SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

CLAHE-Enhanced Heterogeneous CNN Ensemble for Chest X-Ray Classification

Nur Atiqah Sia Abdullah[1,2], Fatin Nabila Abd Razak[3] and Nur Ida Aniza Rusli[4,5*]

[1]Center of Computing Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2]Knowledge and Software Engineering Research Group (KASERG), Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[3]English Language Department, Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[4]Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia
[5]Center of Computing Sciences, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Negeri Sembilan, Kampus Kuala Pilah, Kuala Pilah, Negeri Sembilan, Malaysia
*Corresponding Author: E-mail: idaaniza@upm.edu.my

Abstract
Currently, there appear to be significant issues when classifying pneumonia based on chest X-ray images. To address this issue, this study introduced a deep learning framework designed for classifying three chest radiograph categories: (i) pneumonia, (ii) COVID-19, and (iii) normal cases. The proposed framework incorporated five primary components: (i) contrast limited adaptive histogram equalisation (CLAHE), (ii) progressive transfer learning, (iii) heterogeneous ImageNet-pretrained convolutional neural network (CNN) backbones, (iv) weighted soft voting, and (v) 10-pass test-time augmentation (TTA). A dataset consisting of 3,227 chest radiographs was structured into a development pool of 2,886 images, along with a held-out test set of 341 images. An 85:15 split was adopted to partition this development pool into training and validation subsets. Using a three-phase transfer-learning procedure, three CNN architectures underwent independent fine-tuning: (i) ResNet50V2, (ii) EfficientNetB3, and (iii) VGG16. These individual models computed accuracy values of 85.34%, 77.42%, and 87.98%, respectively. An observed accuracy of 90.03% was depicted by a post hoc accuracy-weighted ensemble utilising TTA (macro-precision = 90.14%, macro-recall = 90.05%, macro-F1 score = 89.83%). Meanwhile, there was a lower recall (78.76%) measured for pneumonia. The ensemble result should also be viewed as an exploratory estimate, with ensemble weights being determined through individual test accuracies.

Keywords
Chest X-ray; Classification transfer learning; Contrast Limited Adaptive Histogram Equalisation; Convolutional Neural Network; Ensemble learning; Pneumonia; Test-time augmentation

1. Introduction

Among certain demographics (children, the elderly, and those with weakened immune systems), a significant contributor to respiratory morbidity and mortality globally is represented by pneumonia [1]. Hence, to ensure effective clinical management and patient outcomes, timely and accurate diagnosis is crucial [2]. Generally, a commonly utilised initial imaging technique for evaluating pneumonia is chest radiography (CXR). For this utilisation, it is attributed to its cost-effectiveness, quick acquisition time, non-invasive nature, and widespread availability in resource-constrained healthcare environments [3]. Nonetheless, it appears to be highly difficult to interpret chest radiographs owing to the several issues (low pulmonary abnormality contrast, overlapping anatomical structures, and significant radiographic presentation variability). As such, inter-observer variability and inconsistent interpretation remain concerning [4], [5].

Typically, there are three aspects mainly determined when conducting radiographic evaluation of pneumonia: (i) pulmonary opacities, (ii) infiltrates, and (iii) consolidations [6]. Nevertheless, these findings do not seem to be exclusively tied to a specific disease, along with intersecting with symptoms of other respiratory conditions. In performing integration, it can also be complicated by other technical and patient-related factors (acquisition setting, exposure, positioning, and image quality variations). Therefore, to aid in radiographic classification and clinical assessment, computer-aided methods have been employed [6], [7].

In medical image analysis, one field that has gained significant attention involves deep learning. For this phenomenon, it can be ascribed to its capacity to learn hierarchical image representations directly from

data [8]. Notably, chest radiograph classification has primarily adopted convolutional neural networks (CNNs), which have presented impressive performance in identifying various thoracic abnormalities (pneumonia) [9–11]. Within this context, transfer learning is also usually utilised. Regarding this selection, it can be ascribed to the pretrained CNNs that can be modified for medical imaging tasks. Specifically, compared to extensive natural-image collections, the availability of labelled datasets is significantly limited.

From a negative viewpoint, there are numerous issues persisting when classifying automated chest radiographs. Regarding the disease-related pattern manifestation, it can be impacted by several parameter variations (image contrast, illumination, acquisition conditions, and anatomical superimposition). Through this process, classification performance is affected [13], [14]. For pneumonia, it also demonstrates considerable heterogeneity in several elements (severity, anatomical distribution, and radio-graphic manifestation), alongside certain imaging features being similar to COVID-19 and other pulmonary conditions [1], [12], [15].

Meanwhile, there are CNN performance variations demonstrated that are influenced by various factors (network architecture, optimisation strategy, dataset composition, and evaluation protocol) [13], [16]. For these differences, they can then complicate the ability to make meaningful comparisons between studies, as the reported performance appears to be highly correlated with several variables (dataset size, preprocessing strategy, model configuration, class distribution, and evaluation design) [13], [17].

To address some of these issues, there are currently two adopted approaches: (i) image enhancement and (ii) ensemble learning. In contrast limited adaptive histogram equalisation (CLAHE), it has been primarily used to improve local contrast, along with preventing excessive contrast amplification. Likewise, to minimise dependence on a single network architecture, it can be attained by using ensemble learning to combine predictions from multiple models. Nonetheless, most studies have examined image enhancement, transfer learning, and ensemble modelling as distinct phases within the classification process. For studies that have explored these elements within a cohesive three-class classification framework for chest radiographs, they have remained limited.

Based on these observed gaps, a deep learning framework was introduced in this study to classify three chest radiograph categories: (i) pneumonia, (ii) COVID-19, and (iii) normal cases. Within this proposed framework, there were then five primary components integrated as follows:

i. CLAHE-based preprocessing
ii. Progressive transfer learning
iii. Three heterogeneous ImageNet-pretrained CNN backbones (ResNet50V2, EfficientNetB3, and VGG16)
iv. Weighted soft voting
v. Test-time augmentation (TTA)

In this study, there were four primary contributions as follows:

i. To improve local contrast prior to CNN feature extraction, a CLAHE-based preprocessing pipeline was employed.
ii. To progressively adapt ResNet50V2, EfficientNetB3, and VGG16 to the chest radiograph clas-sification task, it was accomplished via a three-phase transfer-learning protocol.
iii. Using both weighted soft voting and 10-pass TTA, the predictions generated by the inde-pendently trained CNN models were integrated at the decision level.
iv. To uncover classification behaviour variations for the chest radiograph categories (pneumonia, COVID-19, and normal radiographs), it was achieved by examining both aggregate and class-specific measures.

For the remainder of this study, it was outlined as follows: In Section 2, previous chest radiograph-related articles concerning deep learning, image enhancement, transfer learning, and ensemble classification were assessed. Both Sections 3 and 4 also explained the proposed framework and its experimental findings, respectively. Finally, this study was concluded in Section 5, along with discussing its limitations and possible future research directions.

2. Related Works

Regarding CXR, it is primarily employed for the preliminary evaluation of pneumonia. In comparison to more advanced imaging modalities (computed tomography), this approach also appears to be more cost-effective, alongside possessing quick acquisition time and widespread accessibility [18]. Generally, there are three aspects determined when performing radiographic evaluation of pneumonia: (i) pulmonary opacities, (ii) infiltrates, and (iii) consolidations. Nevertheless, these manifestations seem to coincide with other thoracic conditions (tuberculosis, pulmonary oedema, atelectasis, and COVID-19 [19], [20]. Another

issue arises in interpreting outcomes, as they are variations occurring in several aspects (image contrast, patient positioning, acquisition protocols, and superimposed anatomical structures) [21].

In automated medical image classification, a prevalent method is deep learning (CNNs). Compared to traditional methods that rely on manually crafted features, hierarchical representations are usually learned by CNNs directly from image data [22], [23]. Typically, fundamental features (edges and local intensity variations) are captured by the early layers. In contrast, more intricate image patterns that can aid in disease classification are encoded by the deeper layers [24–26]. In one article by Kermany et al. [24], they reported that paediatric pneumonia classification could benefit from deep learning. Rajpurkar et al. [25] also examined the ChestX-ray14 dataset, in which a DenseNet-based system was created (CheXNet). All these studies then revealed that there is a continued interest in the analysis of chest radiographs using CNNs.

Currently, there are multiple CNN architectures (VGGNet, ResNet, DenseNet, Inception, Mo-bileNet, and EfficientNet) that have been explored [22], [23], [27], [28]. In particular, transfer learning remains the most preferred choice owing to its smaller size compared to extensive natural-image datasets. Regarding certain metrics (model depth, parameter count, feature representation, and computational requirements), distinct trade-offs have also been demonstrated with various architectures. For feature reuse, it can be promoted by DenseNet via dense connectivity [25], [29]. On the contrary, MobileNet is primarily utilised for computational efficiency [27]. In optimising deeper networks, it can also be achieved by ResNet through residual connections. Meanwhile, to adjust certain network parameters (depth, width, and input resolution), EfficientNet mainly uses a compound scaling technique [28]. Nev-ertheless, for all these architectures across datasets and experimental settings, there are still significant relative performance variations observed [28], [30].

Using image preprocessing, chest radiographs can be effectively classified. Currently, these chest radiographs have been evaluated through multiple techniques. Notable examples include resizing, normalisation, denoising, intensity scaling, histogram equalisation, adaptive histogram equalisation, and CLAHE [31–33]. Specifically, when utilising conventional histogram equalisation, improved global contrast is observed. Nonetheless, image noise seems to be amplified, along with modifying the original intensity distribution [33], [34]. Concerning adaptive histogram equalisation, it also functions by focusing on local regions of an image. In contrast, the noise in areas that are relatively uniform can be in-advertently enhanced. Based on these issues, CLAHE remains favourable. When employed, local contrast amplification can then be limited through a predefined clipping threshold [33], [35]. As such, there have been increasing utilisation rates depicted for CLAHE within various chest X-ray preprocessing pipelines [31], [33], [35].

If ensemble learning is employed, predictions from various classifiers can be effectively integrated. At present, there are numerous approaches (majority voting, probability averaging, weighted soft voting, stacking, bagging, and boosting) being used [23], [36], [37]. To create ensemble systems, there are also varying architectural features demonstrated by medical-image classification-related models. Thus, the final decision can be collectively informed. For this outcome, it is mainly attained by utilising the probability estimates from multiple independently trained classifiers. Regarding pneumonia-related chest radiograph classification studies, there was one article by Kundu et al. [23] that combined GoogleNet, ResNet-18, and DenseNet-121. To achieve this outcome, it was accomplished via the weighted averaging method. Concerning an ensemble involving DenseNet169, MobileNetV2, and a Vision Transformer, it was then established by Mabrouk et al. [36]. Another work by Pandya et al. [37] examined stacking and soft-voting combinations. In the study, MobileNetV2, InceptionV3, and ResNet152V2 were employed. From all these studies, it can then be inferred that for chest radiograph classification, employing diverse model combinations remains practical.

From an unfavourable perspective, there remain methodology-related concerns. Regarding various critical elements (dataset size, image provenance, class composition, preprocessing procedures, partitioning strategy, and evaluation protocol), they have appeared to present significant variability in most articles [11], [13], [30]. Moreover, external validation was considered insufficient. Based on these observations, there are difficulties in making conclusions. Notably, the generalisability of findings across different institutions and acquisition contexts can be quite limited. In the evaluation process, there are also several issues (class imbalance, limited model interpretability, and class-specific sensitivity variation) depicted. Likewise, individual assessments have been carried out in both image enhancement and ensemble learning. Specifically, within a three-class chest radiograph classification setting, frameworks containing the previously mentioned components (contrast enhancement, progressive fine-tuning of heterogeneous CNN architectures, and prediction-level ensemble inference) have remained underexplored [31], [36], [39]. Hence, there is a need to address this issue, and this missing framework was then examined in this study.

3. Methodology

3.1 Dataset

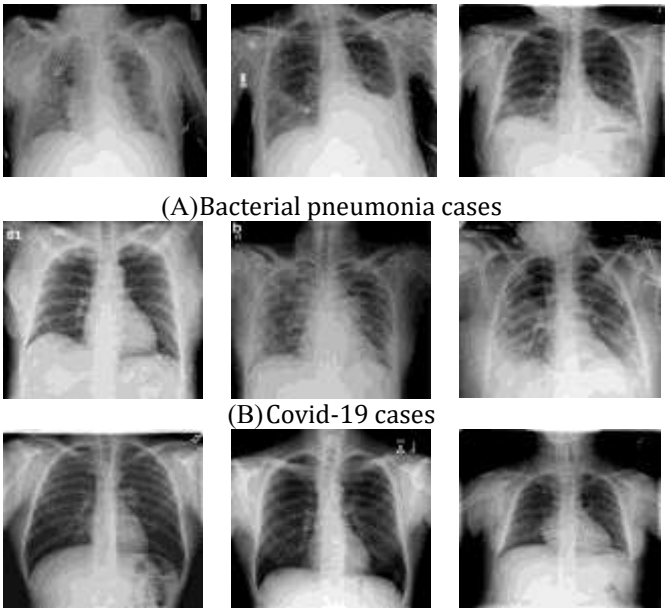
In this study, a publicly available repository [40] was adopted for the employed dataset. Particularly, there were 3,227 frontal chest radiographs recorded. For these radiographs, they were then categorised into three diagnostic categories (COVID-19, normal, and pneumonia). Using the source dataset, a development directory was also created. Notably, there were 2,886 images, alongside a separate test directory (341 images). Consequently, the number of images within the development pool for COVID-19, normal, and pneumonia were computed to be 962, 959, and 965, respectively.

In order to partition the 2,886 development images into training and validation subsets during model development, the validation_split = 0.15 option of the Keras ImageDataGenerator was used. Accordingly, there were approximately 85% and 15% of the development images employed for the model fitting and validation, respectively. Across all three CNN architectures, a fixed random seed of 42 was also used when constructing the generators. For this process, it was carried out to ensure consistency in the partitioning. In Table 1, the experimental distribution is then tabulated.

Table 1. Distribution Summary of the Chest Radiographs.

Subset	COVID-19	Normal	Pneumonia	Total
Training	818	816	821	2,455
Validation	144	143	144	431
Test	116	112	113	341
Total	**1,078**	**1,071**	**1,078**	**3,227**

For the test set, it was kept distinct from the development images, alongside being not involved in the training-validation partition. Meanwhile, a three-class task (COVID-19, normal, and pneumonia chest radiographs) definition was adopted for the classification problem. Figure 1 then depicts the representative examples.



(A)Bacterial pneumonia cases

(B)Covid-19 cases

(C)Normal cases

Figure 1. The representative chest radiographs from the (a) pneumonia, (b) COVID-19, and (c) normal classes.

3.2. Data Preprocessing

Regarding the chest radiographs, a preprocessing pipeline that included four main components was utilised: (i) CLAHE, (ii) image resizing, (iii) backbone-specific preprocessing, and (iv) image augmen-tation. For each image, it then underwent a conversion from RGB to the CIELAB colour space. Using a clip limit of 2.0 and an 8 × 8 tile grid, CLAHE was also implemented on the luminance (L) channel (see Figure. 2).

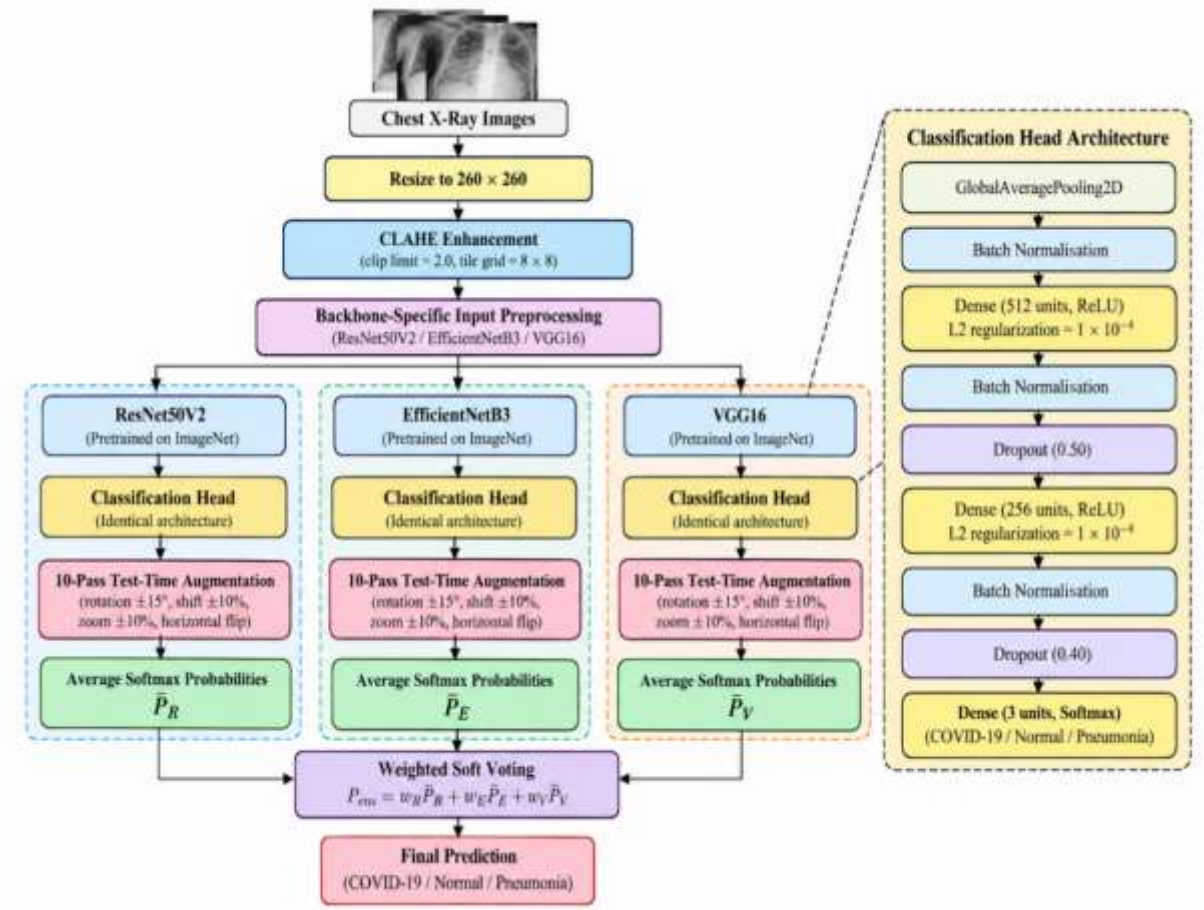


Figure 2. An architecture of the independently trained CNN classifiers used in the ensemble framework.

Subsequently, the improved L channel was merged with the original chromatic channels before being transformed back to RGB. For these images, they were also resized to 260 × 260 pixels. In this study, all tested models (ResNet50V2, EfficientNetB3, and VGG16) adopted this common input resolution. Concerning the preprocessing function provided by the TensorFlow/Keras implementation for each pretrained architecture, it was then applied accordingly. For the development-data generator, the following data augmentation-related aspects were employed:

i. Random rotation of up to 25°
ii. Horizontal and vertical translation of up to 0.20 of the image dimensions
iii. Shear variation up to 0.15
iv. Zoom up to 0.25
v. Horizontal flipping
vi. Brightness variation between 0.75 and 1.25.

In the newly introduced pixels, they were populated utilising the nearest-neighbour interpolation. Likewise, to create both the training and validation subsets, it was accomplished using the identical generator configuration. Resultantly, stochastic augmentation was utilised on validation samples throughout model development, alongside the backbone-specific preprocessing operation. Regarding the independently stored test set, it only underwent processing through CLAHE and backbone-specific preprocessing during standard individual-model evaluation. This study trained all the models using a batch size of 32.

3.3 CNN Ensemble Framework

In this study, there were three ImageNet-pretrained CNN architectures employed as heterogeneous base learners (ResNet50V2, EfficientNetB3, and VGG16). For each backbone, the original ImageNet classification layer was then removed by setting include_top = False. Subsequently, in each pretrained feature extractor, an identical task-specific classification-head architecture was attached independently. In each backbone, the feature maps generated were also processed through a GlobalAveragePooling2D layer before conducting batch normalisation.

Processing was carried out for the resulting representation via a dense layer containing 512 units with Rectified Linear Unit (ReLU) activation and L2 regularisation of 1×10^{-4}. For this process, it was followed by batch normalisation (dropout rate = 0.50). A second dense layer containing 256 units with ReLU activation and identical L2 regularisation was also applied. This process was then followed by batch normalisation (dropout rate = 0.40). Using a final three-unit softmax layer (COVID-19, normal, and pneumonia), classification was finally performed. In these three CNNs, independent training was conducted, in which their intermediate feature representations were neither concatenated nor fused in any manner. Instead, ensemble integration was performed at the probability-output level after individual model training.

3.4 Progressive Training Protocol

In this study, all CNNs underwent a consistent three-phase transfer-learning procedure. For the number of trainable backbone layers, it was gradually increased. Conversely, the initial learning rate was decreased in subsequent phases. In Table 2, the training configuration is then summarised.

Table 2. Summary of the Three-Phase Transfer-Learning Configuration

Phase	Maximum Epochs	Initial Learning Rate	Loss Function	Backbone Status	AdamW Weight Decay
Phase 1: Head training	25	1×10^{-3}	Categorical cross-entropy with label smoothing, $\varepsilon = 0.1$	Frozen	1×10^{-4}
Phase 2: Partial fine-tuning	20	1×10^{-5}	Categorical cross-entropy with label smoothing, $\varepsilon = 0.1$	Top 100 layers trainable	1×10^{-4}
Phase 3: Full fine-tuning	15	1×10^{-6}	Focal loss, $\gamma = 2$, $\alpha = 0.25$	Fully trainable	1×10^{-5}

During Phase 1, the ImageNet-pretrained backbone was entirely frozen. Only the task-specific classification head underwent training. For Phase 1, it was set up to run for a maximum of 25 epochs using the AdamW optimiser. On the contrary, the backbone in Phase 2 was adjusted to be trainable, along with all layers were kept frozen (excluding the uppermost 100). Using an initial learning rate of 1×10^{-5}, training then continued for a maximum of 20 epochs. In Phase 3, the full backbone was also made trainable, along with performing end-to-end fine-tuning for a maximum of 15 epochs (initial learning rate of 1×10^{-6}). The loss function was then changed from smoothed categorical cross-entropy to focal loss ($\gamma = 2$, $\alpha = 0.25$). All phases utilised a CosineDecayRestarts learning-rate schedule. For the first decay interval, it also corresponded to five training epochs ($t_{\text{mul}} = 2.0$, $m_{\text{mul}} = 0.9$, minimum learning-rate fraction = 1×10^{-7}).

To monitor validation accuracy, early stopping was employed. Therefore, the restoration of the best-performing weights during each phase could be facilitated. For Phases 1, 2, and 3, the patience values of 7, 8, and 6 epochs were also adopted, respectively. From the training subset, class weights were then calculated through the balanced weighting procedure provided by Scikit-learn and incorporated during model optimisation.

3.5 TTA and Ensemble Inference

In this study, all CNNs underwent a consistent three-phase transfer-learning procedure. For the number of trainable backbone layers, it was gradually increased. Conversely, the initial learning rate was decreased in subsequent phases. In Table 2, the training configuration is then summarised. The TTA from the training subset was utilised during ensemble inference after individually training the CNNs. In each model, every test radiograph was examined under 10 stochastic transformations (rotations of up to 15°, horizontal and vertical translations of up to 0.10, zoom variation of up to 0.10, and horizontal flipping). Let $p_m^{(t)}(x)$ then denote the probability vector produced by model m for image x during TTA pass t. Hence, the TTA-averaged prediction for model m can be expressed as follows:

$$\bar{p}_m(x) = \frac{1}{T}\sum_{t=1}^{T} p_m^{(t)}(x) \quad (1)$$

where $T = 10$ denotes the number of TTA passes.

For the three TTA-averaged probability vectors, they are then combined via weighted soft voting as follows:

$$p_{\text{ens}}(x) = \sum_{m=1}^{M} w_m \, \bar{p}_m(x) \tag{2}$$

where M = 3 represents the three CNN classifiers (ResNet50V2, EfficientNetB3, and VGG16).

The ensemble weights are normalised such that

$$\sum_{m=1}^{M} w_m = 1 \tag{3}$$

Regarding the final predicted class, it is also determined by choosing the class that exhibits the highest ensemble probability as follows:

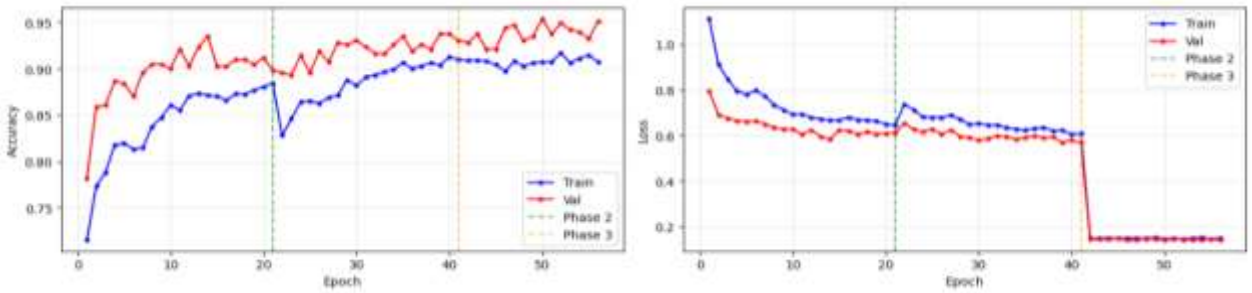
$$\hat{y} = \arg\max_{c} p_{\text{ens},c}(x) \tag{4}$$

where c is one of the three output classes (COVID-19, normal, and pneumonia).
In the experiment conducted, the ensemble weights appeared to be proportional to the individual test accuracies achieved without TTA. Therefore, ResNet50V2, EfficientNetB3, and VGG16 were assigned normalised weights of approximately 0.340, 0.309, and 0.351, respectively. For the ensemble analysis, it also served as a post hoc exploratory evaluation (not a fully independent generalisation estimate). This assumption was based on the weights being derived from the performance on the same test split used for ensemble assessment. As such, when analysing the ensemble results, this limitation was taken into account.

4. Results and Discussions

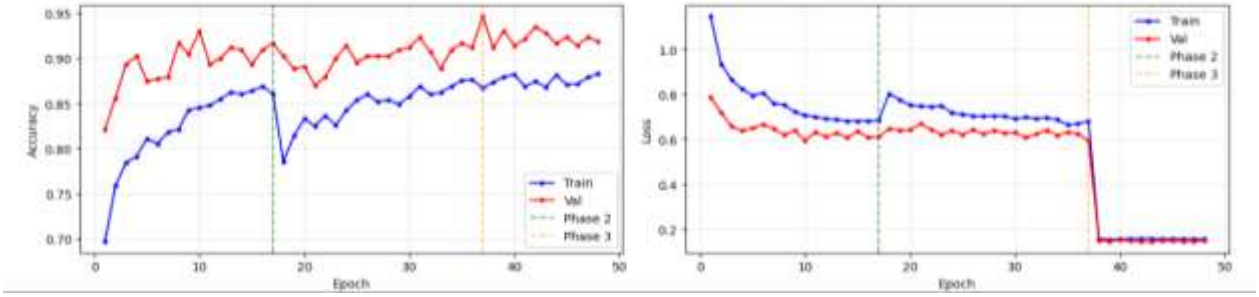
4.1 Individual Backbone Results

In Figures 3 to 5, the training and validation accuracy of ResNet50V2, EfficientNetB3, and VGG16 across the three progressive training phases are demonstrated, alongside their loss trajectories.



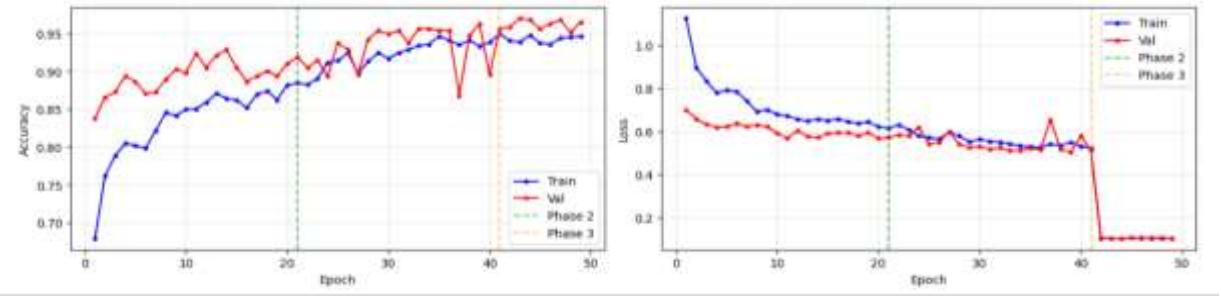
(a) Training and validation accuracy (b) Training and validation loss

Figure 3. Training and validation performance of ResNet50V2 across the three training phases.



(a) Training and validation accuracy (b) Training and validation loss

Figure 4. Training and validation performance of EfficientNetB3 across the three training phases.



(a) Training and validation accuracy (b) Training and validation loss

Figure 5. Training and validation performance of VGG16 across the three training phases.

Regarding the peak validation accuracy, it is computed in Table 3 by identifying the highest validation accuracy achieved during the three training phases. Consequently, the highest peak validation accuracy (96.98%) appeared to be achieved by VGG16. For this outcome, it was followed by ResNet50V2 (95.36%) and EfficientNetB3 (94.66%). Based on these findings, comparably high peak validation accuracy values were reported by all three models. Specifically, the highest observed value was recorded for VGG16. Nevertheless, there was a need to understand these values within the validation procedure framework used in this study, in which the validation generator contained the stochastic augmentation. As such, the results seemed to reflect the highest performance level noted during training (not a separate generalisation assessment).

Table 3. Summary of the Peak Validation Accuracy for the Individual CNN Backbones.

CNN Backbone	Peak Validation Accuracy
ResNet50V2	95.36%
EfficientNetB3	94.66%
VGG16	96.98%

Table 4. Summary of the Test Accuracy for the Individual CNN Backbones

CNN Backbone	Test Accuracy
ResNet50V2	85.34%
EfficientNetB3	77.42%
VGG16	87.98%

4.2 Ensemble Results

Using 10-pass TTA and test-accuracy-weighted soft voting, the individual CNN probability outputs were integrated. Regarding ResNet50V2, EfficientNetB3, and VGG16, they were then given weights of approximately 0.340, 0.309, and 0.351, respectively. Accordingly, an observed test accuracy of 90.03% was measured when evaluated on the 341 radiographs (see Table 5). Figure. 6 then portrays the confusion matrix.

Table 5. Summary of the Individual and Ensemble Test Accuracy.

CNN Backbone	Test Accuracy
ResNet50V2	85.34%
EfficientNetB3	77.42%
VGG16	87.98%
Proposed Ensemble + TTA	90.03%

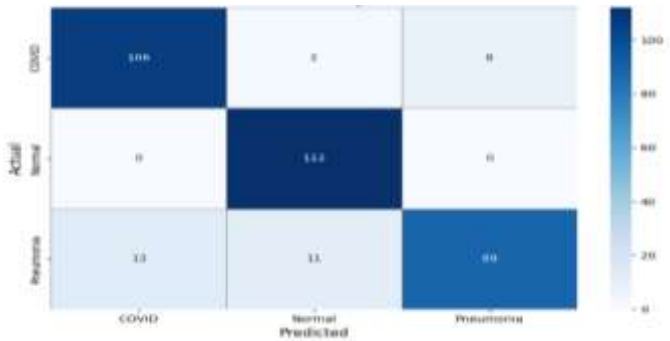


Figure 6. The confusion matrix of the CNN ensemble with 10-pass TTA.

In the case of COVID-19, 106 out of 116 images appeared to be correctly classified (recall = 91.38%). There were also eight COVID-19 images that were categorised under the pneumonia class. Conversely, two COVID-19 images were classified as normal. Concerning all the 112 normal radiographs, they were also correctly classified (recall = 100%) for this test split. Lastly, 89 out of 113 pneumonia images were accurately classified (recall = 78.76%). There were also 13 and 11 pneumonia images that were classified as COVID-19 and normal, respectively. From the class-specific results, pneumonia continued to be the most difficult category (even considering the high overall accuracy achieved). Notably, there was a significant number of images identified as pneumonia-positive that were categorised into one of the other two classes. For this finding, it was made based on the lower pneumonia recall. Resultantly, it could be inferred that a complete understanding of the classifier's behaviour could be fully provided by aggregate accuracy alone.

In comparison to the most robust individual model (VGG16), a higher observed accuracy of 2.05 percentage points was demonstrated by the ensemble produced. For this ensemble, it also achieved macro- precision, macro-recall, and macro-F1 scores of 90.14%, 90.05%, and 89.83%, respectively. Nonetheless, it seemed that this disparity should not be construed as an impartial enhancement assessment relative to VGG16. In this study, the ensemble weights were based on the accuracies of the individual models on the identical test split. Regarding this split, it was later utilised to evaluate the ensemble performance. Hence, it was better to assume the ensemble result as an indication that the model outputs could be effectively combined within the reported post hoc configuration (not that the ensemble would inevitably attain the same enhancement on previously unseen external data).

4.3 Comparison with Existing Studies

In order to provide context for the observed results, a comparative analysis was carried out between the proposed framework and other studies that focused on three-class classification (pneumonia, COVID-19, and normal chest radiographs) (see Table 6). It was also noted that this comparison was conducted to offer context (not a definitive model quality ranking). For this assumption, it could be attributed to the parameter variations (datasets, sample sizes, preprocessing strategies, partitioning schemes, and evaluation protocols) in these works.

Table 6. Comparison summary with Selected Three-Class Chest X-Ray Classification Studies.

Model/Method	Accuracy	Precision	Recall	F1-Score
CLAHE + ResNet50V2/EfficientNetB3/VGG16 ensemble + TTA [this study]	90.03%	0.9014	0.9005	0.8983
U-Net segmentation + DenseNet201 classifier [41]	78.7%	79.1%	78.7%	78.1%
DCGAN data augmentation + Ensemble CNN + MobileNet [42]	96.97%	96.97%	96.97%	96.88%
Ensemble CNN + DenseNet [43]	89.73%	89.75%	89.73%	89.73%
CNN + ResNet50[44]	93.94%	94.32%	94.10%	93.96%

In this study, the proposed ensemble configuration revealed accuracy, macro-precision, macro-recall, and macro-F1 scores of 90.03%, 90.14%, 90.05%, and 89.83%, respectively. Thus, it was suggested that there was a uniform overall performance based on the resemblance of the aggregate measures. Nevertheless, these figures appeared to mask the previously mentioned lower recall for pneumonia. Regarding the U-Net segmentation and DenseNet201 approach proposed by Bassi and Attux [41], they also documented lower aggregate values. Nonetheless, compared to this study (held-out split from the same overall data source), the study included external generalisation testing. Based on this observation, the lower reported performance should not be viewed merely as an indication of a less robust methodology.

In Wubineh et al's work [42], an accuracy of 96.97% was revealed when utilising DCGAN-based augmentation and a CNN-MobileNet approach. Alternatively, only 66 authentic test images were adopted in the study. If a single classification error occurred, it could then alter the observed accuracy by approximately 1.52 percentage points. Hence, there was a greater sensitivity of the reported result to the test-set composition. Concerning the ensemble CNN-DenseNet approach established by Chitradevi et al. [43], they computed an accuracy of 89.73%. For this value, it appeared to be numerically close to the 90.03% observed in this study. Nevertheless, one should refrain from interpreting the difference as indicative of significant superiority, as there were no comparisons performed for a shared test set, confidence intervals, or paired statistical comparison.

In another article by Ankushavali et al. [44], a 93.94% accuracy was observed for the modified ResNet50 approach. Once more, a direct quality comparison of intrinsic methods seemed to be hindered owing to the differences in several components (image provenance, dataset composition, preprocessing, and experimental design). Regarding the computational cost of the proposed framework, it should also be considered. Therefore, there were 10 TTA passes performed independently for each of the three CNN models. Through this process, a maximum of 30 forward passes per radiograph prior to ensemble fusion was demonstrated. Compared to standard single-model inference, this outcome appeared to require significantly more computational resources, along with restricting its practical application in resource-constrained or latency-sensitive environments.

Based on all these findings, it could be concluded that similar performance levels were depicted by the proposed framework and those documented in earlier works. Nonetheless, to draw robust comparative conclusions, it could only be accomplished by ensuring that four critical aspects were present as follows:

i. Uniform patient-independent dataset
ii. Identical preprocessing
iii. Common performance measures
iv. Formal statistical testing

4.4 Limitations

In this study, there were several methodological limitations that should be considered when interpreting the findings. Initially, the same stochastic augmentation configuration was employed for both the validation and training generator (despite a separate validation subset being utilised). As such, the validation performance appeared to be assessed using dynamically transformed images (not on a fixed preprocessing-only validation set). Another constraint entailed the individual model accuracies achieved on the test set being primarily adopted for computing the ensemble weights. To examine ensemble performance, the same test split was then applied. Hence, a post hoc model configuration was implied. Specifically, the 90.03% ensemble accuracy did not serve as a fully independent test estimate.

Across all three training phases, peak validation accuracy was reported. Notably, the best weights were effectively restored by early stopping independently within each phase. Through this process, the highest observed validation performance during training appeared to be represented by the reported peak validation value. Conversely, this finding might not necessarily align with the precise final weights evaluated on the test set. In this study, a single-source dataset was also employed, alongside missing external or multi-centre validation. Therefore, to determine patient-level independence, it was not possible to be achieved based on the experimental information available. For the controlled ablation experiments that would isolate the specific contributions of CLAHE, progressive fine-tuning, TTA, and ensemble fusion, it was also not incorporated in this study. Accordingly, the enhancements noted in the final configuration could not be attributed independently to any single component.

5. Conclusion

In this study, a classification framework for a three-class chest radiograph was successfully established. For this proposed framework, it contained five primary components (CLAHE-based contrast enhancement, progressive transfer learning, heterogeneous pretrained CNN backbones, weighted soft voting, and TTA). Accordingly, peak validation accuracy values of 95.36%, 94.66%, and 96.98% were recorded for ResNet50V2, EfficientNetB3, and VGG16, respectively. Conversely, these values appeared to be 85.34%, 77.42%, and 87.98% when assessed independently on the distinct test split without TTA. Within the established experimental framework, VGG16 also demonstrated the best individual test performance.

Regarding the post hoc weighted ensemble with 10-pass TTA, it produced an observed test accuracy of 90.03% (macro-precision = 90.14%, macro-recall = 90.05%, and macro-F1 score = 89.83%). Compared to VGG16, this accuracy value was larger by 2.05 percentage points. Nonetheless, it seemed that this enhancement should be regarded as exploratory (not an independent ensemble generalisation assessment). For this assumption, it was made based on the ensemble weights being derived from the accuracies of individual models on the same test set.

From the class-specific analysis, it was noted that the most difficult category was represented by pneumonia. In comparison to COVID-19 (91.38%) and normal radiographs (100%), the recall for pneumonia on the evaluated test split was computed to be 78.76%. What could then be concluded was that class-specific performance remained a critical assessment, alongside aggregate accuracy. In this study, there were also several methodological constraints. For these limitations, they entailed stochastic augmentation-based validation data and ensemble parameter selection via test performance. There was

also no external validation conducted, along with the individual components of the framework were not assessed through controlled ablation. Regarding the ensemble, it also necessitated as many as 30 forward passes for every radiograph. Hence, there could be a significantly increased inference cost compared to a single CNN.

Concerning future studies, independent test cohorts that involve distinctly separated training and non-augmented validation should be examined (patient-level partitioning). From the validation data prior to the final evaluation on the test set, ensemble weights should also be chosen solely. Other research-related aspects that must be explored also include external multi-centre validation, repeated experimental runs, confidence intervals, statistical comparison of paired predictions, calibration analysis, class-specific sensitivity and specificity, and controlled ablation. For explainable AI techniques, their involvement may also support the understanding of model predictions, along with bolstering their potential relevance to clinical decision-support applications.

Acknowledgement: The authors gratefully acknowledge the Ministry of Higher Education Malaysia for funding this research through the Fundamental Research Grant Scheme – Early Career (FRGS-EC), Grant No. FRGS-EC/1/2024/ICT04/UITM/02/2. The authors also extend their sincere appreciation to the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Negeri Sembilan, and Universiti Teknologi MARA Shah Alam for their valuable support and assistance throughout this research. Additionally, the authors acknowledge the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, for its institutional support during the preparation and publication of this work.

References

[1] Torres, A., Cilloniz, C., Niederman, M. S., Menéndez, R., Chalmers, J. D., Wunderink, R. G., and van der Poll, T., 2021. Pneumonia. *Nature Reviews Disease Primers*, 7, Article 25. https://doi.org/10.1038/s41572-021-00259-0.

[2] Kanwal, K., Asif, M., Khalid, S. G., Liu, H., Qurashi, A. G., and Abdullah, S., 2024. Current diagnostic techniques for pneumonia: A scoping review. *Sensors*, 24(13), Article 4291. https://doi.org/10.3390/s24134291.

[3] Cook, A. E., Garrana, S. H., Martínez-Jiménez, S., and Rosado-de-Christenson, M. L., 2022. Imaging patterns of pneumonia. *Seminars in Roentgenology*, 57(1), pp. 18–29. https://doi.org/10.1053/j.ro.2021.10.005.

[4] Voigt, G. M., Thiele, D., Wetzke, M., Weidemann, J., Parpatt, P.-M., Welte, T., Seidenberg, J., Vogelberg, C., Koster, H., Rohde, G. G. U., Härtel, C., Hansen, G., and Kopp, M. V., 2021. Interobserver agreement in interpretation of chest radiographs for pediatric community acquired pneumonia: Findings of the pedCAPNETZ-cohort. *Pediatric Pulmonology*, 56, pp. 2676–2685. https://doi.org/10.1002/ppul.25528.

[5] Brooks, D., Wright, S. E., Beattie, A., McAllister, N., Anderson, N. H., Roy, A. I., Gonsalves, P., Yates, B., Graziadio, S., Mackie, A., Davidson, J., Gopal, S. V., Whittle, R., Zahed, A., Barton, L., Elameer, M., Tuckett, J., Holmes, R., Sutcliffe, A., Santamaria, N., and Simpson, A. J., 2024. Assessment of the comparative agreement between chest radiographs and CT scans in intensive care units. *Journal of Critical Care*, 82, Article 154760, pp. 1–8.

[6] Xie, Y., Zhu, B., Jiang, Y., Zhao, B., and Yu, H., 2025. Diagnosis of pneumonia from chest X-ray images using YOLO deep learning. *Frontiers in Neurorobotics*, 19, Article 1576438. https://doi.org/10.3389/fnbot.2025.1576438.

[7] Alshanketi, F., Alharbi, A., Kuruvilla, M., Mahzoon, V., Siddiqui, S. T., Rana, N., and Tahir, A., 2025. Pneumonia detection from chest X-ray images using deep learning and transfer learning for imbalanced datasets. *Journal of Imaging Informatics in Medicine*, 38, pp. 2021–2040. https://doi.org/10.1007/s10278-024-01334-0.

[8] Li, M., Jiang, Y., Zhang, Y., and Zhu, H., 2023. Medical image analysis using deep learning algorithms. *Frontiers in Public Health*, 11, Article 1273253. https://doi.org/10.3389/fpubh.2023.1273253.

[9] Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., and Ilono, P., 2025. Deep convolutional neural networks in medical image analysis: A review. *Information*, 16(3), Article 195. https://doi.org/10.3390/info16030195.

[10] Nasser, A. A., and Akhloufi, M. A., 2023. A review of recent advances in deep learning models for chest disease detection using radiography. *Diagnostics*, 13(1), Article 159. https://doi.org/10.3390/diagnostics13010159.

[11] Siddiqi, R., and Javaid, S., 2024. Deep learning for pneumonia detection in chest X-ray images: A comprehensive survey. *Journal of Imaging*, 10(8), Article 176. https://doi.org/10.3390/jimaging10080176.

[12] Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V., and Kaur, M., 2021. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *Journal of Biomolecular Structure and Dynamics*, 39(15), pp. 5682–5689. https://doi.org/10.1080/07391102.2020.1788642.
[13] Vyas, R., and Pandey, R. K., 2026. Deep learning for pneumonia detection from X-ray: A systematic review of models, datasets, and clinical translation challenges. *Biomedical Signal Processing and Control*, 115, Article 109272. https://doi.org/10.1016/j.bspc.2025.109272.
[14] Ayalew, A. M., Salau, A. O., Abeje, B. T., and Enyew, B., 2022. Detection and classification of COVID-19 disease from X-ray images using convolutional neural networks and histogram of oriented gradients. *Biomedical Signal Processing and Control*, 74, Article 103530. https://doi.org/10.1016/j.bspc.2022.103530.
[15] Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V., and Kaur, M., 2021. Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *Journal of Biomolecular Structure and Dynamics*, 39(15), pp. 5682–5689. https://doi.org/10.1080/07391102.2020.1788642.
[16] Bressem, K. K., Adams, L. C., Erxleben, C., Hamm, B., Niehues, S. M., and Vahldiek, J. L., 2020. Comparing different deep learning architectures for classification of chest radiographs. *Scientific Reports*, 10, Article 13590. https://doi.org/10.1038/s41598-020-70479-z.
[17] Khan, M. A., Kadry, S., Zhang, Y.-D., Akram, T., Sharif, M., Rehman, A., and Saba, T., 2021. Prediction of COVID-19 pneumonia based on selected deep features and one-class kernel extreme learning machine. *Computers and Electrical Engineering*, 90, Article 106960. https://doi.org/10.1016/j.compeleceng.2020.106960.
[18] Pohl, O., 2024. Radiological assessment of pneumonia: Role of imaging modalities and interpretation challenges. *Journal of Clinical Respiratory Medicine*, 8(1), Article 190. https://doi.org/10.35841/aajcrm-8.1.190.
[19] Hanfi, S. H., Lalani, T. K., Saghir, A., McIntosh, L. J., Lo, H. S., and Kotecha, H. M., 2021. COVID-19 and its mimics: What the radiologist needs to know. *Journal of Thoracic Imaging*, 36(1), pp. W1–W10. https://doi.org/10.1097/RTI.0000000000000554.
[20] Wang, G., Liu, X., Shen, J., et al., 2021. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral, and COVID-19 pneumonia from chest X-ray images. *Nature Biomedical Engineering*, 5, pp. 509–521. https://doi.org/10.1038/s41551-021-00704-1.
[21] Wang, T., Nie, Z., Wang, R., Xu, Q., Huang, H., Xu, H., Xie, F., and Liu, X.-J., 2023. PneuNet: Deep learning for COVID-19 pneumonia diagnosis on chest X-ray image analysis using Vision Transformer. *Medical & Biological Engineering & Computing*, 61, pp. 1395–1408. https://doi.org/10.1007/s11517-022-02746-2.
[22] Rahman, T., Chowdhury, M. E. H., Khandakar, A., Islam, K. R., Islam, K. F., Mahbub, Z. B., Kadir, M. A., and Kashem, S., 2020. Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray. *Applied Sciences*, 10(9), Article 3233. https://doi.org/10.3390/app10093233.
[23] Kundu, R., Das, R., Geem, Z. W., Han, G.-T., and Sarkar, R., 2021. Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PLOS ONE*, 16(9), Article e0256630. https://doi.org/10.1371/journal.pone.0256630.
[24] Kermany, D. S., Goldbaum, M., Cai, W., et al., 2018. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), pp. 1122–1131. https://doi.org/10.1016/j.cell.2018.02.010.
[25] Rajpurkar, P., et al., 2017. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint*, arXiv:1711.05225. https://doi.org/10.48550/arXiv.1711.05225.
[26] Chouhan, V., Singh, S. K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R., and de Albuquerque, V. H. C., 2020. A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Applied Sciences*, 10(2), Article 559. https://doi.org/10.3390/app10020559.
[27] Al Reshan, M. S., Gill, K. S., Anand, V., Gupta, S., Alshahrani, H., Sulaiman, A., and Shaikh, A., 2023. Detection of pneumonia from chest X-ray images utilizing MobileNet model. *Healthcare*, 11(11), Article 1561. https://doi.org/10.3390/healthcare11111561.
[28] Ucan, M., Kaya, B., Aygun, O., Kaya, M., and Alhajj, R., 2025. Comparison of EfficientNet CNN models for multi-label chest X-ray disease diagnosis. *PeerJ Computer Science*, 11, Article e2968. https://doi.org/10.7717/peerj-cs.2968.
[29] Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q., 2017. Densely connected convolutional networks. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.
[30] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K., 2018. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint*, arXiv:1807.00431. https://doi.org/10.48550/arXiv.1807.00431.

[31] Bhardwaj, P., and Kaur, A., 2023. Impact of image enhancement methods on lung disease diagnosis using X-ray images. *International Journal of Information Technology*, 15, pp. 3521–3526. https://doi.org/10.1007/s41870-023-01409-1.

[32] Wang, Y., Guo, Y., Wang, Z., Yu, L., Yan, Y., and Gu, Z., 2024. Enhancing semantic segmentation in chest X-ray images through image preprocessing: ps-KDE for pixel-wise substitution by kernel density estimation. *PLOS ONE*, 19(6), Article e0299623. https://doi.org/10.1371/journal.pone.0299623.

[33] Saifullah, S., and Dreżewski, R., 2023. Modified histogram equalization for improved CNN medical image segmentation. *Procedia Computer Science*, 225, pp. 3021–3030. https://doi.org/10.1016/j.procs.2023.10.295.

[34] Nia, S. N., and Shih, F. Y., 2024. Medical X-ray image enhancement using global contrast-limited adaptive histogram equalization. *International Journal of Pattern Recognition and Artificial Intelligence*, 38(12), Article 2457010. https://doi.org/10.1142/S0218001424570106.

[35] Han, J., Choi, B., Kim, J. Y., and Lee, Y., 2025. BO-CLAHE enhancing neonatal chest X-ray image quality for improved lesion classification. *Scientific Reports*, 15, Article 4931. https://doi.org/10.1038/s41598-025-88451-0.

[36] Mabrouk, A., Díaz Redondo, R. P., Dahou, A., Abd Elaziz, M., and Kayed, M., 2022. Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks. *Applied Sciences*, 12(13), Article 6448. https://doi.org/10.3390/app12136448.

[37] Pandya, A., Pandya, K., and Yadav, H., 2026. Ensemble deep learning framework for pneumonia detection in chest X-rays using the RSNA dataset. In *Information Systems for Intelligent Systems*, edited by Iglesias, A., Shin, J., Bhatt, N., and Joshi, A., Vol. 1748, pp. 106–114. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-032-12990-1_11.

[38] Yanar, E., Hardalaç, F., and Ayturan, K., 2025. PELM: A deep learning model for early detection of pneumonia in chest radiography. *Applied Sciences*, 15(12), Article 6487. https://doi.org/10.3390/app15126487.

[39] Anumula, S. K., Tamilmani, V., Singh, A. A., Rajendran, D., and Namburi, V. D., 2026. Deep learning–based CNN model for automated detection of pneumonia from chest X-ray images. In *Innovative Computing and Communications*, edited by Hassanien, A. E., Anand, S., Jaiswal, A., and Kumar, P., Vol. 2045, pp. 360–377. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-032-30310-3_31.

[40] Balaazi, A., 2024. COVID-pneumonia detection: Expert-selected X-ray dataset (COVIDx-CXR). *Figshare*, Dataset. https://doi.org/10.6084/m9.figshare.25917340.

[41] Bassi, P. R. A. S., and Attux, R., 2022. COVID-19 detection using chest X-rays: Is lung segmentation important for generalization? *Research on Biomedical Engineering*, 38, pp. 1121–1139. https://doi.org/10.1007/s42600-022-00242-y.

[42] Wubineh, B. Z., Rusiecki, A., and Halawa, K., 2026. Viral pneumonia detection during the COVID-19 pandemic using deep learning and DCGAN-based data augmentation. *Scientific Reports*, 16, Article 22644. https://doi.org/10.1038/s41598-026-52523-6.

[43] Chitradevi, T. N., Swathypriyadharsini, P., Dhanusri, M., Nandhitha, K., Hemalatha, G., and Dhavamani, S., 2026. Transfer learning-based explainable CNN model for COVID-19 and pneumonia classification using chest X-rays. *Proceedings of the 2026 International Conference on Computing Networks and Inventive Communication Technologies (ICCNCT)*, Erode, India, pp. 1–6. https://doi.org/10.1109/ICCNCT68477.2026.11590114.

[44] Das, S., Ankushavali, M., Sharma, N., Badhani, R., and Kharola, A., 2026. Improving COVID-19 detection using transfer learning on chest X-ray images with a modified ResNet50 architecture. *Proceedings of the 2026 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, pp. 1–6. https://doi.org/10.1109/SCEECS68810.2026.11429822.