GAN-Enhanced multimodal fusion and ensemble learning for imbalanced chest X-Ray classification

Aissa Snani [a,1,*], Mohammed Tarek Khadir [a,2], Andri Pranolo [b,3], Modawy Adam Ali Abdalla [c,4]

[a] LABGED Laboratory, Computer Sciences Department, Badji Mokhtar University of Annaba, Annaba 23000, Algeria
[b] Department of Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta 55166, Indonesia
[c] Department of Electrical and Electronic Engineering, College of Engineering Science, Nyala University, Nyala 63311, Sudan
[1] aissa.snani@univ-annaba.dz; [2] khadir@labged.net; [3] andri.pranolo@tif.uad.ac.id; [4] brojacter88@yahoo.com
* corresponding author

ARTICLE INFO

Article history
Received May 9, 2025
Revised June 12, 2025
Accepted July 2, 2025
Available online August 31, 2025

Keywords
Class imbalanced
GAN-based augmentation
Multimodal fusion
Multiclass CXR classification
Ensemble learning

ABSTRACT

Chest X-ray (CXR) classification tasks often suffer from severe class imbalance, resulting in biased predictions and suboptimal diagnostic performance. To address this challenge, we propose an integrated framework that combines high-fidelity data augmentation using Generative Adversarial Networks (GANs), ensemble learning via hard and soft voting, and multimodal feature fusion. The method begins by partitioning the majority class into multiple subsets, which are individually balanced through GAN-generated synthetic images. Deep learning models, specifically DenseNet201 and EfficientNetV2B3, are trained separately on each balanced subset. These models are then combined using ensemble voting to improve robustness. Additionally, features extracted from the most performant models are fused and used to train traditional classifiers such as Logistic Regression, Multilayer Perceptron, CatBoost, and XGBoost. Evaluations on a publicly available CXR dataset demonstrate consistent improvements across key metrics, including accuracy, precision, recall, F1-score, AUROC, AUPRC, MCC, and G-mean. This framework shows superior performance in multiclass scenarios.

1. Introduction

The advancement of machine learning (ML), and particularly deep learning (DL), has significantly transformed clinical decision-making by automating diagnostic processes and improving the accuracy and efficiency of medical assessments. DL models trained on large-scale datasets have been successfully applied in domains such as cardiology and oncology, facilitating early detection and enabling personalized healthcare strategies [1]. These models excel at analyzing complex, high-dimensional medical data, thereby supporting informed and timely clinical decisions [2], [3].

However, a persistent challenge in medical data analysis is class imbalance, as the underrepresentation of certain disease categories often results in biased model predictions and poor generalizability [4]. Furthermore, obtaining large, high-quality annotated medical datasets is resource-intensive and depends heavily on expert annotation [5], posing barriers to the development of robust disease classification systems [6]. Although progress has been made in medical imaging, traditional techniques such as resampling or basic data augmentation often fail to address imbalance, particularly with high-dimensional clinical images, effectively. These methods tend to simplify data distributions and generate unrepresentative samples for minority classes [7]. In contrast, recent advances in GANs have shown

promise in producing realistic synthetic data that enhances minority class representation and improves model performance [8], [9]. Yet, many approaches using GANs focus exclusively on data generation without integrating them with advanced classification strategies such as ensemble learning or multimodal feature fusion, both of which are essential for improving model robustness and generalizability. To overcome these limitations, we propose a novel framework that combines GAN-based augmentation with ensemble voting and improved feature fusion. The dataset is divided into three subsets, where GANs generate additional samples for the minority classes. We implement both soft and hard voting strategies to improve decision robustness and introduce a multimodal feature fusion mechanism that integrates learned representations from multiple submodels into a unified embedding for final classification. This approach is evaluated on a CXR dataset including COVID-19, Viral Pneumonia, and Normal cases. The results reveal substantial improvements in class balance and diagnostic accuracy, confirming the effectiveness of combining synthetic data generation with ensemble classification and feature integration. To the best of our knowledge, this is the first study to integrate GAN-based data augmentation with ensemble feature fusion for multiclass CXR classification. The main contributions of this work are summarized as follows:

- A new method employing GANs to generate realistic synthetic data for minority classes, effectively reducing class imbalance.
- State-of-the-art models (e.g., EfficientNet and DenseNet) for high-quality feature extraction and classification in imbalanced datasets.
- Integration of multiple submodels in an ensemble framework, leveraging their strengths to improve performance and robustness.
- Outputs from multiple submodels are combined into a unified feature representation, allowing diverse models to complement each other in medical imaging tasks.
- A comprehensive case study on CXR images demonstrates the effectiveness of the proposed framework in enhancing diagnostic accuracy across imbalanced classes.

Previous studies have explored multiple strategies to address class imbalance in medical imaging. Malygina *et al.* and Qin *et al.* [10], [11] leveraged GANs to generate synthetic samples, which improved accuracy but risked lacking the morphological diversity essential in clinical data, potentially limiting generalizability. Kothawade *et al.* [12] proposed a submodular mutual information-based active learning framework to enhance minority-class representation, although it is annotation-intensive and may overlook subtle disease progression. Yeung *et al.* [13] introduced the Unified Focal Loss to penalize misclassification of the minority class, showing promise in binary tasks; however, its effectiveness in complex, noisy multiclass CXR datasets remains uncertain. Fan and Bu [14] adopted transfer learning using ImageNet-pretrained models for lung disease classification, yet domain mismatch between natural and medical images may prevent optimal feature learning, especially in diverse clinical environments. Beyond single-model approaches, ensemble learning has shown strong potential for improving robustness and accuracy. For instance, Jangam *et al.* [15] achieved high performance using stacking ensembles, though such methods are often computationally intensive and less suitable for resource-constrained environments. Kaleem [16] enhanced multiclass detection through advanced ensemble architectures, yet interpretability remains limited due to increased complexity. Habib *et al.* [17] combined CheXNet and VGG-19 with oversampling for binary classification, showing gains but limited flexibility. In contrast, our work introduces a dual-strategy framework that fuses GAN-based augmentation with ensemble voting and feature fusion, aiming to tackle imbalance more comprehensively while enhancing transparency and modularity. Unlike prior efforts focused on isolated solutions, our framework is explicitly designed for multiclass scenarios and scalable clinical integration.

2. Method

2.1. Residual Block

Neural network performance is generally proportional to its depth, as deeper architectures tend to achieve superior feature representation and improved accuracy. However, excessively deep networks often encounter the challenge of gradient degradation, where performance initially improves but subsequently

declines due to vanishing or exploding gradients. To mitigate this issue, ResNet [18] introduces residual blocks, which incorporate identity mappings to facilitate stable feature propagation throughout the network. A key component of this architecture is the convolutional layer within each residual block, which plays a crucial role in feature extraction. The mathematical operation governing this convolutional process is formally defined by Eqs (1) and (2).

$$y_l^j = f(z_l^j) \tag{1}$$

$$z_l^j = \sum_{i \in M_j} x_{l-1}^i \times k_{l,ij} + b_l^j \tag{2}$$

where z_l^j represents the output feature map at layer l, f(·) is the ReLU activation function, x_{l-1}^i is the feature map from the (l-1)th layer, Mj denotes a subset of the input feature map, and $k_{l,ij}$ is the 3 × 3 convolutional kernel matrix at layer l, applied without a bias term. The convolutional operation is followed by batch normalization and the ReLU activation function. The architecture and functionality of the residual block, specifically its single convolutional layer, are illustrated in Fig. 1(a).

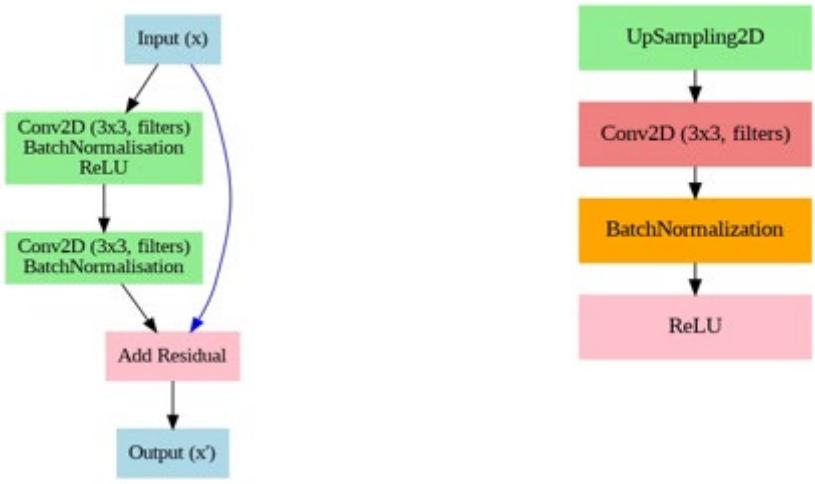


Fig. 1. Illustration of key network components: (a) Residual Block, and (b) Upsampling Block

2.2. Upsampling Block

The upsampling block is responsible for enhancing the spatial dimensions of the feature map, playing a crucial role in generating high-resolution output images from low-resolution inputs. It begins with an UpSampling2D layer, which expands the feature map by replicating rows and columns using a form of nearest-neighbor interpolation. This is followed by a 3 × 3 convolutional layer, which integrates and smooths the feature map. The convolved feature map then undergoes batch normalization, which stabilizes training by normalizing activations. Finally, a ReLU activation function is applied to introduce nonlinearity. Nonlinearity is illustrated in this upsampling module of Fig. 1(b).

2.3. Generative Adversarial Networks And Proposed Architecture

GANs, a prominent sub-category of generative models, were first introduced by Goodfellow *et al.* [19]. A GAN framework consists of two primary components: a generator and a discriminator (also referred to as a critic). The generator is trained to synthesize realistic fake images that can deceive the discriminator, while the discriminator is trained to distinguish between real and generated (fake) images accurately. This adversarial training continues until an equilibrium is reached, at which point the generator and the discriminator perform optimally.

The proposed GAN architecture, depicted in Fig. 2, is designed to produce high-resolution, realistic, and diverse data samples for augmentation. The generator takes a random noise vector of size 128 and projects it through a dense layer to produce a small 4×4 feature map. This map is progressively upsampled,

first to 8×8, then to 16×16, 32×32, 64×64, and finally 128×128. At each stage, convolutional layers refine the features, followed by batch normalization and ReLU activation, with the number of channels gradually decreasing from 256 down to 16 to ensure a smoother, higher-quality output. Ultimately, two ResNet blocks are applied to refine the image's structure and texture further, aided by skip connections that help stabilize the training process.

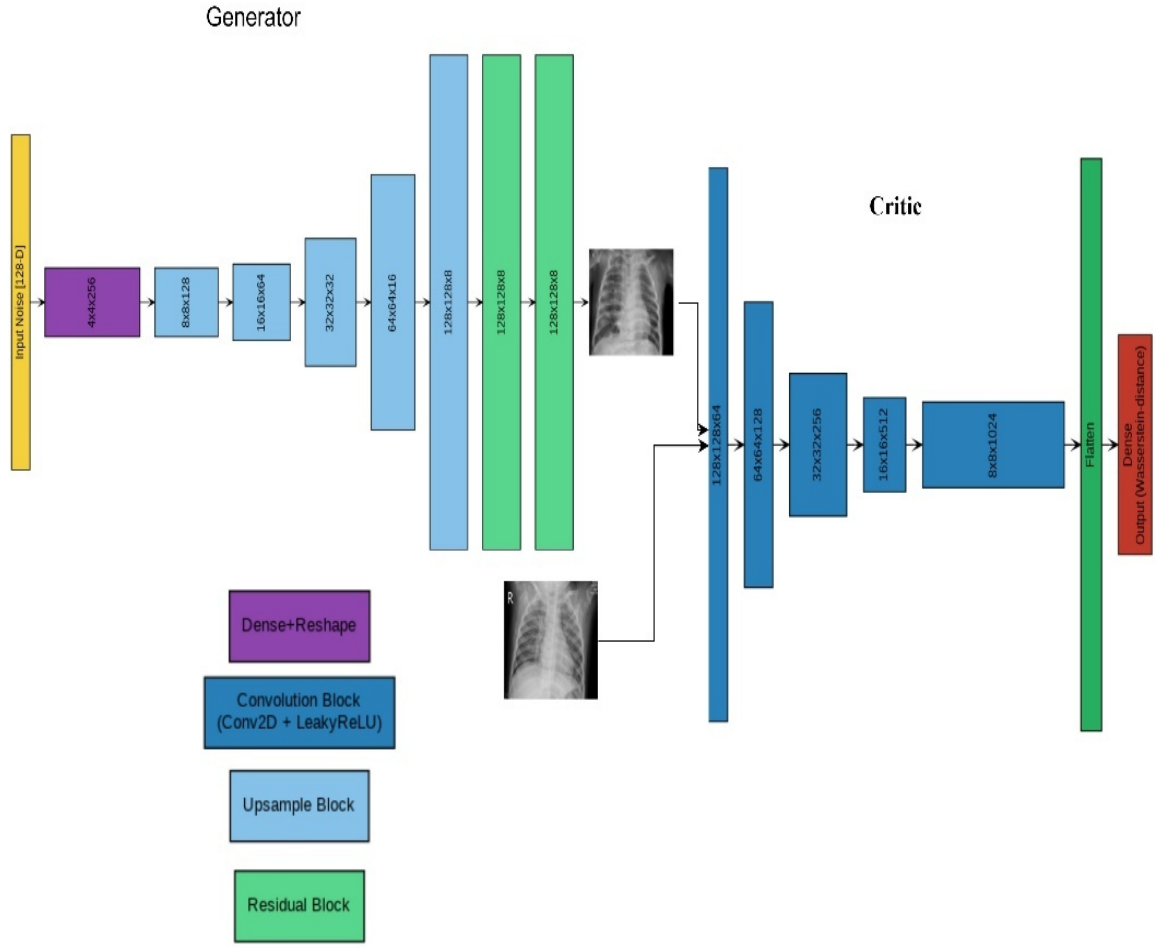


Fig. 2. Proposed generator and critic architecture for generating CXR images

The final output layer produces a single-channel 128×128 image. In contrast, the critic performs a downsampling operation to evaluate the authenticity of images. Beginning with an input image of size 128×128, it applies successive convolutional layers that progressively reduce the spatial resolution to 64×64, 32×32, 16×16, and finally, 8×8, while increasing the depth of the feature maps from 64 to 1024. The flattened output is passed through a dense layer to yield a scalar that approximates the Wasserstein distance, which is used to assess the quality of generated images. Minimizing the distance between the real distribution and the generated distribution, as formulated in [20], can be unstable. To stabilize training, the WGAN-GP loss [21] is adopted. Let G be the generator and D the critic; if the input image is x, then the output of D on x is D(x). The generator G receives a noise vector z ~ pz (e.g., from a standard normal or uniform distribution) and outputs a generated image xN = G(z). Real images follow the distribution pr. Instead of the Jensen-Shannon divergence, WGAN-GP relies on the Wasserstein-1 distance, which is more robust for training when pr (real distribution) and pg (generated distribution) lie on low-dimensional manifolds. The WGAN-GP framework introduces a gradient penalty to enforce soft Lipschitz constraints on D. Specifically, the generator and critic losses are defined by Eqs. (3) and (4).

$$L_G(x_N) = \min_G \left(-E_{x_N \sim p_g}[D(x_N)]\right) \tag{3}$$

$$L_D(x; x_N; x_O) = \min_D \left(E_{x_N \sim p_g}[D(x_N)] - E_{x \sim p_r}[D(x)] + \lambda E_{x_O \sim p_{x_O}} \left[\left(|\nabla_{x_O} D(x_O)|_2 - 1 \right)^2 \right] \right) \quad (4)$$

where $\lambda > 0$ is a balancing coefficient, and x_O is a linear interpolation between the real image x and x_N is the generated image.

$$x_O = \alpha x + (1 - \alpha) x_N, \ \alpha \sim U[0,1] \quad (5)$$

2.4. Deep Learning Models

We employed two pre-trained deep learning models: DenseNet201 and EfficientNetV2B3. DenseNet201, a densely connected convolutional network, enhances feature propagation and reuse by creating direct connections between each layer and all preceding layers within dense blocks, thereby improving gradient flow and parameter efficiency [22]. EfficientNetV2B3, a scalable model, applies a compound scaling method to balance depth, width, and resolution, and integrates fused MBConv layers to boost training speed and computational efficiency [23]. Both models effectively extract hierarchical features, capturing low-level patterns such as edges and textures while progressively learning high-level semantic representations capabilities essential for tackling the complexities of medical imaging tasks [24]. To adapt these models to our task, we removed their original classification layers and appended custom task-specific layers. These included a Global Average Pooling (GAP) layer for dimensionality reduction, a Dropout layer to reduce overfitting [25], a Dense layer with ReLU activation for feature transformation, and a final Dense layer with softmax activation for classification. We trained the modified architecture end-to-end on our dataset. We adopted a fine-tuning strategy by initializing training with a low learning rate, which allowed the pre-trained weights to adapt gradually while optimizing the newly added layers

2.5. Machine Learning Models

We integrated machine learning (ML) and deep learning (DL) approaches to leverage their complementary strengths in medical image analysis. Traditional ML algorithms such as Logistic Regression (LR), Multilayer Perceptron (MLP), Categorical Boosting (CatBoost), and Extreme Gradient Boosting (XGBoost) performed effectively on structured datasets, particularly when provided with high-quality feature representations. Below, we briefly describe the methods we adopted in this study, including those used for comparative analysis.

- LR: Logistic Regression models the conditional probability for a class k given an input vector $x \in R^d$ using the softmax function:

$$P(y = k \mid x) = \frac{\exp(w_k^T x + b_k)}{\sum_{j=1}^{K} \exp(w_j^T x + b_j)} \quad (6)$$

  where wk ∈ R^d and bk ∈ R are the weight vector and bias for class k, and K is the number of classes. DL features replace the input x with a rich, high-dimensional representation $\phi(x)$ extracted from a pre-trained network.

- MLP: An MLP is a feed-forward neural network that models complex relationships by composing multiple nonlinear transformations. For a network with L layers, the transformation in layer l is given by.

$$a^{(l)} = \sigma\left(W^{(l)} a^{(l-1)} + b^{(l)}\right), \quad l = 1,2,\ldots,L \quad (7)$$

  where $a^{(0)}$ = x (the input), $W^{(l)}$ is the weight matrix, $b^{(l)}$ is the bias vector, and $\sigma^{(.)}$ is a nonlinear activation function (e.g., ReLU, Sigmoid, or Tanh).

- CatBoost: CatBoost is a gradient-boosted decision tree (GBDT) algorithm optimized for handling categorical features. Its objective function combines a loss term with a regularization term.

$$\mathcal{L} = \sum_{i=1}^{N} l\left(y_i, f(x_i)\right) + \Omega(f) \quad (8)$$

where ℓ is a loss function (e.g., cross-entropy or mean squared error), f(xi) is the prediction, for instance, i, and Ω(f) penalizes model complexity. A distinctive feature of CatBoost is its handling of categorical data via target-based statistics, which transform categorical variables c into numerical features:

$$\tilde{c} = \frac{\sum_{j \in \mathcal{J}(c)} y_j + a}{|\mathcal{J}c(c)| + b} \tag{9}$$

where $(\mathcal{J}(c))$ is the set of indices corresponding to category $c, a, and\ b$ are smoothing parameters.

Additionally, CatBoost employs an ordered boosting algorithm to minimize prediction shift by ensuring that the model only utilizes past data to predict future data, thereby effectively mitigating overfitting.

- XGBoost: XGBoost is another gradient-boosting framework that sequentially builds an ensemble of decision trees [26]–[28]. The model's objective is formulated as:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{10}$$

where $(\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x_i}))$ is the ensemble prediction, fk represents an individual tree, and $(\Omega(f_k) = \gamma T + \frac{1}{2}\lambda|\boldsymbol{w}|^2)$ is a regularization term that penalizes the complexity of the tree (with T being the number of leaves and γ and λ being regularization hyperparameters). In multiclass settings, the softmax function is employed for probability estimates, and the gradient and Hessian for each tree are computed to perform a second-order Taylor expansion of the loss function, enabling efficient optimization:

$$\mathcal{L} \approx \sum_{i=1}^{N} \left[g_i f(x_i) + \frac{1}{2} h_i f(x_i)^2 \right] + \Omega(f) \tag{11}$$

where (g_i) and (h_i) are the first and second derivatives of the loss with respect to the prediction $(\hat{y}_i)$. XGBoost's ability to handle class imbalance and its support for parallel tree construction make it a robust choice in various complex scenarios.

2.6. Ensemble Voting

Ensemble learning is a robust ML methodology that enhances predictive accuracy by integrating multiple models, commonly referred to as base classifiers. The fundamental idea is to leverage the collective decision-making of these models to produce a more reliable and accurate outcome than any individual model could achieve. Two prominent approaches are used: hard voting and soft voting. The first method, hard voting, determines the final class label $(\hat{y})$ based on the class (k) that receives the maximum number of votes among all models. Mathematically, this can be expressed as:

$$\hat{y} = \arg\max_k \sum_{i=1}^{N} I(\hat{y}_i = k) \tag{12}$$

where $(I(\hat{y}_i = k))$ is an indicator function that equals 1 if the (i) −th model predicts the class (k), and 0 otherwise. Here, (N) represents the total number of models in the ensemble. Hard voting relies on plurality, selecting the class that garners the most support among the base classifiers. The second method, soft voting, utilizes the predicted probabilities provided by the base classifiers. Instead of considering only the most frequently predicted class, soft voting aggregates the probability distributions of all models, allowing more confident predictions to have greater influence. The final class label $(\hat{y})$ is derived by summing the predicted probabilities $(P_{i,k})$ for each class (k) across all (N) models and selecting the class with the highest aggregated probability:

$$\hat{y} = \arg\max_k \sum_{i=1}^{N} P_{i,k} \tag{13}$$

where $(P_{i,k})$ denotes the predicted probability of class (k) by the (i) −th model, and (N) represents the total number of models in the ensemble.

3. Experiment And Proposed Pipelines

All experiments were executed in Python. Image preprocessing tasks, including resizing and dataset partitioning, were executed locally on a Windows 10 system featuring an Intel Core i7-12700H CPU (4.7 GHz maximum), 32 GB of DDR5 RAM, and an NVIDIA RTX A2000 GPU (8 GB VRAM). Compute-intensive phases, encompassing GAN training and subsequent model assessment, were executed in Kaggle's cloud computing platform, which offers an NVIDIA Tesla P100 GPU via TensorFlow and Keras. Fig. 3 provides a comprehensive overview of the implemented methodology, encompassing data preprocessing, class balancing, model training, and feature fusion. The pipeline starts with raw CXR image acquisition and proceeds through five key phases: (1) dataset preprocessing, (2) data division and GAN balancing, (3) training the classifier models, (4) ensemble learning (Approach one), and (5) multimodal feature fusion (Approach two).

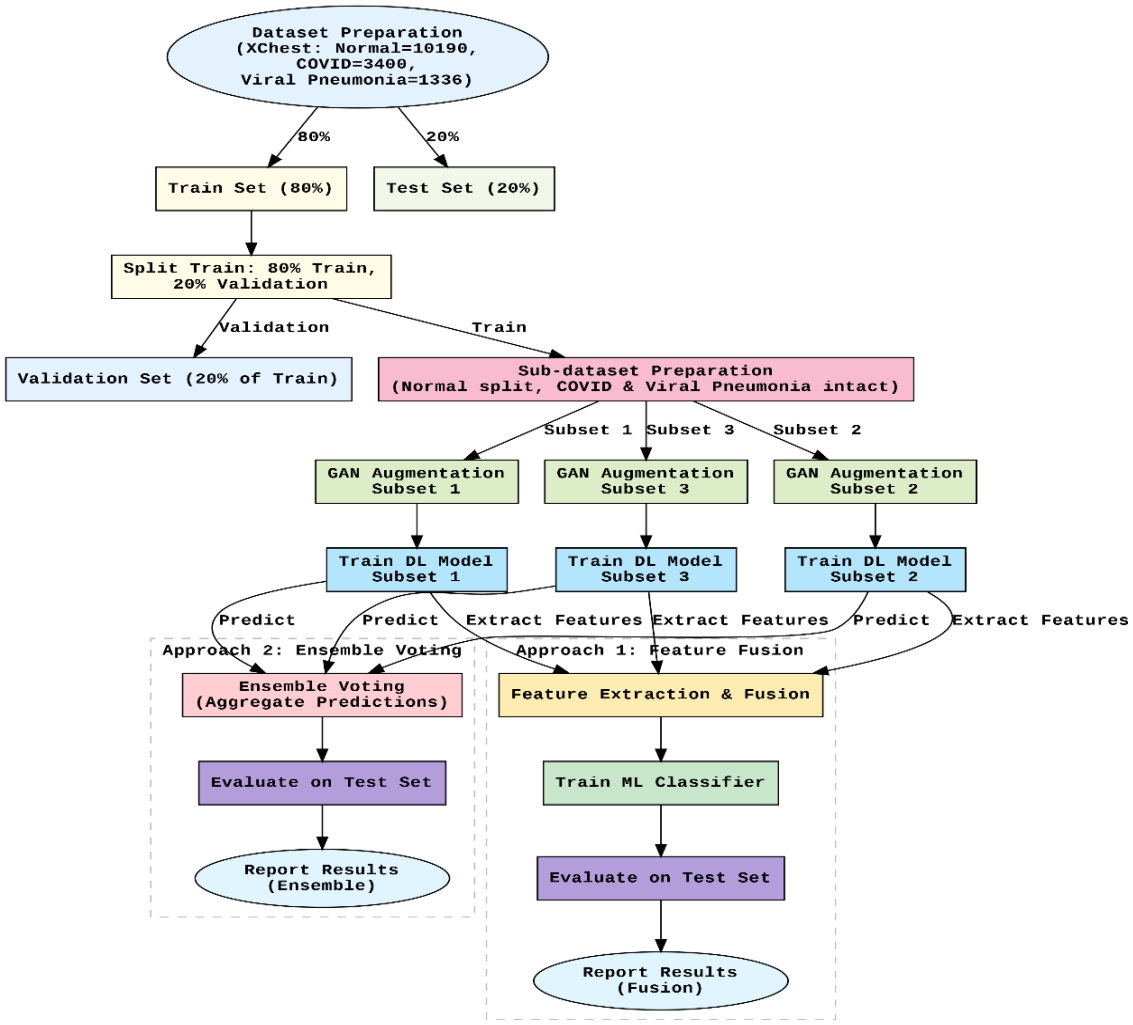


Fig. 3. Overview of the proposed methodology for imbalanced CXR classification. The pipeline begins with GAN-based data balancing, followed by two classification approaches incorporating multimodal feature fusion and ensemble learning

3.1. Dataset Description And Preprocessing

We utilized a publicly available dataset from the Kaggle platform (https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database), which consists of CXR images categorized into three classes: COVID-19, Normal, and Viral Pneumonia. The original dataset contained 3,616 COVID-19 images, 10,192 Normal images, and 1,345 Viral Pneumonia images, offering both diversity and scale. However, we observed a notable class imbalance, with a significantly higher number of Normal cases compared to the other two categories. To reduce computational

overhead, we resized the original CXR images (299 × 299 pixels) to 128 × 128 during data preprocessing. We then used the ImageHash library to detect and remove duplicate images, thereby improving data quality and minimizing redundancy. After deduplication, the final class distribution comprised 3,400 COVID-19 images, 10,190 Normal images, and 1,337 Viral Pneumonia images. We split the dataset into training (80%) and testing (20%) subsets. We further divided the training set into training and validation splits using an 80:20 ratio.

3.2. Data Division And GAN Balancing

To address class imbalance in the training dataset, we randomly divided the majority class (Normal) into three partitions. If a subset lacked samples for certain classes, we supplemented it with additional samples from the other partitions to match the number of COVID-19 samples (2,176). The Viral Pneumonia and COVID-19 classes were kept unchanged across all subsets to maintain consistency. This approach resulted in more balanced training subsets for model development. To further enhance balance, we used the GAN architecture from Section 3.3 to generate Viral Pneumonia images, ensuring equal representation of all classes in each subset. We optimized the parameters for the WGAN-GP loss functions in Eq. (3) and (4) through iterative training and parameter selection. During GAN training, we drew the generator's input from a standard normal distribution to produce grayscale images of size 128 × 128. We scaled both real and generated samples to the range of [-1, 1] before passing them to the discriminator. We used a Leaky-ReLU activation function with a slope of 0.2 to alleviate potential "dying ReLU" issues. We initialized the weights using a normal distribution and applied normalization to ensure stable training. We applied the Adam optimizer [26], with $\beta 1 = 0.0$, $\beta 2 = 0.99$, and a gradient penalty coefficient $\lambda = 10$, consistent with established best practices [20]. We used a slower learning rate of 0.0002 to ensure gradual and stable convergence throughout the training process. We trained the model for 1000 epochs with a batch size of 64, providing sufficient iterations for the GAN to converge effectively without overfitting. After training, we employed the generator to synthesize Viral Pneumonia images, which we then used to augment and balance each subdataset. Fig. 4 shows the distribution of the original training dataset after division and subsequent GAN-based balancing.

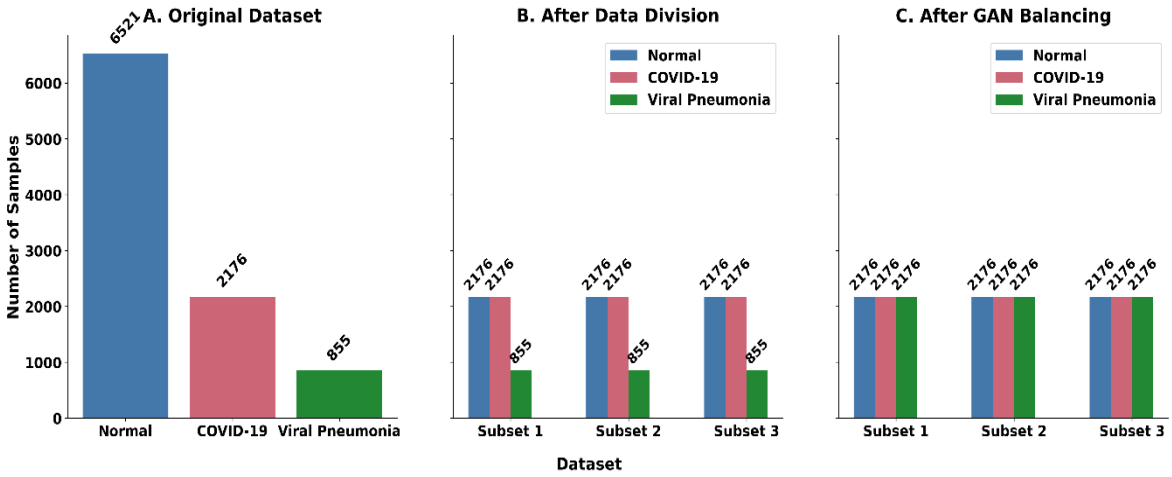


Fig. 4. Progression of data balancing strategies Left: Original dataset with an imbalanced class distribution (Normal: 6521, COVID-19: 2176, Viral Pneumonia: 855). Middle: After data division, the majority class (Normal) is split into three subsets, each containing 2176 samples, while COVID-19 and Viral Pneumonia remain unchanged (2176 and 855, respectively). Right: After GAN balancing, the Viral Pneumonia class is augmented to 2176 samples per subdataset

3.3. Training And Classifier Models

We fine-tuned two CNN architectures: D_Net (DenseNet201) and E_Net (EfficientNetV2B3), for 60 epochs using a batch size of 64, following the methodology described in Section 3.4. We initialized the learning rate at 1×10^{-3} and progressively decreased it at epochs 20, 30, and 40 by factors of 0.1, 0.01, and 0.001, respectively. After epoch 50, we stabilized the learning rate at 0.5×10^{-3}. To dynamically adjust the learning rate, we used a ReduceLROnPlateau callback, which applied a 31.6% reduction (a factor of

0.1) after five epochs without improvement. No cooldown period was used, and we set the minimum learning rate to 0.5×10^{-6}. We employed the Adam optimizer throughout the training process. We first trained the models on the original (imbalanced) dataset to obtain the baseline models: DNet_Orig and ENet_Orig. Then, we trained on three split subsets: Non-GAN-balanced subsets (DNet_S1, DNet_S2, DNet_S3 for D_Net and ENet_S1, ENet_S2, ENet_S3 for E_Net) and GAN-balanced subsets (DNet_GS1, DNet_GS2, DNet_GS3 for D_Net and ENet_GS1, ENet_GS2, ENet_GS3 for E_Net).

3.3.1. Ensemble Voting (Approach One)

After training the classifier models, we utilized ensemble voting on the three sub-trained models to improve overall performance, as detailed in Section 3.6. We used Eq. (12) for Soft Voting (SV) and Eq. (13) for Hard Voting (HV). Ensemble voting was applied to two types of subsets: GSV and GHV denote ensembles of sub-models trained on GAN-based data augmentation, where DNet_GSV and DNet_GHV refer to the DNet model, and ENet_GSV and ENet_GHV refer to the E_Net model. For models trained on non-GAN subsets, which utilize the original imbalanced data, the ensembles include DNet_SV and DNet_HV for D_Net, as well as ENet_SV and ENet_HV for E_Net.

3.3.2. Training On Multimodal Feature Fusion (Approach Two)

Feature-level fusion was explored by extracting vectors from the global pooling layers of trained submodels (both GAN-balanced and non-GAN). These vectors, which capture salient image representations, were concatenated into a single feature vector:

$$F_{fused} = F_1 \oplus F_2 \oplus F_3 \tag{14}$$

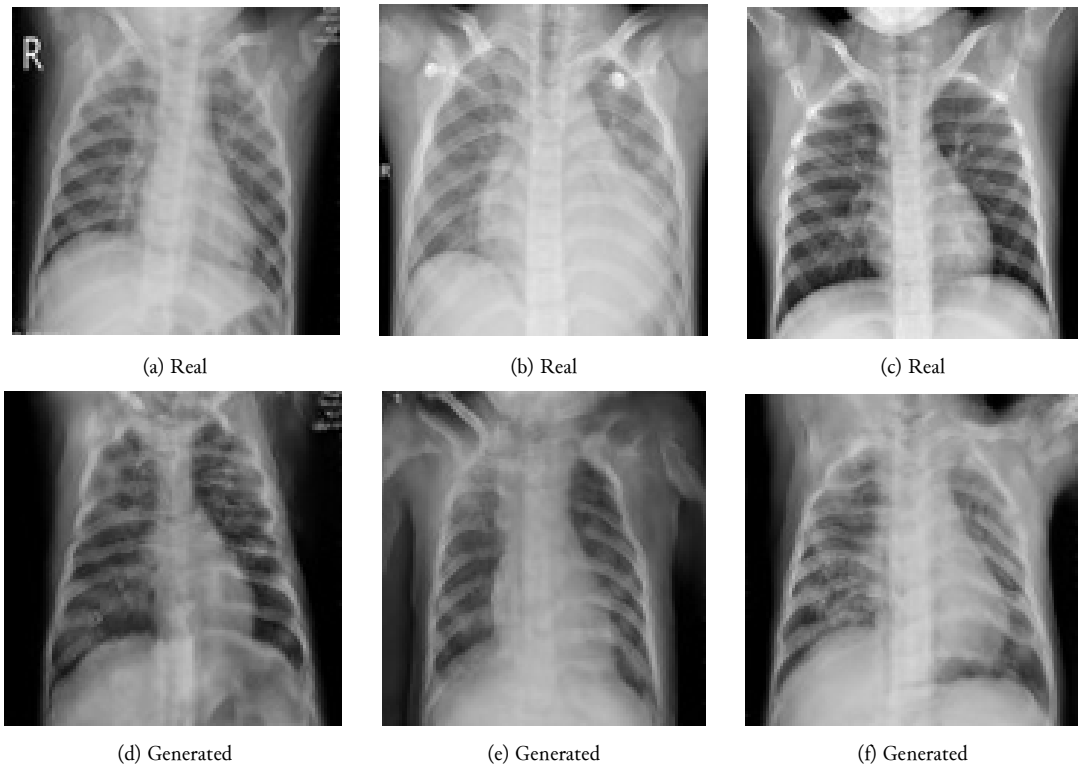
where F_1, F_2, and F_3 represent feature vectors obtained from three submodels, and $\oplus$ denotes concatenation. A final standardization step ensured the uniform contribution of all features before classification.

$$F_{scaled} = \frac{F_{fused}-\mu}{\sigma} \tag{15}$$

We applied feature fusion to the optimal sub-dataset for this method, which we selected based on validation accuracy. The extracted features were concatenated and standardized using Eqs (14) and (15). These processed features were then used to train four ML classifiers, implemented with the scikit-learn library using default parameters, except for specific configurations we trained LR for up to 1000 iterations to ensure convergence; we configured XGBoost for multiclass classification with objective="multi: softmax" and num_class=3; we defined the MLP as a neural network with two hidden layers (128 and 64 neurons), ReLU activation, and trained it for 500 iterations; and we trained CatBoost with 500 boosting iterations, a learning rate of 0.1, and a maximum tree depth of 6. We duplicated the entire pipeline across DNet, ENet, and their hybrid variants: DNet_SF, ENet_SF, DNet_GSF, and ENet_GSF, where SF denotes submodel feature fusion based on submodels trained on the original subdataset, and GSF represents feature fusion from submodels trained on the GAN-augmented subdataset.

4. Results and Discussion

We evaluated the quality of GAN-generated images using the Fréchet Inception Distance (FID) [29] and the Multi-Scale Structural Similarity Index (MS-SSIM) [30]. FID measures the distance between the feature distributions of real and synthetic images, where lower values indicate higher fidelity. MS-SSIM assesses perceptual similarity, with values closer to 1 reflecting greater visual resemblance. As shown in Table 1, the FID and MS-SSIM scores remained similar across the three subdatasets, indicating that the GAN consistently generated high-quality and perceptually realistic CXR images. Fig. 5 illustrates this comparison, showing real images in the top row and corresponding synthetic samples in the bottom row, which highlight the realism and diversity of the generated images. After augmentation, we balanced each subdataset with 2,176 samples per class (Normal, COVID-19, Viral Pneumonia), while keeping the validation and testing sets unchanged.



(a) Real (b) Real (c) Real

(d) Generated (e) Generated (f) Generated

Fig. 5. Sample real (top row) and generated (bottom row) CXR images of Viral pneumonia

Table 1. FID Scores for Generated Images Across Subdatasets

Subdataset	Subdataset 1	Subdataset 2	Subdataset 3
FID Score	72.588	72.757	72.348
MS SSIM	0.546	0.542	0.548

4.1. Evaluation metrics

Although accuracy is commonly used to evaluate classifier performance, it can overemphasize majority classes in imbalanced datasets and thus serves as an unreliable standalone metric [31], [32]. To establish a more robust evaluation framework, particularly for imbalanced classification tasks, we consider multiple complementary metrics:

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

$$Precision\ (Prec) = \frac{TP}{TP+FP} \tag{17}$$

$$Sensitivity\ (Sens) = \frac{TP}{TP+FN} \tag{18}$$

$$Specificity\ (Spec) = \frac{TN}{TN+FP} \tag{19}$$

$$F1-Score\ (F1) = \frac{2 \cdot Prec \cdot Sens}{Prec+Sens} \tag{20}$$

$$\mathrm{G-Mean} = \sqrt{\mathrm{Spec} \times Sens} \tag{21}$$

$$MCC = \frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{22}$$

where TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) are derived from the confusion matrix, in addition to threshold-dependent metrics such as accuracy and F1 score, we employ the Matthews Correlation Coefficient (MCC), which considers all four elements of

the confusion matrix and quantifies the correlation between predicted and actual labels. To further evaluate classifier behavior across varying thresholds, we analyze curve-based metrics. The Receiver Operating Characteristic (ROC) curve, evaluated by the Area Under the Curve (AUC), illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR). However, recent studies suggest that AUC may overestimate performance in highly imbalanced datasets [33], [34]. Therefore, we also emphasize the Area Under the Precision-Recall Curve (AUPRC), which is more informative for imbalanced classification tasks [35]. By directly assessing the trade-off between precision and recall, AUPRC offers a more reliable estimate of a model's ability to identify minority-class instances while minimizing false positives correctly [35].

4.2. Classification Results Based On Balanced And Imbalanced Subdatasets

This section evaluates the effectiveness of GAN-based synthetic data balancing at the subsataset level for DNet and ENet architectures. Table 2 and Table 3 show that this approach significantly improves classifier performance by harmonizing the precision-sensitivity trade-off and increasing sensitivity for minority classes. The observed gains in G-mean, F1 score, MCC, and AUROC are statistically significant (paired t-test, $p < 0.05$), confirming the method's effectiveness. In DNet models, GAN balancing increased accuracy from 0.978 to 0.984 and F1 score from 0.972 to 0.978 (+0.6%) in DNet_B1, and improved precision by 1.6% (0.964 to 0.980) in DNet_B2. All GAN-augmented DNet variants reached an AUROC of 0.999, with DNet_B2 and DNet_B3 achieving AUPRC of 0.997. MCC rose to 0.966.

Table 2. Classification results of DNet-based models on subdatasets with and without balancing

	Model	Acc	Prec	Sens	F1	Spec	G-Mean	MCC	AUROC	AUPRC
Without GAN Balancing	DNet_S1	0.978	0.967	0.977	0.972	0.987	0.982	0.954	0.998	0.995
	DNet_S2	0.973	0.964	0.973	0.968	0.984	0.979	0.945	0.998	0.996
	DNet_S3	0.979	0.967	0.978	0.972	0.987	0.982	0.955	0.998	0.996
With GAN Balancing	DNet_GS1	**0.984**	0.975	**0.982**	0.978	**0.990**	**0.986**	**0.966**	0.998	0.993
	DNet_GS2	0.983	**0.980**	0.976	0.978	0.988	0.982	0.965	**0.999**	**0.997**
	DNet_GS3	0.982	0.975	0.980	0.977	0.989	0.984	0.963	0.999	0.997

ENet models showed even stronger improvements. ENet_GS3 increased sensitivity by 1.6% (0.965 to 0.981), specificity by 0.2% (0.983 to 0.985), and reached the highest accuracy (0.985) and F1-Score (0.979, +0.8%). All augmented ENet variants reached AUROC of 0.999, with AUPRC rising to 0.997 in ENet_GS2. MCC reached 0.968.

Table 3. Classification results of ENet-based models on subdatasets with and without balancing

	Model	Acc	Prec	Sens	F1	Spec	G-Mean	MCC	AUROC	AUPRC
Without GAN Balancing	ENet_S1	0.975	0.967	0.966	0.966	0.980	0.973	0.943	0.997	0.988
	ENet_S2	0.976	0.955	0.980	0.967	**0.987**	0.983	0.950	0.998	0.995
	ENet_S3	0.980	0.978	0.965	0.971	0.983	0.974	0.957	0.997	0.994
With GAN Balancing	ENet_GS1	0.983	0.978	0.975	0.976	0.980	0.980	0.963	0.999	0.995
	ENet_GS2	0.983	**0.981**	0.973	0.977	0.980	0.980	0.964	0.999	0.997
	ENet_GS3	**0.985**	0.978	**0.981**	**0.979**	0.985	**0.985**	**0.968**	**0.999**	**0.997**

To compare GAN-based augmentation with traditional methods, we evaluated DNet on the imbalanced subdataset 2 (DNet_S2), comparing GAN-augmented (DNet_GS2) against SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), and geometric oversampling. As depicted in Fig. 6, GAN-augmented DNet_GS2 achieved the highest AUPRC (0.997), outperforming both the baseline (AUPRC 0.996) and other oversampling methods (SMOTE/Geometric: 0.996/0.994 AUPRC). These findings highlight GAN's superiority in generating synthetic data for class balancing, offering better precision-sensitivity trade-offs than conventional oversampling techniques.

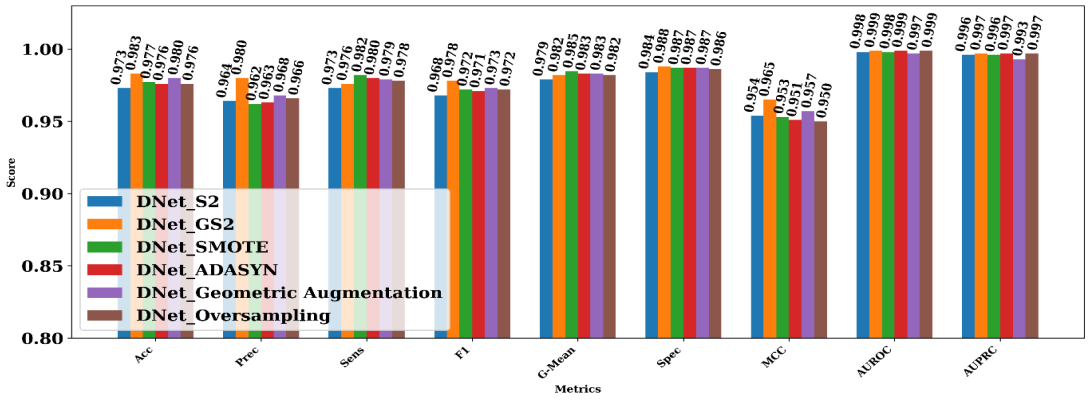


Fig. 6. Performance comparison of GAN with different oversampling methods

4.3. Ablation Study Of Ensemble Voting Strategies With Balanced And Imbalanced Subsets

To investigate the contributions of ensemble voting strategies and GAN-based data augmentation the DNet model achieved an F1 score of 0.972 and a sensitivity of 0.978. At the same time, ENet obtained 0.973 and 0.970, respectively, reflecting performance limitations induced by skewed class distributions (Table 4 and Table 5, "Original" rows). SV consistently outperformed HV in both architectures. For DNet (Table 4), SV elevated F1-Score by +0.3%, G-Mean by +0.3%, and MCC by +0.4% relative to HV.

Table 4. Classification results for DNet using the original imbalanced dataset, an ensemble of submodels without GAN balancing, and an ensemble with GAN balancing

	Model	Acc	Prec	Sens	F1	Spec	G-Mean	MCC	AUROC	AUPRC
Original	Enet	0.979	0.976	0.970	0.973	0.983	0.977	**0.955**	0.997	0.992
Ensemble voting	SV	0.986	0.979	0.983	0.981	0.990	0.987	0.970	0.999	0.996
	HV	0.983	0.975	0.978	0.977	0.988	0.983	0.964	0.983	0.980
GAN+ Ensemble voting	GSV	**0.990**	**0.985**	**0.987**	**0.986**	**0.994**	**0.990**	**0.979**	**0.999**	**0.998**
	GHV	0.990	0.985	0.985	0.985	0.993	0.989	0.978	0.989	0.987

For ENet (Table 5), SV enhanced F1 score by +0.4%, sensitivity by +0.5%, and AUPRC by +1.6%. These gains are attributed to SV's confidence-weighted aggregation, which refines boundary decisions, particularly critical under class imbalance. Integrating GAN-generated samples further improved performance for both voting strategies. In DNet (Table 4), GSV increased precision by +2.0% (from 0.967 to 0.987), MCC by +2.1% (from 0.955 to 0.976), and achieved an AUPRC of 0.998. In ENet (Table 5), GSV delivered a +0.9% precision gain, +1.7% sensitivity improvement, and +1.1% higher AUPRC compared to GHV. Although both GSV and GHV surpassed non-GAN ensembles, GSV consistently demonstrated superior performance in key metrics, including MCC and F1-Score.

Table 5. Classification results for ENet using the original imbalanced dataset, an ensemble of submodels without GAN balancing, and an ensemble with GAN balancing

	Model	Acc	Prec	Sens	F1	Spec	G-Mean	MCC	AUROC	AUPRC
Original	Dnet	0.979	0.967	0.978	0.972	0.986	0.982	0.955	0.998	0.995
Ensemble voting	SV	0.982	0.973	0.983	0.978	0.989	0.986	0.963	0.999	0.998
	HV	0.981	0.972	0.978	0.975	0.988	0.983	0.959	0.983	0.979
GAN+ Ensemble voting	GSV	**0.989**	**0.987**	**0.983**	**0.985**	**0.991**	**0.987**	**0.976**	**0.999**	**0.998**
	GHV	0.988	0.984	0.983	0.984	0.991	0.987	0.975	0.993	0.993

4.4. Ablation Study Of Feature Fusion And Machine Learning Classifiers On Balanced And Imbalanced Subsets

Integrating GAN-augmented multimodal feature fusion resulted in significant improvements in classification performance for both DNet and ENet architectures (Table 6 and Table 7). These findings highlight the synergy between GAN-driven minority-class augmentation and feature fusion, which jointly address class imbalance and enhance the representation of discriminative features. For DNet, the

baseline DNet_Orig demonstrated strong accuracy (0.978–0.980), but revealed notable gaps between sensitivity (0.967–0.976) and specificity (0.984–0.986), suggesting room for improvement in class separation. Multimodal feature fusion without GAN balancing (DNet_SF) yielded small gains in F1-Score (e.g., from 0.974 to 0.983 with MLP) and MCC (from 0.958 to 0.973 with MLP). Still, it preserved a statistically significant sensitivity-specificity imbalance ($p < 0.05$). In contrast, the GAN-augmented DNet_GSF model markedly narrowed this gap. For XGBoost, specificity increased from 0.984 to 0.994 (+1.0%), and sensitivity rose from 0.969 to 0.991 (+2.2%), thereby significantly narrowing the sensitivity-specificity gap. All DNet_GSF classifiers also achieved near-perfect AUROC values of 0.999, and AUPRC reached 0.998 for LR, XGBoost, and CatBoost, indicating clear improvements over DNet_Orig.

Table 6. Classification results of models using DNet feature extraction from the original dataset (DNet_Orig) (single model) versus feature fusion from submodels, both without (DNet_SF) and with GAN balancing (DNet_GSF), evaluated across various classifiers (LR, XGBoost, MLP, CatBoost).

	Model	Acc	Prec	Sens	F1	Spe	G-Mean	MCC	AUROC	AUPRC
Original DNet_Orig	LR	0.978	0.973	0.967	0.970	0.984	0.975	0.954	0.998	0.994
	XGBoost	0.978	0.973	0.969	0.971	0.984	0.976	0.954	0.998	0.994
	MLP	0.980	0.972	0.976	0.974	0.986	0.981	0.958	0.996	0.992
	CatBoost	0.980	0.974	0.972	0.973	0.985	0.979	0.957	0.998	0.995
submodels feature fusion DNet_SF	LR	0.984	0.984	0.973	0.981	0.986	0.979	0.967	0.999	0.997
	XGBoost	0.986	0.986	0.978	0.983	0.988	0.983	0.971	0.999	0.996
	MLP	0.986	0.984	0.979	0.981	0.989	0.984	0.973	0.998	0.995
	CatBoost	0.987	0.984	0.978	0.981	0.989	0.984	0.970	0.999	0.997
GAN+ submodels feature fusion DNet_GSF	LR	0.991	0.987	0.985	0.986	0.993	0.989	0.980	0.999	0.998
	XGBoost	**0.992**	0.988	**0.991**	**0.989**	**0.994**	**0.993**	**0.983**	**0.999**	**0.998**
	MLP	0.992	**0.989**	0.988	0.989	0.994	0.991	0.983	0.999	0.997
	CatBoost	0.991	0.987	0.989	0.988	0.994	0.991	0.982	0.999	0.998

Similarly, the baseline ENet_Orig performed slightly worse than DNet_Orig, with sensitivity (0.959 to 0.965) and MCC (0.945 to 0.952) values lagging behind. Feature fusion without GAN balancing (ENet_SF) improved certain metrics, such as XGBoost sensitivity (from 0.959 to 0.980, a 2.1% increase) and MCC (from 0.945 to 0.975). However, it still exhibited discrepancies in specificity and sensitivity. By contrast, the ENet_GSF approach nearly aligned sensitivity (0.986–0.987) and specificity (0.993–0.994). For MLP specifically, specificity increased from 0.982 to 0.994 (+1.2%) and sensitivity from 0.965 to 0.987 (+2.2%), substantially reducing the sensitivity-specificity gap. As with DNet_GSF, all ENet_GSF classifiers attained MCC values reaching 0.983, AUROC values of 0.999, and an AUPRC of 0.998, confirming their robustness and improved balanced classification performance.

Table 7. Classification results of models using ENet feature extraction from the original dataset (ENet_Orig) (single model) versus feature fusion from submodels, both without (ENet_SF) and with GAN balancing (ENet_GSF), evaluated across various classifiers (LR, XGBoost, MLP, CatBoost)

	Model	Acc	Prec	Sens	F1	Spe	G-Mean	MCC	AUROC	AUPRC
Original DNet_Orig	LR	0.977	0.978	0.965	0.971	0.982	0.973	0.952	0.997	0.992
	XGBoost	0.974	0.971	0.959	0.965	0.980	0.970	0.945	0.995	0.989
	MLP	0.976	0.969	0.965	0.967	0.982	0.973	0.948	0.994	0.987
	CatBoost	0.976	0.976	0.964	0.970	0.981	0.972	0.950	0.997	0.992
submodels feature fusion DNet_SF	LR	0.987	0.985	0.978	0.981	0.989	0.984	0.972	0.999	0.997
	XGBoost	0.988	0.986	0.980	0.983	0.991	0.986	0.975	0.999	0.997
	MLP	0.986	0.984	0.978	0.981	0.989	0.984	0.971	0.998	0.996
	CatBoost	0.987	0.984	0.978	0.981	0.989	0.984	0.971	0.999	0.995
GAN+ submodels feature fusion DNet_GSF	LR	0.991	0.987	0.987	0.987	0.993	0.990	0.980	0.999	0.998
	XGBoost	0.991	0.989	0.986	0.987	0.993	0.990	0.982	0.999	0.997
	MLP	**0.992**	**0.990**	**0.987**	**0.989**	**0.994**	**0.991**	**0.983**	**0.999**	**0.998**
	CatBoost	0.990	0.987	0.987	0.987	0.993	0.990	0.980	0.999	0.998

4.5. Comparison With State-Of-The-Art Methods

Fig. 7 comprehensively compares the proposed method with several state-of-the-art approaches on a multiclass CXR dataset that encompasses COVID-19, Viral pneumonia, and Normal classes. Win *et al.* [36] employ an ensemble strategy combining multiple high-performing CNNs through majority voting, effectively capturing complementary feature representations to enhance decision robustness. Verma *et al.* [37] integrate class-weighted loss functions within a VGG16 backbone to mitigate bias toward overrepresented classes. Chamseddine *et al.* [38] apply SMOTE (Synthetic Minority Over-sampling Technique) in conjunction with DenseNet201, thereby augmenting minority class samples to enhance classifier sensitivity. Mohan *et al.* [39] propose a custom CNN architecture enhanced through data augmentation pipelines, achieving better generalization under skewed class distributions. Expanding beyond traditional CNN architectures, Nahiduzzaman *et al.* [40] introduce a hybrid pipeline that leverages a lightweight CNN for initial feature extraction, the Pearson Correlation Coefficient (PCC) for dimensionality reduction, and the Extreme Learning Machine (ELM) for rapid, low-complexity classification. This approach demonstrates that simplicity and efficiency can yield competitive performance. In contrast, Wang *et al.* [41] take a model-free approach, proposing a semantic-powered few-shot learning framework that aligns radiology report semantics with image features using a Report Image Explanation Cell (RIEC). Their method, further enhanced by a multi-task collaborative diagnosis strategy (MCDS), not only performs well with minimal annotated data but also provides strong interpretability, though it relies heavily on the availability and quality of radiology reports.As shown in Fig. 7, our approach demonstrates superior Accuracy, F1-score, and AUC [36]–[41]

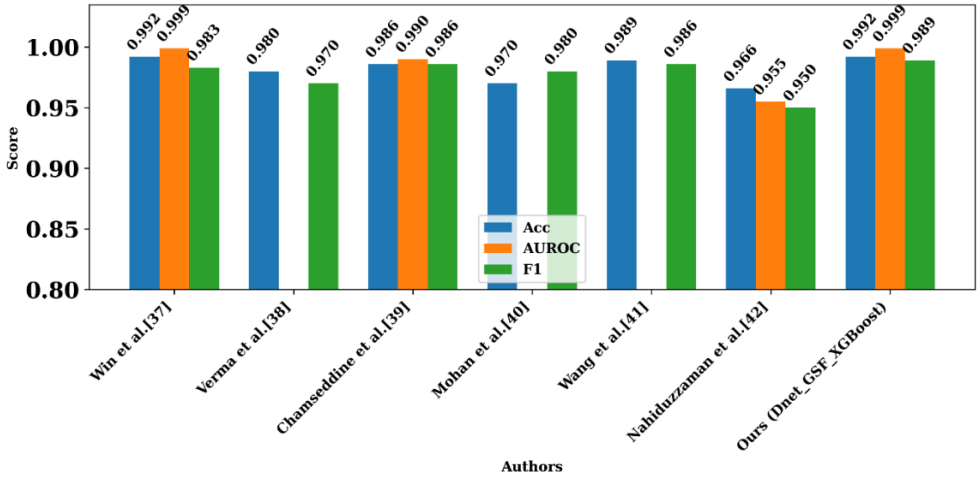


Fig. 7. Comparison of state-of-the-art methods with our proposed approach on imbalanced multiclass CXR data

4.6. Deployment Considerations and Explainable AI

To support interpretability and foster clinical trust, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the salient regions within CXR images most strongly influencing the CNN model's predictions. These class-discriminative heatmaps highlight areas contributing to positive classifications for COVID-19 and viral pneumonia, as illustrated in Fig. 8. While Grad-CAM visualizations offer valuable preliminary insights into potential disease-specific biomarkers, their clinical validity remains contingent upon expert annotation, which is currently limited. Nonetheless, these visualizations serve as a bridge between AI model reasoning and clinical interpretability, potentially guiding future radiological and pathological investigations. We tested our models on the Kaggle platform equipped with an NVIDIA Tesla P100 GPU (16 GB VRAM). Using a batch of five images, the ensemble model achieved an average inference time of 34.32 seconds, while the multimodal fusion model recorded 0.0021 seconds. These values reflect the actual performance of our system under the specified GPU environment. From a theoretical and future-oriented perspective, the system we designed is intended to support integration into clinical environments through a modular architecture that could eventually adhere to healthcare interoperability standards such as HL7 and FHIR. This includes potential

interaction with hospital information systems via secure APIs, automated report generation, and integration with Electronic Health Records (EHRs). Furthermore, the models could be adapted for deployment on mobile devices or edge computing platforms, enabling real-time point-of-care diagnostics in diverse clinical settings.

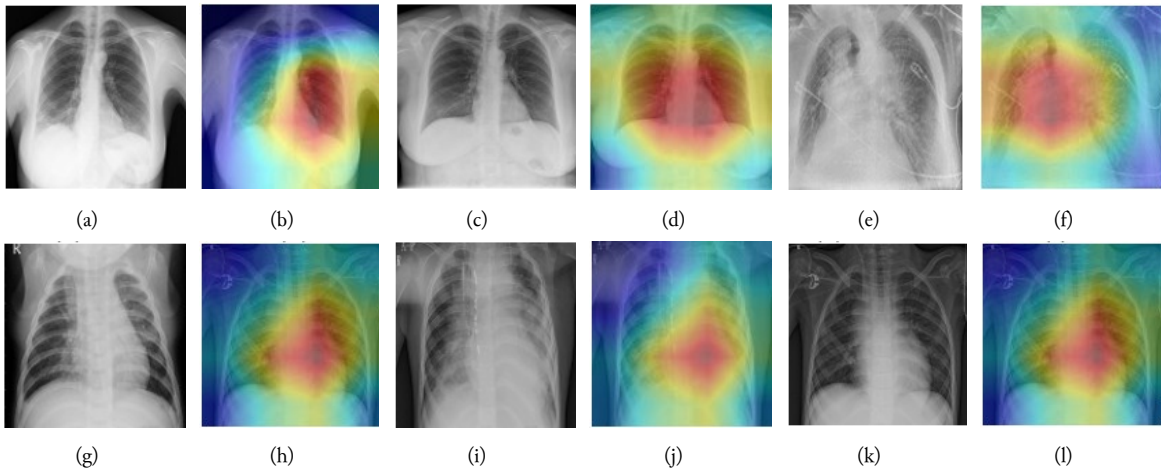


Fig. 8. Grad-CAM-based localization of disease-relevant regions using DenseNet201 activations for correctly classified cases. The first row (a–f) displays CXR from the COVID-19 class, while the second row (g–l) presents cases of viral pneumonia. Each pair includes the original image (left) and the Grad-CAM overlay (right), highlighting regions that contributed most to the model's decision. The activated areas correspond to radiological features such as bilateral opacities and localized infiltrates, offering insight into how DenseNet201 distinguishes between the two conditions based on deep feature representations

4.7. Discussion

This study proposes a comprehensive framework to address class imbalance in CXR classification, integrating GAN-based augmentation, ensemble voting, and multimodal feature fusion. Our approach demonstrated significant improvements in classification performance and generalizability. Specifically, soft voting outperformed hard voting by effectively leveraging classifier confidence, while feature fusion using multiple submodels enhanced interclass separability and mitigated overfitting. These results build upon prior research advocating for synthetic oversampling in imbalanced medical datasets. For example, Abbas *et al.* [42] reported performance gains using synthetic COVID-19 CXRs in binary classification tasks, though their scope was limited. Our work extends this to multiclass settings and incorporates ensemble learning and feature integration, aligning with previous research that emphasizes multidomain feature integration for robust classification in imbalanced datasets [43], [44].

Nonetheless, limitations persist. The reliance on GAN-generated samples raises concerns regarding the authenticity and variability of synthetic pathologies. These samples may not capture subtle or rare morphological features, which could potentially impair robustness in diverse clinical environments. Moreover, the evaluation was limited to a single publicly available dataset, which restricted generalizability and increased the risk of dataset-specific overfitting. Ethical considerations must also be acknowledged. Although synthetic data generation offers privacy advantages, improper validation may introduce bias or undermine clinician trust. Transparency regarding data provenance and responsible use guidelines will be critical moving forward. To aid model interpretability, we employed Grad-CAM to visualize class-discriminative regions; however, in the absence of annotated biomarkers, these visualizations could not be clinically validated. Future studies should incorporate expert radiologist review to assess whether the model's attention aligns with established diagnostic markers. Beyond algorithmic performance, the clinical utility of such a framework merits further exploration. In high-volume or resource-limited settings, automated triage using AI models can reduce diagnostic delays and optimize referral workflows, particularly for conditions such as COVID-19 or atypical pneumonias. Moreover, automation may alleviate radiologists' workload and reduce operational costs, supporting scalable and equitable healthcare delivery. A further limitation is the lack of expert validation for model predictions.

While quantitative metrics such as MCC and F1-score provide useful benchmarks, they are insufficient for assessing clinical reliability. Radiologists did not review our framework, and no inter-reader or intra-reader agreement was measured. Addressing this gap will be essential for clinical adoption. Future research will focus on external validation using multi-institutional and prospective datasets to evaluate the robustness of the model under dataset shifts. Real-world clinical trials with embedded feedback and integration into radiology workflows are also needed to assess inference latency and practical utility. Additionally, we plan to explore advanced methods, such as self-supervised learning, transformer-based architectures, and hybrid radiomics–deep learning pipelines, to enhance performance, interpretability, and generalization further.

5. Conclusion

This study presented a CXR classification framework that addresses class imbalance through GAN-based data augmentation, multimodal feature fusion, and ensemble learning. The framework reduces bias and enhances reliability by generating synthetic samples for underrepresented classes and controlling the influence of dominant ones. Experimental results show that soft voting ensembles and multimodal fusion, when combined with GAN-based augmentation, each independently improve diagnostic performance. This results in high accuracy, balanced sensitivity and precision, and strong F1-scores and MCC values. Future work will focus on improving GAN architectures and exploring advanced imbalance mitigation strategies, such as contrastive and meta-learning, to further enhance classification outcomes. Validation on multi-institutional datasets will assess generalizability, and lightweight models will be considered to reduce computational demands. The proposed framework demonstrates the effectiveness of combining data-driven augmentation with multimodal feature fusion and ensemble methods for addressing class imbalance in medical imaging.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] J. W. Pickering, "Machine learning for decision-making in cardiology: a narrative review to aid navigating the new landscape," *Rev. Española Cardiol. (English Ed.*, vol. 76, no. 8, pp. 645–654, Aug. 2023, doi: 10.1016/j.rec.2023.02.009.

[2] Deepak Kumar, Priyanka Pramod Pawar, Hari Gonaygunta, Geeta Sandeep Nadella, Karthik Meduri, and Shoumya Singh, "Machine learning's role in personalized medicine & treatment optimization," *World J. Adv. Res. Rev.*, vol. 21, no. 2, pp. 1675–1686, Feb. 2024, doi: 10.30574/wjarr.2024.21.2.0641.

[3] C. Srinivas *et al.*, "Deep Transfer Learning Approaches in Performance Analysis of Brain Tumor Classification Using MRI Images," in *Journal of Healthcare Engineering*, Mar. 2022, vol. 2022, pp. 1–17, doi: 10.1155/2022/3264367.

[4] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[5] M. Galanty *et al.*, "Assessing the documentation of publicly available medical image and signal datasets and their impact on bias using the BEAMRAD tool," *Sci. Rep.*, vol. 14, no. 1, p. 31846, Dec. 2024, doi: 10.1038/s41598-024-83218-5.

[6] N. Sourlos *et al.*, "Recommendations for the creation of benchmark datasets for reproducible artificial intelligence in radiology," *Insights Imaging*, vol. 15, no. 1, p. 248, Oct. 2024, doi: 10.1186/s13244-024-01833-2.

[7] D. Ueda *et al.*, "Fairness of artificial intelligence in healthcare: review and recommendations," *Jpn. J. Radiol.*, vol. 42, no. 1, pp. 3–15, Jan. 2024, doi: 10.1007/s11604-023-01474-3.

[8] J. Zhu, Z. Ye, M. Ren, and G. Ma, "Transformative skeletal motion analysis: optimization of exercise training and injury prevention through graph neural networks," *Front. Neurosci.*, vol. 18, p. 1353257, Mar. 2024, doi: 10.3389/fnins.2024.1353257.

[9] A. Makhlouf, M. Maayah, N. Abughanam, and C. Catal, "The use of generative adversarial networks in medical image augmentation," *Neural Comput. Appl.*, vol. 35, no. 34, pp. 24055–24068, Dec. 2023, doi: 10.1007/s00521-023-09100-z.

[10] T. Malygina, E. Ericheva, and I. Drokin, "Data Augmentation with GAN: Improving Chest X-Ray Pathologies Prediction on Class-Imbalanced Cases," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11832 LNCS, Springer, Cham, 2019, pp. 321–334, doi: 10.1007/978-3-030-37334-4_29.

[11] X. Qin, F. M. Bui, H. H. Nguyen, and Z. Han, "Learning From Limited and Imbalanced Medical Images With Finer Synthetic Images From GANs," *IEEE Access*, vol. 10, pp. 91663–91677, 2022, doi: 10.1109/ACCESS.2022.3202560.

[12] S. Kothawade, A. Savarkar, V. Iyer, G. Ramakrishnan, and R. Iyer, "CLINICAL: Targeted Active Learning for Imbalanced Medical Image Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13559 LNCS, Springer, Cham, 2022, pp. 119–129, doi: 10.1007/978-3-031-16760-7_12.

[13] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Comput. Med. Imaging Graph.*, vol. 95, p. 102026, Jan. 2022, doi: 10.1016/j.compmedimag.2021.102026.

[14] R. Fan and S. Bu, "Transfer-Learning-Based Approach for the Diagnosis of Lung Diseases from Chest X-ray Images," *Entropy*, vol. 24, no. 3, p. 313, Feb. 2022, doi: 10.3390/e24030313.

[15] E. Jangam, A. A. D. Barreto, and C. S. R. Annavarapu, "Automatic detection of COVID-19 from chest CT scan and chest X-Rays images using deep learning, transfer learning and stacking," *Appl. Intell.*, vol. 52, no. 2, pp. 2243–2259, Jan. 2022, doi: 10.1007/s10489-021-02393-4.

[16] S. Kaleem, A. Sohail, M. U. Tariq, M. Babar, and B. Qureshi, "Ensemble learning for multi-class COVID-19 detection from big data," *PLoS One*, vol. 18, no. 10, p. e0292587, Oct. 2023, doi: 10.1371/journal.pone.0292587.

[17] N. Habib, M. M. Hasan, M. M. Reza, and M. M. Rahman, "Ensemble of CheXNet and VGG-19 Feature Extractor with Random Forest Classifier for Pediatric Pneumonia Detection," *SN Comput. Sci.*, vol. 1, no. 6, p. 359, Nov. 2020, doi: 10.1007/s42979-020-00373-y.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[19] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 1–11, [Online]. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, vol. 2017-Janua, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[22] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proceedings of Machine Learning Research*, 2021, vol. 139, pp. 10096–10106, [Online]. Available at: https://proceedings.mlr.press/v139/tan21a.html.

[23] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, "Dense Convolutional Network and Its Application in Medical Image Analysis," *Biomed Res. Int.*, vol. 2022, no. 1, p. 2384830, Jan. 2022, doi: 10.1155/2022/2384830.

[24] T. Zhou, Q. Cheng, H. Lu, Q. Li, X. Zhang, and S. Qiu, "Deep learning methods for medical image fusion: A review," *Comput. Biol. Med.*, vol. 160, p. 106959, Jun. 2023, doi: 10.1016/j.compbiomed.2023.106959.

[25] Y. Liu, Y. Li, Z. Xu, X. Liu, H. Xie, and H. Zeng, "Guided Dropout: Improving Deep Networks Without Increased Computation," *Intell. Autom. Soft Comput.*, vol. 36, no. 3, pp. 2519–2528, Mar. 2023, doi: 10.32604/iasc.2023.033286.

[26] M. A. A. Abdalla *et al.*, "Machine learning-based residential load demand forecasting: Evaluating ELM, XGBoost, RF, and SVM for enhanced energy system and sustainability," *Sci. Inf. Technol. Lett.*, vol. 6, no. 1, pp. 1–15, 2025. [Online]. Available at: https://pubs2.ascee.org/index.php/sitech/article/view/1866.

[27] M. Abdallah, B. Mohammadi, H. Nasiri, O. M. Katipoğlu, M. A. A. Abdalla, and M. M. Ebadzadeh, "Daily global solar radiation time series prediction using variational mode decomposition combined with multi-functional recurrent fuzzy neural network and quantile regression forests algorithm," *Energy Reports*, vol. 10, pp. 4198–4217, Nov. 2023, doi: 10.1016/j.egyr.2023.10.070.

[28] M. A. A. Abdalla, W. Min, W. Bing, A. M. Ishag, and B. Saleh, "Double-layer home energy management strategy for increasing PV self-consumption and cost reduction through appliances scheduling, EV, and storage," *Energy Reports*, vol. 10, pp. 3494–3518, Nov. 2023, doi: 10.1016/j.egyr.2023.10.019.

[29] I. Bakurov, M. Buzzelli, R. Schettini, M. Castelli, and L. Vanneschi, "Structural similarity index (SSIM) revisited: A data-driven approach," *Expert Syst. Appl.*, vol. 189, p. 116087, Mar. 2022, doi: 10.1016/j.eswa.2021.116087.

[30] N. B. Bynagari, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *Asian J. Appl. Sci. Eng.*, vol. 8, no. 1, pp. 25–34, Apr. 2019, doi: 10.18034/ajase.v8i1.9.

[31] X. Si, L. Wang, W. Xu, B. Wang, and W. Cheng, "Highly Imbalanced Classification of Gout Using Data Resampling and Ensemble Method," *Algorithms*, vol. 17, no. 3, p. 122, Mar. 2024, doi: 10.3390/a17030122.

[32] M. Kim and K.-B. Hwang, "An empirical evaluation of sampling methods for the classification of imbalanced data," *PLoS One*, vol. 17, no. 7, p. e0271260, Jul. 2022, doi: 10.1371/journal.pone.0271260.

[33] C. Mosquera, L. Ferrer, D. H. Milone, D. Luna, and E. Ferrante, "Class imbalance on medical image classification: towards better evaluation practices for discrimination and calibration performance," *Eur. Radiol.*, vol. 34, no. 12, pp. 7895–7903, Jun. 2024, doi: 10.1007/s00330-024-10834-0.

[34] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big Data," *J. Big Data*, vol. 10, no. 1, p. 42, Apr. 2023, doi: 10.1186/s40537-023-00724-5.

[35] S. Ayoub, Y. Gulzar, J. Rustamov, A. Jabbari, F. A. Reegu, and S. Turaev, "Adversarial Approaches to Tackle Imbalanced Data in Machine Learning," *Sustainability*, vol. 15, no. 9, p. 7097, Apr. 2023, doi: 10.3390/su15097097.

[36] 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," *Appl. Sci.*, vol. 11, no. 22, p. 10528, Nov. 2021, doi: 10.3390/app112210528.

[37] 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," *J. Med. Phys.*, vol. 47, no. 1, pp. 57–64, Jan. 2022, doi: 10.4103/jmp.jmp_100_21.

[38] 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," *Appl. Soft Comput.*, vol. 129, p. 109588, Nov. 2022, doi: 10.1016/j.asoc.2022.109588.

[39] 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," *Discov. Artif. Intell.*, vol. 4, no. 1, p. 20, Mar. 2024, doi: 10.1007/s44163-024-00110-x.

[40] M. Nahiduzzaman *et al.*, "Detection of various lung diseases including COVID-19 using extreme learning machine algorithm based on the features extracted from a lightweight CNN architecture," *Biocybern. Biomed. Eng.*, vol. 43, no. 3, pp. 528–550, Jul. 2023, doi: 10.1016/j.bbe.2023.06.003.

[41] Y. Wang *et al.*, "Semantic-Powered Explainable Model-Free Few-Shot Learning Scheme of Diagnosing COVID-19 on Chest X-Ray," *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 12, pp. 5870–5882, Dec. 2022, doi: 10.1109/JBHI.2022.3205167.

[42] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network," *Appl. Intell.*, vol. 51, no. 2, pp. 854–864, Feb. 2021, doi: 10.1007/s10489-020-01829-7.

[43] S.-T. Hsieh and Y.-A. Cheng, "Multimodal feature fusion in deep learning for comprehensive dental condition classification," *J. X-Ray Sci. Technol. Clin. Appl. Diagnosis Ther.*, vol. 32, no. 2, pp. 303–321, Mar. 2024, doi: 10.3233/XST-230271.

[44] H. Jia and H. Lao, "Deep learning and multimodal feature fusion for the aided diagnosis of Alzheimer's disease," *Neural Comput. Appl.*, vol. 34, no. 22, pp. 19585–19598, Nov. 2022, doi: 10.1007/s00521-022-07501-0.