i)$ represents the Conv2D-Batch Normalization residual mapping, and $W_s x$ denotes the shortcut connection. The layer-wise relationships within the residual blocks are described by Equations (3) and (4):

$$x_{l-1} = H(x_{l-2}) \tag{3}$$

$$x_l = ReLU(F(x_{l-1}) + x_{l-2}) \tag{4}$$

These equations illustrate the strategic combination of skip connections with convolutional transformations. An optimized variant of the residual block further enhances efficiency by structuring layers as a sequence of 1×1−3×3−1×1 operations, with Batch Normalization and ReLU applied before each convolutional step. These modifications help mitigate vanishing gradient issues and stabilize deep network training. The final layer of the generator consists of a 3×3 convolution with a tanh activation function, producing a 128×128×1 grayscale image normalized within the range [−1,1]. The discriminator (or critic) processes input images, whether real or generated, through convolutional layers to extract hierarchical features and provide evaluative feedback. It starts with convolutional layers that progressively increase the number of filters (64,128,256,512,1024), each applying a Leaky ReLU activation function Equation (5) to introduce controlled non-linearity:

$$\text{LeakyReLU(x)} = \begin{cases} \text{x}, & \text{x} \geq 0 \\ \alpha\text{x}, & \text{x} < 0 \end{cases} \tag{5}$$

Where α is a slight positive slope. The critic incorporates a mini-batch discrimination layer to stabilize training further, which computes inter-sample similarities using learnable kernels, encouraging diversity in generated outputs. The architecture concludes with a dense layer that outputs a single scalar value, representing the critic's evaluationn n of the input image. Traditional GANs use the Jensen-Shannon divergence to measure the distance between distributions. However, this approach has inherent limitations, particularly when distributions have disjointed supports [32]. To address this, Gulrajani et al. [33] proposed the Wasserstein GAN with Gradient Penalty (WGAN-GP), which replaces the original loss function with the Wasserstein distance and introduces a gradient penalty term to enforce the Lipschitz constraint. This modification improves training stability and mitigates mode collapse by ensuring the critic function behaves smoothly. The WGAN-GP loss function is formulated as follows:

$$L_D^{WGAN-GP} = \underbrace{E_{\tilde{y}\sim p_g}[D(\tilde{y})] - E_{y\sim p_r}[D(y)]}_{L_D^{WGAN}} + \gamma \cdot \underbrace{E_{\hat{y}\sim p_{\hat{y}}}\left(\left(|\nabla_{\hat{y}} D(\hat{y})|_2 - 1\right)^2\right)}_{\text{Gradient Penalty}} \tag{6}$$

Where (γ) is the gradient penalty coefficient, and ($\hat{y}$) is a random interpolation between real and generated samples, defined as:

$$\hat{y} = \delta\tilde{y} + (1-\delta)y, \quad \delta \sim U(0,1) \tag{7}$$

III.2 DENSENET-201

Densely connected convolutional networks (DenseNet) [34] introduce a novel connectivity pattern where each layer receives inputs from all preceding layers. This architecture enhances gradient flow, encourages feature reuse, and optimizes learning efficiency while mitigating the vanishing gradient problem. The mathematical representation of dense connectivity is given by:

$$x_l = H([x_0, x_1, \dots, x_{l-1}]) \tag{8}$$

Where $[x_0, x_1, \dots, x_{l-1}]$ represents the concatenation of all previous feature maps, and $H(\cdot)$ denotes a composite function consisting of Batch Normalization, ReLU activation, and convolutional operations. A key parameter in DenseNet is the growth rate (k), which determines the number of new feature maps added by each layer. Given an initial feature map count (k_0), the total number of feature maps after l layers in a dense block is given by:

$$k_0 + l \times k \tag{9}$$

By systematically controlling the expansion of feature maps, DenseNet achieves parameter efficiency and effective feature propagation without introducing excessive redundancy. In our experiments, we employ DenseNet-201, a variant with 201 layers.

III.3 EFFICIENTNETV2-B3

The EfficientNet family [35] uses an innovative scaling method that carefully adjusts the model's depth, width, and input resolution to get high accuracy while using less computing power. The relationships governing this scaling approach are given by:

$$depth \propto \alpha^{\phi}, \quad width \propto \beta^{\phi}, \quad resolution \propto \gamma^{\phi}, \tag{10}$$

Where α, β, γ are scaling coefficients determined through grid search, and ϕ is a scaling factor that adjusts these parameters in a balanced manner. Building upon the success of EfficientNet, EfficientNetV2 introduces two major architectural enhancements that improve efficiency and accuracy. Fused inverted bottleneck layers streamline convolutional operations, leading to faster training and reduced inference latency. Progressive learning strategy Initially, the model is trained on lower-resolution images to accelerate convergence, followed by higher-resolution images in later stages to refine fine-grained feature extraction. For our experiments, we utilize EfficientNetV2-B3, a variant that strikes an optimal balance between accuracy, efficiency, and model size. This architecture is particularly well-suited for medical image classification, as it delivers high performance while maintaining computational efficiency.

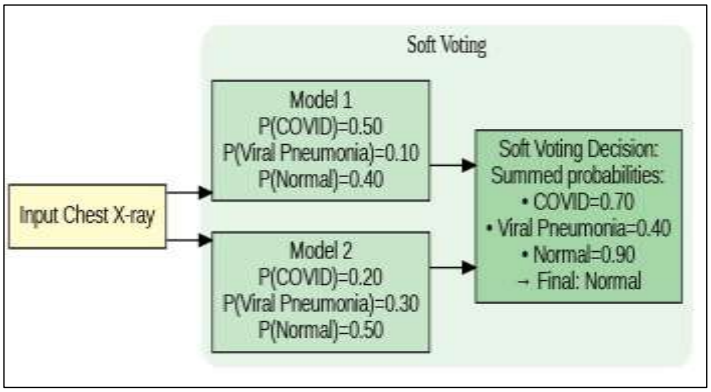

Figure 2: Ensemble learning using majority soft voting.
Source: Authors, (2025).

III.4 ENSEMBLE LEARNING

Ensemble learning has been empirically shown to enhance classification accuracy and mitigate overfitting compared to individual models. This improvement arises from aggregating predictions across multiple classifiers, increasing robustness and generalization. We employed soft voting in this study because it can leverage probabilistic confidence scores [36]. By aggregating the predicted probabilities from both models, soft voting ensures that the final decision reflects the relative confidence of each classifier rather than treating them as equally certain. Figure 2 depicts ensemble learning using soft majority voting methods.

IV. EXPERIMENTAL AND RESULTS

IV.1 DATASET DESCRIPTION

The dataset, available as an open source on Kaggle(https://www.kaggle.com/datasets/tawsifurrahman/Covid19-radiography-database), a collaborative creation and public contribution from a team comprising members of Qatar University, Doha, and the University of Dhaka, Bangladesh. This joint effort also engaged contributors from Pakistan and Malaysia, working collaboratively with medical professionals. In its second version, the dataset comprises a comprehensive collection of 15,153 CXR images, distributed across categories

as 10,192 Normal, 3,616 COVID-19, and 1,345 viral pneumonia CXRs. Figure 4 illustrates the initial class distribution, emphasizing the dominance of the Normal class relative to COVID-19 and Viral Pneumonia cases.

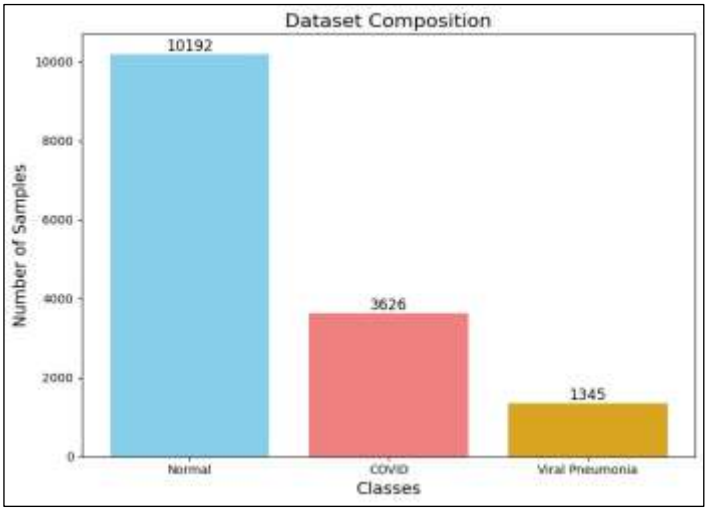

Figure 4: Overview of the class distribution in the dataset.
Source: Authors, (2025).

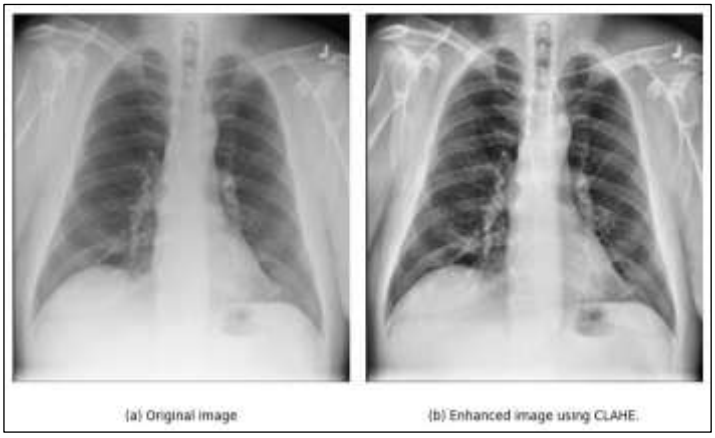

Figure 5: Image preprocessing using CLAHE.
Source: Authors, (2025).

IV.2 DATA PREPROCESSING

Medical images often suffer from noise and low contrast, necessitating preprocessing for deep learning applications. Contrast-Limited Adaptive Histogram Equalization (CLAHE) [37] enhances local contrast and detail, leading to improved predictions. It refines traditional Histogram Equalization by applying Adaptive Histogram Equalization (AHE) to separate regions of the image, while the "contrast-limited" feature prevents noise amplification. Figure 5 presents examples of CLAHE-enhanced images. To clean the dataset, we utilized the ImageHash library to remove duplicate images. The final dataset consists of 3,400 COVID-19, 10,190 normal, and 1,337 viral pneumonia images, divided into 80% for training and 20% for testing. To optimize computational resources, all images were resized to 128×128×3.

IV.3 BALANCING DATASET

To address class imbalance in the dataset, a GAN based data augmentation approach was applied to the minority classes: viral pneumonia and Covid19. The goal was to generate synthetic images for these classes until their distribution closely matched the normal class's. This augmentation was performed exclusively on the training dataset, ensuring that the test dataset remained unchanged. The proposed GAN model (Section III.1) was trained separately on the viral pneumonia and COVID-19 datasets, with the loss function defined in Equation (6) (Section III.1). Training was conducted over 1,000 epochs using the Adam optimizer with a learning rate of 0.0001. A batch size of 64 was used to ensure stable convergence, and the Leaky ReLU activation function was applied with a positive slope (α) of 0.2. The gradient penalty weight (γ) was set to 10 to enforce the Lipschitz constraint. Following training, the quality and realism of the generated images were quantitatively evaluated using the Frechet Inception Distance (FID) metric (Equation (11)). The FID measures the similarity between real and generated image distributions based on deep feature statistics extracted from the Inception network.

$$FID = |\mu_r - \mu_g|^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right) \tag{11}$$

where (μ_r) and (μ_g) represent the mean feature vectors of the real and generated data distributions, respectively, while (Σ_r) and denote their corresponding covariance matrices. A lower FID score indicates that the generated images are more similar to real images.

IV.4 TRAINING AND CLASSIFICATION

Two pre-trained CNN architectures, DenseNet201 and EfficientNetV2B3 (Section III.2, III.3), were implemented to classify the targeted dataset. To adapt these models to the specific classification task, their original fully connected classification layers were removed and replaced with a series of task-specific layers. These modifications included a Global Average Pooling (GAP) layer to reduce feature dimensionality, a Dropout layer to mitigate overfitting, a Dense layer with ReLU activation to introduce non-linearity and a final Dense layer with softmax activation for multiclass classification. The modified architectures, integrating the pre-trained base models with task-specific layers, were trained end-to-end using the Adam optimizer with a categorical cross-entropy loss function. A batch size of 32 was used, and different learning rate schedules (0.001, 0.01, and 0.005) were explored to optimize performance. Data augmentation techniques were applied during the training phase to improve dataset diversity and minimize the risk of model overfitting. These augmentations included image rotation by 10% and random horizontal flipping to introduce variability in the input space while preserving essential visual features. Each model was trained for 100 epochs to ensure adequate convergence. Following training, the final prediction was obtained using a soft voting ensemble of the two models, as described in Section III.4.

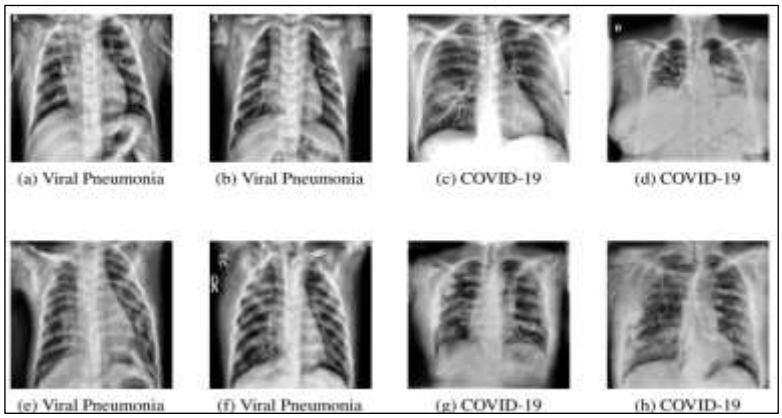

Figure 6: Sample real (top row) and generated (bottom row) CXR images for Viral pneumonia and COVID-19 cases.
Source: Authors, (2025).

IV.5 EVALUATION METRICS

Evaluating the performance of a classification model requires multiple metrics, depending on the specific task and context. These metrics provide insights into the model's strengths and weaknesses, enabling a comprehensive assessment of its effectiveness. In classification tasks, the model's predictions can be categorized into four possible outcomes, True Positive (TP) occurs when the model correctly identifies an instance as belonging to the actual class, such as correctly classifying a COVID-19 asCovid19. False Positive (FP) occurs when the model incorrectly classifies an instance into a category it does not belong to, for example, misclassifying a normal as COVID-19. True Negative (TN) occurs when the model correctly identifies an instance as not belonging to a specific class, such as correctly classifying a Normal as non-COVID-19. False Negative (FN) occurs when the model fails to recognize a true positive instance, for instance, misclassifying a COVID-19 as normal or Viral pneumonia. These results are typically summarized in confusion matrix, which provides a detailed breakdown of the model's classification performance. Several key evaluation metrics can be derived from this matrix to assess different aspects of the model's effectiveness. The fundamental performance metrics are defined as follows :

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{12}$$

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

$$Recall\ (Sensitivity) = \frac{TP}{TP+FN} \tag{14}$$

$$Specificity = \frac{TN}{TN+FP} \tag{15}$$

$$F1\ Score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{16}$$

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{17}$$

While accuracy provides a general overview of model performance, it may be misleading in cases where the dataset is imbalanced, meaning one class significantly outnumbers the others. The F1 score serves as a harmonic mean of precision and sensitivity, offering a balanced measure particularly useful for imbalanced datasets, as it considers both false positives and false negatives. However, a more robust metric for handling imbalanced datasets is the MCC. Unlike accuracy and F1 score, MCC takes into account all four elements of the confusion matrix (TP, TN, FP, FN) and provides a correlation coefficient between the predicted and actual classifications.

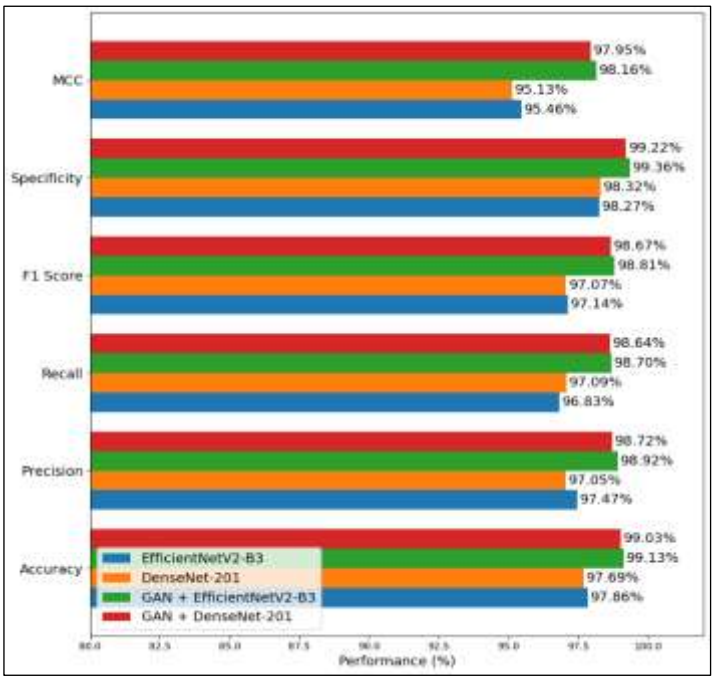

Figure 7: Classification Performance of Normal, Pneumonia, and COVID-19 Cases with and without GAN Balancing for Individual Classifiers.
Source: Authors, (2025).

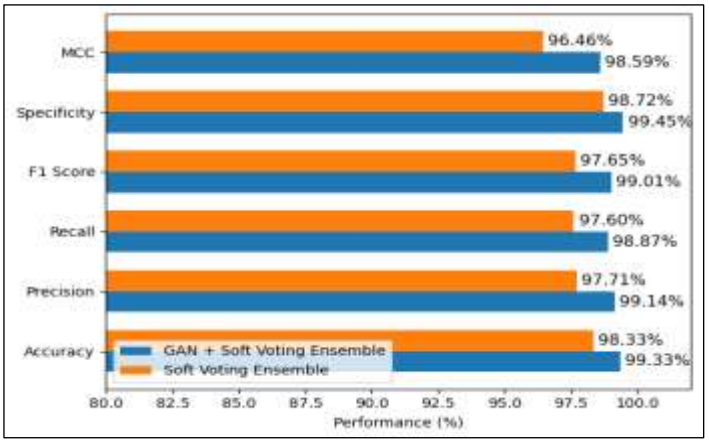

Figure 8: Classification Performance of Normal, Pneumonia, and COVID-19 Cases with and without GAN Balancing for Ensemble Classifiers.
Source: Authors, (2025).

IV.6 RESULTS AND DISCUSSION

The study aimed to address class imbalance in medical image classification by employing GANs for synthetic data augmentation. Through rigorous preprocessing, duplicate images were removed to prevent redundancy, and CLAHE was applied to enhance contrast and texture details. The dataset was then split into training and testing subsets. The training set consisted of 8,752 Normal, 2,720 Covid, and 1,070 Viral pneumonias (minority class), while the test set included 2,188 normal, 680 Covid, and 267 Viral pneumonia. The GAN-generated images significantly improved minority-class recall and reduced prediction bias. The FID scores of 62 for Viral pneumonia and 87 for the minority pathology confirmed the high visual fidelity of these synthetic samples, ensuring that they preserved clinically relevant features. Figure 6 illustrates both real and GAN-generated images produced by our architecture. When incorporated into model training, these images significantly enhanced performance. The baseline models, EfficientNetV2-B3 and DenseNet-201, initially achieved high overall accuracy (97.86% and 97.69%, respectively), but exhibited moderate recall scores for the minority classes (97%), reflecting a bias toward the majority class. Following GAN augmentation, GAN+EfficientNetV2B3 achieved 99.13% accuracy and 98.70% recall, while GAN+DenseNet201 reached 99.03% accuracy and 98.64% recall. As seen in Figure 7, the MCC, which measures classification balance,

increased from 95.46% and 95.13% in the baseline models to 98.16% and 97.95%, indicating a significant reduction in class bias. To further enhance generalization, a soft voting ensemble combining EfficientNetV2B3 and DenseNet201 was implemented. The ensemble model without augmentation reached 98.33% accuracy and an F1-score of 97.65% (Figure 7). However, when trained with GAN-augmented samples, the GAN+Soft Voting Ensemble achieved 99.33% accuracy, 98.87% recall, and an MCC of 98.59%, demonstrating the synergy between GAN augmentation and ensemble learning. The high specificity of 99.45% suggests that the model effectively reduced false positives while maintaining robust sensitivity for minority pathology cases.

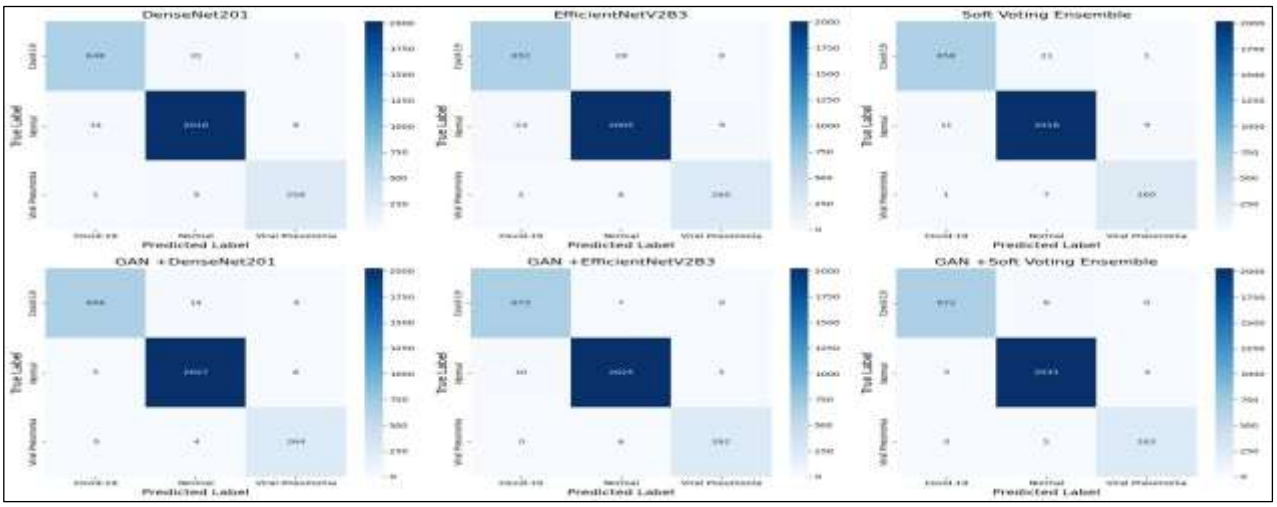

Figure 9: Confusion matrices showing classification performance for different models and ensemble approaches, with and without GAN-based augmentation. The bottom row (GAN-enhanced models with ensemble learning) minimizes misclassification errors, especially for the minority class.
Source: Authors, (2025).

For the classification performance of different models and ensemble approaches, we present the confusion matrices in Figure 9. These matrices depict the number of correctly and incorrectly classified instances for each class, providing a detailed assessment of model performance. We computed the Grad-CAMs of the CNN models and superimposed these activation maps onto CXRs to identify regions indicative of COVID-19 and Viral pneumonia related biomarkers. The regions highlighted in red correspond to areas with the greatest influence on the classification decision. When a CXR is classified as COVID-19 positive or as Viral Pneumonia, the red-highlighted regions indicate the areas that contributing most significantly to the prediction, suggesting the likely location of the disease. Figure 10 and 11 presents Grad-CAM visualizations for COVID-19 and Viral pneumonia CXRs were correctly classified by the model.

The comparison with different oversampling techniques, presented in Figure 12, highlights the limitations of interpolation-based methods such as ADASYN and SMOTE, which often fail to generate samples that reflect clinically meaningful features and fine-grained image details used by the model for its predictions. While ADASYN combined with DenseNet201 improved accuracy to 98.16%, its recall (96.90%) remained lower than that achieved by GAN+DenseNet201 (98.64%), suggesting that GAN-generated data provided more realistic and diagnostically meaningful variations than traditional resampling methods. Similarly, SMOTE+DenseNet201 achieved a 97.10% F1-score but still fell short of the performance observed with GAN augmentation. These findings emphasize that interpolation-based techniques often fail to preserve complex pathological patterns, whereas GANs effectively capture the heterogeneity of medical conditions, leading to improved classification performance.

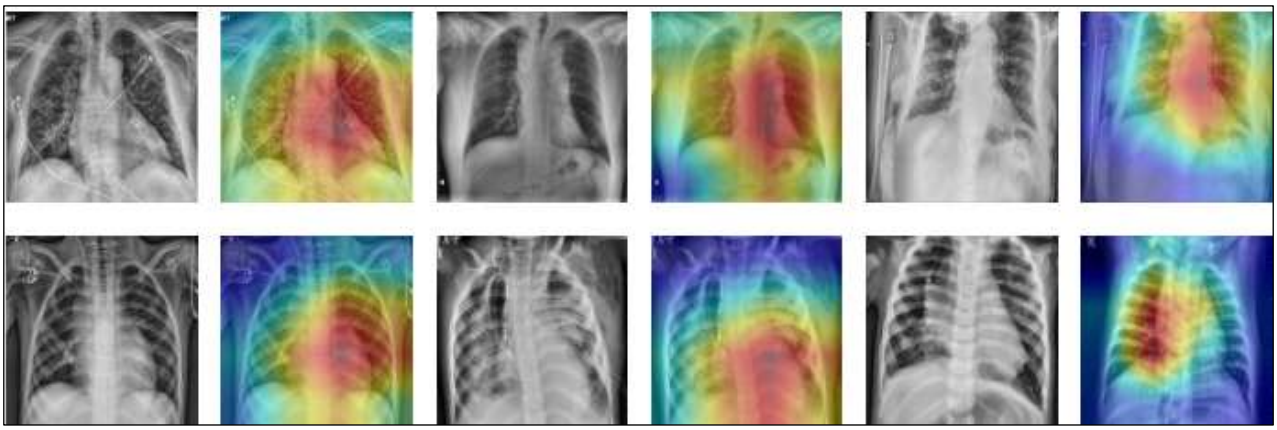

Figure 10**:** Grad-CAM visualizations from EfficientNetV2-B3. The first row shows activation maps for COVID-19 positive CXRs, and the second row shows activation maps for Viral pneumonia cases, highlighting the regions most influential in the classification decisions.
Source: Authors, (2025).

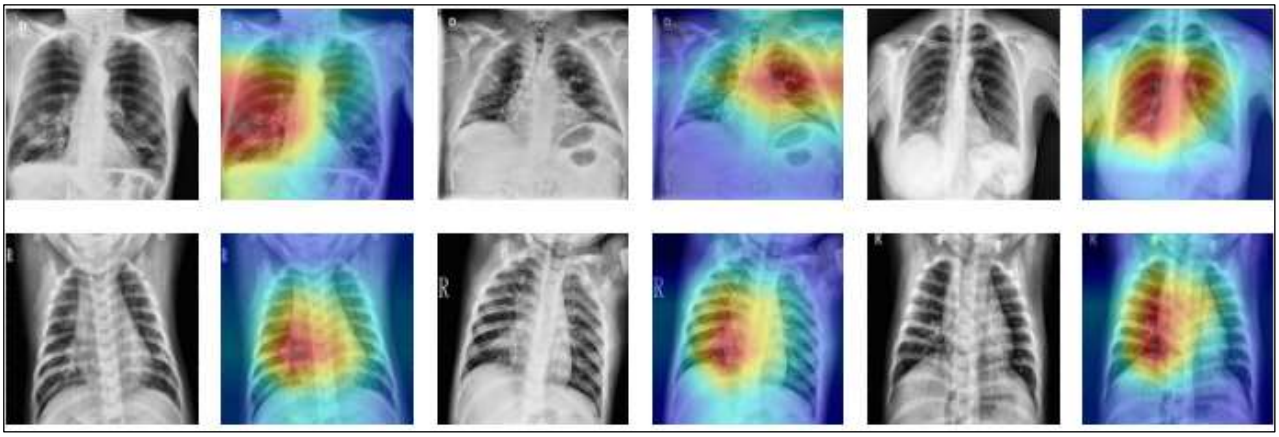

Figure 11**:** Grad-CAM visualizations from DenseNet-201. The first row shows activation maps for COVID-19 positive CXRs, and the second row shows activation maps for Viral pneumonia cases, highlighting the regions most influential in the classification decisions.
Source: Authors, (2025).

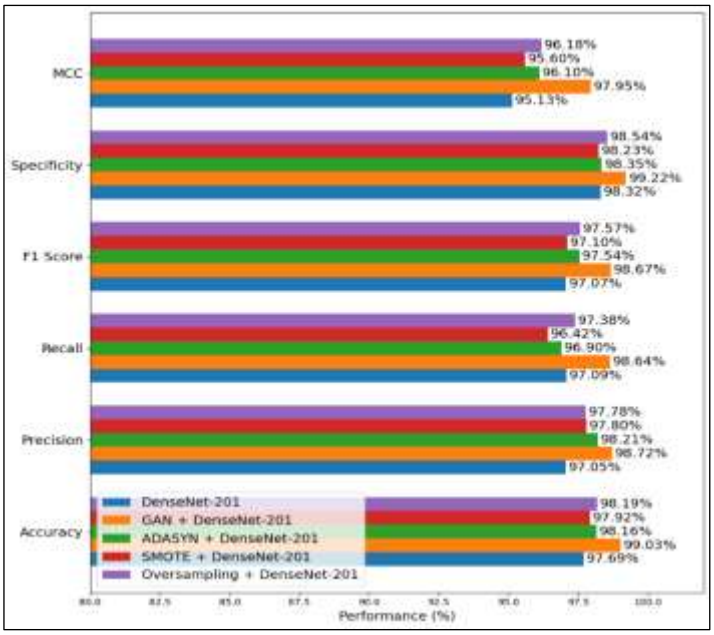

Figure 12: comparison with oversampling methods.
Source: Authors, (2025).

Table 1: Comparison of Classification Performance in Previous Studies vs. Our Proposed GAN-Based Ensemble Approach (DenseNet201 + EfficientNetV2B3).

Study	**Method**	**Accuracy (%)**	**F1-score (%)**
Rajaraman et al. [38]	Ensemble of iteratively pruned deep learning models	99.01	—
Win et al. [28]	Ensemble classifier with voting strategies	99.23	98.30
Verma et al. [39]	Class weights + VGG16	98.00	97.00
Chamseddine et al. [30]	DenseNet201 + SMOTE	98.64	98.63
Mohan et al. [40]	Custom CNN + Data Augmentation	97.00	98.00
Beghoura et al. [41]	Improved CovidConvLSTM	98.78	98.78
Ours	**GAN + Ensemble Learning (DenseNet201 + EfficientNetV2B3)**	**99.33**	**99.01**

Source: Authors, (2025).

A comparative analysis with prior studies tackling class imbalance in medical imaging, summarized in Table 1, further supports the superiority of the proposed method. Previous approaches, such as DenseNet201 + SMOTE (98.64% accuracy, 98.63% F1-score) by Chamseddine et al. [30] and custom CNN + data augmentation (97.00% accuracy, 98.00% F1-score) by Mohan et al. [40], demonstrated moderate success in mitigating class imbalance. However, the proposed GAN+Ensemble Learning (DenseNet201 + EfficientNetV2B3) method outperformed all previous approaches, achieving the highest recorded accuracy 99.33%. This result highlights the effectiveness of GAN-augmented training in improving classification fairness beyond existing methodologies. The findings reinforce that GAN-based augmentation along with ensemble of the soft voting is a powerful tool for addressing class imbalance in deep learning-based medical image classification. While these results are promising, several limitations must be acknowledged. The models were developed and validated using the same publicly available dataset, which may not fully capture the diversity of real-world clinical data. Although the models demonstrated high accuracy on the internal testing set, their performance may not generalize to external datasets, potentially indicating overfitting. Thus, further validation using external, real-time clinical datasets is necessary to assess robustness and

generalizability. Additionally, the dataset lacked clinical metadata such as patient age, gender, and symptoms (e.g., cough, fever), limiting the ability to perform inferential statistical analyses. Incorporating such information in future studies could reduce dependence on extensive labeling and enhance model interpretability. Overall, this study highlights GAN-based augmentation combined with a soft-voting ensemble as a powerful and scalable approach for achieving balanced, high-accuracy performance in medical image classification, offering considerable promise for improved diagnostic efficacy in clinical settings. Enhancing diagnostic performance through the integration of clinical metadata remains a focus of our future research.

V. CONCLUSIONS

This research demonstrated the efficacy of GAN-based data augmentation with ensemble voting of DenseNet201 + EfficientNetV2B3 in mitigating the adverse effects of class imbalance on COVID-19 and Viral pneumonia detection in CXR images. Through CLAHE preprocessing and the generation of high-fidelity synthetic samples, both DenseNet201 and EfficientNetV2-B3 models experienced a marked increase in recall and reduced classification bias. Notably, the soft-voting ensemble combining these two architectures achieved an overall accuracy of 99.33%, an F1 score of 99.01%, and an MCC of 98.59%, underscoring its robust discriminative power. The enhanced performance highlights the capacity of GAN-based approaches to capture complex, clinically relevant features and produce synthetic images that meaningfully expand minority-class representation. However, relying on a single, publicly available dataset may limit the generalizability of these findings, and potential constraints arise from the computational demands associated with both GAN training and ensemble methodologies. Extending this approach to diverse, multi-institutional datasets could further validate its robustness while exploring more sophisticated generative models such as StyleGAN or CycleGAN may enhance image fidelity and variability. Moreover, self-supervised or semi-supervised strategies could reduce dependence on extensive labeling.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Aissa Snani, Mohammed Tarek Khadir.
Methodology: Aissa Snani.
Investigation: Aissa Snani.
Discussion of results: Aissa Snani, Mohammed Tarek Khadir and Modawy Adam Ali Abdalla.
Writing – Original Draft: Aissa Snani.
Writing – Review and Editing: Aissa Snani, Mohammed Tarek Khadir and Modawy Adam Ali Abdalla.
Resources: Aissa Snani.
Supervision: Mohammed Tarek Khadir and Modawy Adam Ali Abdalla.
Approval of the final text: Aissa Snani, Modawy Adam Ali Abdalla.

VII. REFERENCES

[1] S. Dawkes and M. O'Reilly, "Chest x-ray interpretation," *British Journal of Cardiac Nursing*, vol. 14, no. 5, pp. 1–9, 2019.

[2] I. D. Mienye and Y. Sun, "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data," *Informatics in Medicine Unlocked*, vol. 25, p. 100690, 2021.

[3] M. Wei, Y. Zhou, Z. Li, and X. Xu, "Class-imbalanced complementary-label learning via weighted loss," *Neural Networks*, vol. 166, pp. 555–565, 2023.

[4] Y. E. I. El-Bouzaidi and O. Abdoun, "Advances in artificial intelligence for accurate and timely diagnosis of COVID-19: A comprehensive review of medical imaging analysis," *Scientific African*, p. e01961, 2023.

[5] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI), 2017, pp. 79–85.

[6] X. Zhao et al., "Characteristics and clinical value of chest CT images in COVID-19 pneumonia," Clinical Radiology, vol. 75, no. 5, pp. 335–340, 2020.

[7] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*, vol. 135, pp. 32–41, 2014.

[8] A. Snani, M. T. Khadir, A. Pranolo, and M. A. A. Abdalla, "GAN-enhanced multimodal fusion and ensemble learning for imbalanced chest X-ray classification," Int. J. Adv. Intell. Informat., vol. 11, no. 3, pp. 514–532, 2025.

[9] A. Onan, "Consensus clustering-based undersampling approach to imbalanced learning," Scientific Programming, vol. 2019, 2019.

[10] N. A. A. Majrashi, "The value of chest x-ray and ct severity scoring systems in the diagnosis of COVID-19: A review," *Frontiers in Medicine*, vol. 9, p. 1076184, 2023.

[11] M. Churruca, E. Martínez-Besteiro, F. Couñago, and P. Landete, "COVID-19 pneumonia: A review of typical radiological characteristics," *World Journal of Radiology*, vol. 13, no. 10, p. 327, 2021.

[12] A. Rehman, M. A. Iqbal, H. Xing, and I. Ahmed, "COVID-19 detection empowered with machine learning and deep learning techniques: A systematic review," *Applied Sciences*, vol. 11, no. 8, p. 3414, 2021.

[13] M. Ozsoz, A. U. Ibrahim, S. Serte, F. Al-Turjman, and P. S. Yakoi, "Viral and bacterial pneumonia detection using artificial intelligence in the era of COVID-19," Res. Sq., 2021.

[14] A. Mangal et al., "COVIDAID: COVID-19 detection using chest X-ray," arXiv:2004.09803, 2020.

[15] M. Mahin, S. Tonmoy, R. Islam, T. Tazin, M. Monirujjaman Khan, S. Bourouis, *et al.*, "Classification of COVID-19 and pneumonia using deep transfer learning," *Journal of Healthcare Engineering*, vol. 2021, 2021.

[16] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest x-ray images using detrac deep convolutional neural network," *Applied Intelligence*, vol. 51, pp. 854–864, 2021.

[17] A. Tiwari, T. S. Sharan, S. Sharma, and N. Sharma, "Deep learning-based automated multiclass classification of chest X-rays into COVID-19, normal, bacterial pneumonia and viral pneumonia," *Cogent Engineering*, vol. 9, no. 1, p. 2105559, 2022.

[18] A. Bashar, G. Latif, G. Ben Brahim, N. Mohammad, and J. Alghazo, "COVID-19 pneumonia detection using optimized deep learning techniques," *Diagnostics*, vol. 11, no. 11, p. 1972, 2021.

[19] Nillmani et al., "Four types of multiclass frameworks for pneumonia classification and validation in X-ray scans using deep learning models," Diagnostics, vol. 12, no. 3, p. 652, 2022.

[20] W. Jin, S. Dong, C. Dong, and X. Ye, "Hybrid ensemble model for differential diagnosis between COVID-19 and common viral pneumonia by chest X-ray radiograph," Comput. Biol. Med., vol. 131, p. 104252, 2021.

[21] V. Gupta et al., "Improved COVID-19 detection with chest X-ray images using deep learning," Multimed. Tools Appl., vol. 81, no. 26, pp. 37657–37680, 2022.

[22] K. Kc, Z. Yin, M. Wu, and Z. Wu, "Evaluation of deep learning-based approaches for COVID-19 classification based on chest X-ray images," Signal Image Video Process., vol. 15, pp. 959–966, 2021.

[23] I. D. Apostolopoulos and T. A. Mpesiana, "COVID-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," Phys. Eng. Sci. Med., vol. 43, pp. 635–640, 2020.

[24] A. I. Khan, J. L. Shah, and M. M. Bhat, "Coronet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Computer methods and programs in biomedicine*, vol. 196, p. 105581, 2020.

[25] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with x-ray images," *Computers in biology and medicine*, vol. 121, p. 103792, 2020.

[26] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 7263–7271.

[27] R. S. Alharbi, H. A. Alsaadi, S. Manimurugan, T. Anitha, and M. Dejene, "Multiclass classification for detection of COVID-19 infection in chest x-rays using cnn," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[28] K. Y. Win, N. Maneerat, S. Sreng, and K. Hamamoto, "Ensemble deep learning for the detection of COVID-19 in unbalanced chest x-ray dataset," *Applied Sciences*, vol. 11, no. 22, p. 10528, 2021.

[29] M. T. Naseem, T. Hussain, C.-S. Lee, and M. A. Khan, "Classification and detection of COVID-19 and other chest-related diseases using transfer learning," *Sensors*, vol. 22, no. 20, p. 7977, 2022.

[30] E. Chamseddine, N. Mansouri, M. Soui, and M. Abed,"Handling class imbalance in COVID-19 chest x-ray images classification: Using smote and weighted loss,"Applied Soft Computing, vol. 129, p. 109588, 2022.

[31] I. Goodfellow et al., "Generative adversarial networks," Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.

[32] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in Int. Conf. Mach. Learn. (ICML), pp. 214–223, PMLR, 2017.

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Adv. Neural Inf. Process. Syst.* (NeurIPS), vol. 30, 2017, pp. 5767–5777.

[34] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), Honolulu, HI, USA, 2017, pp. 4700–4708.

[35] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.* (ICML), vol. 139, 2021, pp. 10096–10106.

[36] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Amsterdam, The Netherlands: Elsevier, 2011.

[37] P. Musa, F. Al Rafi, and M. Lamsani, "A review: Contrast-limited adaptive histogram equalization (CLAHE) methods to help the application of face recognition," in *Proc. 3rd Int. Conf. Informatics Comput.* (ICIC), Palembang, Indonesia, 2018, pp. 1–6.

[38] S. Rajaraman, J. Siegelman, P. Alderson, L. Folio, and S. Antani, "Iteratively pruned deep learning ensembles for COVID-19 detection in chest x-rays," *IEEE Access*, vol. 8, pp. 115041–115050, 2020.

[39] D. K. Verma, G. Saxena, A. Paraye, A. Rajan, A. Rawat, and R. K. Verma, "Classifying COVID-19 and viral pneumonia lung infections through deep convolutional neural network model using chest x-ray images," *Journal of Medical Physics*, vol. 47, no. 1, pp. 57–64, 2022.

[40] G. Mohan, M. M. Subashini, S. Balan, and S. Singh, "A multiclass deep learning algorithm for healthy lung, COVID-19, and pneumonia disease detection from chest x-ray images," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 20, 2024.

[41] I. Beghoura, M. Benssalah, and F. Sbargoud, "An improved CovidConvLSTM model for pneumonia–COVID-19 detection and classification," Biomed. Eng.: Appl., Basis Commun., vol. 37, no. 3, p. 2550019, 2025.