applied sciences

MDPI

Article

Robust Multiclass Pneumonia Classification via Multi-Head Attention and Transfer Learning Ensemble

Shenghua Rao, Zhuo Zeng and Jiemeng Zhang *

School of Mathematics and Physics, Wuhan Institute of Technology, Wuhan 430205, China; raoshenghua@stu.wit.edu.cn (S.R.); 15527853616@163.com (Z.Z.)
* Correspondence: zhangjiemeng@wit.edu.cn; Tel.: +86-133-4988-0709

Abstract

Pneumonia is an acute respiratory infection caused by pathogens such as bacteria or viruses, and accurate early diagnosis is critical for reducing mortality. Chest X-ray (CXR) imaging serves as a conventional diagnostic tool. However, radiographic features of pneumonia often overlap with those of other pulmonary diseases and are subject to inter-observer variability. Traditional Convolutional Neural Network (CNN) models tend to capture redundant information during feature extraction, and single pre-trained models often exhibit limited generalization in multiclass classification tasks. This study proposes a multi-model ensemble learning framework based on multi-head attention mechanism. Firstly, the three pre-trained backbones—DenseNet-121, ResNet-50, and VGG-19—were fine-tuned through transfer learning by replacing their classification heads, adapting pooling layers, and optimizing the fully connected layers. Secondly, feature maps extracted from these tuned backbones were concatenated and fused using a multi-head attention mechanism; the fused representation was then refined by two consecutive multi-head attention layers and finally passed to a fully connected classifier to produce the ensemble prediction. Three task sets were constructed from a public Kaggle dataset: binary classification (normal vs. pneumonia), three-class classification (normal, COVID-19, viral pneumonia), and four-class classification (normal, lung opacity, viral pneumonia, COVID-19), achieving accuracies of 91.67%, 93.79%, and 90.60%, respectively. The results demonstrate that the proposed multi-head attention-based ensemble framework offers significant advantages for pneumonia multiclass classification, particularly by maintaining high recall and robustness in more complex scenarios such as four-class differentiation, indicating its potential as a clinical decision-support tool. Future work will involve expanding the dataset and evaluating the model's generalizability across additional disease categories.

Keywords: transfer learning; ensemble learning; multi-head mechanism

Academic Editor: Pedro Couto

Received: 25 September 2025
Revised: 22 October 2025
Accepted: 23 October 2025
Published: 25 October 2025

Citation: Rao, S.; Zeng, Z.; Zhang, J. Robust Multiclass Pneumonia Classification via Multi-Head Attention and Transfer Learning Ensemble. *Appl. Sci.* **2025**, *15*, 11426. https://doi.org/10.3390/app152111426

1. Introduction

The lungs are among the most important organs in the body, and lung diseases often have a substantial impact on overall health. Pneumonia is an infection of the lungs caused by pathogens and can be broadly classified as infectious or non-infectious. Infectious pneumonia is further subdivided into cases caused by bacteria, viruses, mycoplasma species, or Chlamydia. The most prevalent pathogens causing viral pneumonia are influenza virus, respiratory syncytial virus (RSV), and SARS-CoV-2. Streptococcus pneumoniae is the most common cause of bacterial pneumonia [1]. Chest X-ray examinations have long been used to image the chest, and radiographs of the head, teeth, and bones are also commonly

performed. This imaging modality helps clinicians identify anatomical abnormalities and skeletal injuries. Although chest radiographs provide relatively low-cost imaging with limited information content, they remain a valuable diagnostic tool for detecting abnormalities and are commonly used in diagnosing pneumonia. However, even for experienced radiologists, diagnosing pneumonia from chest X-ray images can be challenging. Radiographic findings in pneumonia are often nonspecific and can be easily confused with those of other diseases, or may reflect other conditions, leading to subjective interpretations and diagnostic variability. Consequently, there is a need to develop computer-assisted diagnostic tools to support radiologists in diagnosing pneumonia.

In recent years, Artificial Intelligence (AI), especially its subfields Machine Learning (ML) and Deep Learning (DL), has demonstrated advantages in visual tasks, playing an increasingly pivotal role especially in the field of materials science [2,3]. AI has been applied to the early auxiliary diagnosis of various diseases, such as lung disease prediction, diabetic retinopathy, and the novel coronavirus (COVID-19) outbreak in 2019 [4]. These applications have also been successfully applied to pneumonia detection based on X-ray images. Given the difficulty of accurately identifying pneumonia from chest X-rays, developing efficient and highly accurate automated diagnostic methods is crucial for achieving early detection and reducing mortality rates. Meanwhile, attention mechanisms are gradually replacing traditional CNN models as a key advancement in current computer vision tasks [5]. Conventional CNNs often exhibit a tendency to uniformly process all features within an input image, an approach that can result in the accumulation of superfluous information and the occurrence of false negative predictions. Conversely, attention mechanisms have the capacity to selectively focus on valuable features and suppress redundant information, thereby enhancing model interpretability and classification performance without a substantial increase in computational costs. Kaya et al. [6] proposed a novel integrated CNN model that combines the model yielding the highest accuracy with the model yielding the lowest false negatives. This is achieved by identifying the optimal CNN model and combination weight ratios, with genetic algorithms (GA) setting the optimal weights. On public datasets, the model achieved an accuracy of 97.23% and an F1 score of 97.45. Lafraxo et al. [7] proposed a novel combined deep learning framework that employs median filters for image enhancement, followed by regularized convolutional neural networks and long short-term memory for feature extraction and classification. This approach ultimately achieved accuracies of 99.91% and 88.86% on the Kermany and RSNA datasets, respectively. Sunil Kumar et al. [8] proposed an innovative ensemble model, Efficient-VGG16, to address the need for rapid and precise classification of patients with confirmed cases of pneumonia due to the novel coronavirus (SARS-CoV-2) using chest X-ray images. This method combines the advantages of two models and compares them with traditional machine learning and transfer learning methods. The study utilised the COVID-Xray-5k dataset and employed a three-training-testing split ratio (80:20, 75:25, 70:30) to validate the model, achieving an accuracy rate of 99.46% and an F1 score of 98.41%. Mamalakis et al. [9] developed a new transfer learning pipeline named DenResCov-19, this integrated pipeline leverages DenseNet-121 and ResNet-50 architectures for chest X-ray image analysis, targeting the classification and detection of pneumonia, COVID-19, tuberculosis, and normal cases. The model attained an AUC of 96.51%, with performance metrics including 87.29% F1-score, 85.28% accuracy, and 89.38% overall recall. The hybrid model proposed by Ukwuoma et al. [10] combined a CNN and a Transformer encoder, using two ensemble models for feature extraction in different scenarios. Utilising the Mendeley dataset and the Chest X-ray-15k dataset, the model attained an overall accuracy of 99.21% and an F1 score of 99.21% for binary classification, and an accuracy of 98.19% and an F1 score of 97.29% for multi-class classification.

In ensemble learning research, Rajasekar et al. [11] proposed a Generative Autoencoder with Attention Mechanism (GAME) that integrates ensemble learning, unsupervised learning, and attention mechanisms. By incorporating attention mechanisms into generative autoencoders, it accurately localizes and extracts features within lung images. Simultaneously, it reduces the demand for high-quality data, ultimately achieving an F1 score of 0.95 and an accuracy of 0.95 on the CheXpert dataset. Meanwhile, Yanar et al. [12] introduced a novel deep learning framework—PELM (Pneumonia Ensemble Learning Model)—which sequentially integrates four high-performance pre-trained models: InceptionV3, ResNet50, VGG16, and Vision Transformer. This approach achieved a recall rate of 91% and an accuracy rate of 96% on a large dataset sourced from four distinct data sets. Prasath et al. [13] enhanced pneumonia detection accuracy by 20.7% and recall by 21.8%. This improvement was achieved through image preprocessing using region-aware neural graph collaborative filtering (RNGCF), feature extraction via wavelet transform, and final optimization with the Hunter Prey Optimization Algorithms (HPOA).

Transfer learning applied to various CNN architectures in an appropriate manner has been demonstrated to enhance the feature extraction abilities of machine learning models. Abbas et al. [14] trained a binary model based on DeTraCResNet18 to detect COVID-19 using a dataset of 196 images (105 COVID-19, 80 Normal, and 11 SARS cases). The model achieved an accuracy of 95.12%, sensitivity of 97.91%, specificity of 91.87%, and an overall accuracy of 93.36%. Lamouadene et al. [15] used a ResNet18 model combined with SVM applying transfer learning on chest X-rays containing 21,165 slides and used different optimizers to obtain classification rates of 94% with Adagrad optimizer, 96% with RMSProp optimizer, and 97% with Adam optimizer. Kurt et al. [16] emphasized the importance of image preprocessing and suggested a semi-automated process to improve the quality of the images, they proposed a transfer learning approach using the EfficientNet model and finally obtained results with 97.93% accuracy by fine-tuning techniques. Montalbo et al. [17] proposed the Fused-DenseNet-Tiny model: a lightweight DCNN model based on a densely connected neural network (DenseNet) truncated and concatenated. Through training transfer learning, and feature fusion, this model achieved an accuracy rate of 97.99%. Hussain et al. [18] introduced CoroDet, a novel CNN for automatic COVID-19 detection using raw chest X-ray and CT images; evaluated against ten existing techniques on a claimed largest X-ray dataset, it achieved accuracies of 99.1% (2-class), 94.2% (3-class) and 91.2% (4-class), outperforming state-of-the-art methods. Gifani et al. [19] posited a pre-trained model ensemble method based on a majority voting strategy. The scheme was trained and evaluated on a CT dataset containing 349 COVID-19 positive and 397 negative cases, ultimately achieving an accuracy rate of 85%. Mostafiz et al. [20] extracted the best features by using minimum redundancy and maximum relevance as well as recursive feature elimination in the mixture of features, and then detected the chest X-rays by using a random forest based bagging method. The overall accuracy was more than 98.5%. Nasiri et al. [21] gathered the features from the X-ray images through DenseNet-169 and used the collected features as inputs to the classification task performed by the XGBoost algorithm, which ultimately achieved accuracies of 98.23% for the two-class task and 89.70% for the three-class task. Aslan et al. [22] used features harvested by CNN model. They then used Bayesian optimisation to determine the hyperparameters of the machine learning algorithm and image segmentation based on ANN, which was applied to the COVID-19 X-ray dataset, achieving an accuracy rate of 96.29% and an F1 score of 94.53%. In another work, Jangam et al. [23] combined VGG-16 with DenseNet-169 to construct stacking ensemble model for COVID-19 detection in individual CT or chest X-ray. Evaluation showed that this hybrid approach performed best in SARS-CoV-2 identification, achieving an accuracy rate of 91.5% and a sensitivity of 95.5%.

In this study, we propose a novel approach that integrates deep learning with transfer learning. Specifically, three pre-trained models—DenseNet-121, ResNet-50, and VGG-19—are fine-tuned by augmenting them with task-specific classification layers. The performance of each model is evaluated using accuracy, precision, recall, and F1-score across binary, three-class, and four-class classification tasks. Furthermore, a multi-head attention mechanism is introduced to effectively fuse the feature representations from these models. This integration leads to an ensemble framework that significantly improves both predictive performance and stability across all evaluation metrics.

The experiments are conducted using three distinct datasets sourced from the public Kaggle repository. The study is structured around three main classification tasks: a binary classification distinguishing normal and pneumonia cases; a three-class classification extending the previous task by incorporating COVID-19 as a separate category; and a four-class classification further differentiating pneumonia into viral and bacterial subtypes. In summary, the principal contributions of this work are summarized as follows:

- We propose an ensemble model leveraging transfer learning based on DenseNet-121, ResNet-50, and VGG-19 to address classification tasks across varied datasets.
- A more extensive and balanced dataset is utilized to undertake more challenging multi-class classification tasks, leading to more stable and reliable model performance.

The remainder of this paper is organized as follows. We describe the datasets used, different pre-trained models, and the proposed ensemble model method in Section 2. Experimental results are presented in Section 3. Finally, discussions and conclusions are provided in Sections 4 and 5.

2. Materials and Methods

This section outlines the data preprocessing procedures employed in our study. It further presents the pre-trained deep learning models and the transfer learning framework adopted, as well as the subsequent ensemble learning method constructed thereafter. Finally, we describe the evaluation metrics used to assess the performance of the proposed models.

2.1. Description of the Dataset

Three different types of chest radiograph datasets are used in the deep learning model proposed in this paper, which are categorized as the binary dataset [24], tertiary dataset [25], and quaternary dataset [26], dataset [24] were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou, dataset [25] was compiled from dataset [24] and various publicly available resources published on the Kaggle website, dataset [26] was created by a group of researchers from Qatar University in Doha, Qatar, and the University of Dhaka in Bangladesh, in collaboration with doctors and partners from Pakistan and Malaysia.wherein dataset [24] denotes the dataset with two types of data (normal, pneumonia), and dataset [25] categorizes the normal chest images, viral pneumonia and COVID-19. Finally, dataset [26] is more diverse as it categorizes normal chest pictures, viral pneumonia, lung opacity images (non-COVID lung infection) and COVID-19. In total, three publicly available datasets were analyzed, with a total of 13,949 normal images, 4308 COVID-19 images, 6473 viral pneumonia images, and 6012 lung opacity images (non-COVID lung infection) obtained prior to data analysis. Figure 1 provides examples.

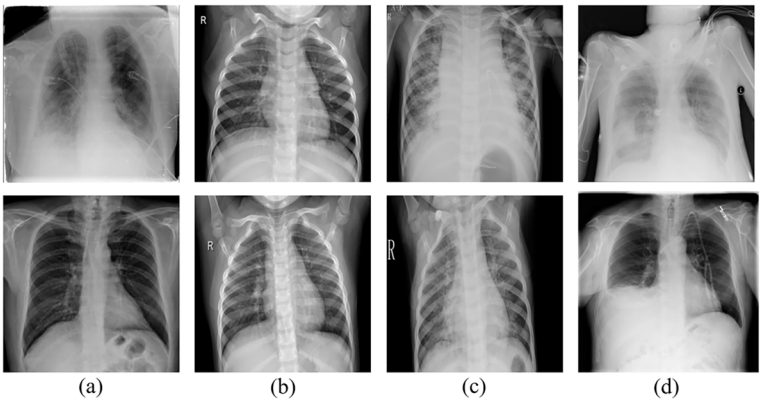

Figure 1. (**a**) COVID, (**b**) normal, (**c**) viral pneumonia, (**d**) lung opacity.

In the data preprocessing step, the chest radiographs exhibit varying resolutions, so all images were resized to 224 × 224 pixels prior to training and augmentation. This resizing step was crucial for ensuring consistency and compatibility with our chosen model. Paths for both training and testing datasets were specified. The data loading function loaded images from all categories in the training set. Stratified sampling was employed during data splitting to keep the proportion of each class in the training/validation/test sets consistent with the overall dataset, thereby preserving minority class representation even in smaller subsets. The total number of training images was determined.

Data augmentation is a necessary part of the deep learning model as it now requires a large amount of data to improve performance. An image data generator was created to implement data augmentation. The augmentation techniques applied included:

- Zooming: Pixel values are normalized to [0, 1] (i.e., multiplied by 1.0/255). Simultaneously, the scaling range dynamically resizes the original image to a ±10% scale variation.
- Random horizontal Flipping: Random mirroring along the vertical axis simulates lateral pose variations, while avoiding vertical flipping to preserve anatomical positioning of the thoracic cavity.
- Random Rotation: Introducing random rotations to the images, control the random rotation amplitude of the image within the range of angles [−15, 15].

By combining these preprocessing and enhancement techniques, we enable our model to adeptly adapt to variations in image quality, size, and perspective, ultimately significantly improving the accuracy of our pneumonia detection capabilities.

2.2. Pre-Trained Deep Learning Model

In this study, the selected pre-trained backbone deep learning models are VGG-19, ResNet-50 and DenseNet-121. Owing to the pre-training on the ImageNet dataset, these backbone networks demonstrate strong representational capabilities in capturing low-level features including spatial structure, rotation invariance, and edge information. All models are compatible with TensorFlow and Keras frameworks. In this study, they serve as base models, each of which is fine-tuned to adapt to the specific characteristics of different target datasets.

VGG-19 [27]: VGG-19 adopts a stacked structure of consecutive 3 × 3 convolution layers: the first two convolution blocks each contain two convolution layers, and the subsequent three convolution blocks each contain four convolution layers, with 2 × 2 maximum pooling following each block. At the end of the last convolution operation, three fully connected layers (FC) and softmax (for output) are added to complete the architecture.

DenseNet [28]: This architecture employs dense connections where each layer receives concatenated feature maps from all preceding layers within the same dense block. Transition layers between blocks use 1×1 convolutions followed by pooling to reduce feature dimensions. A global average pooling layer is typically applied after the final dense block but before the softmax classifier, effectively replacing fully connected layers while maintaining classification accuracy.

ResNet [29]: The milestone model proposed by He Kaiming's team in 2015 solves the gradient vanishing problem in deep networks through residual learning. The core component is the residual block: input data bypasses the convolution layer through a shortcut connection and is directly added to the output, enabling the network to learn 'differences' rather than direct mappings, thereby stabilising the training of thousand-layer networks. The ResNet-50 model was employed in this study.

2.3. Transfer Learning

Transfer learning is a machine learning technique that utilizes pre-trained models that have been trained for a specific problem and adapts them to new tasks by means of fine-tuning. This technique can effectively reduce the model training time and improve the generalization ability. Consider CNN as an example: training a model from scratch usually requires large-scale labeled datasets and substantial computing resources, whereas transfer learning leveraging pre-trained weights significantly accelerates convergence and enhances performance, especially with limited data. A typical example is the benchmark model in the ImageNet image recognition task, which was pre-trained on over a million images. In the initial stage of the feature extraction experiment, we removed the network head or the final layer of the pre-trained model, which was originally pre-trained on the ImageNet dataset. This step is important because pre-trained models are optimized for different classification tasks. Removing the classifier head discards the weights and biases associated with the original class scores, and the removed portion is replaced with newly initialized layers suitable for the target task. The architectural modifications of the VGG-19 pre-trained model primarily involves the following core improvements: adjust the average pooling size to 4×4, adjust the fully connected network dimensions to 512 and 256, and set the dropout layer parameter to 0.5, which is shown in Figure 2. The final layer sets different classification heads for the four categories of normal, COVID-19, viral pneumonia, and lung opacity according to the classification requirements of the dataset. The fine-tuning of the other two models is roughly the same as that of VGG-19.

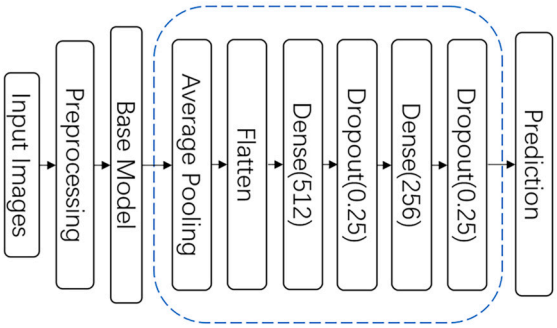

Figure 2. Fine-tuning process.

2.4. Multi-Head Attention Mechanism

The attention mechanism is a central concept in deep learning that mimics human visual attention and forms a core component of the transformer architecture. In essence, it is a process of weighting and summing the 'Values' based on the 'Query' and 'Key', and then redistributing the weights and generating the final output. Meanwhile, in the

self-attention layer, a scaled dot-product attention mechanism is used to form the attention function, mathematically speaking the 'Key' dimension d_k and the 'Value' dimension d_v are fed into the network. The similarity between 'Query' and 'Key' vectors is computed using the dot product and divided by $\sqrt{d_k}$, while using softmax to obtain the weight of each value. Use scaled dot-product attention to compute the self-attention output formula as

$$Attention(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{1}$$

Multi-head attention, based on self-attention, computes multiple self-attention heads in parallel, enabling the model to learn diverse features from different subspaces and to capture multiple dependencies. By employing multiple attention heads, the model can flexibly capture both local and global dependencies in the input sequence, which improves the performance and representation ability. Moreover, the multi-head attention mechanism does not raise computational costs but allows the model to capture relevant features across multiple representation subspaces. which improves the model's perceptual ability. Figure 3 shows the structural diagram of Scaled dot-product attention and multi-head attention mechanisms, and the calculation formula is as follows:

$$Multihead(Q,K,V) = Concat(head_1,\ldots,head_h)W^0 \tag{2}$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), W_i^Q \in R^{d_{model}\times d_q}, W_i^K \in R^{d_{model}\times d_k}, W_i^V \in R^{d_{model}\times d_v}, W^O \in R^{hd_i\times d_{model}} \tag{3}$$

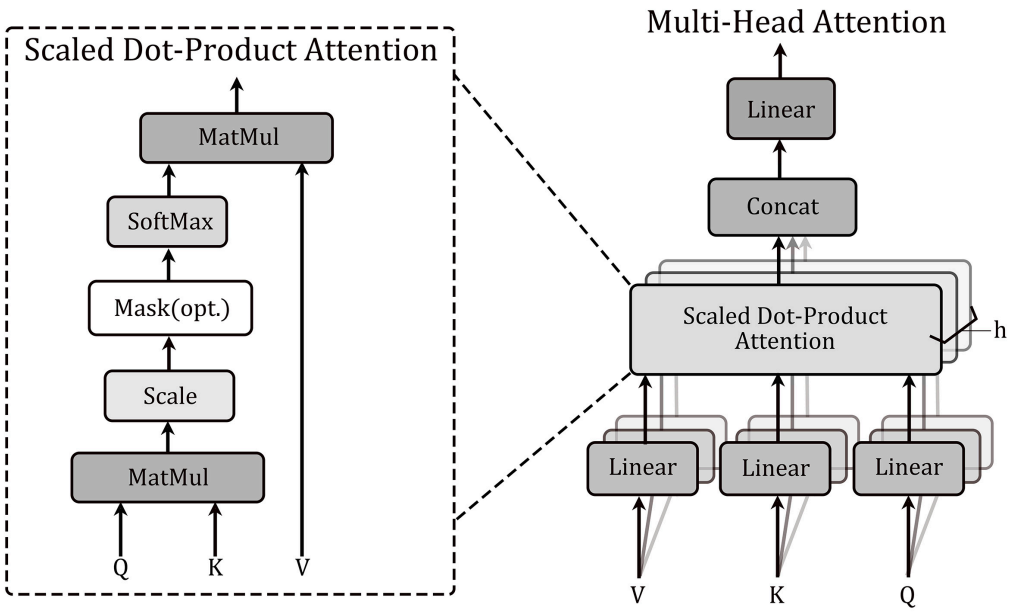

Figure 3. Scaled dot-product attention and the multi-head attention mechanism.

2.5. Multi-Model Ensemble Learning

In fact, deep learning networks are nonlinear models that offer considerable flexibility when training on small or sparse datasets. They are fine-tuned using random algorithms, and each training session involves some changes to the weights, which causes the neural network to produce different predictions for the results, resulting in high variance. To reduce this high variance generated during the training of deep neural networks, ensemble learning techniques can be used to learn from two, three, or more different deep neural network models. These different neural networks are then combined to predict the final results. To address the demand for model optimization within the field of pneumonia

recognition in medical image analysis, the majority of existing studies focus on performance improvement of a single deep convolutional architecture. In contrast, the systematic exploration of multi-model collaborative frameworks specifically for pneumonia recognition is still insufficient. In this study, we propose an ensemble learning approach based on a multi-head attention mechanism to analyze datasets from multiple sources, fuse features from several pre-trained models, and construct a hierarchical feature fusion mechanism to enhance classification robustness and performance. The aim is to achieve better feature extraction and overall performance gains.

In the ensemble model, the workflow begins with data preprocessing, followed by fine-tuning each of three pre-trained models. Then, the features corresponding to the final convolutional layers of each model are extracted. The feature vectors from multiple models are first concatenated to form a new feature sequence. Next, the first multi-head attention layer is introduced, which uses self-attention mechanism with multiple attention heads to extract important information from the overall features. To facilitate deep learning, residual connections are used to add the attention outputs to the input features. This architectural choice mitigates vanishing gradients and helps preserve information during training. After applying layer normalization and dropout, the model's training performance is further enhanced and overfitting is reduced. The model's output is subsequently flattened and reshaped to serve as input to the second multi-head attention layer, where the same sequence of operations is repeated. Finally, the resulting representations are passed through a dense (fully connected) layer and a softmax activation to yield the final classification results, as shown in Figure 4.

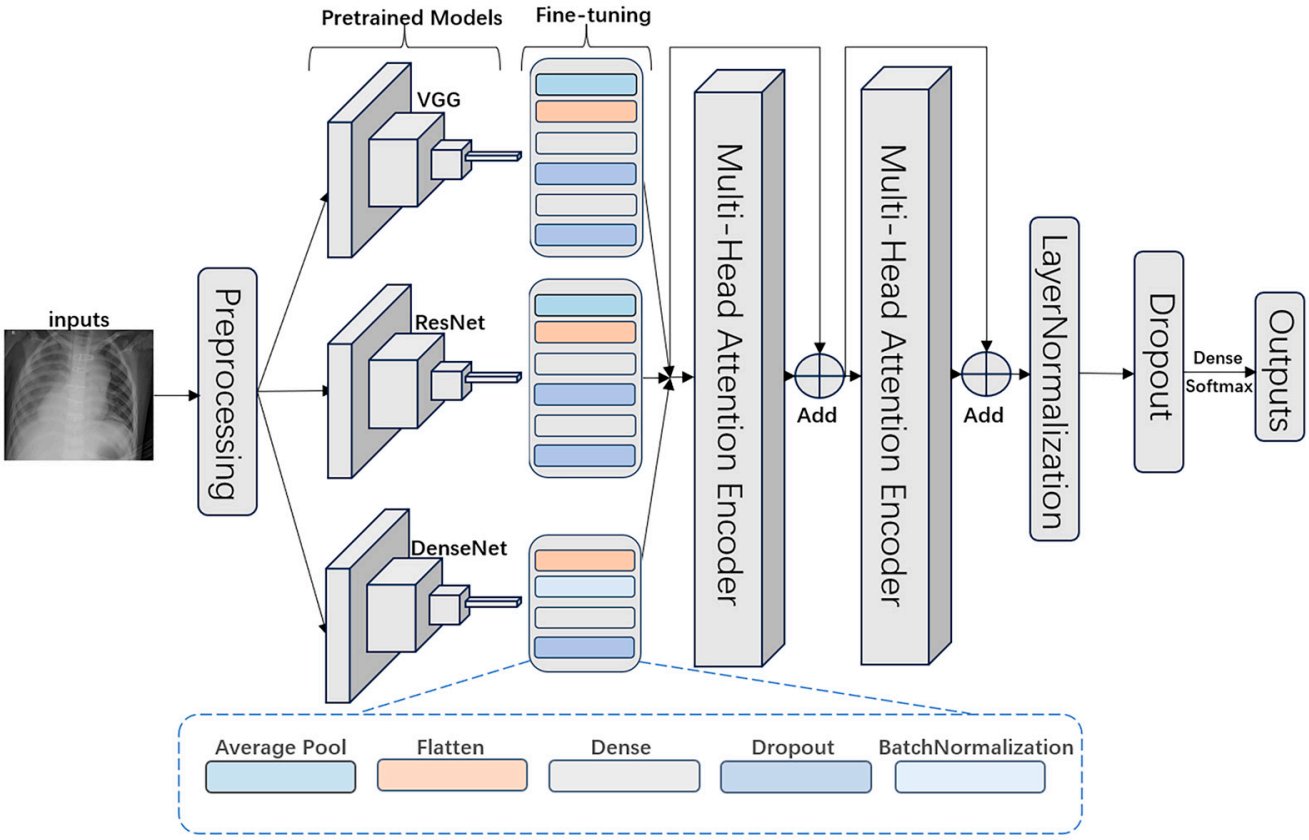

Figure 4. The proposed ensemble model.

2.6. Performance Evaluation Metrics

A range of performance evaluation metrics is employed to assess the proposed model, including accuracy, recall, precision, and the F1 score, and the specific formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

In the equation above, *TP* stands for true positives (cases that are truly positive and correctly predicted as positive by the model), *TN* stands for true negatives (cases that are truly negative and correctly predicted as negative), *FP* stands for false positives (cases that are actually negative but mistakenly predicted as positive), and *FN* stands for false negatives (cases that are actually positive but mistakenly predicted as negative). They form the basis for metrics such as accuracy, recall, specificity, and precision, which help assess overall correctness, the model's ability to detect positive cases, and its tendency to produce false alarms.

3. Results

3.1. Experimental Setup

We used Python (3.12.4), TensorFlow (2.6.0), Keras (2.6.0), Sklearn (1.5.2), OpenCV (4.10), matplotlib (3.7.5), Pandas (2.2.2) and NumPy (1.23.4) libraries. We trained all models for 20 epochs while optimizing using the Adam optimizer with a learning rate of 0.00001 and a Batch size of 16. All experiments were run on a pc with the following hardware specifications: A Lenovo laptop in China featuring an Intel(R) Core(TM) i7-12800HX@2.00 GHz and an NVIDIA GeForce RTX 3060 graphics processing unit (GPU). We ran experiments for binary and multi-class classification to evaluate the performance under two different settings. The data split rate is 80%, 10%, and 10% for the training set and 80%, 10%, and 10% for the test and validation sets. It should be noted that all models were trained for 20 epochs using the Adam optimizer with a learning rate of 0.0001, a batch size of 16, and classification cross-entropy. Each epoch took approximately 137 s to complete given the training dataset and computational setup, resulting in a speed of 689 ms/step.

Meanwhile, the estimated FLOPS for the VGG-19 network is approximately 39.03 GFLOPs, the estimated FLOPS for the ResNet-50 network is approximately 3.86 GFLOPs, the estimated FLOPS for the DenseNet-121 network is approximately 2.81 GFLOPs, and the estimated FLOPS for the proposed ensemble model is approximately 46.02 GFLOPs.

3.2. Testing Binary Classification

We tested all the models used in this study on a binary classification dataset and present the results in Table 1. The VGG-19 model achieved an accuracy of 91.03%, precision of 92.51%, recall of 88.63%, and an F1 score of 90.02%. ResNet-50 yielded 87.02% accuracy, 89.61% precision, 83.38% recall, and an F1 score of 85.16%. DenseNet-121 obtained 88.30% accuracy, 88.99% precision, 85.94% recall, and an F1 score of 87.06%. The ensemble model produced an accuracy of 91.67%, precision of 92.19%, recall of 90.00%, and an F1 score of 90.89%. Overall, among the single pre-trained networks, VGG-19 performed best. The ensemble provided a modest but meaningful improvement over the individual models—

particularly in recall and F1—resulting in the best-balanced performance for this binary classification task.

Table 1. Binary classification results.

Model	Accuracy	Precision	Recall	F1-Score
VGG19	0.9103	0.9251	0.8863	0.9002
ResNet-50	0.8702	0.8961	0.8338	0.8516
DenseNet121	0.8830	0.8899	0.8594	0.8706
Ensemble	0.9167	0.9219	0.9000	0.9089

3.3. Testing Multiple Classification

We tested the model used in this study on three-class classification tasks and four-class classification tasks, and the results are shown in Tables 2 and 3, respectively. Overall, the outcomes are similar. VGG-19 performs less well on multi-class tasks: it recognizes positive-class instances relatively well in the three-class task, but its performance drops markedly in the four-class task, indicating weaker adaptability as the number of classes increases. ResNet-50 performs slightly below DenseNet-121 on multi-class tasks but shows the smallest accuracy decline as classes increase and generalizes better than the other models. However, ResNet-50's recall in the three-class task is lower than that of the other models, which raises the risk of missing positive-class instances. The DenseNet-121 network model demonstrates the most balanced performance and overall superiority in multi-class classification tasks, maintaining leading levels of accuracy and F1 scores. However, with an increase in the number of categories, the decline in its accuracy exhibits the most pronounced trend among all candidate models; furthermore, its stability when applied to complex tasks still leaves room for improvement. In contrast, the proposed ensemble network model combines the advantages of the three models while maintaining high robustness. It is not affected by poorly trained models in classification tasks and maintains the highest recall rate, indicating that the proposed ensemble model can handle multi-category classification tasks without missing or misclassifying instances. In summary, the proposed ensemble model demonstrates excellent adaptability to complex tasks, balanced performance metrics, and stability.

Table 2. Three classification results.

Model	Accuracy	Precision	Recall	F1-Score
VGG19	0.9006	0.9016	0.9445	0.9134
ResNet-50	0.9325	0.9348	0.9219	0.9278
DenseNet121	0.9433	0.9384	0.9576	0.9452
Ensemble	0.9379	0.9243	0.9599	0.9389

Table 3. Four classification results.

Model	Accuracy	Precision	Recall	F1-Score
VGG19	0.7841	0.8027	0.7721	0.7737
ResNet-50	0.9027	0.9243	0.9046	0.9118
DenseNet121	0.9017	0.9246	0.9111	0.9175
Ensemble	0.9060	0.9072	0.9145	0.9094

Meanwhile, Figure 5 displays the confusion matrices of the test ensemble model across different classification tasks. It can be observed that the model proposed in this study achieved favorable results under the classification challenges posed by three datasets.

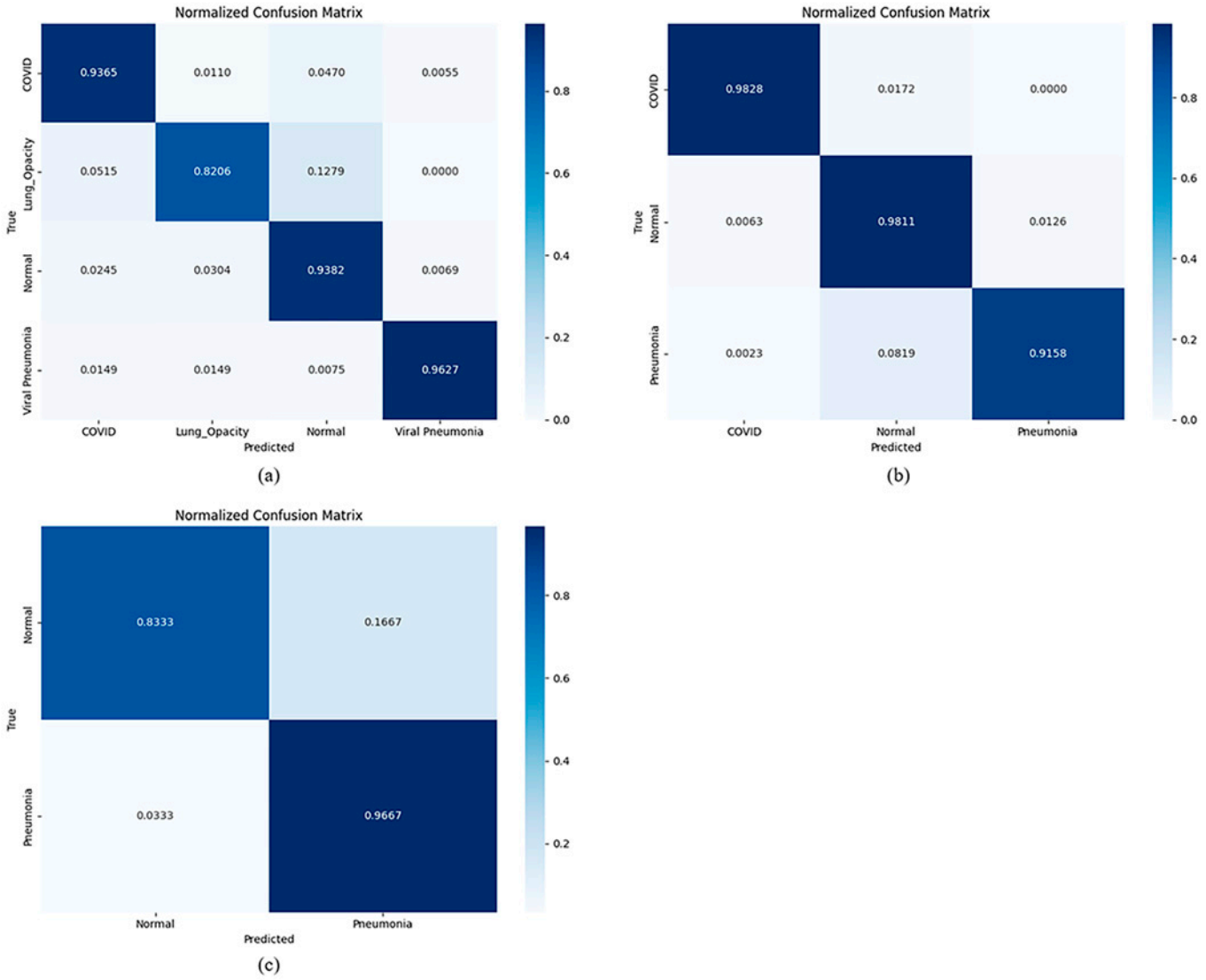

Figure 5. The proposed ensemble model confusion matrix, (**a**) Four-Class Confusion Matrix, (**b**) Three-Class Confusion Matrix, (**c**) Binary Classification Confusion Matrix.

3.4. Ablation Study

In the field of machine learning, ablation study involves removing or temporarily disabling certain modules or components to compare whether the results differ, thereby demonstrating the contribution of these modules to the overall experiment. The primary purpose of ablation study is to identify which modules are sufficiently important.

In this study, we removed the multi-head attention mechanism from the model to determine whether this module influenced the results. The findings are presented in Table 4. After removing the multi-head attention mechanism, comparing the ensemble model formed by concatenating the three transfer models with the original model reveals that the proposed method achieves significant improvements in classification performance, particularly demonstrating superior recall and F1-score metrics. This enhancement is more pronounced in multi-class classification tasks, validating the effectiveness of innovative components such as the introduced attention mechanism. This approach is particularly well-suited for practical applications demanding high recall rates.

Table 4. Comparison of Ablation Study Results.

Dataset	Method	Accuracy	Precision	Recall	F1-Score
Binary classification	Proposed	0.9167	0.9219	0.9000	0.9089
	Transfer + Concatenate	0.8734	0.9109	0.8329	0.8533
Three-class classification	Proposed	0.9379	0.9243	0.9599	0.9389
	Transfer + Concatenate	0.9317	0.9248	0.9422	0.9328
Four-class classification	Proposed	0.9060	0.9072	0.9145	0.9094
	Transfer + Concatenate	0.8653	0.9217	0.8225	0.8613

4. Discussion

We compared our work with recent literature and present the results in Table 5. The numbers in the first column represent the two-class, three-class, and four-class classifications. The results show that the integrated model proposed in this study performs well on different types of datasets because it achieves good results on multiple large and balanced datasets. We have reason to believe that the model we propose is meaningful.

Table 5. Comparison of our approach with state-of-the-art methods for Pneumonia detection.

Classes	Reference	Dataest	Techniques	Accuracy%
Two Class	[14]	349 COVID-19, 397 non-COVID,	EfficientNets	85
	[16]	500 pneumonia and 500 normal,	DenseNet169 + XGBoost	98.23
	Proposed	4537 pneumonia and 1319 normal	Ensemble	91.67
Three Class	[8]	692 COVID-19, 1900 normal and 5128 pneumonia,	Ensemble	94.64
	[9]	80 normal, 105 COVID-19, 11 SARS,	DeTraC	93.1
	[16]	125 COVID-19, 500 pneumonia and 500 normal,	DenseNet169 + XGBoost	89.70
	Proposed	692 COVID-19, 1900 normal and 5128 pneumonia	Ensemble	93.79
Four Class	[8]	10192 normal, 3616 COVID-19, 6012 lung opacity and 1345 viral pneumonia,	Ensemble	86.53
	[10]	10192 normal, 3616 COVID-19, 6012 lung opacity and 1345 viral pneumonia,	ResNet18-SVM	86.18
	Proposed	10192 normal, 3616 COVID-19, 6012 lung opacity and 1345 viral pneumonia	Ensemble	90.60

This study shows that the ensemble model is based on the multi-head attention mechanism. By integrating three pre-trained networks, the model allows users to flexibly combine prediction models according to the quality of datasets or pre-trained models. By extracting deep features from multiple training models, it can effectively ensure the independence of features during the integration process. Comparisons among Tables 1–3 indicate that the ensemble model generally outperforms other deep CNN architectures, exhibiting higher accuracy, recall, and robustness, and it also demonstrates strong generalization. Ensemble modeling combines three pre-trained deep CNNs, enabling the meta-learner to determine predictions based on distinct features learned by the base models. Deep features extracted from the three trained models ensure feature independence within the ensemble process. Unlike traditional ensemble learning models that aggregate final predictions from base models to generate ensemble forecasts, deep ensemble learning models utilize deep feature

vectors or feature maps from trained models to train shallow or deep meta-learners. We plan to augment the dataset with additional data resources. For future work, we also plan to explore other deep learning models and compare their performance with the models employed in this study.

Typically, multi-head attention mechanisms are applied within a single model (e.g., in Vision Transformers) to capture complex relationships and patterns among features in different regions of the same X-ray image. In our research, the multi-head attention mechanism is employed as a meta-classifier within the ensemble framework. Its input is not the raw pixels or local features of the original image, but rather the high-level abstract features extracted by multiple pre-trained models (such as DenseNet-121, VGG-19, ResNet-50, etc.). Its core task is to dynamically learn and balance the importance and interrelationships among these heterogeneous features from different architectures. Simultaneously, it incorporates residual connections by linking the attention outputs to the inputs through residual connections, combined with LayerNormalization and Dropout, which facilitates stable training and preserves original feature information. This design enables the attention module to refine/reconstruct the fused features without losing the original discriminative power.

5. Conclusions

As detailed in Table 5, we evaluate our method alongside the latest top-performing techniques. Although we have compared the performance of the proposed model with existing literature (as shown in Table 5), caution must be exercised when interpreting these cross-sectional comparisons. A significant limitation lies in the fact that the cited studies employed datasets that differ in scale, category distribution, and image sources.

These variations introduce inherent methodological limitations to direct comparisons of accuracy metrics. For instance, a model trained on a small, balanced dataset may report high accuracy that fails to generalize to more challenging, large-scale datasets closer to real-world distributions. This also represents an unresolved standard in current deep learning research. There is currently no unified large-scale public dataset available. Moreover, the objectives of research are not entirely consistent. Therefore, the comparisons in Table 5 should be viewed as indicative trends, aiming to situate our work within a broader academic context rather than asserting absolute superiority. We strongly recommend that future research be conducted on unified public benchmark datasets to enable fairer and more meaningful comparisons.

Using chest X-rays to predict COVID-19 can prevent the disease from spreading in the chest and detect the virus more quickly. In this study, we used transfer learning to train, test, and validate three widely deep learning algorithms. We tested DenseNet-121, VGG-19, and ResNet-50 as pre-trained models to classify whether chest X-rays can predict COVID-19, prevent the disease from spreading in the chest, and detect the virus more quickly. In this study, we used transfer learning to train, validate, and test three popular deep learning algorithms. We tested DenseNet-121, VGG-19, and ResNet-50 as pre-trained models for classifying CXR images of pneumonia. The results showed that the DenseNet-121 pretrained model achieved the best performance. Additionally, this study introduced pneumonia datasets collected from three different sources. After data augmentation, the dataset consisted of 13,949 normal images, 6012 images of lung opacity, 6473 images of viral pneumonia, and 4308 chest X-ray images of COVID-19. Furthermore, the proposed multi-head attention-based ensemble model demonstrated consistent performance across different datasets and achieved strong training outcomes. The results can assist medical experts in early identification of pneumonia types from chest X-rays, thereby supporting faster clinical decision-making. Future work will focus on expanding the dataset if more

open-source data becomes available, as well as incorporating chest X-ray images of other thoracic diseases.

Author Contributions: Conceptualization, S.R. and Z.Z.; methodology, S.R. and Z.Z.; software, S.R.; validation, S.R.; writing—original draft preparation, S.R.; writing—review and editing, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 12101469.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study has been publicly released on Kaggle (URL: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data; https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia/data; https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database/data), accessed on 14 September 2024.

Acknowledgments: The authors thank the players of the team who participated in the study.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
VGG	Visual Geometry Group
DenseNet	Densely Connected Convolutional Networks
ResNet	Residual Neural Network

References

1. Pneumonia | CDC. Available online: https://www.cdc.gov/pneumonia/index.html (accessed on 14 September 2024).
2. Asif, S.; Wenhui, Y.; Rehman, U.S.; Ul-Ain, Q.; Amjad, K.; Yueyang, Y.; Jinhai, S.; Awais, M. Advancements and prospects of machine learning in medical diagnostics: Unveiling the future of diagnostic precision. *Arch. Comput. Methods Eng.* **2025**, *32*, 853–883. [CrossRef]
3. Chong, S.S.; Ng, Y.S.; Wang, H.Q.; Zheng, J.C. Advances of machine learning in materials science: Ideas and techniques. *Front. Phys.* **2024**, *19*, 13501. [CrossRef]
4. Paules, C.I.; Marston, H.D.; Fauci, A.S. Coronavirus Infections-More Than Just the Common Cold. *JAMA* **2020**, *323*, 707–708. [CrossRef]
5. Fu, J.; Li, W.; Du, J.; Huang, Y. A multiscale residual pyramid attention network for medical image fusion. *Biomed. Signal Process. Control* **2021**, *66*, 102488. [CrossRef]
6. Kaya, M.; Çetin-Kaya, Y. A novel ensemble learning framework based on a genetic algorithm for the classification of pneumonia. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108494. [CrossRef]
7. Lafraxo, S.; El Ansari, M.; Koutti, L. A new hybrid approach for pneumonia detection using chest X-rays based on ACNN-LSTM and attention mechanism. *Multimed. Tools Appl.* **2024**, *83*, 73055–73077. [CrossRef]
8. Kumar, S.; Kumar, H. Efficient-VGG16: A Novel Ensemble Method for the Classification of COVID-19 X-ray Images in Contrast to Machine and Transfer learning. *Procedia Comput. Sci.* **2024**, *235*, 1289–1299. [CrossRef]
9. Mamalakis, M.; Swift, A.J.; Vorselaars, B.; Ray, S.; Weeks, S.; Ding, W.; Clayton, R.H.; Mackenzie, L.S.; Banerjee, A. DenResCov-19: A deep Transfer learning network for robust automatic classification of COVID-19, pneumonia, and tuberculosis from X-rays. *Comput. Med. Imaging Graph.* **2021**, *94*, 102008. [CrossRef] [PubMed]
10. Ukwuoma, C.C.; Qin, Z.; Heyat, M.B.B.; Akhtar, F.; Bamisile, O.; Muaad, A.Y.; Addo, D.; Al-Antari, M.A. A hybrid explainable ensemble transformer encoder for pneumonia identification from chest X-ray images. *J. Adv. Res.* **2023**, *48*, 191–211. [CrossRef] [PubMed]
11. Rajasekar, E.; Chandra, H.; Pears, N.; Vairavasundaram, S.; Kotecha, K. Lung image quality assessment and diagnosis using generative autoencoders in unsupervised ensemble learning. *Biomed. Signal Process. Control* **2025**, *102*, 107268. [CrossRef]

12. Yanar, E.; Hardalaç, F.; Ayturan, K. PELM: A Deep Learning Model for Early Detection of Pneumonia in Chest Radiography. *Appl. Sci.* **2025**, *15*, 6487. [CrossRef]
13. Prasath, J.; Prabu, S.; Mayil, V.V.; Saini, S. Optimized double transformer residual super-resolution network-based X-ray images for classification of pneumonia identification. *Knowl. Based Syst.* **2025**, *311*, 113037.
14. Abbas, A.; Abdelsamea, M.M.; Gaber, M.M. Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Appl. Intell.* **2021**, *51*, 854–864. [CrossRef] [PubMed]
15. Lamouadene, H.; El Kassaoui, M.; El Yadari, M.; El Kenz, A.; Benyoussef, A.; El Moutaouakil, A.; Mounkachi, O. Detection of COVID-19, lung opacity, and viral pneumonia via X-ray using machine learning and deep learning. *Comput. Biol. Med.* **2025**, *191*, 110131. [CrossRef]
16. Kurt, Z.; Işık, Ş.; Kaya, Z.; Anagün, Y.; Koca, N.; Çiçek, S. Evaluation of EfficientNet models for COVID-19 detection using lung parenchyma. *Neural Comput. Appl.* **2023**, *35*, 12121–12132. [CrossRef]
17. Montalbo, F.J.P. Diagnosing Covid-19 chest x-rays with a lightweight truncated DenseNet with partial layer freezing and feature fusion. *Biomed. Signal Process. Control* **2021**, *68*, 102583. [CrossRef]
18. Hussain, E.; Hasan, M.; Rahman, M.A.; Lee, I.; Tamanna, T.; Parvez, M.Z. CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images. *Chaos Solitons Fractals* **2021**, *142*, 110495. [CrossRef]
19. Gifani, P.; Shalbaf, A.; Vafaeezadeh, M. Automated detection of COVID-19 using ensemble of Transfer learning with deep convolutional neural network based on CT scans. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 115–123. [CrossRef]
20. Mostafiz, R.; Uddin, M.S.; Reza, M.M.; Rahman, M.M. Covid-19 detection in chest X-ray through random forest classifier using a hybridization of deep CNN and DWT optimized features. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 3226–3235. [CrossRef]
21. Nasiri, H.; Hasani, S. Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost. *Radiography* **2022**, *28*, 732–738. [CrossRef]
22. Aslan, M.F.; Sabanci, K.; Durdu, A.; Unlersen, M.F. COVID-19 diagnosis using state-of-the-art CNN architecture features and Bayesian Optimization. *Comput. Biol. Med.* **2022**, *142*, 105244. [CrossRef] [PubMed]
23. Jangam, E.; Barreto, A.A.D.; Annavarapu, C.S.R. Automatic detection of COVID-19 from chest CT scan and chest X-Rays images using deep learning, Transfer learning and stacking. *Appl. Intell.* **2022**, *52*, 2243–2259. [CrossRef] [PubMed]
24. Binary Classification Dataset. Available online: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia/data (accessed on 14 September 2024).
25. Ternary Classification Dataset. Available online: https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia/data (accessed on 14 September 2024).
26. Quaternary Classification Dataset. Available online: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database/data (accessed on 14 September 2024).
27. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
28. Huang, G.; Liu, Z.; Weinberger, Q.K.; van der Maaten, L. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.