Apply a CNN-Based Ensemble Model to Chest-X Ray Image-Based Pneumonia Classification

Ngoc Ha Pham [1,2,*] and Giang Son Tran [1]

[1] Information, Communication and Technology Laboratory, University of Science and Technology of Hanoi, Hanoi, Vietnam
[2] Information and Communication Technology Department, FPT University, Hanoi, Vietnam
Email: hapn10@fe.edu.vn (N.H.P.); tran-giang.son@usth.edu.vn (G.S.T.)
*Corresponding author

***Abstract*—Pneumonia commonly results from a lung ailment that leads to irritation and harm to the lungs. A chest X-ray is one of the most effective imaging techniques for detecting pneumonia, but diagnosing and treating it can be difficult due to its similarity to other lung conditions. To improve the accuracy of classifying X-ray images, we suggest using an ensemble model in our research that combines deep Convolutional Neural Network (CNN) architectures. The suggested approach classifies the input image as having pneumonia or not by extracting data features using an ensemble of three CNN models. The comparison involves using a single CNN model and a combination of CNN models to evaluate the ensemble architecture. This work evaluates the InceptionResNetV2, DenseNet201, and VGG16 ensemble. The suggested ensemble algorithm provides comparatively positive classification results with an accuracy of almost 95%, outperforming previous ensemble models and improving the average F1-Score by 3% compared to the single model approach.**

***Keywords*—pneumonia, chest X-ray, ensemble learning, deep learning, convolutional neural network**

I. Introduction

Pneumonia can present with mild to life-threatening symptoms, posing significant risks, particularly to individuals who are over 65, have pre-existing health conditions or weakened immune systems, newborns, or young toddlers. In 2019, over 700,000 kids who are younger than five years old lost their lives due to pneumonia, making it the leading cause of death for children. Pneumonia contributes to 14% of all child deaths; the highest number of cases is found in South Asia and West and Central Africa [1]. Vietnam ranks among the top 15 countries with a significant number of children impacted by pneumonia, seeing around 4,000 annual deaths out of 2.9 million cases [2]. At the end of 2019, the World Health Organization (WHO) determined that the COVID-19 pandemic was caused by a novel coronavirus strain leading to characteristic pneumonia. By May 2023, the pandemic had resulted in approximately 6.8 million deaths and 676 million infections worldwide. In Vietnam, there have been 11,624,000 confirmed COVID-19 cases and 43,206 deaths, making Vietnam the country with the highest number of confirmed cases in Southeast Asia and the 13th highest globally. Hanoi is the hardest hit, with 1,649,654 cases and 1,238 deaths, followed by Ho Chi Minh City, with 628,736 cases and 20,476 deaths.

The pandemic has significantly affected significant sectors such as manufacturing, economics, medicine, and education. Viral diseases like COVID-19 pose a severe threat to public health, and the virus's rapid, widespread, and frequent mutations make prevention, detection, control, and treatment more complex. One of the best ways to diagnose pneumonia and COVID-19 is by using a chest X-ray. An electromagnetic radiation type called an X-ray has a shorter wavelength, more incredible energy than ultraviolet light, and the ability to pass through solid materials. Using a chest X-ray, radiologists can locate and gauge the severity of an infection by examining the lungs for white patches known as infiltrates. The X-ray scans displayed in Fig. 1 illustrate the difference between a normal lung and a lung infected with pneumonia. White patches on the lung X-ray are indicative of a pneumonia case, setting it apart from a typical case. Such a diagnosis necessitates X-ray reading expertise and understanding. Diagnosing pneumonia using X-rays can be a laborious and less precise process because several other illnesses, including lung cancer or an infection, can produce comparable picture opacities. Radiologists and doctors urgently need computer-aided diagnosis methods to assist in reducing pneumonia mortality, especially in children.

A. Background and Motivation

The pandemic has seen technology play a crucial role in battling COVID-19. In the age of 5G, Beyond 5G (B5G), and 6G, along with medical cloud services, mobile applications, and Artificial Intelligence (AI), advancements in bioinformatics have created unique opportunities for virus informatics research. These developments are essential for comprehensively modeling virus biology at a systems level. Deep learning methods to treat medical conditions have become increasingly popular. Deep learning methods to treat medical conditions have become increasingly popular. Most current approaches in these areas rely on deep learning methods that mimic how people learn and acquire specific types of information.

Manuscript received January 15, 2024; revised April 24, 2024; accepted July 30, 2024; published November 8, 2024.

Computer-assisted diagnostic systems utilizing Convolutional Neural Networks (CNNs) are increasingly prevalent and significantly advance image processing. A Convolutional Neural Network (CNN) is designed to mimic biological processes' receptive fields. A Convolutional Neural Network (CNN) is designed to mimic biological processes' receptive fields. Well-known CNN architectures include ResNet [3], DenseNet [4], AlexNet [5], GoogLeNet [6], and others. For X-ray image processing, researchers have proposed several computer algorithms, including image segmentation tasks with U-Net [7], image classification tasks with VGG16 [8], and image detection tasks with Faster R-CNN [9], YOLOv3 [10], and Mask R-CNN [11].

B. Proposed Work

This research mainly focuses on automatically classifying X-ray pictures as pneumonia using an ensemble model. This strategy, which combines several conventional CNN techniques, not only performs better in accuracy but also in processing efficiency. The stacking method involves creating, training, and evaluating a meta-model compared to other models. This meta-model consists of three CNN models: DenseNet201 [12], VGG16, and InceptionResNetV2 [13]. These tests illustrate that a layered approach outperforms a single CNN-based model when applied to the Chest X-ray dataset, reassuring the audience about its practicality.

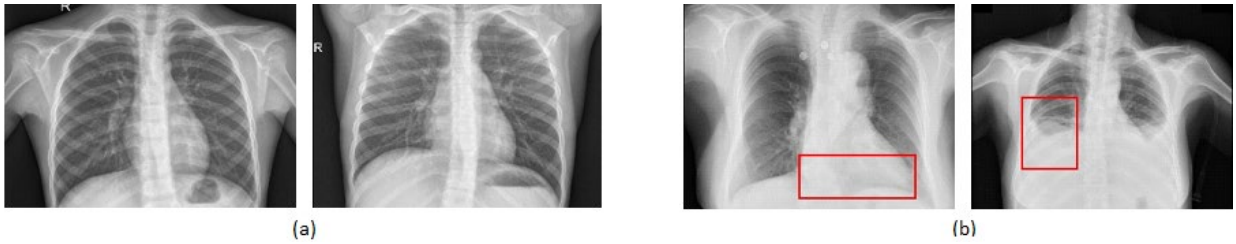


Fig. 1. Examples of images from the dataset provided by Kermany [21] are shown. The images in (a) represent normal cases, while the images in (b) represent pneumonia cases.

C. Contribution

The work we have done has made important contributions that can be summarized as follows:

- We present a deep learning function integration technique to classify pneumonia images. This approach combines the benefits of ensemble learning and deep learning, potentially increasing the entire model's performance by leveraging multiple CNN models. (R1)
- The model's accuracy depends on the number of deep learning models assembled to create ensembles. (R2)
- Compared to a single novel methodology such as ViT-B16 in the pneumonia picture dataset, the combined deep learning CNN models yield better classification results. (R3)

The following paper is organized in the following manner: Section II recaps the different networks important to this study. Section III explains a detailed description of the proposed ensemble network. Section IV displays the experimental findings on performance and assesses each situation. In Section V, the conclusion and the direction of future development are explored.

II. Related Work

Because deep learning is developing quickly and large datasets are available, radiologists now have easier access to artificial intelligence alternatives. As a result, numerous research studies focusing on X-ray imaging have produced positive outcomes when compared to the general performance of CNN models. Deep learning models in the modern era have been improved to specifically categorize pneumonia based on chest X-ray images. Another crucial strategy to enhance pneumonia prediction performance is by employing ensemble techniques. The dataset for the RSNA challenge includes 30,000 frontal view chest X-rays and was utilized by Shih *et al.* [14] from a pool of 112,000 open images in the CXR-8 dataset [15]. Kundu *et al.* [16] used decision scores from three CNN models to develop an ensemble framework that generates a weighted average ensemble: ResNet-18, GoogLeNet, and DenseNet-121. This framework's results on the RSNA challenge dataset were 86.86%, 87.02%, 86.89%, and 86.95%, respectively, for accuracy, sensitivity, precision, and F1-Score. The two pneumonia X-ray datasets used to obtain these results are accessible to the public. Harsh, Bhatt, and colleagues [17] investigated an ensemble network of three CNN models. The computational expense was kept low while maintaining accuracy and other metrics by utilizing three models with different kernel sizes. Convolutional neural networks served as the inspiration for all these algorithms. We can enhance performance by modifying the architecture versions of most single CNN models. An effective approach is to employ an ensemble, which amalgamates the strengths of multiple high-performing models to tackle a regression or classification problem, yielding results superior to any individual model. An *et al.* [18] suggested creating an attention ensemble using deep CNN models such as EfficientNetB0, and DenseNet121. Li *et al.* [19] proposed an ensemble learning strategy for the radiographic categorization of pneumonia. The VGG16 algorithm is used along with layered generalization ensemble learning to create a cascade classifier. The study focused on classifying individuals with newly diagnosed coronavirus pneumonia, those with recurrent pneumonia, and healthy individuals. Gaur *et al.* [20] proposes a combination of VGG16, InceptionV3, and EfficientNetB0. The dataset

used in the study was created by compiling X-ray images from various public sources. It included images of viral pneumonia, coronavirus-related pneumonia, and normal pneumonia. The outcomes indicate that the suggested method yielded a top-notch model, achieving an overall accuracy of 92.93% and a sensitivity of 94.79%, specifically for COVID-19.

Previous studies demonstrate that ensemble models outperform individual models in both machine learning and deep learning contexts. Moreover, the research highlights that data augmentation improves performance and mitigates the risk of overfitting. Using a pre-trained model for pneumonia has achieved high accuracy through transfer learning. This study utilizes pre-trained models to form an ensemble by employing the CNN model, building upon these discoveries.

III. Materials and Methods

A. Dataset

The chest X-ray dataset from Kermany *et al.* [21] was utilized in this research study. The research used anterior-posterior chest X-ray images from the Guangzhou Women and Children's Medical Center, Guangzhou. The X-rays were taken as part of routine clinical care. For the analysis, all radiographs underwent quality control screening, with low-quality or unreadable scans being excluded. Two experienced doctors assessed the diagnoses in the images, and these assessments were confirmed to train the AI system. A third experienced doctor also examined the evaluation set to rectify any possible grading mistakes. The dataset consists of three directories: train, test, and val, and contains 5,856 JPEG X-ray images divided into two groups (Pneumonia/Normal), with 5,216 allocated for training and 624 designated for testing. During preprocessing, the images are resized from their original dimensions, which vary (e.g., 1,344×600, 1,272×1144), to a standardized size for consistency in model implementation and computation. The dimensions of each image are adjusted to 224×224 pixels with three color channels (red, green, and blue). Following resizing, data augmentation techniques are applied to expand the dataset.

The number of samples used for training and testing is detailed in Table I. Our objective is to create a validation set. We accomplish this by taking the original training dataset and performing a simple stratified split, allocating 75% for actual training, 15% for validation, and 10% for testing.

TABLE I. Number of Samples Training and Testing

Class	Training Set	Testing Set	Validation Set	Total
Normal	1,341	234	8	1,583
Pneumonia	3,875	390	8	4,273
Total	5,216	624	16	5,856

B. Preprocessing Image

Resizing the X-ray images was a critical step in data preprocessing, as it was necessary to accommodate the varying image input sizes required by different algorithms. For the base models, images were resized to 224 by 224 pixels. All images were normalized according to the norms established by the trained model. Additionally, we employed data augmentation techniques to expand a relatively small dataset, as our working database was not extensive. Improving the available data can enhance the performance of deep learning models rather than collecting new ones. Table II presents the augmentation parameters used for picture preprocessing in the study.

TABLE II. Augmentation Parameters

Parameters	Values
Rescale	1/255
Shear range	10
Zoom_range	0.1
Horizontal_flip	True
Brightness_range	(0.5, 1.0)
Width_shift_range	0.1
Rotation_range	20

C. Hyperparameters

To improve the models' capability to extract features, we began by selecting the top-performing deep CNN architecture from existing ones. The process of selecting involved training each architecture with the Adam optimizer, and a learning rate of 0.001 was implemented, then choosing the model that performed best. For this binary classification task, we used the Sigmoid activation function. Hyperparameter tuning was essential for improving model performance. In our study, we fine-tuned the models with various combinations of optimizers, learning rates, and other hyperparameters, as detailed in Table III, to determine the optimal configuration for achieving the highest validation accuracy.

TABLE III. Hyperparameter Values for Training

Hyperparameters	Values
Input activation function	ReLU
Output activation function	Sigmoid
Optimizer	ADAM
Initial learning rate	0.001
Learning rate decay	0.2
Number of epochs	50
Batch size	32

The loss function, binary cross-entropy, compares the predicted and actual outputs to measure the discrepancy. Its value ranges from 0 to 1 and is determined by Eq. (1).

$$Loss = \sum_{i=0}^{output\ size} y_i \times \log(y_i) + (1 - y_i) \times \log(1 - y_i) \quad (1)$$

D. Evaluation Metrics

In this work, we employ a confusion matrix, a common table that illustrates the low-level classification model's performance at a basic level on a test set. This matrix shows the percentage of positively predicted cases that were observed (True Positive, TP), the percentage of negatively predicted cases that were observed (True Negative, TN), and the percentage of cases that were incorrectly classified as positively or negatively (False Positive, FP, and False Negative, FN).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

E. Overview of Ensemble Learning

In this work, we combine various models based on the reasonable assumption that different models have varying capacities and can efficiently carry out subtasks. Consequently, when these models are utilized properly, they can create a strong ensemble model that outperforms the individual models alone. Generally speaking, there are two kinds of ensemble techniques characterized by gathering professional judgment to improve prediction accuracy and dependability during problem resolution. The initial instance consists of obtaining characteristics from images utilizing different CNN models. Utilizing a variety of machine learning algorithms, the retrieved features are integrated and applied to classification tasks. However, to pursue this method, we must meet certain requirements, such as undergoing two separate training processes and using complex algorithms. In the second approach, a mathematical model is utilized to combine the predictions made by the model. Examining how the ensemble system achieves accurate data classification through the aggregation of other models' precise predictions serves as an illustration of this approach. It is also feasible to employ ensemble learning for other purposes, such as data fusion and feature selection.

Ensemble approaches fall into three primary categories: stacking, boosting, and bagging. Bagging, also known as Bootstrap aggregating, involves training multiple base models independently and in parallel on various subsets of the training data in ensemble learning. The bagging classifier makes the final prediction by combining the predictions of all base models through majority voting. In regression models, the final prediction is generated by averaging the predictions from all base models, which is referred to as bagging regression. Boosting's training procedure is identical to bagging's; however, it takes place in a different order. By following along in this order, each model in this series will learn how to correct the errors in the model that came before it (i.e., the data that the prior model predicted incorrectly).

F. Stacked Convolutional Neural Network

Stacked generalization [22] is an ensemble method where a new model is trained to integrate optimal predictions from various existing models. The basic stacking model is usually separated into two levels: level-0 models and level-1 models. Level-0 models, also called Base-Models, derive their predictions for the level-1 model directly from the dataset, while level-1 models, also called Meta-Models, derive their predictions from the level-0 base models. Algorithm 1 presents a brief overview of Stacked generalization [22]. The stacking ensemble's primary advantage lies in its potential to harness the capabilities of numerous effective models for addressing both regression and classification challenges. Additionally, it aids in the development of a superior model with predictions that surpass the performance of each model.

Algorithm 1. Stacking Algorithm

Input: $D = \{(x_i, y_i) | \ x_i \in X, y_i \in Y\}$
Output: An ensemble classifier H
1: **Step1:** Learn first-level classifiers
2: for $t \leftarrow 1$ to T do
3: Learn a base classifier h_t based on D
4: **Step 2:** Construct new data set from D
5: for $i \leftarrow 1$ to m do
6: Construct a new data set that contains $\{x_i^{new}, y_i\}$, where $x_i^{new} = \{h_i(x_i) \ for \ j = 1 \ to \ T\}$
7: **Step 3:** Learn a second-level classifier
8: Learn a new classifier h^{new} based on the newly constructed data set
9: **Return** $H_{(x)} = h^{new}(h_1(x), h_2(x), \ldots, h_T(x))$

The stacking method proposes that different CNN sub-models capture nonlinear discriminative features and semantic image representations at varying levels. Therefore, a stacked ensemble CNN model is expected to be highly generalized and accurate.

Fig. 2 illustrates this stacked convolutional neural network. Bagging decreases the variance of weak learners while boosting lowers their bias. Using an ensemble learning technique called stacking can greatly improve the predictive performance of machine learning models. Through the combination of the forecasts made by several fundamental models, stacking can reduce bias and variance, increase model diversity, and improve the interpretability of the final prediction.

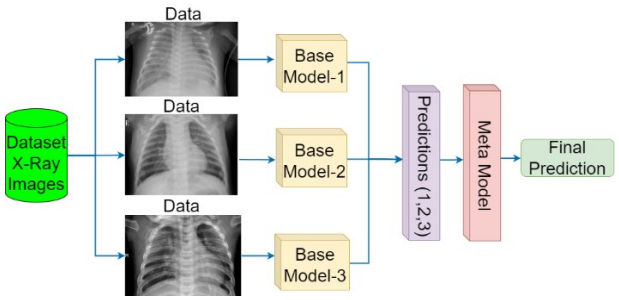


Fig. 2. Representing stacking is the method of classification of pneumonia images.

G. Methodology

The stacked CNN approach integrates multiple CNN models to maximize performance. The core idea involves feeding the output of one CNN model into another in a stacked formation. Each CNN model extracts features from the images, and by combining these models, the overall system can learn more complex and abstract features compared to using a single CNN model alone. An overview of our proposed framework is given in Figs. 3 and 4, which enhances the performance of pneumonia

image classification by the ensemble of three CNN models with diverse architectures, following the ensemble principles. Combining three different models, we propose an ensemble deep learning strategy that, compared to earlier approaches, helps boost deep learning prediction accuracy for pneumonia and reduces the misclassification error rate. The three CNN models are selected from InceptionResNetV2, MobileNetV2, DenseNet169, ResNet50V2, DenseNet201, and VGG16 models. These CNN architectures are trained with various image features; for example, one model might be trained to recognize edges, while another focuses on textures. Each model uses the Adam optimizer and focal loss function. The stacked model, which includes InceptionResNetV2, MobileNetV2, DenseNet169, ResNet50V2, DenseNet201, and VGG16, learns a complex and diverse set of image features by employing different architectures and hyperparameters. In the end, the CNNs in the stack process the image, and their results are combined before being input into a fully connected layer for the ultimate classification or regression task. The specific method of combining the CNN results depends on the unique approach of the CNN stacking algorithm. This approach allows for the development of a more accurate model by eschewing the use of a single model.

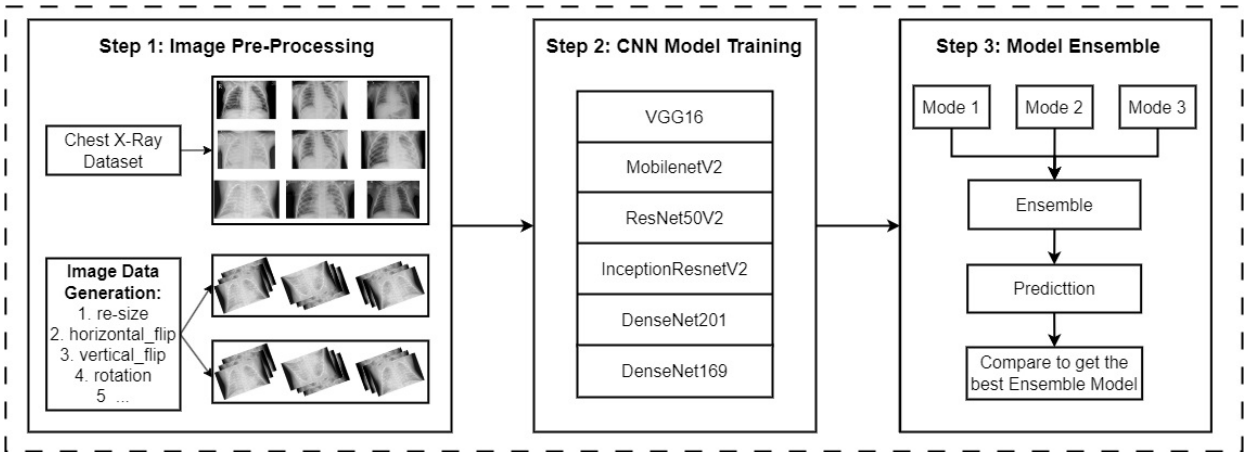


Fig. 3. The architecture of our proposed system consists of three main stages: step 1) preprocessing, step 2) model training, and step 3) model ensemble to get the best result.

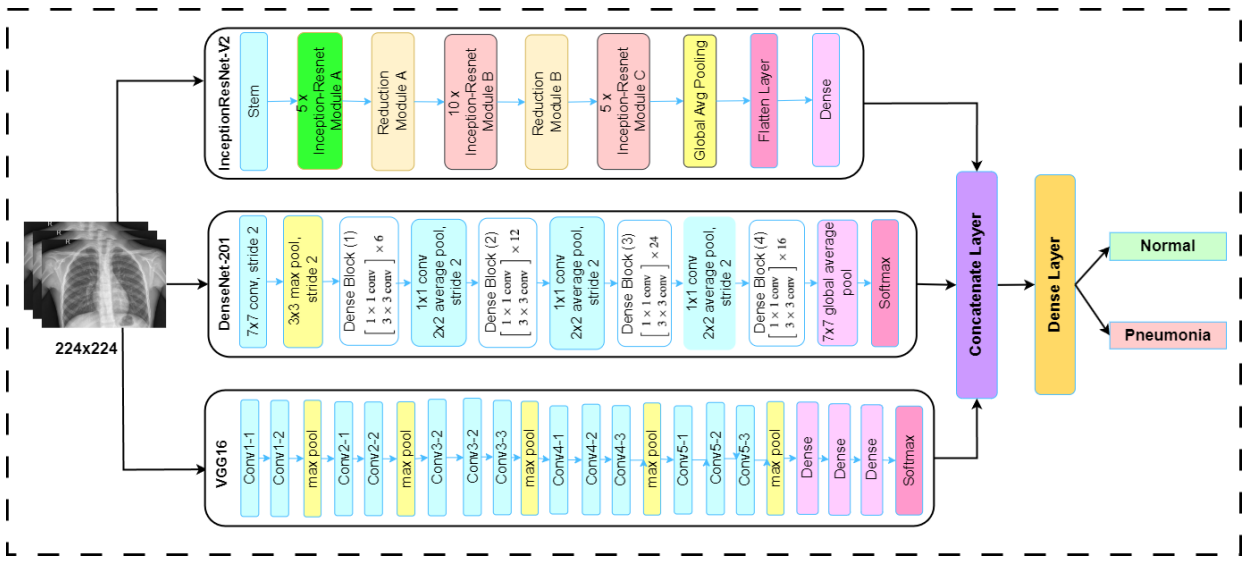


Fig. 4. The image shows the overall stacking strategy with the three effective models.

IV. Result and Discussion

A. Implementation Details

We applied the suggested technique in our environment on Debian GNU/Linux 12 with PyTorch 1.1.3 on Python 3.7 and CUDA 12.1 for GPU acceleration. The deep learning, machine learning, and image processing toolboxes were used to train the CNN-based models. The computational setup, which is a local server, consists of an Intel Xeon 2620v3 (6 cores, 12 threads @ 2.4GHz), 64GB of DDR4 memory, and an NVIDIA GeForce RTX 3090 with 12GB of memory. After initializing with weights for training images, each ensemble model underwent 50 training epochs. Test reports were recorded at the 20th, 30th, and 50th epochs, measuring accuracy, loss, precision, recall, and F1-Score. In the final stage, the results of the

meta-model classifying pneumonia images were compared to identify the top-performing ensemble model, as shown in Fig. 4.

B. Experimental Results of Single Models

As can be seen in the following section, the precision, recall, and F1-Score metrics were used to assess each model's performance. The performance of individual CNN models during training, in terms of loss and accuracy, is displayed in Fig. 5.

C. Experimental Results of Ensemble Models

In Step 3, we attempt to ensemble two or three CNN models during training. Figs. 6 and 7, which show the ensembles of three CNN models, present the ensemble models' loss, accuracy, and confusion matrix results. Using the chest X-ray dataset (Kermany [23]) for training, each model's performance was evaluated based on performance metrics, to identify the best-performing ensemble model (Fig. 7).

In our research, we investigated the use of both individual and combined learning models for the categorization of multiple classes of X-ray images to determine the most effective approach based on various criteria. We observed that the F1-Score is used in cases where False Negatives (FN) and False Positives (FP) have a significant impact, while accuracy is valued when True Positives (TP) and True Negatives (TN) are more crucial. When classes are evenly distributed, accuracy is suitable, but the F1-Score is more appropriate for imbalanced classes.

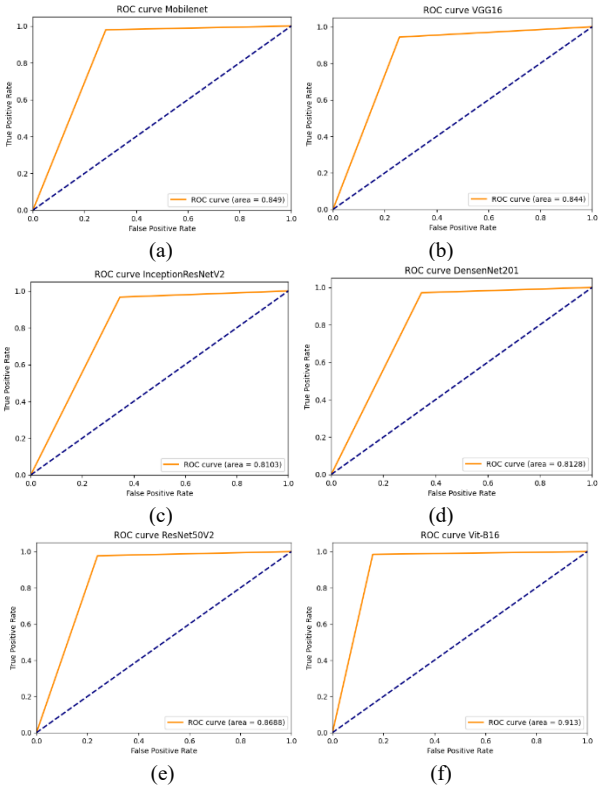
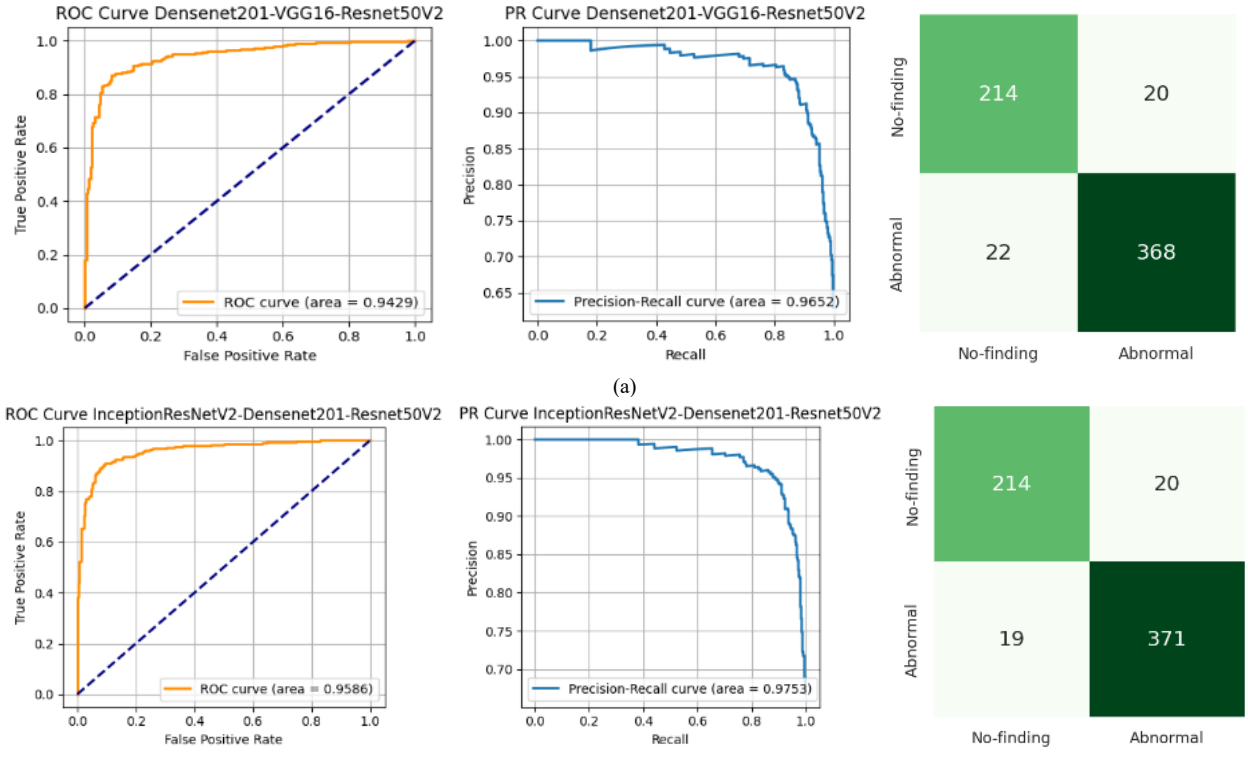


Fig. 5. The ROC curve of single CNN methods. a) MobilenetV2 model; b) VGG16 model; c) InceptionResNetV2 model; d) DenseNet201 model; e) ResNet50 model; and f) ViT-B16.model.



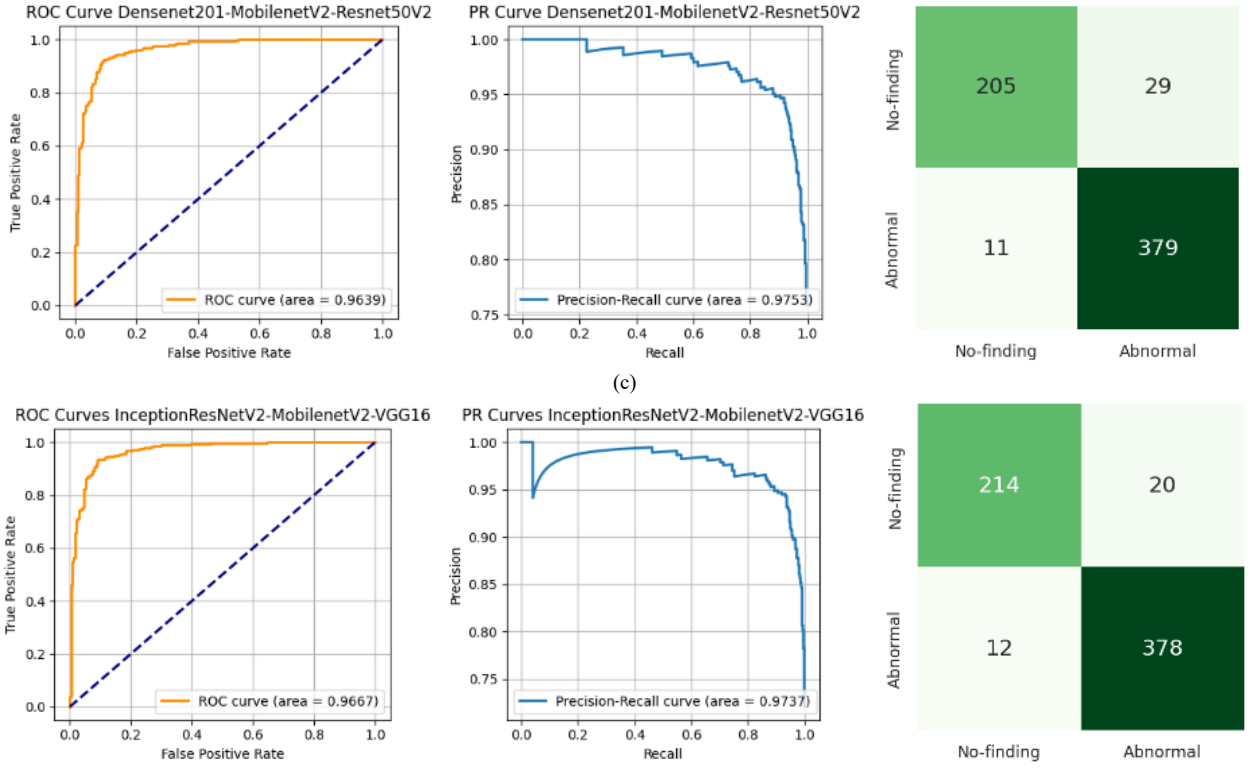
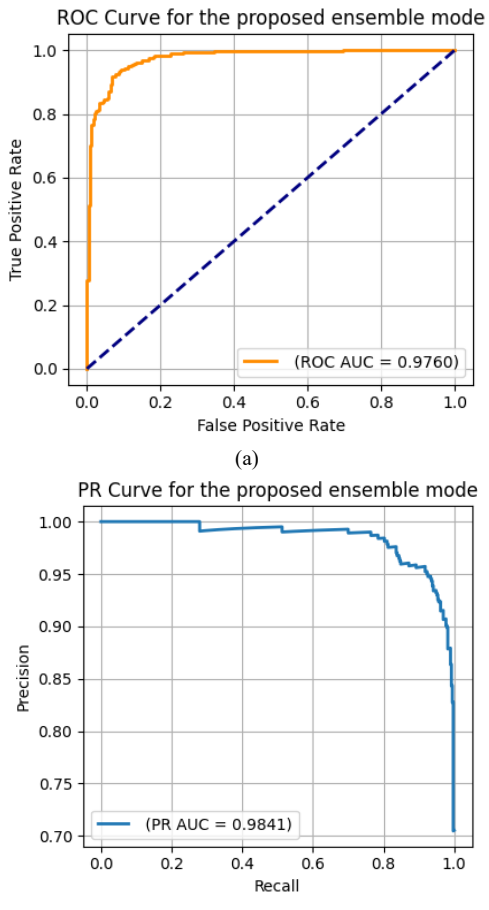


Fig. 6. The ROC curve, Precision-Recall curve, and confusion matrix of combining three CNN methods. (a) DenseNet201, VGG16, and ResNet50V2; (b) InceptionResNetV2, DenseNet201, and ResNet50V2; (c) DenseNet201, MobilenetV2, and ResNet50V2; (d) InceptionResNetV2, MobilenetV2, and VGG16.



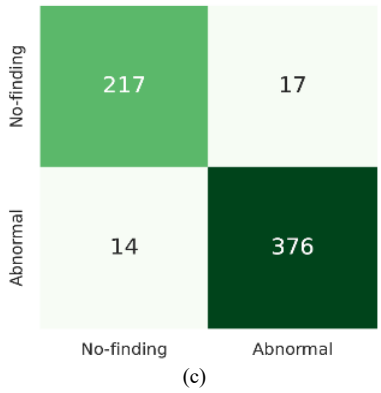
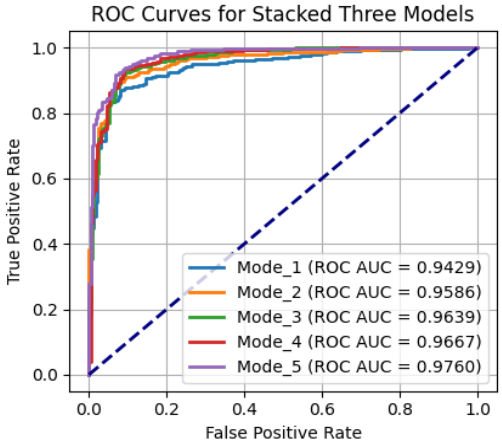


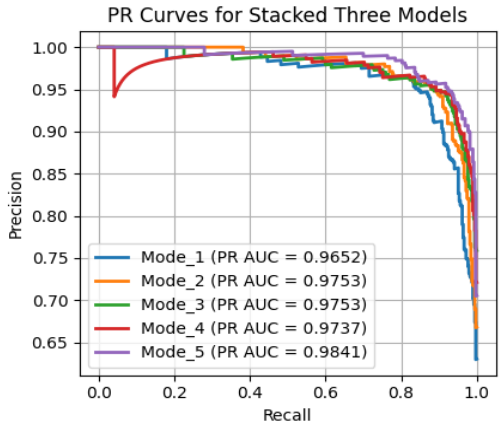
Fig. 7. The ROC curve (a), Precision-Recall curve (b), and confusion matrix (c) of combining the InceptionResNetV2, DenseNet201, and VGG16 models.

Our method demonstrated high sensitivity, allowing us to accurately classify pneumonia-positive cases. To further analyze the performance of our proposed method, we evaluated the confusion matrix and Receiver Operating Characteristic (ROC) curve (including the AUC's confidence interval) for each fold, as shown in Figs. 6–8. Illustrated in Fig. 8 are the ROC and Precision-Recall curves compared for ensemble CNN models (InceptionResNetV2, MobileNetV2, DenseNet169, ResNet50V2, DenseNet201, and VGG16). The graph reveals that the ensemble model from InceptionResNetV2, DenseNet201, and VGG16 achieved the highest Area under the ROC Curve (AUC) among the classifiers. Additionally, all the models have an AUC greater than 0.9,

indicating they all perform satisfactorily in classifying emotional identification.



(a)



(b)

Fig. 8. Comparison the ROC (a) and Precision-Recall (b) curves of stacking three CNN methods: Model_01 (InceptionResNetV2, DenseNet201, ResNet50V2 modes); Model_02 (DenseNet201, MobilenetV2, ResnetV2 modes); Model_03 (InceptionResNetV2, MobilenetV2, VGG16 modes); Model_04 (DenseNet201, VGG16, ResnetV2 modes); Model_05 (InceptionResNetV2, DenseNet201, VGG16 modes).

TABLE IV. COMPARISON OF TRAINING TIME FOR DIFFERENT MODELS

Models	Total params	Training time (s)	Training time (s)
MobilenetV2	2,422,081	8,245	24
VGG16	14,780,481	20,145	116
InceptionResNetV2	54,533,601	21,496	130
DenseNet201	18,568,001	20,707	112
ResNet50V2	23,827,201	15,515	65
InceptionResNetV2 & VGG16	69,618,209	35,200	202
MobilenetV2 & ResNet50V2	26,722,369	15,757	63
InceptionResNetV2&ResNet50V2	78,867,681	28,875	262
DenseNet169 & MobilenetV2	15,700,609	18,322	82
DenseNet201, VGG16, ResNet50V2	57,800,577	26,800	127
nceptionResNetV2, DenseNet201, ResNet50V2	97,688,865	36,886	262
DenseNet201, MobilenetV2, ResNet50V2	45,543,553	27,264	186
InceptionResNetV2, MobilenetV2, VGG16	72,208,993	36,893	161
InceptionResNetV2, DenseNet201, VGG16	**88,439,393**	**26,012**	**135**

Table IV highlights the differences in training and testing durations between single CNN models and stacking learning methods. Training single CNN models is usually quicker because only one model is involved. In contrast, ensemble CNN models take much longer to train, we must train each model separately. The complexity and depth of each model further extend the training time. Despite the longer training times and higher resource consumption, ensemble CNN models deliver superior performance compared to single CNN models. Our proposed method is more cost-effective than similarly complex ensemble models.

D. Discussion

The confusion matrix reveals that our model generates very few false negatives and false positives, especially for pneumonia cases compared to the normal dataset. Reducing incorrect diagnoses is crucial for pneumonia cases. Fig. 8 illustrates the diagnostic effectiveness of the proposed model, demonstrating its strong capability to distinguish pneumonia from chest X-ray images. The ROC curve illustrates the stability of the stacked CNN models, with our current proposed model achieving an AUC of 0.98, an average sensitivity of 92.73%, and a specificity of 96.41% in classifying the images into normal and pneumonia categories.

Table V compares the training and validation performance of MobileNetV2, VGG16, ResNet50V2, DenseNet169, DenseNet201, InceptionResNetV2, ViT-B16, and the stacking learning approaches. The quantitative findings show a clear upward trend. The lowest-performing model, MobileNetV2, achieved a training accuracy of 87.50% and an F1-Score of 90.46%. The highest-performing model, an ensemble of InceptionResNetV2, DenseNet201, and VGG16, achieved a training accuracy of 95.03% and an F1-Score of 96.04%. Fig. 9 displays the accuracy achieved for individual and group CNN models throughout this investigation. The results' accuracy was enhanced by the suggested approach.

TABLE V. THE OVERALL TESTING ACCURACY AND THE AVERAGES OF PRECISION, RECALL, AND F1-SCORE

Models	Accuracy	Precision	Recall	F1-Score
MobilenetV2	87.50%	86.45%	94.87%	90.46%
VGG16	88.94%	93.73%	88.21%	90.89%
InceptionResNetV2	90.87%	89.18%	97.18%	93.01%
DenseNet201	91.19%	92.58%	92.82%	92.70%
ResNet50V2	91.19%	93.51%	92.31%	92.90%
ViT-B16	93.75%	95.24%	92.31%	93.75%
InceptionResNetV2, VGG16	92.36%	93.88%	94.46%	94.12%
MobilenetV2, ResNet50V2	92.47%	93.86%	94.10%	93.98%
InceptionResNetV2, ResNet50V2	92.47%	94.32%	93.59%	94.95%
DenseNet169, MobilenetV2	92.79%	91.97%	96.92%	94.38%
DenseNet201, VGG16, ResNet50V2	93.27%	94.85%	94.36%	94.60%
InceptionResNetV2, DenseNet201, ResNet50V2	93.75%	94.88%	95.13%	95.01%
DenseNet201, MobilenetV2, ResNet50V2	94.39%	94.49%	96.67%	95.56%
InceptionResNetV2, MobilenetV2, VGG16	94.87%	94.97%	96.92%	95.94%
InceptionResNetV2, DenseNet201, VGG16	**95.03%**	**95.67%**	**96.41%**	**96.04%**

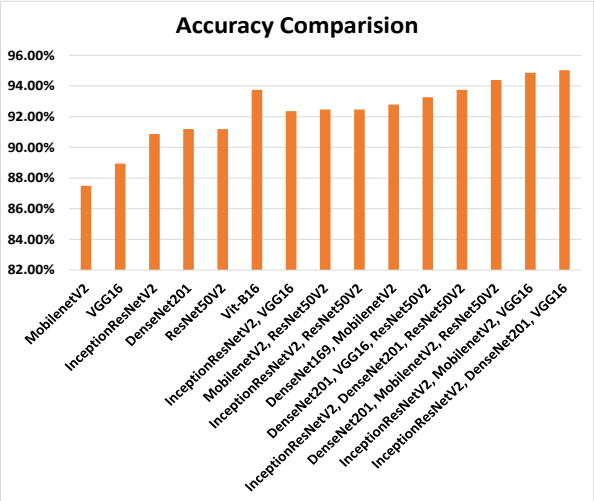


Fig. 9. The overall performance of single and stacking CNN models.

The studies conducted on images of pneumonia are listed in Table VI. Li *et al.* [19] classified pneumonia images with 93.57% accuracy by combining ensemble learning and VGG16, while Gaur *et al.* [20] offered a customized EfficientNetB0 with test accuracies of 92.93 %. H. Sharma *et al.* [24] presented a tailored CNN pneumonia classification model with an accuracy of 90.07%. A DenseNet-121-based pneumonia detection technique with 93.4% accuracy was presented by Salehi *et al.* [25].

TABLE VI. THE OVERVIEW OF OTHER METHODS AND ACCURACY RESULTS TOWARD PNEUMONIA CLASSIFICATION

Author	Methods	Accuracy
Gaur *et al.* [20]	EfficientNetB0	92.93 %
Sharma *et al.* [24]	CNN model_1	90.70 %
Sharma and Guleria [27]	VGG16 with Neural Networks	92.15 %
Salehi *et al.* [25]	DenseNet-121	93.40 %
Li *et al.* [19]	Ensemble learning with VGG16	93.57 %
Mabrouk *et al.* [26]	DenseNet169, MobileNetV2, and ViT	93.91 %
Saraiva *et al.* [28]	Customize CNN	94.40 %
The proposed method	**InceptionResNetV2, DenseNet201, and VGG16**	**95.03%**

Mabrouk *et al.* [26] used an ensemble of deep CNN and ViT, which produced good results with a score of 93.91%. Sharma *et al.* [27] and Saraiva [28] proposed customized CNNs, which had test accuracies of 92.15% and 94.4%, respectively. These findings, however, demonstrate worse performance than the proposed approach, which has an accuracy of 95.03%.

V. CONCLUSION AND FUTURE WORK

We utilize ensemble learning models in this study that are built upon finely adjusted stacking models (InceptionResNetV2, DenseNet201, and VGG16) to classify chest X-ray images into normal and pneumonia. Consequently, we discovered that for a single CNN model, InceptionResNetV2 yields the highest recall and F1-Score (97.18% and 93.01%, respectively), whereas DenseNet201 and ResNet50V2 yield the highest accuracies (91.19%). The accuracy of CNN model ensembles with 92.36%, 92.47%, 92.79%, 93.27%, 93.75%, 94.39%, 94.87%, and 95.05%, respectively, outperforms DenseNet201 and ResNet50V2 in this regard. Ensembles comprising two and three CNN models also produced F1-Scores of 94.12%, 93.98%, 94.95%, 94.38%, 94.60%, 95.01%, 95.56%, 95.94%, and 96.04%, surpassing the performance of individual models. This outcome demonstrates how we can enhance the performance of the overall model by integrating the benefits of several CNN models (R1). Additionally, ensembles of three CNN models (from InceptionResNetV2, MobilenetV2, DenseNet169, ResNet50V2, DenseNet201, and VGG16) outperform ensembles of two models (from InceptionResNetV2, MobilenetV2, DenseNet169, ResNet50V2, DenseNet201, and VGG16), with accuracies over 93% compared to less than 92.8%, respectively. Training accuracy and the F1-Score were both higher for ensembles of two CNN models compared to individual models. The ensembles achieved over 92.3% accuracy and over 94% F1-Score, while the single models achieved less than 91.2% accuracy and less than 93% F1-Score, respectively. Evidence supports the claim that the total number of deep learning models impacts the accuracy of the model used to build the ensembles (R2). The proposed method outperforms the Vision Transformer, as three combined models achieve training accuracy of 94.39%, 94.87%, and 95.03% compared to an accuracy of 93.75%. This indicates that the collective CNN models generate superior classification outcomes compared to modern technologies like Vision Transformer (R3).

For future work, we suggest a hybrid model named Pneu-Conv-ViT that enhances the performance of pneumonia image classification by combining a Vision Transformer with a backbone model. This backbone model is an ensemble of four CNN models with various architectures. Unlike earlier approaches, we propose a hybrid solution that can help reduce the misclassification error rate and improve deep learning prediction accuracy for pneumonia. The backbone model, consisting of the four CNN models, functions as a meta-model following ensemble principles. Furthermore, we could improve performance using larger datasets and more sophisticated feature extraction methods.

DATA ACCESS

The data is openly available in a public repository that issues datasets with doi:10.17632/rscbjbr9sj.2

CONFLICT OF INTEREST

The authors reported no potential conflict of interest.

AUTHOR CONTRIBUTIONS

G.S.T. and N.H.P. conducted the research. N.H.P. designed and performed the experiments derived the models, and analyzed the data. In consultation with G.S.T., N.H.P. wrote the manuscript; all authors approved the final version.

ACKNOWLEDGMENT

This work is supported by the Information and Communication Technology Lab, University of Science and Technology of Hanoi, and the Department of Information and Communication Technology, FPT University, Vietnam.

REFERENCES

[1] WHO (World Health Organization). (May 2024). Pneumonia in children. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[2] M. M. Thao and B. T. Chinh. (May 2024). Pneumonia: Causes, symptoms, diagnosis and treatment. [Online]. Available: https://vnvc.vn/viem-phoi/

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778.

[4] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017. doi: 10.1145/3065386

[6] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 1–9.

[7] A. Abedalla, M. Abdullah, M. Al-Ayyoub, and E. Benkhelifa, "Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures," *Peer J. Comput. Sci.*, vol. 7, p. e607, Jun. 2021.

[8] Z.-P. Jiang, Y.-Y. Liu, Z.-E. Shao, and K.-W. Huang, "An improved VGG16 model for Pneumonia image classification," *Applied Sciences*, vol. 11, no. 23, Jan. 2021.

[9] R. Sa *et al.*, "Intervertebral disc detection in X-ray images using faster R-CNN," in *Proc. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2017, pp. 564–567.

[10] S. Yao, Y. Chen, X. Tian, R. Jiang, and S. Ma, "An improved algorithm for detecting pneumonia based on YOLOv3," *Applied Sciences*, vol. 10, no. 5, Jan. 2020.

[11] Z. Ghomi *et al.*, "Segmentation of COVID-19 pneumonia lesions: A deep learning approach," *Medical Journal of The Islamic Republic of Iran (MJIRI)*, vol. 34, no. 1, pp. 1216–1222, Feb. 2020.

[12] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 2261–2269.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA: AAAI Press, Feb. 2017, pp. 4278–4284.

[14] G. Shih *et al.*, "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible Pneumonia," *Radiology: Artificial Intelligence*, vol. 1, no. 1, e180041, Jan. 2019.

[15] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3462–3471.

[16] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar, "Pneumonia detection in chest X-ray images using an ensemble of deep learning models," *PLOS ONE*, vol. 16, no. 9, e0256630, Sep. 2021.

[17] H. Bhatt and M. Shah, "A convolutional neural network ensemble model for Pneumonia detection using chest X-ray images," *Healthcare Analytics*, vol. 3, 100176, Nov. 2023.

[18] Q. An, W. Chen, and W. Shao, "A deep convolutional neural network for Pneumonia detection in X-ray images with attention ensemble," *Diagnostics*, vol. 14, no. 4, Jan. 2024.

[19] X. Li, W. Tan, P. Liu, Q. Zhou, and J. Yang, "Classification of COVID-19 chest CT images based on ensemble deep learning," *Journal of Healthcare Engineering*, vol. 2021, e5528441, Apr. 2021.

[20] L. Gaur, U. Bhatia, N. Z. Jhanjhi, G. Muhammad, and M. Masud, "Medical image-based detection of COVID-19 using deep convolution neural networks," *Multimedia Systems*, vol. 29, no. 3, pp. 1729–1738, Jun. 2023.

[21] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and chest X-ray images for classification," *Mendeley Data*, vol. 2, Jan. 2018.

[22] D. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1990.

[23] C. C. Aggarwal, *Data Classification: Algorithms and Applications*, 1st ed. Chapman & Hall/CRC, 2014.

[24] H. Sharma, J. S. Jain, P. Bansal, and S. Gupta, "Feature extraction and classification of chest X-ray images using CNN to detect Pneumonia," in *Proc. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Jan. 2020, pp. 227–231.

[25] M. Salehi, R. Mohammadi, H. Ghaffari, N. Sadighi, and R. Reiazi, "Automated detection of Pneumonia cases using deep transfer learning with paediatric chest X-ray images," *British Journal of Radiology*, vol. 94, no. 1121, 20201263, May 2021.

[26] A. Mabrouk, R. P. Díaz Redondo, A. Dahou, M. Abd Elaziz, and M. Kayed, "Pneumonia detection on chest X-ray images using ensemble of deep convolutional neural networks," *Applied Sciences*, vol. 12, no. 13, Jan. 2022.

[27] S. Sharma and K. Guleria, "A deep learning based model for the detection of Pneumonia from chest X-ray images using VGG-16 and neural networks," *Procedia Computer Science*, vol. 218, pp. 357–366, Jan. 2023

[28] A. Saraiva *et al.*, "Models of learning to classify X-ray images for the detection of Pneumonia using neural networks," in *Proc. the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, Prague, Feb. 2019, pp. 76–83.